Departament de Llenguatges i Sistemes Informàtics
Universitat Jaume I

Distribuciones de Clases No Balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje

TESIS DOCTORAL

Presentada por:
Vicente García Jiménez
Dirigida por:
Dr. Ramón A. Mollineda Cárdenas
Dr. José Salvador Sánchez Garreta

Castellón, Febrero de 2010

No quiero convencer a nadie de nada. Tratar de convencer a otra persona es indecoroso, es atentar contra su libertad de pensar o creer o de hacer lo que le dé la gana. Yo quiero sólo enseñar, dar a conocer, mostrar, no demostrar. Que cada uno llegue a la verdad por sus propios pasos, y que nadie le llame equivocado o limitado. (¡Quién es quién para decir "esto es así", si la historia de la humanidad no es más que una historia de contradicciones y de tanteos y de búsquedas?)...

Jaime Sabines.

A mis padres.

A mis amigos.

A la memoria de Don Ignacio Juárez
y del Dr. Ricardo Barandela.

Resumen

El problema de las distribuciones no balanceadas de datos entre clases ha recibido una considerable atención en disciplinas como el Aprendizaje Automático, el Reconocimiento de Formas y la Minería de Datos. Un conjunto de dos clases se dice que es no balanceado si una de las clases (mayoritaria) está significativamente más representada que la otra (minoritaria). Este tipo de complejidad de los datos puede conducir, en métodos tradicionales de clasificación, a aprendizajes sesgados en perjuicio de la clase minoritaria que, usualmente, contiene los casos de mayor interés. Esta deficiencia es particularmente importante en aplicaciones reales donde el coste por clasificar erróneamente los ejemplos de la clase menos representada es muy elevado.

Actualmente, es posible identificar tres grandes líneas de investigación sobre dicha problemática. En este sentido, la presente tesis doctoral aborda el problema del desbalance desde las perspectivas de evaluación del aprendizaje, solución y análisis de la complejidad de los datos.

En el caso de las métricas de efectividad, se propone una nueva forma de evaluar los clasificadores que solventa una posible desventaja de las métricas tradicionales, las cuales no consideran la magnitud ni el sentido de la diferencia entre las tasas de aciertos por clase. Diversos análisis sobre la nueva métrica, denominada IBA, empleando medidas de correlación de Pearson y cambios en la matriz de confusión, la confirman como una medida robusta y eficaz en entornos desbalanceados.

Dentro del campo de las soluciones para tratar el desbalance, se proponen nuevos esquemas para reducir/incrementar la talla de las clases mayoritaria/minoritaria, respectivamente. Asimismo, se presenta una solución algorítmica para la regla del vecino más cercano con el objetivo de favorecer la clasificación de muestras de la clase minoritaria. Este análisis no sólo se centra en el beneficio que podría representar cada una de las soluciones, sino también en cómo estas se ven afectadas por la naturaleza del clasificador, el sesgo de los datos y las métricas de efectividad.

Finalmente, la tercera línea de investigación abordada en esta tesis se centra en analizar los efectos sobre el aprendizaje que produce la presencia en los datos de complejidades como el desbalance y solapamiento. En particular, se estudian dos escenarios. En el primero, denominado homogéneo, las clases mayoritaria y minoritaria mantienen esta misma relación en la región de solapamiento. En el segundo, por el contrario, la clase minoritaria se convierte en la más representada en dicha región. Los resultados muestran que el desbalance local en la región de solapamiento puede ser más determinante que el desbalance global y que la talla de dicha región.

Abstract

The class imbalance problem has received considerable attention from areas such as Machine Learning, Pattern Recognition and Data Mining. A two-class data set is said to be imbalanced (or skewed) when one of the classes (the minority one) is heavily under-represented in comparison to the other class (the majority one). This type of complexity may lead to biased learning to the detriment of the minority class, which usually contains the most interesting cases. This issue is particularly important in real-world applications where it is costly to misclassify examples from the minority class.

Nowadays, it is possible to identify three main lines of research related to this problem. In this sense, this Ph.D. Thesis addresses the problem of class imbalance from the perspective of learning assessment, solution and analysis of the complexity of the data.

With respect to effectiveness metrics, a new way to evaluate classifiers is proposed, that solves a potential disadvantage of traditional metrics, which do not consider the magnitude and direction of the difference between the recognition rates per class. Several analysis of the new metric, called IBA, using Pearson correlation measures and changes in the confusion matrix, confirm it as a robust and effective choice in unbalanced environments.

Within the field of solutions to address the imbalance, new schemes to reduce/increase the size of the majority/minority class are proposed. In addition, an algorithm based on the nearest neighbor rule is introduced with the aim of favouring the correct classification of minority class samples. This analysis does not only focus on the intrinsic benefits of these solution methods, but also on how they are affected by the classifier nature, the skewed data and the effectiveness metrics.

The third line of research analyzes the effects on learning produced by some complexities such as data imbalance and overlap. In particular, we study two scenarios. In the first, called homogeneous, the majority and minority classes keep this relation in the overlap region. In the second, on the contrary, the minority class becomes the most represented in that region. The results show that the local imbalance in the overlap region can be more relevant than the global imbalance and the size of the region.

Agradecimientos

A mis amigos les adeudo la paciencia de tolerarme las espinas más agudas; los arrebatos de humor, la negligencia, las vanidades, los temores y las dudas...
Alberto Cortez

Esta memoria de tesis doctoral fue posible gracias al apoyo, comprensión y ayuda de muchas personas que en su momento, contribuyeron de diversas formas a la culminación de dicho trabajo. A todas ellas les dedico esta tesis y las siguientes líneas pretendiendo expresar mi más humilde agradecimiento esperando no olvidarme de alguna.

En primer lugar, mi más sincero (y eterno) agradecimiento a mis directores de tesis, José Salvador Sánchez Garreta y Ramón Alberto Mollineda Cárdenas, quienes con mucha paciencia, buen humor y gran profesionalismo han conseguido que hubiera terminado este trabajo. Asimismo por el gran apoyo y confianza que siempre me brindaron no sólo a nivel profesional sino también a nivel personal. También deseo agradecer a todos los que conforman el grupo de Visión por Ordenador, así como a mis profesores de Doctorado del Departamento de Lenguajes y Sistemas Informáticos. Por supuesto, mis agradecimientos van también a los profesores de Ingeniería y Maestría del Instituto Tecnológico de Villahermosa y del Instituto Tecnológico de Toluca, respectivamente, quienes me han dado las herramientas necesarias y el apoyo para continuar y culminar mi carrera académica. En especial a la Lic. Dulce León, al Ingeniero José R. Hernández Samberino, al Ing. Claudio Bastiani y al Dr. Eduardo Gasca.

Desde el lado personal y como parte de la fe, la educación y la tradición que me ha inculcado mi familia deseo agradecer a Dios, principalmente en situaciones difíciles a nivel personal, familiar y económico en las que de alguna u otra manera siempre salí avante.

A mis padres, Ma. del Carmen Jiménez y Nazario García, por todo el cariño, comprensión y apoyo que me han brindado a pesar de la distancia y de las largas ausencias que esto implica. Ustedes son el motivo principal que me impulsa a seguir y culminar mis proyectos profesionales y este trabajo no es más que un ejemplo de ello. A mis abuelitos, tíos y primos quienes en su momento han sido también como unos padres y hermanos. En particular a la familia Félix García.

Deseo expresar unas palabras de cariño y agradecimiento a mis grandes amigos y amigas a quienes admiro, respeto y considero como mis hermanos(as), y que como tal, en diversas situaciones han sabido llamarme la atención, aconsejarme y por supuesto darme su apoyo incondicional. Por supuesto como hermanos también hemos tenido nuestras grandes diferencias y como tal hemos sabido arreglarlas. En mi etapa de doctorado: Dino, Luis, Patricio, Graciela, Soledad, Roberto, Jerónimo, Mónica, Sergio, Sandra, Mauro, Bea, Israel, Mitzi, Hebert, Arturo y Alejandro. De mi etapa de trabajo en Gobierno del Estado de México y maestría: Adriana, Nacho, Marco A., Alex, Octavio, Rosa María, Socorro, Elizabeth, Martín, Lupillo, Nuri, Car, Inés

Gonzalez, Leti, Marina, Sonia, Beto, Tere, David, Sahid Becerril, Jorge Becerril, Juan Embriz, Rene Ortega, Monserrat Salgado, Karla Contreras y Vero Díaz. De mi tierra natal Tabasco: Cesar, Abraham, Martha, Marco, Palmi, mamá Palmi, Don Wilo, Wilo, Pepe, José Santana, Jesús, Ayde, Sandi, Ricardo, Perla, Inti y Mary.

A las siguientes familias quienes me adoptaron como un miembro más: García Álvarez, Juárez Domínguez, Garrido Vázquez, Cruz Gutiérrez, Ojeda Ramírez y Jiménez Gordillo.

Finalmente, en el plano institucional, deseo agradecer a diversas entidades como la Universitat Jaume I, el Ministerio de Educación y Ciencia, el Instituto Tecnológico de Toluca y el Gobierno del Estado de México, las cuales me han apoyado a través de diversos proyectos: MIPRCV, CICyT, COMECyT y CONACyT.

Gracias a todos.

Índice general

Resumen II

Abstract IV

Nomenclatura XX

1. Introducción 1
- 1.1. Marco general . . . 1
- 1.2. Motivación . . . 3
- 1.3. Objetivos . . . 4
- 1.4. Organización de la tesis . . . 5

2. Fundamentos teóricos 7
- 2.1. El reconocimiento de formas . . . 7
- 2.2. Sistema de reconocimiento de formas . . . 7
  - 2.2.1. Adquisición y preprocesamiento de datos . . . 8
  - 2.2.2. Representación de los datos . . . 9
    - 2.2.2.1. Instancia y atributo . . . 9
    - 2.2.2.2. Extracción y selección de atributos . . . 9
  - 2.2.3. Clasificación . . . 10
    - 2.2.3.1. Aprendizaje . . . 10
    - 2.2.3.2. Selección del modelo, estimación y evaluación . . . 11
- 2.3. Paradigmas del reconocimiento de formas . . . 12
  - 2.3.1. Aproximación estadística o geométrica . . . 13
    - 2.3.1.1. Teoría de decisión de Bayes . . . 14
    - 2.3.1.2. Regla de mínimo coste/riesgo de Bayes . . . 14
  - 2.3.2. Aproximación sintáctica o estructural . . . 15
  - 2.3.3. Redes neuronales artificiales . . . 15
  - 2.3.4. Correspondencia de plantillas . . . 16
- 2.4. Clasificadores . . . 16
  - 2.4.1. Naïve Bayes . . . 17
  - 2.4.2. Máquinas de vectores soporte . . . 18
  - 2.4.3. Red neuronal perceptrón multicapa . . . 20
  - 2.4.4. Redes de funciones de base radial . . . 22
  - 2.4.5. Árboles de decisión . . . 23
  - 2.4.6. Clasificadores basados en vecindad . . . 24
- 2.5. Complejidad de los datos . . . 25
  - 2.5.1. Ruido . . . 26
  - 2.5.2. Solapamiento . . . 26
  - 2.5.3. Talla de los datos . . . 26

2.5.4. Desbalance entre clases 27

3. Conjuntos de datos desbalanceados 29
3.1. Introducción 29
3.2. Naturaleza de los conjuntos desbalanceados 31
3.3. Las líneas de investigación 32
3.3.1. Soluciones para tratar el desbalance 32
3.3.1.1. Técnicas de remuestreo 33
3.3.1.1.1. Sobremuestreo. 33
3.3.1.1.2. Submuestreo. 35
3.3.1.1.3. Otras alternativas. 39
3.3.1.2. Métodos de clasificación en entornos no balanceados 39
3.3.1.2.1. Aprendizaje sensible al coste. 40
3.3.1.2.2. Algoritmos de clasificación con sesgo hacia la clase minoritaria. 41
3.3.1.2.3. Clasificadores de una clase. 41
3.3.1.3. Combinación de estrategias 42
3.3.1.3.1. Sistemas de múltiple clasificación. 43
3.3.1.3.2. Combinación de técnicas de remuestreo. . . 45
3.3.2. Evaluación de la efectividad de los clasificadores 46
3.3.2.1. Curvas *ROC* 47
3.3.2.1.1. Curvas *ROC* con *precision* y $posfrac$. . . 48
3.3.2.2. Área bajo la curva *ROC* 49
3.3.2.2.1. AUCPREC y Weighted-AUC. 50
3.3.2.3. Media geométrica 50
3.3.2.4. Medida f 50
3.3.2.5. Media geométrica de *tvp* y *precision* 51
3.3.2.6. Precisión optimizada 51
3.3.3. Relación con otras complejidades 51

4. Índice de precisión balanceada 53
4.1. Introducción 53
4.2. Una nueva medida de efectividad 55
4.2.1. Índice de dominio 56
4.2.2. Gráfico de precisión balanceada 56
4.2.3. Índice de precisión balanceada 57
4.2.4. Un ejemplo teórico 58
4.3. Análisis teórico de IBA 60
4.3.1. Análisis por correlación de Pearson 60
4.3.2. Análisis de las propiedades invariantes de las medidas 62
4.4. Experimentos en bases de datos reales 64
4.4.1. Análisis de resultados experimentales 66
4.5. Conclusiones 68

5. Técnicas para tratar el problema del desbalance 69
5.1. Introducción . . . 69
5.2. Soluciones a nivel de datos: algoritmos de submuestreo . . . 70
5.2.1. Edición de Wilson . . . 70
5.2.2. Subconjunto selectivo modificado . . . 71
5.2.3. Propuesta de un algoritmo de submuestreo . . . 72
5.2.4. Experimentos . . . 73
5.2.4.1. Análisis de resultados . . . 75
5.3. Soluciones a nivel de datos: algoritmos de sobremuestreo . . . 77
5.3.1. SMOTE basado en vecinos envolventes . . . 80
5.3.2. Experimentos y análisis de resultados . . . 81
5.4. Submuestreo vs. sobremuestreo . . . 83
5.4.1. Experimentos y análisis de resultados . . . 84
5.5. Soluciones a nivel de algoritmos . . . 88
5.5.1. Distancia ponderada . . . 88
5.5.2. Experimentos y análisis de resultados . . . 89
5.6. Combinación de estrategias . . . 89
5.6.1. Edición de Wilson con distancia ponderada . . . 90
5.6.1.1. Experimentos y análisis de resultados . . . 90
5.6.2. Solución a nivel de datos más solución a nivel de algoritmos . 92
5.6.2.1. Experimentos y análisis de resultados . . . 92
5.7. Conclusiones preliminares . . . 92

6. Análisis del comportamiento de algoritmos de clasificación en conjuntos de datos con solapamiento y desbalance 95
6.1. Introducción . . . 95
6.2. $k-$NN: un escenario infrecuente de desbalance y solapamiento . . . 96
6.3. Experimentos y análisis de resultados . . . 97
6.3.1. Desbalance homogéneo . . . 98
6.3.1.1. Comportamiento del algoritmo k-NN . . . 99
6.3.1.2. k-NN frente a otros modelos de clasificación . . . 101
6.3.2. Desbalance heterogéneo . . . 101
6.3.2.1. Comportamiento del algoritmo k-NN . . . 103
6.3.2.2. k-NN frente a otras aproximaciones . . . 106
6.3.3. Experimentos en bases de datos con un enorme ratio de desbalance . . . 108
6.4. Conclusiones preliminares . . . 110

7. Conclusiones y trabajos futuros 111
7.1. Contribuciones . . . 111
7.1.1. Las métricas de efectividad . . . 111
7.1.2. Soluciones para tratar el desbalance . . . 112
7.1.3. Complejidad de los datos . . . 113
7.2. Direcciones para trabajos futuros . . . 113

7.3. Publicaciones resultantes 114
7.3.1. Publicaciones en revistas internacionales 114
7.3.2. Publicaciones en congresos internacionales y nacionales 114

A. Clasificación sobre conjuntos originales **117**

B. Interpretación de los links de Tomek **121**

C. Tablas de pruebas $t-$student **123**

Referencias bibliográficas **193**

Índice de tablas

2.1. Paradigmas del reconocimiento de formas. 13

3.1. Matriz de confusión para un problema de dos clases 46

4.1. Ejemplo teórico en resultados de clasificación (tvp, tvn) hipotéticos, y valores de medidas de efectividad (los mejores resultados han sido resaltados). 60
4.2. Un ejemplo de una lista original (T) y dos listas alteradas (P1 y P2) para un desbalance de 1:3. Los elementos en negrita representan las muestras etiquetadas erróneamente. 61
4.3. Propiedades invariantes. '+' y '–' indican la invariancia y variancia, respectivamente. 64
4.4. Conjuntos de datos usados en los experimentos. 65
4.5. Resultados experimentales (en negrita mejores valores para cada métrica). 67

5.1. Prueba $t-$student sobre los resultados promedios de clasificación por $1-$NN en las 17 bases de datos originales y submuestreadas: métricas de efectividad $IBA_{0.1}$ y Acc. 76
5.2. Índice de efectividad empleando $IBA_{0.1}$ sobre los algoritmos de submuestreo. 78
5.3. Ratios de desbalance en conjuntos submuestreados 80
5.4. Ratios de desbalance y porcentajes de reducción empleando edición de Wilson con distancia ponderada (EWW^{-}). 92

A.1. Resultados de clasificación sobre el conjunto de datos Breast. 117
A.2. Resultados de clasificación sobre el conjunto de datos Ecoli. 117
A.3. Resultados de clasificación sobre el conjunto de datos German. 117
A.4. Resultados de clasificación sobre el conjunto de datos Glass. 118
A.5. Resultados de clasificación sobre el conjunto de datos Haberman. . . . 118
A.6. Resultados de clasificación sobre el conjunto de datos $Laryngeal_2$. . . 118
A.7. Resultados de clasificación sobre el conjunto de datos $Letter_2$. 118
A.8. Resultados de clasificación sobre el conjunto de datos Optdigits. . . . 118
A.9. Resultados de clasificación sobre el conjunto de datos Pendigits. . . . 118
A.10. Resultados de clasificación sobre el conjunto de datos Phoneme. . . . 119
A.11. Resultados de clasificación sobre el conjunto de datos Pima. 119
A.12. Resultados de clasificación sobre el conjunto de datos Satimage. . . . 119
A.13. Resultados de clasificación sobre el conjunto de datos Scrapie. 119
A.14. Resultados de clasificación sobre el conjunto de datos Segmentation. 119
A.15. Resultados de clasificación sobre el conjunto de datos Spambase. . . 119

A.16.Resultados de clasificación sobre el conjunto de datos Vehicle. 120
A.17.Resultados de clasificación sobre el conjunto de datos Yeast. 120

C.1. Prueba $t-$student sobre resultados de clasificación por la regla del vecino más cercano con distancia Euclídea y distancia ponderada según las métricas *tvp*, *tvn*, *Acc*, f_1, $IBA_{0.1}$ y Gm 123
C.2. Prueba $t-$student del clasificador $1-$NN en términos de *tvp* y *tvn* . . 124
C.3. Prueba $t-$student del clasificador $1-$NN en términos de *Acc* y f_1 . . 125
C.4. Prueba $t-$student del clasificador $1-$NN en términos de $IBA_{0.1}$ y Gm 126
C.5. Prueba $t-$student del clasificador $7-$NN en términos de *tvp* y *tvn* . . 127
C.6. Prueba $t-$student del clasificador $7-$NN en términos de *Acc* y f_1 . . 128
C.7. Prueba $t-$student del clasificador $7-$NN en términos de $IBA_{0.1}$ y Gm 129
C.8. Prueba $t-$student del clasificador $13-$NN en términos de *tvp* y *tvn* . 130
C.9. Prueba $t-$student del clasificador $13-$NN en términos de *Acc* y f_1 . 131
C.10.Prueba $t-$student del clasificador $13-$NN en términos de $IBA_{0.1}$ y Gm 132
C.11.Prueba $t-$student del clasificador J48 en términos de *tvp* y *tvn* . . . 133
C.12.Prueba $t-$student del clasificador J48 en términos de *Acc* y f_1 134
C.13.Prueba $t-$student del clasificador J48 en términos de $IBA_{0.1}$ y Gm . 135
C.14.Prueba $t-$student del clasificador J48 en términos de *tvp* y *tvn* . . . 136
C.15.Prueba $t-$student del clasificador MLP en términos de *Acc* y f_1 . . . 137
C.16.Prueba $t-$student del clasificador MLP en términos de $IBA_{0.1}$ y Gm 138
C.17.Prueba $t-$student del clasificador NBC en términos de *tvp* y *tvn* . . 139
C.18.Prueba $t-$student del clasificador NBC en términos de *Acc* y f_1 . . . 140
C.19.Prueba $t-$student del clasificador NBC en términos de $IBA_{0.1}$ y Gm 141
C.20.Prueba $t-$student del clasificador RBF en términos de *tvp* y *tvn* . . 142
C.21.Prueba $t-$student del clasificador RBF en términos de *Acc* y f_1 . . . 143
C.22.Prueba $t-$student del clasificador RBF en términos de $IBA_{0.1}$ y Gm 144
C.23.Prueba $t-$student del clasificador SVM en términos de *tvp* y *tvn* . . 145
C.24.Prueba $t-$student del clasificador SVM en términos de *Acc* y f_1 . . . 146
C.25.Prueba $t-$student del clasificador SVM en términos de $IBA_{0.1}$ y Gm 147
C.26.Prueba $t-$student del clasificador $1-$NN en términos de *tvp* y *tvn*: bases de datos moderadamente desbalanceadas 148
C.27.Prueba $t-$student del clasificador $1-$NN en términos de *Acc* y f_1: bases de datos moderadamente desbalanceadas 149
C.28.Prueba $t-$student del clasificador $1-$NN en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 150
C.29.Prueba $t-$student del clasificador $7-$NN en términos de *tvp* y *tvn*: bases de datos moderadamente desbalanceadas 151
C.30.Prueba $t-$student del clasificador $7-$NN en términos de *Acc* y f_1: bases de datos moderadamente desbalanceadas 152
C.31.Prueba $t-$student del clasificador $7-$NN en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 153
C.32.Prueba $t-$student del clasificador $13-$NN en términos de *tvp* y *tvn*: bases de datos moderadamente desbalanceadas 154

C.33. Prueba $t-$student del clasificador 13−NN en términos de Acc y f_1: bases de datos moderadamente desbalanceadas 155
C.34. Prueba $t-$student del clasificador 13−NN en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 156
C.35. Prueba $t-$student del clasificador J48 en términos de tvp y tvn: bases de datos moderadamente desbalanceadas 157
C.36. Prueba $t-$student del clasificador J48 en términos de Acc y f_1: bases de datos moderadamente desbalanceadas 158
C.37. Prueba $t-$student del clasificador J48 en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 159
C.38. Prueba $t-$student del clasificador MLP en términos de tvp y tvn: bases de datos moderadamente desbalanceadas 160
C.39. Prueba $t-$student del clasificador MLP en términos de Acc y f_1: bases de datos moderadamente desbalanceadas 161
C.40. Prueba $t-$student del clasificador MLP en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 162
C.41. Prueba $t-$student del clasificador NBC en términos de tvp y tvn: bases de datos moderadamente desbalanceadas 163
C.42. Prueba $t-$student del clasificador NBC en términos de Acc y f_1: bases de datos moderadamente desbalanceadas 164
C.43. Prueba $t-$student del clasificador NBC en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 165
C.44. Prueba $t-$student del clasificador RBF en términos de tvp y tvn: bases de datos moderadamente desbalanceadas 166
C.45. Prueba $t-$student del clasificador RBF en términos de Acc y f_1: bases de datos moderadamente desbalanceadas 167
C.46. Prueba $t-$student del clasificador RBF en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 168
C.47. Prueba $t-$student del clasificador SVM en términos de tvp y tvn: bases de datos moderadamente desbalanceadas 169
C.48. Prueba $t-$student del clasificador SVM en términos de Acc y f_1: bases de datos moderadamente desbalanceadas 170
C.49. Prueba $t-$student del clasificador SVM en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas 171
C.50. Prueba $t-$student del clasificador 1−NN en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 172
C.51. Prueba $t-$student del clasificador 1−NN en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 173
C.52. Prueba $t-$student del clasificador 1−NN en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 174
C.53. Prueba $t-$student del clasificador 7−NN en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 175
C.54. Prueba $t-$student del clasificador 7−NN en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 176

C.55. Prueba $t-$student del clasificador 7−NN en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 177
C.56. Prueba $t-$student del clasificador 13−NN en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 178
C.57. Prueba $t-$student del clasificador 13−NN en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 179
C.58. Prueba $t-$student del clasificador 13−NN en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 180
C.59. Prueba $t-$student del clasificador J48 en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 181
C.60. Prueba $t-$student del clasificador J48 en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 182
C.61. Prueba $t-$student del clasificador J48 en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 183
C.62. Prueba $t-$student del clasificador MLP en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 184
C.63. Prueba $t-$student del clasificador MLP en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 185
C.64. Prueba $t-$student del clasificador MLP en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 186
C.65. Prueba $t-$student del clasificador RBF en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 187
C.66. Prueba $t-$student del clasificador RBF en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 188
C.67. Prueba $t-$student del clasificador RBF en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 189
C.68. Prueba $t-$student del clasificador SVM en términos de tvp y tvn: bases de datos fuertemente desbalanceadas 190
C.69. Prueba $t-$student del clasificador SVM en términos de Acc y f_1: bases de datos fuertemente desbalanceadas 191
C.70. Prueba $t-$student del clasificador SVM en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas 192

Índice de figuras

2.1. Ciclo de un sistema de reconocimiento de formas. Las flechas en gris indican que el sistema puede conllevar la repetición de ciertas tareas. 8
2.2. Ciclo de un sistema automático de reconocimiento con aprendizaje supervisado. 11
2.3. Modelo canónico de un modelo de clasificación. 17
2.4. Ejemplo de un problema de dos clases en un espacio bidimensional. Los vectores soporte (puntos en negro) definen el margen máximo de separación entre dos clases. 18
2.5. Esquema de una neurona artificial. Modelo compuesto por un vector de pesos $w_1 \dots w_n$, equivalente a las conexiones sinápticas, y por el umbral θ de activación. 21
2.6. Modelo genérico de una red MLP con una capa oculta. 21
2.7. Modelo de una red neuronal RBF. 23
2.8. Ejemplo de un árbol de decisión para un problema de dos clases. . . 23
2.9. Clasificador basado en la regla de los k vecinos más cercanos. 25

3.1. Espacio de una curva ROC 48
3.2. Área bajo la curva ROC de un clasificador discreto 49

4.1. (a) Conjunto de datos con un desbalance 1:8 y (b) Resultados de clasificación. 54
4.2. Gráfico de precisión balanceada: el punto óptimo (*dominance* $= 0$ y $Gm^2 = 1$) que da lugar al rectángulo con área máxima. 57
4.3. Región que contiene las combinaciones posibles de *dominance* y Gm^2. 57
4.4. Función IBA. 58
4.5. Función IBA_α. 59
4.6. Diagramas de coeficientes de correlación de Pearson entre las medidas de efectividad consideradas. 63

5.1. Submuestreo por EW^- + SSM^-: (a) Conjunto de entrenamiento original, (b) Conjunto de datos resultado de EW^-, (c) Conjunto de datos después de EW^- + SSM^-. 74
5.2. Ranking de los algoritmos de submuestreo empleando los clasificadores 1−NN, 7−NN, 13−NN, J48, MLP, NBC y SVM. 79
5.3. Ranking de los algoritmos de sobremuestreo empleando los clasificadores 1−NN, 7−NN, 13−NN, J48, MLP, NBC y SVM. 83
5.4. Ranking de los algoritmos de submuestreo y sobremuestreo sobre bases de datos con un alto grado de desbalance empleando los clasificadores 1−NN, 7−NN, 13−NN, J48, MLP, NBC y SVM. 86

5.5. Ranking de los algoritmos de submuestreo y sobremuestreo sobre bases de datos con un bajo nivel de desbalance empleando los clasificadores 1−NN, 7−NN, 13−NN, J48, MLP, NBC y SVM. 87
5.6. Ranking de los clasificadores 1−NN, 7−NN, 13−NN y 1−NN+DP sobre los conjuntos originales empleando seis métricas de efectividad. 89
5.7. Ranking de las dos versiones de submuestreo empleando los clasificadores 1−NN, 7−NN, 13−NN, J48, MLP, NBC y SVM. 91
5.8. Ranking de la combinación de 1−NN+DP con los conjuntos submuestreados empleando seis métricas de efectividad. 93

6.1. Dos casos diferentes de desbalance homogéneo: a) 0 % de solapamiento y b) 60 % de solapamiento 98
6.2. Desbalance homogéneo: evaluación de la efectividad de los clasificadores $1, 3, 7, 9$ y 13-NN según las métricas (a) *tvp*, (b) *tvn*, (c) *precision*, (d) *Gm*, (e) *Gpr*, (f) *AUC*, (g) $IBA_{0.1}$ y (h) *Acc* 100
6.3. Desbalance homogéneo: evaluación de la efectividad de los clasificadores $1, 3, 9$-NN, NBC, MLP, RBF y J48 según las métricas (a) *tvp*, (b) *tvn*, (c) *precision*, (d) *Gm*, (e) *Gpr*, (f) *AUC*, (g) $IBA_{0.1}$ y (h) *Acc* 102
6.4. Dos casos diferentes de desbalance heterogéneo: [75–100] y [80–100]. El primer caso representa una situación balanceada en la región de solapamiento. Para el último caso, en dicha región, la clase mayoritaria aparece menos representada que la clase minoritaria 103
6.5. Desbalance heterogéneo: evaluación de la efectividad de los clasificadores $1, 3, 7, 9$ y 13-NN según las métricas (a) *tvp*, (b) *tvn*, (c) *precision*, (d) *Gm*, (e) *Gpr*, (f) *AUC*, (g) $IBA_{0.1}$ y (h) *Acc* 105
6.6. Desbalance heterogéneo: evaluación de la efectividad de los clasificadores $1, 3, 9$-NN, NBC, MLP, RBF y J48 según las métricas (a) *tvp*, (b) *tvn*, (c) *precision*, (d) *Gm*, (e) *Gpr*, (f) *AUC*, (g) $IBA_{0.1}$ y (h) *Acc* 107
6.7. Alto nivel de desbalance homogéneo: evaluación de la efectividad de los clasificadores $1, 3, 9$-NN, NBC, MLP, RBF y J48 según las métricas (a) *tvp*, (b) *tvn*, (c) *precision*, (d) *Gm*, (e) *Gpr*, (f) *AUC*, (g) $IBA_{0.1}$ y (h) *Acc* 109

B.1. Ejemplo para comparar diferentes versiones de los links de Tomek. . 121

Índice de algoritmos

1. SMOTE 34
2. One–sided selection 36
3. Edición de Wilson (EW) 71
4. Edición de Wilson sobre clase mayoritaria (EW^{-}) 71
5. Subconjunto Selectivo Modificado (SSM) 72
6. Submuestreo por edición y condensado 73
7. ncn–SMOTE 81

Nomenclatura

Ω_i	Región de decisión i
ω_i	Clase i
$\mathbf{x}$	Muestra, ejemplo o instancia
$d(x, y)$	Distancia Euclídea entre dos puntos $\mathbf{x}$ y $\mathbf{y}$
dominance	Índice de dominio
g_i	Función discriminante
K	Número de particiones del conjunto de datos
k	Número de vecinos más cercanos
L	Número de clases
N	Número total de ejemplos o muestras
n	Dimensionalidad
T	Conjunto de muestras etiquetadas
x_i	Atributo i
AUC	Area under curve ROC
BAG	Balanced accuracy graph
EW	Edición de Wilson
GG	Grafos de Gabriel
Gm	Media geométrica
CNN	Condensado de Hart
IBA	Índice de precisión balanceada o Index of balanced accuracy
MLP	Multilayer perceptron
NBC	Naïve Bayes classifier
NCN	Vecindad del centroide más cercano
RNG	Grafos de vecindad relativa
NN	Nearest neighbours

OP Optimized precision

OSS One–sided selection

RBF Radial basis function

ROC Receiver operating characteristic

RUS Submuestreo aleatorio

SMOTE Synthetic minority over-sampling technique

SSM Subconjunto selectivo modificado

SVM Support vector machine

tvp Tasa de falsos negativos

tfp Tasa de falsos positivos

TMK Links de Tomek

tvn Tasa de verdaderos negativos

tvp Tasa de verdaderos positivos

Capítulo 1

Introducción

En este primer capítulo, se introduce el tema central de la presente tesis doctoral, así como algunos conceptos que se emplearán a lo largo de la memoria. En particular, se describen la problemática a tratar, las principales líneas de investigación actuales y el ámbito en el cual se enmarca el trabajo llevado a cabo. Finalmente, se exponen los objetivos de la investigación y la estructura que se ha adoptado para este estudio.

1.1. Marco general

El proceso de *aprendizaje* llevado a cabo por los seres vivos y, en particular, por los seres humanos, es un tema objeto de amplios estudios por parte de la psicología y la biología. Dada su complejidad, actualmente existen diversas teorías y modelos que intentan explicar este proceso.

Una de las definiciones más generales establece que el aprendizaje es la adquisición de conocimiento por medio del estudio, la experiencia y la asociación de ideas, siguiendo principios como la causalidad y la contigüidad espacial y temporal. Esta habilidad permite a los seres humanos, entre otras cosas, la posibilidad de reconocer a una persona por su silueta, voz o escritura, leer caracteres escritos, decidir por el olor o el color si una fruta está madura, identificar variaciones en notas musicales y determinar el estilo al que pertenecen.

Diversas teorías apuntan que, debido a la existencia de una enorme cantidad de información, el ser humano busca simplificar tal multitud mediante la *categorización* o *clasificación*, el cual es un proceso dinámico que ordena o dispone por *clases* o *grupos* objetos del mundo real, y que empieza al nacer y continúa a lo largo de toda la vida. Un ejemplo de esto es el trabajo realizado por Carl Von Linné, a quien se le conoce como el padre de la taxonomía. Car Von Linné propone el empleo de una nomenclatura binomial con el propósito de poder nombrar con precisión todas las especies animales y vegetales, sirviéndose únicamente de dos términos: el género y la especie.

En la categorización, al igual que en el aprendizaje, una de las preguntas existentes es, ¿cómo se lleva a cabo este proceso? Dos teorías apuntan a que éste se realiza en dos formas: *perceptiva* y *conceptual*. En la primera, la agrupación de la información se produce considerando rasgos físicos de semejanza. En la segunda, empleamos conocimientos o teorías previas que se tienen sobre el mundo real. Categorizamos perceptivamente al reconocer formas geométricas, colores, texturas o materiales que constituyen los objetos, mientras que al identificar la función de los objetos, lo hacemos de forma conceptual. En ambos casos se requiere tener un conocimiento previo

(aprendizaje).

Esta capacidad del ser humano de poder categorizar de forma perceptual o conceptual es empleada en casi todas las actividades de la vida. Dos ejemplos son la industria y la medicina. En el primer caso, es frecuente el uso de personas en tareas de control de calidad; por ejemplo, en la selección de tapones de corcho, los operarios cualificados consideran aspectos como la porosidad, la densidad y el color. En el caso de la medicina, en particular para la detección temprana del cáncer de mama, un técnico radiólogo emite una clasificación numérica después de interpretar la mamografía.

En un esfuerzo por diseñar y construir máquinas capaces de imitar el reconocimiento llevado a cabo por el ser humano, múltiples disciplinas han surgido como una respuesta a tal necesidad. Entre éstas, podemos mencionar el reconocimiento de formas, el aprendizaje automático y la minería de datos, las cuales se apoyan en otras ciencias como las matemáticas, la estadística, la biología y la psicología.

La aplicación de estas disciplinas puede abarcar dominios como la bioinformática, la identificación biométrica, la clasificación de documentos, el análisis de imágenes, la recuperación/ordenación en bases de datos, el reconocimiento de voz, la teledetección, la seguridad informática y la automatización industrial. Esto hace que los sistemas puedan ser empleados en áreas tan diferentes como la medicina, la industria, la agricultura y la banca, entre otras.

El éxito de los sistemas automáticos de reconocimiento ha sido ampliamente documentado en la literatura. Sin embargo, para ello, se ha tenido que sortear una cantidad de dificultades que van desde la elección del modelo hasta el coste de almacenamiento y procesamiento de los datos. En este sentido, diversos estudios han demostrado analíticamente que el éxito o el fracaso en el diseño de un sistema de reconocimiento de formas depende fuertemente de la talla y de la complejidad que presenta el conjunto de datos de aprendizaje o entrenamiento.

A modo de ejemplo, en ciertas áreas de aplicación, se ha observado que los datos podrían no estar uniformemente distribuidos entre las distintas categorías o clases. A este tipo de complejidad, en la que alguna de las clases presenta un número de muestras significativamente mayor que las otras, se le conoce como el *problema de las clases desbalanceadas*[1]. La naturaleza de dicho desequilibrio puede deberse a razones intrínsecas del propio problema, o al elevado coste de adquirir los datos correspondientes a ciertas clases.

La efectividad de los métodos tradicionales suele verse afectada al aprender de conjuntos de datos no balanceados. Muchos algoritmos aprenden muy bien de las clases más representadas en detrimento de las minoritarias, lo cual da lugar a clasificadores con bajas tasas de aciertos sobre estas últimas. La situación se torna crítica cuando las clases con menos ejemplos son las de mayor interés para el problema, por ejemplo, enfermedades raras, transacciones fraudulentas, defectos en sistemas

[1] Aunque en la lengua castellana no existe el término desbalance ni ninguna de sus derivaciones, dada su proximidad al término anglosajón "imbalance", su aparente poca ambigüedad y su comodidad de uso, en la presente memoria se emplean para cualificar todo lo referente a la problemática de los conjuntos no balanceados.

de control de calidad, etc. En tales escenarios, el clasificador sería ineficaz al cometer frecuentemente errores en el reconocimiento de muestras de las clases minoritarias, a los cuales suelen asociarse altos costes en términos humanos y/o económicos.

El tema central de esta tesis es el estudio del problema de las clases desbalanceadas y de sus efectos sobre los procesos de aprendizaje y clasificación, en el contexto de las disciplinas de Reconocimiento de Formas, Aprendizaje Automático y Minería de Datos.

Estas tres áreas científicas, aunque con distinta denominación, suelen coincidir en muchos de sus objetivos. Considérese la definición de Aprendizaje Automático ofrecida por Alpaydin (2004), en la cual se presenta como "*el desarrollo de programas informáticos que optimizan un criterio de rendimiento usando ejemplos o experiencias pasadas*". Por su parte, Duda y otros (2001) definen el Reconocimiento de Formas como "*la acción de tomar una decisión basada en la categoría del ejemplo, teniendo un conocimiento a priori*". En cuanto a la Minería de Datos, Witten y Frank (2005) puntualizan que es "*el proceso que permite resolver problemas mediante el análisis de ejemplos presentes en las bases de datos*". Como se puede observar, estas tres definiciones coinciden en la necesidad de un conocimiento *a priori* para poder emitir una decisión.

1.2. Motivación

Un caso particular del problema de las clases desbalanceadas es aquel que se define para dos clases, de las que una (minoritaria) contiene un número de muestras significativamente menor que la otra clase (mayoritaria). En este contexto, las clases minoritaria y mayoritaria suelen representar la presencia y ausencia, respectivamente, de casos poco comunes, por lo que también son conocidas como clases positiva y negativa.

En la literatura, pueden encontrarse algunos ejemplos de problemas reales de dos clases desbalanceadas. En el área de la minería de datos, por ejemplo, Fawcett y Provost (1997) desarrollan un sistema para la detección de llamadas telefónicas fraudulentas, cuyos costes alcanzan los cientos de millones de dólares por año, ocasionando grandes inconvenientes a clientes y a proveedores de servicios. En esta aplicación, los autores manejan un conjunto de datos formado por un 20 % de casos de fraude (clase positiva) y un 80 % de no fraude (clase negativa). Además, establecen que el coste por decidir erróneamente que el teléfono del cliente ha sido clonado (falso positivo) es de 5 dólares, mientras que a la situación de no detectar una llamada fraudulenta (falso negativo) se le asigna un coste de 0.40 centavos por minuto. Otro ejemplo, perteneciente al área de la medicina, ha sido discutido en Cohen y otros (2006), donde la base de datos usada para monitorizar y detectar infecciones en hospitales está compuesta por un 11 % de pacientes infectados (casos positivos) y un 89 % de sanos (casos negativos). Un tercer ejemplo consiste en un sistema de detección de manchas de petróleo en el océano partir de imágenes de satélites (Kubat y otros, 1998), cuyo conjunto de datos contiene un 4 % de casos positivos (presencia

de manchas) y un 96 % de negativos.

Como se ha comentado antes, los conjuntos de datos que presentan esta particularidad constituyen un reto para las técnicas tradicionales de aprendizaje. Generalmente, estos métodos tienden a optimizar medidas globales de efectividad, lo que ocasiona que el sistema "menosprecie" las muestras positivas dada su escasa representatividad en el conjunto de datos. El inconveniente de dichas medidas es que son incapaces de distinguir entre el error de clasificar como negativa una muestra positiva y viceversa.

Por otra parte, el desbalance añade un nuevo factor a la complejidad de los datos que, al combinarse con otros como el solapamiento, incrementan la dificultad en su análisis.

Esta problemática motiva tres líneas fundamentales de investigación en el campo del desbalance, las cuales han sido consideradas en esta tesis:

1. Soluciones a nivel de algoritmos y datos.
2. Medidas de evaluación de la efectividad del clasificador en entornos desbalanceados.
3. Análisis de la relación entre el desbalance y otras complejidades de los datos, así como su incidencia en la calidad del proceso de aprendizaje.

1.3. Objetivos

El objetivo central de esta tesis es el estudio del problema de distribuciones binarias desbalanceadas desde las perspectivas de solución, evaluación del aprendizaje y análisis de la complejidad de los datos.

Del enunciado anterior, se derivan los siguientes tres objetivos particulares:

1. Estudio y desarrollo de medidas de evaluación del aprendizaje idóneas para entornos desbalanceados.
2. Estudio y desarrollo de técnicas de preprocesamiento de conjuntos desbalanceados, con el objetivo de lograr aprendizajes más efectivos.
3. Análisis de la interacción entre el desbalance y el solapamiento, así como de su influencia sobre los procesos de aprendizaje.

En el primer punto, se propone una nueva métrica de evaluación de la efectividad de un clasificador para problemas de 2 clases no balanceadas, la cual se compara con la tasa de exactitud global (*accuracy*) y con un conjunto de medidas que son comúnmente empleadas en estos contextos.

Con respecto al segundo objetivo, es bien conocido que el comportamiento de un clasificador depende significativamente de la complejidad de los datos presentados al sistema. Una práctica común previa al aprendizaje es el preprocesamiento de los datos con la finalidad de reducir su complejidad. En esta tesis, se presentan

varios métodos de remuestreo dirigidos a balancear las cantidades de instancias en ambas clases, los cuales se dividen en técnicas de submuestreo y sobremuestreo aplicables a las clases mayoritaria y minoritaria, respectivamente. Posteriormente, se emplean diversos clasificadores para evaluar la calidad de los conjuntos de datos resultantes en relación a los conjuntos originales. Por otra parte, se propone una solución algorítmica al problema del desbalance basada en la regla del vecino más próximo, la cual no requiere modificar las distribuciones originales de los datos.

Finalmente, el tercer objetivo se centra en el análisis de la influencia del desbalance sobre la efectividad de los métodos de aprendizaje, así como en su interacción con otras fuentes de complejidad como el solapamiento. Los experimentos demuestran que, en ciertos escenarios, el desbalance global es mucho menos determinante que el tipo y la magnitud del solapamiento.

1.4. Organización de la tesis

En función de los objetivos enumerados anteriormente, la tesis se ha organizado en siete capítulos. En particular, consta de una introducción, dos capítulos que abarcan temas referentes a métodos, herramientas y estado del arte, tres capítulos que conforman el cuerpo principal con las aportaciones de esta tesis y, finalmente, un último capítulo con las principales conclusiones. Además, se han incluido una serie de anexos con diversa información complementaria.

El Capítulo 1 (Introducción) muestra una visión general de la problemática tratada en esta memoria, presenta las diferentes líneas de investigación en las que se ha trabajado y detalla los objetivos de esta tesis.

En el Capítulo 2 (Fundamentos teóricos), se introducen brevemente diversos aspectos conceptuales referentes al reconocimiento de formas que constituyen los fundamentos teóricos de esta tesis. Entre ellos, cabe mencionar la arquitectura general de un sistema de reconocimiento de formas, los principales paradigmas de esta disciplina, numerosos clasificadores que se usarán posteriormente en los experimentos y algunas reflexiones referentes a las complejidades que pueden presentar los conjuntos de datos.

El Capítulo 3 (Conjuntos de datos desbalanceados) contiene una revisión de la literatura acerca del problema del desbalance enmarcado en las tres líneas de investigación que trata esta tesis: soluciones, métricas y complejidad de los datos. Para cada una de ellas, se citan los trabajos más importantes que se han llevado a cabo.

Los Capítulos 4, 5 y 6 están dedicados a la presentación y discusión de las aportaciones realizadas en este tesis para cada uno de los objetivos planteados, incluyendo sus resultados experimentales.

Finalmente, el Capítulo 7 resume las principales conclusiones a las que se han arribado. Asimismo, se enumeran las aportaciones realizadas y se plantean las futuras líneas de investigación a seguir como posible extensión de esta tesis.

Los Anexos A, B y C presentan, respectivamente, la información referente a

los resultados de clasificación sobre conjuntos originales del Capítulo 4, una breve discusión acerca de la definición original de los links de Tomek (1976b) y la versión descrita en el trabajo de Kubat y Matwin (1997) y las tablas correspondientes a la prueba $t-$student realizada sobre los resultados del Capítulo 5, las cuales sirven para realizar comparaciones fundamentadas entre los diferentes esquemas de solución.

Capítulo 2

Fundamentos teóricos

Este capítulo tiene por objetivo exponer ciertas generalidades del reconocimiento de formas y, en particular, del modelo estadístico, así como algunos contenidos teóricos necesarios para el posterior desarrollo de la presente tesis.

2.1. El reconocimiento de formas

El *Reconocimiento de Formas* es un campo de estudio interdisciplinar que comenzó a desarrollarse en la década de 1960 y que se nutre de avances provenientes de ciencias como las matemáticas, la estadística, la ingeniería, la inteligencia artificial, la informática, la psicología y la fisiología, entre otras. Su campo de aplicación puede abarcar áreas muy diversas, tales como el reconocimiento automático de caracteres, el diagnóstico médico, el análisis/extracción de información de grandes bases de datos, etc.

El objetivo de un sistema automático de reconocimiento, descripción, clasificación o agrupamiento es emitir una decisión, teniendo como base un conocimiento *a priori*. Esta información acumulada se compone de descripciones de objetos conocidos en forma de representaciones vectoriales o estructuradas, denominadas *ejemplos* o *muestras*[1], que suelen permitir distinguir total o parcialmente un objeto de otro. Por lo general, aquellos ejemplos que presentan propiedades comunes o similares se ordenan o agrupan en una misma *clase*/*categoría*.

En la presente tesis, los temas a tratar se enmarcan en un problema de clasificación y, por tanto, desde el punto de vista que nos ocupa, el objetivo de un sistema de reconocimiento de formas es asignar un ejemplo de entrada a una de las múltiples categorías existentes.

2.2. Sistema de reconocimiento de formas

Un *sistema de reconocimiento de formas* está compuesto por una serie de procesos orientados a la transformación de señales o datos en experiencias o entidades con significado. Considerando su función, estos procesos pueden organizarse dentro de alguna de las tres etapas esenciales de las que consta un sistema automático de reconocimiento de formas:

1. Adquisición y preprocesamiento de los datos.

[1] Los términos ejemplo, instancia y muestra serán empleados indistintamente a lo largo de la presente tesis para describir una observación, un dato o un objeto.

2. Representación de los datos.

3. Clasificación.

Este esquema general, a su vez, agrupa en cada una de sus etapas diferentes actividades de vital importancia: colección de datos, selección/extracción de características, elección del modelo, entrenamiento y evaluación.

La Figura 2.1 muestra los componentes del ciclo de un sistema de reconocimiento de formas, cuyo inicio parte de un problema del mundo real a resolver y finaliza con la obtención de la solución a dicho problema.

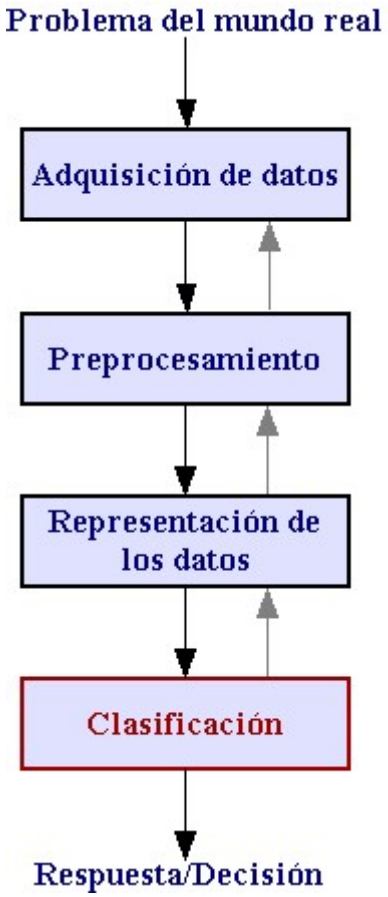


Figura 2.1: *Ciclo de un sistema de reconocimiento de formas. Las flechas en gris indican que el sistema puede conllevar la repetición de ciertas tareas.*

2.2.1. Adquisición y preprocesamiento de datos

La adquisición de datos tiene por objetivo captar, mediante sensores (por ejemplo, un micrófono o una cámara), los objetos del mundo real y proporcionar representaciones primarias de éstos, para su posterior utilización en el proceso de clasificación.

La dificultad de esta actividad radica en su dependencia de las características y limitaciones de los dispositivos, como pueden ser el ancho de banda, la resolución, la sensibilidad y la latencia, entre otros.

El módulo de preprocesamiento efectúa, en general, tareas como segmentar la región de interés, eliminar ruido, suprimir grises por umbral, así como realizar otras operaciones que podrían contribuir a obtener una representación compacta del ejemplo.

2.2.2. Representación de los datos

El objetivo en este módulo es llevar a cabo una transformación de la representación primaria de un objeto a otra que resulte más apropiada para un sistema de clasificación. Este proceso incluye la *selección/extracción de características*, con el objetivo de reducir la dimensionalidad del conjunto de datos sin perjudicar la efectividad del clasificador. En ciertas aplicaciones, como etapa final, se incluye un proceso de *normalización* de los datos.

2.2.2.1. Instancia y atributo

Se denominará *instancia* a una unidad de información mediante la cual se representan objetos del mundo real, como caracteres escritos, símbolos, dibujos, imágenes, objetos tridimensionales, firmas, huellas dactilares, espectrogramas, etc.

Considerando la naturaleza del problema y, por tanto, el método de interpretación escogido, la representación de los objetos se puede dividir en dos grandes grupos:

No estructurada. El objeto se representa por un conjunto de subobjetos sin relación entre ellos. Un ejemplo típico es un vector o matriz en los que no importa cómo se distribuyen sus *componentes*.

Estructurada. El objeto se representa mediante un conjunto de subobjetos más un conjunto de relaciones entre ellos (una estructura). Normalmente, las relaciones son de contigüidad espacial y/o sucesión: cadenas, árboles, grafos, son algunos ejemplos de este tipo de representación.

En el caso de la primera representación, los componentes que conforman el vector reciben el nombre de *atributos*. Si los valores de éstos son *ordinales* o *nominales*, se dice que los atributos son del tipo *cualitativo* o *categórico*. Por el contrario, si son continuos o discretos, son definidos como *cuantitativos* o *numéricos*.

2.2.2.2. Extracción y selección de atributos

Los métodos de *extracción/selección* de atributos permiten identificar y eliminar información que pudiera ser redundante y/o irrelevante. Este tipo de proceso consiste en encontrar, a partir de los atributos originales, un nuevo espacio de representación

que optimice un funcional que, en muchos casos, se refiere a la efectividad de los propios algoritmos de aprendizaje/clasificación.

La diferencia entre extracción y selección radica en que, en la primera de las aproximaciones, cada nueva característica es una combinación o función de los atributos originales, mientras que en la segunda se eligen aquellos atributos originales que proporcionan mejor información al sistema.

Los beneficios que se obtienen en esta etapa se reflejan en una disminución de la dimensionalidad y, por consiguiente, en una reducción del coste computacional del clasificador en términos espacial y temporal. Adicionalmente, al eliminar información irrelevante o con ruido, es posible también reducir el error del clasificador.

2.2.3. Clasificación

La fase de clasificación es una de las etapas esenciales dentro del diseño de un sistema automático de reconocimiento de formas. En él, se interpretan las muestras provenientes del módulo de representación. Sin embargo, debido a que, en general, no se conoce la descripción exacta de las muestras, el sistema es *entrenado a priori* para ajustarse a las características propias del problema de reconocimiento. A este proceso de adquirir e integrar conocimiento a un sistema de clasificación a partir de ejemplos (del mundo real), se le conoce como *aprendizaje* o *entrenamiento*.

2.2.3.1. Aprendizaje

El aprendizaje en un sistema automático de clasificación se realiza siguiendo algunas de las tres aproximaciones metodológicas existentes:

Aprendizaje supervisado. En este caso, se dispone de un conjunto de muestras con etiquetas de clase, las cuales constituyen agrupamientos predefinidos.

Aprendizaje no supervisado. Las muestras no presentan etiquetas de clase, de modo que la estimación de éstas se lleva a cabo siguiendo criterios de proximidad o similitud que, bajo características comunes, agrupan las muestras en una misma clase.

Aprendizaje semi-supervisado. Enfoque que combina las dos aproximaciones anteriores. Parte de un conjunto (generalmente pequeño) de muestras etiquetadas que, progresivamente, puede ir ampliándose mediante el uso de un conjunto (probablemente grande) de nuevos ejemplos sin etiquetar. El reto es usar datos etiquetados y no etiquetados para mejorar la construcción del modelo.

El aprendizaje supervisado, aproximación que se sigue en esta tesis, implica una serie de actividades repetitivas, tales como *la selección del modelo*, *la estimación de parámetros* y *la evaluación del sistema*. En la Figura 2.2, se muestra el subproceso iterativo que ajusta un modelo a un problema, como parte del esquema de un sistema de reconocimiento de formas.

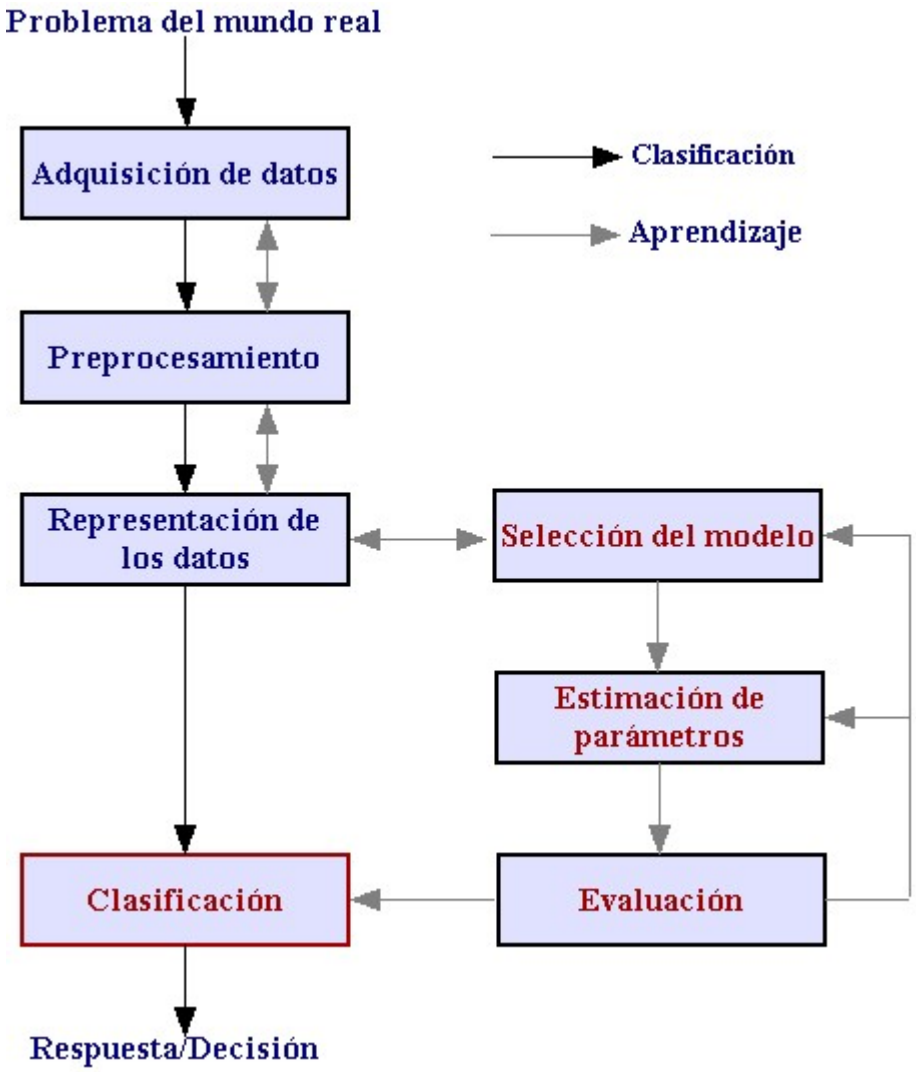


Figura 2.2: *Ciclo de un sistema automático de reconocimiento con aprendizaje supervisado.*

2.2.3.2. Selección del modelo, estimación y evaluación

Dado un problema, existe una gran diversidad de métodos que pueden aplicarse. Habitualmente, se realizan estudios que permiten evaluar la idoneidad de cada método con el fin de seleccionar aquellos que mejor se ajusten a las características del problema.

Se han propuesto diversos criterios, tests y métodos para evaluar y seleccionar los modelos. La aproximación más usual está basada en la estimación del error sobre un conjunto de validación extraído de los datos de entrenamiento.

La estimación de los parámetros de los algoritmos de clasificación se realiza a partir de otro conjunto, distinto al de validación, extraído también de los datos de entrenamiento. Posteriormente, se estima el error empleando el conjunto de vali-

dación. Una vez que el modelo ha sido seleccionado, es posible calcular su error esperado mediante el uso de un conjunto independiente de *test*.

Por otra parte, ocurre que en la vida real sólo se cuenta con un conjunto finito de muestras, lo cual limita la cantidad de datos para entrenamiento, validación y prueba. En este sentido, se han propuesto diversas metodologías para mejorar la utilización de conjuntos de datos pequeños. Las principales alternativas pueden resumirse en los siguientes métodos:

- *Hold-out*. Llamado también *método H*, reserva una cantidad para prueba y otra para entrenamiento. En la práctica, es bastante común emplear una tercera parte para prueba y las dos terceras partes restantes para entrenamiento.
- *Validación cruzada* (*Cross–validation*). Conocido como *método* π o *rotación*, genera aleatoriamente una partición en K bloques de talla N/K. El entrenamiento se lleva a cabo empleando $K-1$ bloques, mientras que el subconjunto restante es empleado como test. Este procedimiento es repetido K veces, eligiendo en cada iteración una parte diferente para prueba. Una extensión a este método es el llamado *stratified cross–validation* con el que, para cada partición, las clases se encuentran distribuidas según sus probabilidades *a priori* en el conjunto original. Por otra parte, para una mejor estimación, el proceso de validación cruzada puede ser repetido P veces.
- *Leave-one-out*. Este método es un caso particular del K fold cross–validation, para $K = N$. Cada conjunto de test está formado por una única muestra, mientras que las $N-1$ restantes constituyen el conjunto de entrenamiento.
- *Bootstrap*. Aleatoriamente se generan, del conjunto de datos original, m conjuntos de cardinalidad N con reemplazo.

Estos mismos métodos son empleados para subdividir la partición de entrenamiento con el objetivo de ajustar o calcular los hiperparámetros de un algoritmo de clasificación.

2.3. Paradigmas del reconocimiento de formas

De acuerdo al tipo de representación utilizada, las aproximaciones al reconocimiento de formas se pueden clasificar en:

1. Aproximación estadística o geométrica.
2. Aproximación sintáctica o estructural.
3. Redes neuronales.
4. Correspondencia entre modelos.

En la Tabla 2.1, se puede observar una breve comparación de estas aproximaciones.

Tabla 2.1: *Paradigmas del reconocimiento de formas.*

Aproximación	Representación	Función de Reconocimiento	Criterio de Evaluación
Estadística o Geométrica	Vector de características	Función discriminante	Error de clasificación
Sintáctica o Estructural	Primitivas	Reglas y gramáticas	Error de aceptación
Redes neuronales artificiales	Muestras, píxeles, características	Función de red	Error cuadrático medio
Correspondencia de modelos	Muestras, píxeles, curvas	Correlación, medidas de distancia	Error de clasificación

2.3.1. Aproximación estadística o geométrica

En la aproximación estadística, cada muestra se representa en términos de un vector de n atributos o características, las cuales pueden verse como un punto en un espacio n-dimensional (generalmente normado métrico). Bajo esta aproximación, un vector de atributos se escribe como $\mathbf{x} = (x_1, \ldots, x_n)^T$, cuyas componentes x_i son los valores resultantes de medir ciertas características de un objeto. Se asume que existen L *clases* o *categorías*, denotadas por $(\omega_1, \ldots, \omega_L)$, y que cada objeto $\mathbf{x}$ está asociado a una de estas clases a través de una variable categórica z de modo que si $z = i$, entonces $\mathbf{x}$ pertenece a la clase ω_i, $i \in \{1, \ldots, L\}$.

El proceso de clasificación puede describirse más formalmente de la siguiente manera. Dado un conjunto de mediciones sobre un nuevo objeto del problema real representadas por un vector $\mathbf{x}$, se desea asignar la muestra $\mathbf{x}$ a una de las L clases posibles, ω_i, $i = 1, \ldots, L$. Una *regla geométrica de decisión* particiona el espacio de representación en L regiones, Ω_i, $i = 1, \ldots, L$. Si una representación $\mathbf{x}$ está en la región Ω_i, entonces se asume que pertenece a la clase ω_i. Las "fronteras" entre las regiones Ω_i reciben el nombre de *fronteras de decisión* o *hipersuperficies de decisión*, las cuales se obtienen por medio de dos posibles estrategias: teoría de decisión y *análisis discriminante*. En la primera de ellas, las fronteras se determinan por las distribuciones de probabilidad de las formas que pertenecen a cada clase. En el segundo caso, se especifica un modo paramétrico (lineal, cuadrático, etc.) de la frontera, sin asumir nada sobre las distribuciones por clase.

2.3.1.1. Teoría de decisión de Bayes

Una de las aproximaciones estadísticas fundamentales en el reconocimiento de formas se refiere a la teoría de decisión basada en la *regla de Bayes*. Esta aproximación establece que el problema de la decisión se puede expresar en términos probabilísticos y que todas las probabilidades relevantes son conocidas.

Formalmente, se asume que cada clase ω_i tiene una probabilidad *a priori* (conocida) $P(\omega_i)$, $i = 1, \ldots, L$ y que $p(\mathbf{x}|\omega_i)$ es la *probabilidad condicional* de una muestra $\mathbf{x}$ dado ω_i. El problema de asignar una muestra $\mathbf{x}$ a una clase ω_i se reduce a calcular las probabilidades *a posteriori* $P(\omega_i|\mathbf{x})$, las cuales pueden ser expresadas en términos de una probabilidad *a priori* y una función de densidad condicional de clase mediante el empleo del teorema de Bayes,

$$P(\omega_i|\mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} \tag{2.1}$$

donde $p(\mathbf{x})$ representa la evidencia,

$$p(\mathbf{x}) = \sum_{i=1}^{L} P(\omega_i)p(\mathbf{x}|\omega_i) \tag{2.2}$$

Por consiguiente, la ecuación 2.1 puede reescribirse de la siguiente forma:

$$P(\omega_i|\mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{\sum_{j=1}^{M} P(\omega_j)p(\mathbf{x}|\omega_j)} \tag{2.3}$$

Finalmente, el clasificador de Bayes elige la clase con la probabilidad *a posteriori* más alta.

El modelo de probabilidad descrito anteriormente es válido para variables con valores *continuos*. Para el caso de valores *discretos*, en lugar de emplear una función de densidad de probabilidad condicional de clases, se emplea una *función de probabilidad condicional de masa de clases* (*pmf*).

2.3.1.2. Regla de mínimo coste/riesgo de Bayes

El riesgo es la posibilidad de cometer un error en el proceso de toma de decisiones y, por ende, de llegar a una conclusión errónea. En muchas de las actividades de la vida real, existe un coste asociado a los errores. Esta teoría ha sido extendida al reconocimiento de formas apoyándose en la teoría bayesiana, donde la resolución final es una combinación de una regla de decisión y el coste que se está dispuesto a aceptar.

El riesgo R de asignar una muestra $\mathbf{x}$ a la clase ω_i, llamado también *riesgo condicional*, se define como

$$R(\omega_i|\mathbf{x}) = \sum_{j=1}^{L} \lambda_{ij} P(\omega_j|\mathbf{x}) \tag{2.4}$$

donde λ_{ij} es el coste de asignar un ejemplo $\mathbf{x}$ a ω_i cuando su verdadera clase es ω_j.

En la práctica, el coste puede ser interpretado en un sentido abstracto aunque, en muchas situaciones, los costes son una combinación de diferentes unidades tales como la monetaria, la temporal, la calidad o la vida.

2.3.2. Aproximación sintáctica o estructural

En esta aproximación, se asume una representación estructurada de los objetos en términos de subobjetos más simples que, a su vez, pudieran estar caracterizados por la interrelación de estructuras aún más simples. Algunos ejemplos son las cadenas de primitivas, los grafos, etc.

El reconocimiento debe basarse entonces, no sólo en identificar los subobjetos del objeto examinado, sino también en analizar la relación entre ellos. Todo reconocedor estructural está compuesto de, al menos, dos etapas: *identificación de subobjetos* e *interpretación de estructura*.

Un ejemplo típico vendría dado por los *métodos gramaticales* o *sintácticos*, en los que cada objeto es representado por un conjunto de símbolos cuya secuencia ordenada proporciona la estructura. Si ésta cumple una serie de reglas especificadas mediante una gramática, significa que pertenece al lenguaje generado por dicha gramática.

2.3.3. Redes neuronales artificiales

Se trata de un enfoque inspirado en las redes neuronales biológicas, aunque posee otras funcionalidades y estructuras de conexión ajenas a esta perspectiva. Las *redes neuronales artificiales* pueden definirse como modelos computacionales, compuestos de unidades procesadoras adaptativas (neuronas artificiales) interconectadas entre sí.

Las modelos de redes neuronales emplean principios de organización tales como aprendizaje, generalización, adaptabilidad, tolerancia a fallos y representación distribuida. Todos ellos convierten el uso de las redes neuronales en una opción especialmente interesante en un gran número de aplicaciones.

El modelo más común de red neuronal artificial es el de las redes multicapa, el cual consta, en su versión más simple, de tres capas: *entrada*, *oculta* y *salida*. Una de las más populares es el *perceptrón multicapa*, la cual puede tener en su organización varias capas ocultas.

Al igual que los clasificadores basados en la aproximación estadística, las redes neuronales artificiales tienen dos etapas: *aprendizaje* y *clasificación*. En la fase de aprendizaje, se usa un conjunto de datos de entrenamiento para determinar los pesos que definen el modelo neuronal. Esta etapa puede incluir un proceso de estimación de parámetros estructurales, por ejemplo, el número de capas ocultas o el número de neuronas en cada capa oculta. Una vez entrenado el modelo, se procede a la fase de prueba, en la que se procesan las muestras de test.

2.3.4. Correspondencia de plantillas

Esta aproximación emplea una operación genérica llamada correspondencia (matching), cuyo objetivo es determinar la similitud entre dos entidades (puntos, curvas o formas). En este método, el ejemplo a ser reconocido se compara con un conjunto de plantillas, teniendo en cuenta que puede haber traslaciones, rotaciones o cambios de escala.

Una medida de similitud comúnmente empleada en esta aproximación es la correlación, la cual puede optimizarse usando el conjunto de entrenamiento.

2.4. Clasificadores

El núcleo de todo sistema de reconocimiento de formas lo constituye el módulo de clasificación, que asigna una muestra $\mathbf{x}$ a una de las L clases existentes. Este módulo suele estar formado por un conjunto de *funciones de decisión* o *funciones discriminantes* $G = \{g_1(\mathbf{x}), \ldots, g_L(\mathbf{x})\}$, que dividen el espacio de representación en L regiones.

En este contexto, un *clasificador* o *regla de clasificación* asigna un vector de características $\mathbf{x}$ a la clase ω_i si

$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad \forall i \neq j, \quad 1 \leq i, j \leq L \tag{2.5}$$

es decir,

$$g_i(\mathbf{x}) = max_j g_j(\mathbf{x}), \quad j = 1, \ldots, L \tag{2.6}$$

Por tanto, un clasificador puede ser representado como muestra la Figura 2.3. Las salidas de las cajas son los valores de las funciones discriminantes y la salida del (máximo) selector es la etiqueta de clase, de acuerdo a la regla de pertenencia máxima.

La naturaleza de las funciones discriminantes se define en base a la aproximación que se haga del problema. Asumiendo que se cuenta con un conjunto de muestras etiquetadas, $T = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, la optimización de un cierto criterio de aprendizaje se lleva a cabo por algunas de las siguientes aproximaciones:

Estadístico-paramétrica. Se asume que las representaciones de los objetos en T siguen ciertas leyes estadísticas, cuyas formas funcionales son conocidas. Las funciones discriminantes se definen en base a estas funciones (densidades) y el aprendizaje consiste en la estimación de los parámetros, adoptando criterios tales como el de máxima verosimilitud.

No estadístico-paramétrica. En esta aproximación, el problema se plantea como uno de decisión geométrica, en el que los parámetros de las funciones se determinan mediante técnicas de optimización, usando como criterio el error empírico de clasificación de la muestra de entrenamiento.

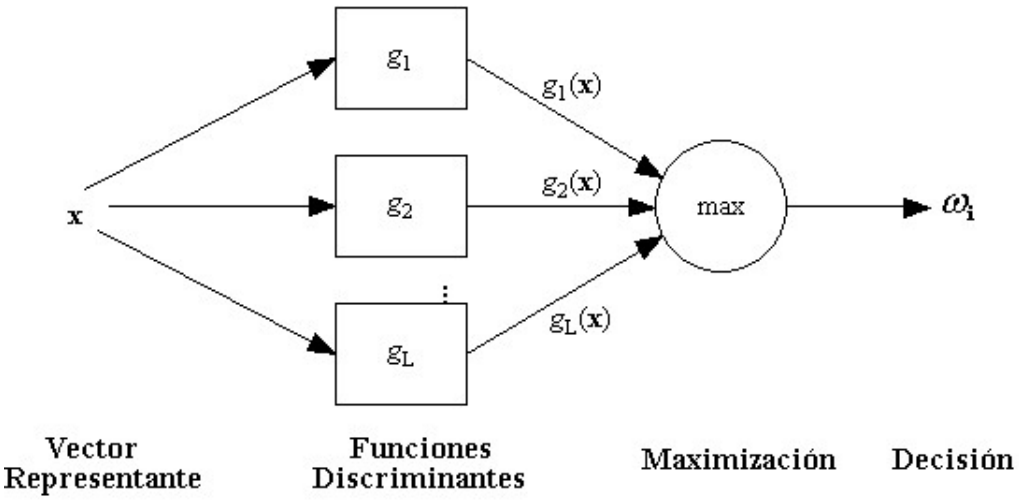


Figura 2.3: *Modelo canónico de un modelo de clasificación.*

Basada en distancias. Es un caso particular no estadístico-paramétrico donde las funciones de decisión se establecen en base a medidas de disimilitud o distancias entre las representaciones de los objetos en T.

En las siguientes secciones, se describen brevemente cada uno de los clasificadores empleados en esta tesis: el naïve Bayes, la regla del vecino más cercano, las máquinas de vectores soporte, el perceptrón multicapa, las redes de funciones de base radial y los árboles de decisión.

2.4.1. Naïve Bayes

Uno de los métodos supervisados que estiman probabilidades *a posteriori* es el algoritmo naïve Bayes. Este clasificador asume, para una muestra $\mathbf{x}$, que sus atributos $x_1, \ldots, x_n$ presentan una independencia condicional dado el valor de la clase, por lo que la probabilidad condicional puede expresarse como el producto de funciones de probabilidad condicional de cada atributo por separado,

$$p(\mathbf{x}|\omega_i) = \prod_{j=1}^{n} p(x_j|\omega_i) \tag{2.7}$$

Usando el teorema de Bayes, la probabilidad *a posteriori* se escribe como,

$$p(\omega_i|\mathbf{x}) = p(\omega_i) \prod_{j=1}^{n} p(x_j|\omega_i) \tag{2.8}$$

Finalmente, el algoritmo naïve Bayes asigna una muestra $\mathbf{x}$ a una de las L clases existentes utilizando la función:

$$\omega^* = argmax_{\omega_j} p(\omega_j) \prod_{i=1}^{n} p(x_i|\omega_j) \tag{2.9}$$

2.4.2. Máquinas de vectores soporte

Las *máquinas de vectores soporte* (*SVM*, *Support Vector Machines*) son sistemas de aprendizaje para problemas de dos clases, pertenecientes al grupo de las funciones discriminantes lineales. Esta técnica, desarrollada a finales de los años 70 y durante la década de los 80, se fundamenta en la *minimización del riesgo estructural* (Cortes y Vapnik, 1995, Vapnik, 1998).

El objetivo de las SVM es encontrar un hiperplano óptimo que divida el espacio de datos en dos regiones. Cada una de estas regiones corresponderá a una de las clases definidas. Para ello, los vectores de entrada son transformados a un nuevo espacio de representación de alta dimensionalidad, en el que las clases se suponen linealmente separables. En este espacio, se construye una función de decisión lineal (hiperplano) con un margen máximo entre los ejemplos de las dos clases, considerando solamente una pequeña cantidad de los datos de entrenamiento llamados *vectores soporte* (Figura 2.4). En términos geométricos, los vectores soporte son aquellos puntos en un espacio n–dimensional que están cercanos a la frontera de decisión y que, por tanto, son los más difíciles de clasificar. La importancia de estos puntos radica en que son los únicos que definen el hiperplano óptimo (Haykin, 1998).

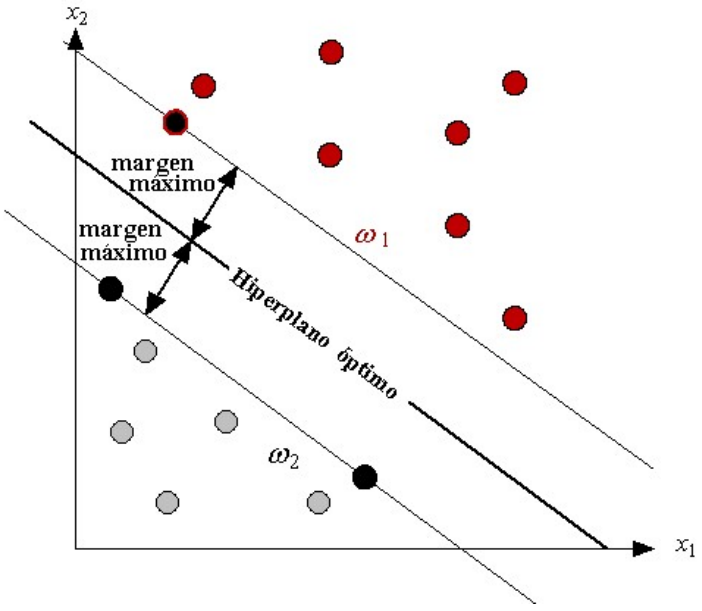


Figura 2.4: *Ejemplo de un problema de dos clases en un espacio bidimensional. Los vectores soporte (puntos en negro) definen el margen máximo de separación entre dos clases.*

Dado un conjunto de entrenamiento de pares muestra-etiqueta $(\mathbf{x}_i, z_i)$, $i =$

$1, \ldots, N$, el mejor hiperplano de separación vendrá dado por la solución a un problema de optimización,

$$\begin{aligned} &min && \tfrac{1}{2}\|\mathbf{w}^2\| \\ &sujeto && z_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{w}_0) \geq +1 \end{aligned} \tag{2.10}$$

donde $\mathbf{w}$ es un vector de pesos ajustables y w_o es un término de generalización. Sin embargo, la anterior ecuación es válida para la situación en la que se asume que los datos son linealmente separables. La adaptación de dicha ecuación a la situación de datos no linealmente separables puede ser reescrita de la siguiente de manera,

$$\begin{aligned} &min && \tfrac{1}{2}\|\mathbf{w}^2\| + C\sum_{i=1}^{N}\xi_i \\ &sujeto && z_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{w}_o) \geq 1 - \xi_i \\ & && \xi_i \geq 0 \end{aligned} \tag{2.11}$$

donde ξ_i son variables de holgura no negativas y C representa el parámetro de penalización que controla la relación entre la maximization del margen y la minimización del error de clasificación.

En general, una máquina de vectores soporte transforma los datos de entrada a otro espacio de características de alta dimensionalidad. Este proceso se lleva a cabo por el empleo de funciones *kernel* que facilitan la transformación de la información de entrada. Estas funciones pueden ser de tipo lineal, polinomial y de base radial (Gaussianas), entre otras.

Las propiedades básicas de las SVM se pueden resumir en los siguientes puntos:

- Inducen un hiperplano lineal en el espacio de atributos, perteneciendo por tanto a la familia de métodos lineales.
- El aprendizaje no lineal con SVM se consigue mediante el uso de las denominadas funciones de kernel, que permiten transformar el espacio de atributos de entrada en un espacio de trabajo de dimensionalidad mucho mayor. Es por esto que las SVM pueden considerarse como un caso particular de los denominados algoritmos basados en kernel. El planteamiento dual de las SVM permite trabajar con estas funciones kernel de manera eficiente sin calcular explícitamente las representaciones de los ejemplos en el espacio de características.
- Entre todos los posibles hiperplanos separadores, las SVM eligen el de margen máximo. Este sesgo inductivo de aprendizaje, que se complementa con la minimización del error en el conjunto de aprendizaje, contribuye a minimizar el riesgo del *sobreentrenamiento* (overfitting).
- El aprendizaje de las SVM representa un problema de optimización con restricciones que se puede resolver usando técnicas de programación cuadrática.

La convexidad garantiza una solución única y las implementaciones actuales permiten una eficiencia razonable para problemas con un elevado número de ejemplos y atributos.

Sin embargo, las SVM también poseen algunas limitaciones:

- Trabajan con datos numéricos y, por tanto, necesitan una codificación especial para emplear datos originalmente nominales.
- Para clasificación no lineal, se necesita proyectar los objetos a un espacio de mayor dimensión. El empleo del producto punto como función kernel requiere de múltiples cálculos. Esta matriz puede tornarse singular al tener muchos atributos poco relevantes o con valores muy pequeños, por lo que la convergencia de este problema se ve fuertemente afectada.
- La utilización de las funciones kernel puede determinar la solución obtenida. En la actualidad, no se cuenta con un consenso acerca de qué función utilizar para cada tipo de problema, dejando a "prueba y error" la elección del tipo de función más adecuada.

2.4.3. Red neuronal perceptrón multicapa

Las *redes neuronales artificiales* (RNA) son un paradigma de aprendizaje y procesamiento automático, cuya base se encuentra en su analogía con el comportamiento y la función del cerebro humano, en particular del sistema nervioso central. Anatómicamente, el sistema está compuesto por redes de *neuronas biológicas* interconectadas, las cuales poseen capacidades de procesamiento y conductividad para producir un estímulo de salida.

En 1943, se propone un modelo abstracto y simple de una neurona artificial como un dispositivo binario (McCulloch y Pitts, 1943). Este modelo posee un umbral de funcionamiento por debajo del cual está inactiva. Asimismo, recibe entradas excitadoras o inhibidoras cuya acción es absoluta: si existe alguna de estas entradas, la neurona permanece activa. El modelo de trabajo es simple: si no existe ninguna entrada inhibidora, se determina la resultante de las entradas excitadoras y, si ésta es mayor que el umbral, la salida es 1, en caso contrario es 0 (vea Figura 2.5).

Teniendo como base los estudios de McCulloch–Pitts, en 1953 se propone una aproximación llamada *perceptrón* (Rosenblatt, 1958). Una de las características de este modelo que más interés despertó fue su capacidad de aprender a reconocer o clasificar objetos. El perceptrón está constituido por un conjunto de sensores de entrada que reciben los ejemplos a reconocer o clasificar, y una neurona de salida que se ocupa de clasificar la muestra en alguna de las dos clases, de acuerdo a un único valor de activación. Sin embargo, dicho modelo era incapaz de converger a buenas soluciones en aquellos problemas con clases no linealmente separables (Minsky y Papert, 1969).

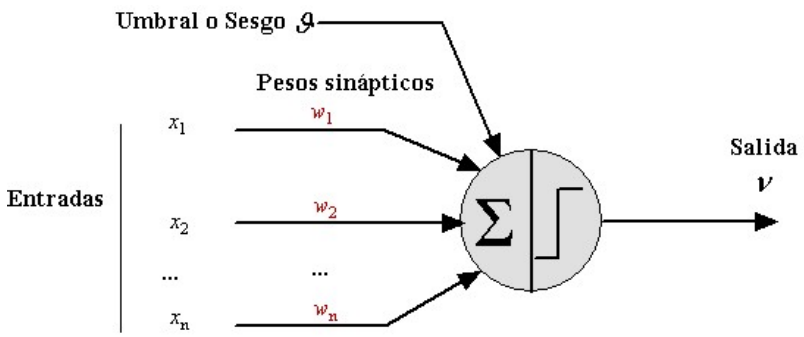


Figura 2.5: *Esquema de una neurona artificial. Modelo compuesto por un vector de pesos $w_1 \dots w_n$, equivalente a las conexiones sinápticas, y por el umbral θ de activación.*

Con la finalidad de solventar las limitaciones presentadas por el perceptrón, surge en 1986 un nuevo modelo llamado *perceptrón multicapa* (*MLP, MultiLayer Perceptron*) (Rumelhart y otros, 1986). La estructura de esta red consta de múltiples capas de neuronas artificiales que le permiten resolver problemas que no son linealmente separables.

La arquitectura más común de una red MLP simple consta de tres capas: entrada, oculta y salida. Cada de estas capas se encuentra totalmente conectada a la siguiente en un sentido unidireccional (ver Figura 2.6). El modelo general admite un número indeterminado de capas ocultas.

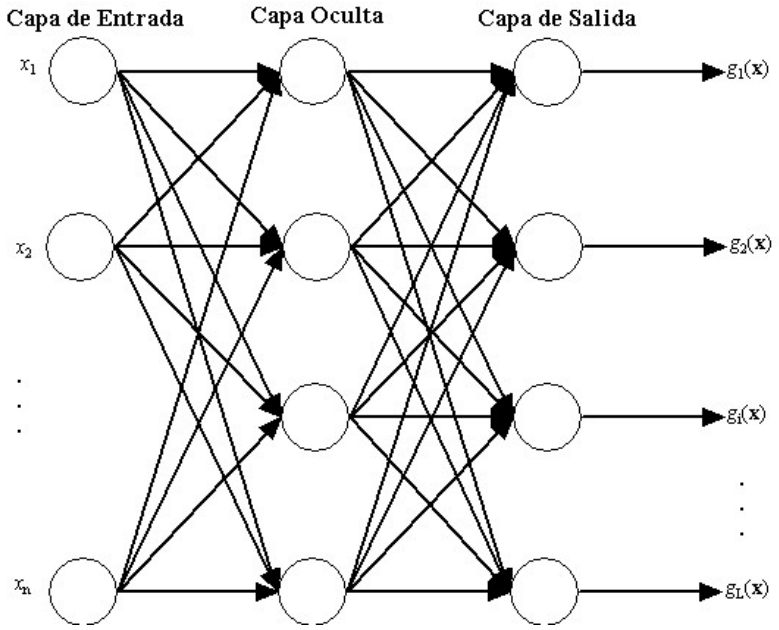


Figura 2.6: *Modelo genérico de una red MLP con una capa oculta.*

El entrenamiento en este tipo de redes se lleva a cabo de manera supervisada

por medio de uno de los algoritmos más populares, conocido como *back–propagation*. Las bases de este algoritmo se encuentran en la regla de aprendizaje de corrección por error (*error-correction learning rule*) (Haykin, 1998).

Básicamente, el algoritmo back–propagation consta de dos fases a través de las diferentes capas de la red:

1. **Fase hacia adelante** (Forward). En esta fase, cada muestra se propaga a través de las diferentes capas de la red hasta generar una salida, la cual es comparada con la salida deseada. Durante esta etapa, los pesos sinápticos se mantienen fijos.

2. **Fase hacia atrás** (Backward). En esta fase, los pesos sinápticos se ajustan de acuerdo a una regla de corrección de error. Específicamente, la *respuesta actual* de la red se substrae de la respuesta deseada para producir una *señal de error*. Esta señal es entonces propagada hacia atrás a través de toda la red, provocando que los pesos sinápticos sean ajustados.

2.4.4. Redes de funciones de base radial

Las *redes de funciones de base radial* (*RBF, Radial Basis Function*), al igual que las MLP, son redes no lineales feedforward. Sin embargo, estas dos redes difieren entre sí en diversos aspectos importantes (Haykin, 1998):

1. Una red RBF (en su estructura más básica) consta de una capa oculta, mientras que una red MLP podría tener más de una.

2. Los modelos de procesamiento en las neuronas ocultas y de salida en una red MLP son iguales. Por el contrario, en una red RBF, los nodos de la capa oculta llevan a cabo un procesamiento diferente.

3. La capa oculta de una red RBF es no lineal, mientras que la capa de salida es lineal. En una red MLP, ambas capas son, por lo general, no lineales (ver Figura 2.7).

4. La función de activación en la capa oculta de la red RBF es una distancia Euclídea entre la muestra de entrada y el centro de un nodo. En el caso de la red MLP, la función de activación calcula el producto interior del ejemplo de entrada y el vector de pesos sinápticos en cada neurona.

5. Las redes MLP construyen una aproximación global, mientras que las redes RBF construyen aproximaciones locales mediante el empleo de funciones kernel (por ejemplo, de tipo Gaussianas).

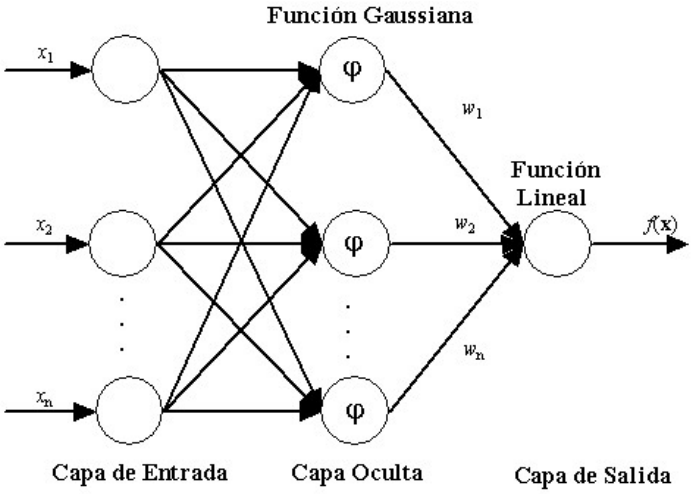


Figura 2.7: *Modelo de una red neuronal RBF.*

2.4.5. Árboles de decisión

Los *árboles de decisión* (*DT, Decision Tree*), también denominados *árboles de clasificación* o de *identificación*, son métodos basados en la estrategia "divide y vencerás". Este tipo de aproximación es un modelo jerárquico perteneciente al grupo de los algoritmos no paramétricos con aprendizaje supervisado. Su estructura consta de nodos de decisión internos que evalúan uno o más atributos, aristas o arcos con los valores de los atributos y hojas terminales que determinan la clasificación (ver Figura 2.8), por lo que también es equivalente a un conjunto de reglas IF–THEN.

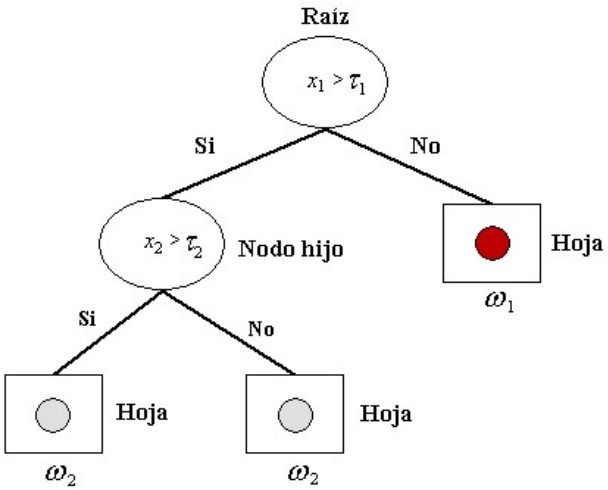


Figura 2.8: *Ejemplo de un árbol de decisión para un problema de dos clases.*

Al igual que cualquier sistema de aprendizaje supervisado, los árboles de clasificación presentan dos etapas de funcionamiento: aprendizaje y clasificación. En la

primera de ellas, se lleva a cabo la construcción del árbol de forma recursiva a partir de un conjunto de entrenamiento, mientras que el objetivo de la segunda etapa es etiquetar una muestra de entrada. La clasificación empieza desde el nodo raíz del árbol evaluando el atributo especificado por ese nodo, y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja, en cuyo caso se devuelve la etiqueta de clase asociada con este nodo.

El proceso de construcción de un DT constituye la fase más compleja y la que determina el resultado final. Este proceso se basa en tres elementos importantes: el *criterio de partición*, la *estrategia de parada* y la *poda*.

El procedimiento para particionar un nodo se basa en cuantificar la pureza del mismo. Algunas de las medidas más conocidas son el coeficiente Gini y la entropía.

Los criterios para detener la construcción de un árbol pueden estar basados en la *pureza del nodo* mediante el empleo de un umbral, o en una *cota de profundidad*, la cual se puede establecer *a priori*. Estos criterios también son denominados reglas de *pre–poda*, al reducir la complejidad del árbol durante su construcción, en contraste con las reglas usuales de post-poda que simplifican el árbol de decisión una vez que éste ha sido construido por completo.

La idea básica de poda asume que los árboles podados (simples) producen mejores resultados de clasificación que los árboles más grandes (complejos), es decir, los árboles podados tendrán más capacidad de generalización.

2.4.6. Clasificadores basados en vecindad

El método de los k-vecinos más cercanos ($k-$NN, *k Nearest Neighbours*) (Fix y Hodges, 1951) forma parte de una familia de técnicas de aprendizaje conocida como aprendizaje basado en ejemplos (instance-based learning). El aprendizaje de estos algoritmos se limita a almacenar en memoria los ejemplos de entrenamiento presentados. La idea básica sobre la que se fundamenta este paradigma establece que los miembros de una población suelen compartir propiedades similares y ciertas características con los individuos que la rodea. Por tanto, la clasificación de una nueva muestra se realiza tomando en cuenta las instancias más próximas del conjunto de entrenamiento de acuerdo a una cierta medida de disimilitud.

Debido al uso de una métrica de distancia, estos algoritmos son conocidos como clasificadores de distancia mínima. Formalmente, sea $T = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ el conjunto de entrenamiento de muestras etiquetadas formado por N instancias, tal que cada $\mathbf{x}_i$ tiene asociada una etiqueta de clase ω_i del conjunto de L clases. Una clasificación por distancia mínima para un nuevo prototipo $\mathbf{y}$ puede definirse como:

$$\mathbf{y} \in \omega_j \quad \Leftrightarrow \quad j \quad = \quad argmin_i d(\mathbf{y}, \mathbf{x}_i) \tag{2.12}$$

donde d es la (pseudo)-métrica definida en T, la cual debe cumplir los cuatro criterios siguientes:

1. No negatividad: $d(\mathbf{y}, \mathbf{x}) \geq 0$.

2. Identidad: $d(\mathbf{y}, \mathbf{x}) = 0 \Leftrightarrow \mathbf{y} = \mathbf{x}$.

3. Simetría: $d(\mathbf{y}, \mathbf{x}) = d(\mathbf{x}, \mathbf{y})$.

4. Desigualdad triangular: $d(\mathbf{y}, \mathbf{x}) \geq d(\mathbf{y}, \mathbf{p}) + d(\mathbf{p}, \mathbf{x})$.

Una medida de similitud muy popular para datos numéricos es la distancia Euclídea,

$$d(\mathbf{y}, \mathbf{x}) = \sqrt[2]{\sum_{j=1}^{n} (y_j - x_j)^2} \tag{2.13}$$

cuyo valor escalar se obtiene a partir del valor de los atributos de las dos instancias.

El método de los k–vecinos más cercanos está considerado como el más básico de los métodos de aprendizaje basados en ejemplos. Si tomamos $k = 1$, una muestra $\mathbf{x}$ se clasificará en la clase asociada a su instancia más cercana. Para valores de $k \geq 2$, la muestra $\mathbf{x}$ se asignará a la clase más representada entre las k instancias más próximas a la muestra (ver Figura 2.9).

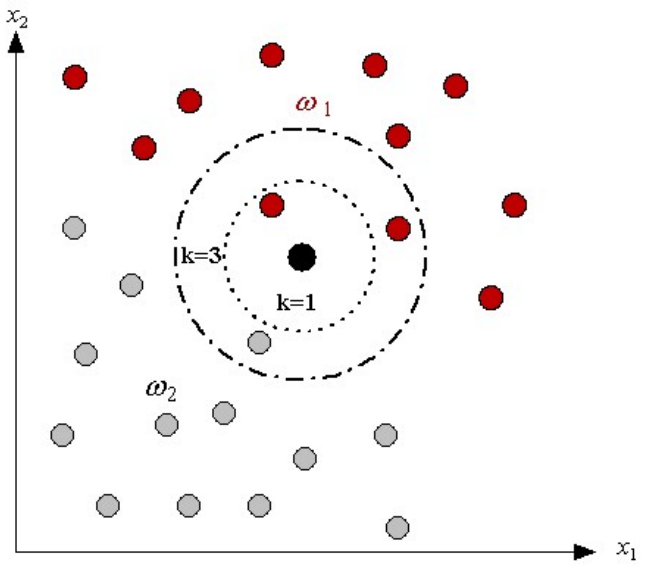


Figura 2.9: *Clasificador basado en la regla de los k vecinos más cercanos.*

2.5. Complejidad de los datos

El éxito de un sistema de reconocimiento estadístico de formas presenta una fuerte dependencia de la complejidad y la calidad del conjunto de datos de entrenamiento. Frecuentemente, este conjunto puede presentar una serie de dificultades, tales como

la presencia de muestras con ruido, el solapamiento entre clases, una cantidad pequeña de instancias y alta dimensionalidad, haciendo que el proceso de aprendizaje se vuelva más complejo.

2.5.1. Ruido

El ruido es una anomalía presente en el conjunto de datos, que puede estar originado por algunos de las siguientes eventos:

1. Imprecisiones durante la colección de datos.
2. Errores en el etiquetado de las muestras.
3. Presencia de atributos adicionales que podrían no ser observables (componentes aleatorios).

Las soluciones a este problema van dirigidas a filtrar (eliminar) instancias de entrenamiento con ruido. Esta estrategia de filtrado consiste en detectar y descartar aquellas muestras que podrían considerarse poco confiables.

2.5.2. Solapamiento

El solapamiento entre clases se puede definir como la región donde, por lo menos, dos clases diferentes, ω_i y ω_j, presentan instancias que comparten información común en algunos de sus atributos. Debido a esto, las muestras en dicha región se consideran poco confiables porque cualquiera de ellas pude pertenecer legítimamente tanto a ω_i como a ω_j.

2.5.3. Talla de los datos

La complejidad en muchos algoritmos de aprendizaje depende del número de atributos, así como del número de instancias. A mayor talla del conjunto de datos, mayores cantidades de recursos requerirá el algoritmo.

La reducción de la talla del conjunto de datos se puede llevar a cabo seleccionando un subconjunto de atributos o ejemplos del conjunto original. Este proceso conlleva una serie de ventajas que se pueden resumir en los siguientes puntos:

- Una reducción en la complejidad de inferencia del algoritmo.
- Cuando un prototipo es innecesario, se reduce el coste de extraerlo.
- Modelos más simples son más robustos sobre pequeños conjuntos de datos.
- Datos con pocos atributos permiten analizar con facilidad el proceso que ocurre en los mismos. Adicionalmente, éstos pueden ser analizados y trazados de forma visual.

2.5.4. Desbalance entre clases

Generalmente, las diferentes aproximaciones al reconocimiento de formas asumen que las clases en el conjunto de datos presentan similares probabilidades *a priori*. Sin embargo, existen aplicaciones reales en las que esta hipótesis no se cumple, dando lugar a ratios entre clases fuertemente sesgados. Esta situación se conoce como el problema del desbalance entre clases.

Para un problema de dos clases, se dice que un conjunto de datos está desbalanceado cuando el número de muestras de una de las clases (*la mayoritaria*) es significativamente muy superior al de la otra clase (*la minoritaria*). Paradójicamente, la clase menos representada suele ser la de mayor importancia para el problema. Consecuentemente, los errores cometidos sobre dicha clase minoritaria suelen tener un coste medio mucho más elevado.

El tema del desbalance entre las clases de un problema constituye precisamente la base de la presente tesis. Por ello, en el próximo capítulo, se abordarán en detalle las principales líneas de investigación que podemos encontrar en la literatura relativa al problema del desbalance.

Capítulo 3

Conjuntos de datos desbalanceados

En este capítulo, se aborda el problema de las distribuciones de clase desbalanceadas en el área del reconocimiento de formas, principalmente en el ámbito del aprendizaje supervisado. En particular, se describen las actuales líneas de investigación y las aportaciones más relevantes en este campo.

3.1. Introducción

El conjunto de datos es un elemento esencial en la etapa de aprendizaje de un sistema de clasificación, detección o localización, el cual representa los objetos del mundo real ya sea de forma vectorial (aproximación estadística) o mediante estructuras (aproximación sintáctica).

A partir del conjunto de datos, es posible construir un clasificador capaz de reconocer y estimar la etiqueta de clase de un nuevo objeto. Idealmente, el conjunto debería estar formado por todos los posibles estados que pueden adoptar los objetos a reconocer, de modo que el clasificador pudiera aprender a predecir correctamente la etiqueta de clase de cada ejemplo. En este sentido, Dietterich (1990) establece que existe una relación entre el número de muestras del conjunto de datos y la capacidad de generalización del clasificador. En general, cualquier incremento en el número de muestras produce una mejora en la exactitud del clasificador. Adicionalmente, algunos investigadores mencionan que si el conjunto de datos presenta un número de muestras por clase aproximadamente igual, el clasificador puede producir predicciones más acertadas (Weiss, 2003). Sin embargo, en muchas aplicaciones reales, no se cumple la hipótesis de contar con un conjunto de datos uniformemente distribuidos entre las clases. Esta particularidad es conocida como el *problema de las clases desbalanceadas o sesgadas*.

El desbalance en un conjunto de datos de dos clases se refiere a la situación en la que una clase (*mayoritaria*) presenta una cantidad notablemente mayor de ejemplos en comparación con la otra clase (*minoritaria*). Usualmente, las clases mayoritarias y minoritarias representan, respectivamente, la ausencia y presencia de situaciones pocos comunes. Algunos ejemplos de este tipo de casos son:

- Detección de fraudes en llamadas telefónicas (Fawcett y Provost, 1997) y de tarjetas de crédito (Chan y Stolfo, 1998).

- Detección de manchas de petróleo en el océano en imágenes de satélites (Kubat y otros, 1998).
- Diagnóstico de fallos en equipos de telecomunicaciones (Weiss y Hrish, 1998, Radivojac y otros, 2003).
- Valoración de créditos bancarios (Huang y otros, 2006).
- Clasificación automática de textos (del Castillo y Serrano, 2004, Tan, 2005, Liu y otros, 2009).
- Filtrado y extracción de información (Lewis y Catlett, 1994).
- Diagnóstico médico de enfermedades raras (Cohen y otros, 2006, Meng y otros, 2008, Sousa y otros, 2008) y de eventos de alto riesgo poco comunes (Taft y otros, 2009).
- Clasificación de páginas web y búsqueda de referencias (Zhang y otros, 2008).

En el problema del desbalance, la clase minoritaria suele representar los casos de mayor interés, es decir, aquellos cuya detección justifican el problema de clasificación (Weiss, 2004). En consecuencia, esta clase se conoce también por el nombre de clase *positiva*, mientras que la clase mayoritaria, que agrupa el resto de casos corrientes, se denomina clase *negativa*.

La presencia del desbalance en los conjuntos de datos, por lo general, ocasiona en los clasificadores tradicionales un aprendizaje sesgado en favor de la clase mayoritaria, lo cual suele producir un bajo rendimiento en la clasificación de la clase minoritaria. Este comportamiento se debe, fundamentalmente, a que la mayoría de los clasificadores, en su etapa de aprendizaje, asumen una distribución balanceada de las clases (Japkowicz y Stephen, 2002). Asimismo, el escaso número de ejemplos positivos que cubren el espacio de representación dificulta el proceso de generalización del clasificador sobre la clase positiva (Weiss, 2004). En la literatura, es posible encontrar diversos estudios de los efectos del desbalance sobre diferentes clasificadores, como la regla del vecino más cercano (Hand y Vinciotti, 2003, Zhang y Mani, 2003), las máquinas de vectores soporte (Wu y Chang, 2003), las redes neuronales en sus distintas aproximaciones (Al-Haddad y otros, 2000, Huang y otros, 2006, Murphey y otros, 2004), los árboles de decisión (Chan y Stolfo, 1998, Estabrooks y otros, 2004, Japkowicz y Stephen, 2002), entre otros (Orriols-Puig y Bernadó-Mansilla, 2005, Xie y Qiu, 2007).

En resumen, los trabajos anteriores muestran que la baja tasa de reconocimiento de los clasificadores tradicionales sobre la clase positiva se debe, principalmente, a los siguientes factores:

1. **Optimización de un índice global.** La mayoría de los algoritmos de clasificación tienen por objetivo maximizar la tasa de aciertos (minimizando la tasa de error global), a la cual la clase positiva contribuye muy poco.

2. **Distribución de las clases.** Los clasificadores convencionales asumen que los datos del conjunto de entrenamiento se distribuyen uniformemente entre las clases.

3. **Costes de clasificación.** Los clasificadores asumen que los errores sobre las diferentes clases tienen los mismos costes.

El problema del sesgo de los datos y sus efectos en diversas aplicaciones de aprendizaje automático y minería de datos es un tema de estudio de gran importancia que se puede encontrar en diversos talleres internacionales (Chawla y otros, 2003a, Dietterich y otros, 2000, Japkowicz, 2000), libros (Chawla, 2005), ponencias en congresos (Gu y otros, 2008) y números especiales de revistas (Chawla y otros, 2004).

3.2. Naturaleza de los conjuntos desbalanceados

La teoría más general sobre conjuntos desbalanceados sugiere que el problema es ocasionado por la diferencia numérica existente entre las tallas de las clases. Sin embargo, otros investigadores afirman que es importante también identificar el origen y, por consiguiente, los tipos de desbalance existentes. En este sentido, Visa y Ralescu (2005) establecen que, considerando la aplicación u origen del problema, el desbalance puede ser de dos tipos:

1. Los datos están naturalmente desbalanceados.

2. Los datos no están naturalmente desbalanceados, pero implica un alto coste el obtener información de la clase positiva.

Dentro del primer punto, se encuentran aplicaciones como la detección de fraudes en tarjetas de crédito o el diagnóstico médico de enfermedades poco comunes. En el segundo caso, se enmarcan aquellas aplicaciones que tienen por objetivo detectar fallos en sistemas electrónicos o de transporte. Algunos ejemplos de esto son los equipos de telecomunicaciones o de sistemas de control aéreo.

Por otra parte, Japkowicz (2001) expone que, tomando en cuenta la distribución de los datos en el espacio de representación, es posible identificar dos tipos de sesgo en los datos:

1. Desbalance entre clases (*between–class imbalance*).

2. Desbalance dentro de una clase (*within–class imbalance*).

Si el número de ejemplos de la clase positiva difiere significativamente del número de ejemplos de la clase negativa, entonces se dice que existe un desbalance entre clases (“clases raras”). Por el contrario, si una clase está fragmentada en diferentes grupos o clusters y éstos, a su vez, no presentan una cantidad similar de ejemplos, el desbalance ocurre dentro de una clase (“casos raros”). Cabe mencionar que esta

última complejidad se puede presentar en las clases mayoritarias. Sin embargo, es más probable que las clases minoritarias tengan una mayor proporción de casos raros.

Cuando el conjunto de datos presenta un desbalance entre clases, los métodos tradicionales tienden a ignorar las clases minoritarias debido a que fueron diseñados para optimizar una medida global de exactitud, la cual no tiene en consideración las distribuciones por clase. Por lo que respecta al desbalance dentro de una clase, en clasificadores simbólicos cuyo aprendizaje de un "concepto" se puede expresar como un conjuntos de reglas que pueden ser vistas como una disyunción, formadas a su vez por definiciones conjuntivas que describen un subconcepto del original, los casos raros tienden a ocasionar pequeños disyuntos (*small disjuncts*) (Holte y otros, 1989), los cuales clasifican correctamente un pequeño número de muestras.

La relación que ambos tipos de desbalance guardan y sus efectos en los algoritmos de aprendizaje han sido analizados por Jo y Japkowicz (2004). Los resultados obtenidos les permiten concluir que la degradación de los clasificadores no está únicamente ocasionada por el desbalance entre clases, sino también por la presencia de pequeños disyuntos.

A pesar de que dichas complejidades son diferentes, Weiss (2004) establece que ambos casos son muy parecidos, por lo que pueden ser tratadas empleando soluciones similares (Prati y otros, 2004b).

3.3. Las líneas de investigación

La mayoría de las investigaciones dentro del problema del desbalance se centran en proponer técnicas que permitan tratar dicho problema desde el punto de vista del clasificador. Sin embargo, diversos trabajos recientes exploran otras dos líneas de investigación. La primera de ellas está dirigida a la *definición de medidas alternativas de evaluación* de la efectividad del clasificador, mientras que la segunda línea estudia la *relación e influencia existentes entre el desbalance y otras complejidades* presentes en los conjuntos de datos.

En las siguientes secciones, se describirán los avances más importantes sobre cada una de estas tres líneas de investigación.

3.3.1. Soluciones para tratar el desbalance

Las técnicas para tratar el desbalance pueden ser catalogadas en dos tipos:

1. Soluciones a nivel de datos.

2. Soluciones a nivel de algoritmos.

En el primer grupo, se encuentran aquellas técnicas cuyo objetivo es equilibrar o balancear las tallas de las clases del conjunto de entrenamiento a través de métodos de remuestreo (*resampling*). Por el contrario, aquellas soluciones que adaptan los

algoritmos de entrenamiento al problema del desbalance caen dentro de la segunda categoría.

Finalmente, es posible encontrar propuestas que emplean soluciones híbridas en un intento de aprovechar las ventajas que ofrece cada grupo.

3.3.1.1. Técnicas de remuestreo

Las soluciones a nivel de datos son métodos que balancean de forma artificial las distribuciones por clase del conjunto de datos original, ya sea mediante *sobremuestreo* (*over–sampling*) de la clase minoritaria o por *submuestreo* (*under–sampling*) de la clase mayoritaria. En ambos casos, las técnicas logran que las clases queden representadas por un número igual (o similar) de muestras.

A pesar de que estás técnicas han demostrado su efectividad al tratar el desbalance, diversos investigadores coinciden al señalar dos posibles inconvenientes:

1. Reducir la clase negativa podría eliminar ejemplos potencialmente útiles.
2. Aumentar la talla de la clase positiva conlleva un incremento de los costes temporal y espacial.

Para ambos casos, la principal crítica radica en que estas técnicas alteran la distribución original de los datos.

3.3.1.1.1. Sobremuestreo. El método más sencillo de *sobremuestreo* es el *aleatorio*, cuyo incremento en la talla de la clase mayoritaria se logra a través de la réplica de instancias seleccionadas al azar dentro de una misma clase (Japkowicz y Stephen, 2002, Ling y Li, 1998). A pesar de ser un algoritmo sencillo y fácil de implementar, la creación de copias exactas de instancias positivas puede ocasionar una alta probabilidad de sobre-ajuste por parte del clasificador.

Con la finalidad de solucionar este inconveniente, diversos investigadores proponen algoritmos *dirigidos* a incrementar la talla de la clase minoritaria de forma "inteligente".

SMOTE. Un método muy popular de sobremuestreo es el algoritmo *SMOTE* (*Synthetic Minority Over-sampling TEchnique*) (Chawla y otros, 2002). Este algoritmo, que responde a una estrategia dirigida, incrementa la talla de la clase minoritaria mediante la incorporación de muestras artificiales positivas generadas por *interpolación*.

La creación de nuevas muestras artificiales requiere, como primer paso, localizar para cada muestra positiva sus k vecinos más cercanos que pertenezcan a su misma clase. Posteriormente, se determina el número de nuevas muestras artificiales que se generarán a partir de cada muestra minoritaria original. Finalmente, la creación de una muestra requiere calcular la diferencia entre el vector de atributos bajo consideración y un vecino más cercano de los k elegidos al azar. El resultado de dicha diferencia se multiplica por un valor aleatorio entre 0 y 1.

El algoritmo SMOTE puede ser escrito como sigue:

Algoritmo 1 SMOTE

Entrada:
$P = (p_1, \ldots, p_{min})$ // Muestras de la clase minoritaria
min // Número de muestras minoritarias
n_s // Número de ejemplos sintéticos a generar por cada muestra minoritaria
k // Número de vecinos más cercanos
Salida:
$Synthetic$ // Conjunto de muestras artificiales
Para $i = 1$ to min **hacer**
Buscar los k vecinos más cercanos de p_i
Para $j = 1$ to n_s **hacer**
Elegir aleatoriamente uno de los k vecinos de p_i, $n(p_i)$
$diff = p_i - n(p_i)$
$gap =$ numAleatorio $(0, 1)$
$NewSample = p_i + gap * diff$
$Synthetic \longleftarrow NewSample$
Fin Para
Fin Para

Al crear los prototipos por interpolación, se tiende a generar fronteras de decisión más generales para la clase minoritaria y, por tanto, el clasificador podría ser capaz de crear reglas de decisión menos específicas.

Borderline–SMOTE. Han y otros (2005) proponen el algoritmo *Borderline–SMOTE* como una extensión a la técnica desarrollada por Chawla y otros (2002).

Teniendo como hipótesis que las muestras que se encuentran más alejadas de la frontera de decisión contribuyen poco a la clasificación, el algoritmo Borderline–SMOTE genera nuevos prototipos a partir de ejemplos positivos cercanos a dicha frontera. Además, estos prototipos que se encuentran en el límite son más susceptibles de ser mal clasificados.

En esencia, la propuesta de Han y otros (2005) conserva todas las características propias del SMOTE, con la diferencia de un paso previo que permite identificar el grupo de muestras positivas próximas a la frontera de decisión. Los resultados muestran que, generalmente, esta propuesta presenta mejores tasas de aciertos en la clase positiva y mayores valores de f_1 que SMOTE, teniendo como clasificador el algoritmo C4.5

LLE–SMOTE En el SMOTE, el proceso de creación de muestras sintéticas se lleva a cabo en el espacio de representación original. Wang y otros (2006) proponen generar las nuevas muestras en un espacio de representación linealmente separable y con baja dimensionalidad. Para ello, emplean un algoritmo llamado LLE (*Locally Linear Embedding*), el cual proyecta un conjunto de datos de alta dimensionalidad

a un espacio de representación de baja dimensionalidad. Es, precisamente, en este nuevo espacio transformado donde se aplica posteriormente el algoritmo SMOTE. Finalmente, los nuevos datos sintéticos de la clase minoritaria son transformados a su espacio de representación original.

Los resultados empleando los clasificadores naïve Bayes, k-NN y SVM muestran que se logra un incremento en los valores del área bajo la curva ROC.

Sobremuestreo basado en agrupamiento. Diversos investigadores proponen emplear técnicas de *agrupamiento* como herramientas útiles para tratar el problema del sesgo entre clases. La aplicación de estos algoritmos pueden ser de dos tipos:

1. Como herramientas que se complementan con algoritmos de sobremuestreo, para determinar la estructura interna del conjunto de datos.

2. Directamente como algoritmos de sobremuestreo.

Del primer punto podemos citar el trabajo de Japkowicz (2001), en el que se propone emplear el algoritmo de clustering k-medias sobre cada clase por separado. Los clusters o grupos resultantes son sobremuestreados aleatoriamente hasta lograr un equilibrio entre clusters. Este algoritmo trabaja bajo la suposición de que el desbalance presente en el conjunto de datos es del tipo dentro de una clase (ver Sección 3.2). En la misma línea, Jo y Japkowicz (2004) muestran que dicha técnica puede emplearse en conjuntos de datos que presentan pequeños disyuntos o un desbalance entre clases. Los resultados que obtienen sugieren que esta técnica es idónea cuando el número de muestras en el conjunto de datos es relativamente pequeño.

Un trabajo similar a los mencionados anteriormente es el presentado por Nickerson y otros (2001), quienes emplean el algoritmo de agrupamiento PDDP (*Principal Direction Divisive Partitioning*) como un método que permite determinar la estructura interna de una clase por la formación de clusters, los cuales son remuestreados hasta que cada uno presente el mismo número de muestras.

La segunda estrategia explora una nueva forma de generar datos sintéticos a través de los algoritmo de clustering, donde los centroides aportan la información necesaria para la generación de los nuevos datos. Dentro de esta propuesta, podemos citar el trabajo de Cohen y otros (2006), cuyos centroides se obtienen por el algoritmo de agrupamiento aglomerativo jerárquico, los cuales son incorporados a la clase positiva.

3.3.1.1.2. Submuestreo. Las técnicas de *submuestreo* intentan equilibrar la distribución del conjunto de datos original mediante la eliminación de muestras de la clase negativa. Una de las primeras propuestas en este sentido corresponde al *submuestreo aleatorio* (Jo y Japkowicz, 2004), el cual consiste en eliminar muestras seleccionadas al azar. Sin embargo, este método presenta la desventaja de descartar posibles ejemplos potencialmente útiles para el clasificador.

Con la finalidad de solventar esta deficiencia, surgen los algoritmos de *submuestreo dirigido* que, apoyados en aproximaciones heurísticas, buscan de manera inteligente seleccionar (eliminar) los casos menos útiles.

One–sided selection. El algoritmo OSS (*One–Sided Selection*) (Kubat y Matwin, 1997) es una de las primeras propuestas de submuestreo dirigido. Este algoritmo reduce la clase mayoritaria al descartar del conjunto original tres tipos de muestras mayoritarias:

1. Aquellas que presentan una etiqueta de clase incorrecta (ruido).
2. Aquellas que están cerca de la frontera de decisión.
3. Aquellas que corresponden a casos redundantes.

El resultado final es un subconjunto de entrenamiento con todas las muestras positivas y una clase mayoritaria (posiblemente) reducida. Dicha reducción se lleva a cabo en dos etapas, valiéndose de conocidos algoritmos heurísticos de filtrado (edición) y condensado. En la primera, se obtiene un subconjunto consistente a partir del conjunto original. Para este fin, se emplea una variante del algoritmo de condensado propuesto por Hart (1968). En la segunda etapa, el algoritmo de los links de Tomek (1976b) elimina aquellas muestras que se encuentran cerca de la frontera de decisión o con etiqueta ruidosa.

El algoritmo 2 resume el procedimiento antes descrito.

Algoritmo 2 One–sided selection

Entrada:

T // Conjunto de entrenamiento original

Salida:

S // Conjunto de entrenamiento submuestreado

Inicialmente, C contiene todas las muestras positivas de S y una muestra negativa seleccionada aleatoriamente

Clasificar T por la regla 1-NN usando los ejemplos de C

Mover todos los ejemplos mal etiquetados a C

Eliminar de C todos las muestras negativas por links de Tomek y almacenar el conjunto resultante en S

Los resultados de aplicar el algoritmo OSS sobre conjuntos desbalanceados muestran el beneficio de este método en términos de media geométrica. Sin embargo, es bien conocido que el algoritmo de Hart presenta múltiples desventajas (Barandela y otros, 2005a, Tomek, 1976b):

- No garantiza la reducción.
- La calidad y talla del subconjunto condensado dependen del orden en que son presentadas las muestras al algoritmo.

- La inicialización aleatoria puede ocasionar que no todas las muestras que están cerca de la frontera de decisión pertenezcan al subconjunto reducido. Asimismo, dicho conjunto pudiera contener muestras redundantes e innecesarias.

Por otra parte, cabe también comentar que Toussaint (1994) demuestra, mediante un sencillo contraejemplo, que el algoritmo de Tomek, a pesar de mantener muestras que están cerca de la frontera, podría derivar en un subconjunto no consistente.

Como una alternativa al algoritmo OSS original, Batista y otros (2000) extienden la aplicación de dicho método a conjuntos de datos con valores simbólicos. Tal propuesta emplea una métrica de distancia llamada *diferencia de valores* (*VDM*, *Value Difference Metric*). Esta métrica calcula la distancia entre dos valores a y b para un determinado atributo x_i, dando como resultado una matriz de distancias para cada atributo, con el número de filas igual al número de clases y el número de columnas igual al número de todos los posibles valores del atributo. La distancia VDM es 0 si los valores a y b del atributo x_i son idénticos.

Neighborhood cleaning rule. A partir del principio del algoritmo OSS, surgen otras alternativas de submuestreo. Una de éstas es el algoritmo denominado regla de limpieza por vecindad (*NCL, Neighborhood Cleaning Rule*) (Laurikkala, 2002). Entre las características principales de este algoritmo se encuentran las siguientes:

1. Trabaja sobre conjuntos de más de dos clases.
2. Utiliza la edición de (Wilson, 1972) como método de reducción.
3. Incorpora un parámetro que controla la cantidad de muestras a eliminar.
4. Trabaja con mezcla de datos por el uso de la métrica diferencia de valores heterogéneos (*HVDM*, *Heterogenous Value Difference Metric*).

El uso de la edición de Wilson evita las desventajas que presenta el algoritmo de Hart sobre el orden de los datos y la sensibilidad al ruido.

Submuestreo por algoritmos genéticos. El *algoritmo genético* es una técnica adaptativa de búsqueda y optimización basada en la teoría de la evolución de Darwin.

La capacidad de los algoritmos genéticos para aprovechar la información acumulada en un espacio de búsqueda, inicialmente desconocido, hace atractiva su aplicación en diversas áreas del reconocimiento de formas. Un ejemplo es la propuesta de Kuncheva y Jain (1999) de editar y seleccionar atributos simultáneamente mediante algoritmos genéticos.

Tomando como base las propuestas de Wilson y de Kuncheva y Jain, Barandela y otros (2005b) proponen tratar el problema del desbalance por la eliminación de muestras de la clase mayoritaria y la selección de características de ambas clases

con el uso de algoritmos genéticos. Adicionalmente, sustituyen en la fórmula de aptitud, la métrica de exactitud global por la media geométrica. En esta misma línea, se encuentran otros trabajos como el de García y Herrera (2009), los cuales definen una nueva función de aptitud que incorpora la media geométrica y el ratio de desbalance.

Submuestreo por distancia. Zhang y Mani (2003) presentan cuatro algoritmos de submuestreo, cuyo método selecciona un pequeño subconjunto de la clase mayoritaria basado en el cálculo de distancias y su vecindad con las muestras positivas:

1. *Nearmiss–1*: selecciona muestras de la clase mayoritaria que estén cerca de alguna de las muestras de la clase minoritaria, siempre y cuando su distancia promedio a los tres vecinos más cercanos de clase minoritaria sea la más pequeña.

2. *Nearmiss–2*: al contrario del método anterior, aquí las muestras son seleccionadas mientras su distancia promedio a los tres vecinos más lejanos de la clase positiva sea la más pequeña.

3. *Nearmiss–3*: selecciona un determinado número de muestras mayoritarias que están cerca de alguna muestra positiva.

4. *Most–distant*: este algoritmo es similar a Nearmiss–2, con la diferencia de seleccionar las muestras cuya distancia promedio es mayor a la distancia a los tres vecinos más cercanos.

Resultados experimentales muestran que el Nearmiss–2 y el submuestreo aleatorio producen un incremento en la tasa de aciertos de la clase positiva.

Submuestreo por clustering. Al igual que en el caso del sobremuestreo, las técnicas de agrupamiento se emplean como herramientas útiles y complementarias para submuestrear la clase mayoritaria.

El uso de los algoritmos de clustering se puede dividir en dos estrategias. En la primera de ellas, se aplica un algoritmo de clustering sobre la clase mayoritaria y los prototipos resultantes (centroides) reemplazan a todas las muestras mayoritarias originales (Cohen y otros, 2006).

La segunda estrategia determina la estructura interna del conjunto de datos al encontrar agrupaciones de datos teniendo en cuenta todas las clases para, posteriormente, submuestrear aleatoriamente cada uno de los grupos formados (Yen y otros, 2006) hasta balancear las distribuciones. Finalmente, el nuevo conjunto de datos balanceado se forma sin tener en cuenta los cluster, es decir, el conjunto de datos vuelve a su estructura original.

3.3.1.1.3. Otras alternativas. Las técnicas de remuestreo logran el equilibrio entre clases por la adición o eliminación de muestras positivas o negativas, respectivamente. En ambos casos, todas las muestras presentan una etiqueta de clase.

Una tercera alternativa para balancear las clases por remuestreo es el *aprendizaje activo* (*active learning*) (Provost, 2003). Esta técnica del área del aprendizaje semi-supervisado selecciona, repetidamente, las muestras más informativas para ser incorporadas al conjunto original y reentrenar el clasificador. La adición de información suplementaria produce una mejora en la efectividad del clasificador. En general, se cuenta con un pequeño conjunto de datos etiquetados y un enorme conjunto de datos sin etiquetar. La adición de muestras se lleva a cabo por la selección de aquellas que proporcionan más información. Posteriormente, el clasificador se vuelve a entrenar con el nuevo conjunto de datos.

Esta técnica de aprendizaje se aplica de igual manera al problema de los conjuntos de datos desbalanceados (Juszczak y Duin, 2003). En este caso, el pequeño conjunto de datos de inicio cuenta únicamente con muestras positivas que son elegidas aleatoriamente. Asimismo, la selección de nuevas muestras se basa en una medida de confidencia con el objetivo de lograr una selección dirigida y evitar la aleatoriedad del método clásico. En esta misma línea, Ertekin y otros (2007) proponen seleccionar muestras de forma iterativa a partir de un pequeño conjunto de datos. Al reducir la búsqueda a pequeñas regiones, el algoritmo consigue una mejora en términos de coste computacional.

Un trabajo que consiste en remuestrear el conjunto de datos por la adición de muestras no sintéticas es el propuesto por Barandela y otros (2003a). La estrategia se apoya en el algoritmo conocido como *descontaminación restringida* (*restricted decontamination*). El objetivo de dicho algoritmo es eliminar algunas muestras y reetiquetar otras. Para el caso del desbalance, esta estrategia se aplica únicamente a la clase mayoritaria, es decir, se eliminan y se cambian etiquetas de muestras negativas, con lo cual se reduce la clase mayoritaria a la vez que se incrementa el tamaño de la clase minoritaria.

3.3.1.2. Métodos de clasificación en entornos no balanceados

Cambiar la distribución *a priori* de las clases no es la única estrategia para mejorar la efectividad del clasificador en entornos desbalanceados. Diversas aproximaciones, conocidas como soluciones a *nivel de algoritmos*, tratan de adaptar las técnicas existentes de clasificación al problema del desbalance presente en el conjunto de datos. Algunas de éstas son:

1. Aprendizaje sensible el coste.
2. Algoritmos de clasificación con sesgo hacia la clase minoritaria.
3. Clasificadores de una clase.

El principal motivo por el que estas estrategias han conseguido popularidad se debe al hecho de evitar cambiar la distribución de los datos. Sin embargo, su mayor

desventaja radica en que, por lo general, las soluciones son específicas para un cierto clasificador, lo cual las hace menos adaptables a diversos problemas.

3.3.1.2.1. Aprendizaje sensible al coste. En muchas aplicaciones reales, existen diferentes tipos de coste asociado a la toma de decisiones (Turney, 2000). En el reconocimiento de formas, el aprendizaje sensible al coste estudia métodos para diseñar y construir clasificadores capaces de utilizar costes asimétricos de errores en sus algoritmos de aprendizaje. El objetivo en la clasificación basado en costes es minimizar el coste del error por la elección de la clase con el menor riesgo condicional.

Los costes de los errores pueden ser descritos por una matriz cuadrada arbitraria λ, donde $\lambda(i, j)$ es el coste de asignar un ejemplo a la clase i cuando en realidad pertenece a la clase j. Para expresar que una correcta clasificación no tiene coste, los elementos de la diagonal principal son iguales a cero (Elkan, 2001).

Un coste de error asimétrico en un problema de dos clases implica que el valor por clasificar erróneamente un ejemplo de una clase es mucho más alto que el coste de clasificar incorrectamente un ejemplo de la otra clase, por lo cual el aprendizaje sensible al coste se ha propuesto como una buena solución al problema del desbalance (Japkowicz y Stephen, 2002, McCarthy y otros, 2005, Weiss, 2004). Maloof (2003) establece que las técnicas de aprendizaje sensible al coste asimétrico pueden emplearse con éxito para tratar el problema de las clases desbalanceadas dado que, en estas situaciones, la clase con pocas muestras suele presentar un coste por error más elevado (por ejemplo, en el caso de la detección de fraudes en tarjetas de crédito (Chan y Stolfo, 1998)).

Las diversas técnicas de aprendizaje basadas en costes que se emplean en el problema del desbalance pueden clasificarse en dos grupos (Ling y Sheng, 2010):

1. Métodos directos (*DCSL*, *Direct Cost-Sensitive Learning*).

2. Métodos de meta-aprendizaje (*CSML*, *Cost–Sensitive Meta-Learning*).

Los métodos DCSL son aquellos que incorporan en el algoritmo de entrenamiento del clasificador la matriz de costes. Entre los algoritmos que se han adaptado se encuentran las redes neuronales (Kukar y Kononenko, 1998), los árboles de decisión (Drummond y Holte, 2000, Ling y otros, 2004, Pazzani y otros, 1994), las máquinas de vectores soporte (Veropoulos y otros, 1999, Yang y otros, 2008), entre otros.

Por su parte, los métodos CSML convierten un clasificador común en uno basado en costes sin la necesidad de modificar su algoritmo de aprendizaje. Este proceso de conversión se obtiene por el *preproceso* del conjunto de entrenamiento original o por el uso de la *matriz de costes* en la *salida del clasificador*.

El preproceso del conjunto de entrenamiento puede realizarse de dos formas distintas: por *ponderación* de las muestras (Elkan, 2001, Zadrozny y otros, 2003) o por *remuestreo* de las clases (Drummond y Holte, 2003, Maloof, 2003). En ambos casos, la modificación se realiza tomando en cuenta los valores de la matriz de costes.

Una estrategia muy popular que cambia la distribución del conjunto de datos es el algoritmo MetaCost (Domingos, 1999). El objetivo es reetiquetar cada muestra de entrenamiento por la estimación del riesgo de Bayes. Para ello, un sistema de múltiple clasificación (por ejemplo, el Bagging) estima la probabilidad de cada ejemplo. Finalmente, se entrena un clasificador no basado en costes con el conjunto de datos reetiquetado.

Por otra parte, algunos métodos de remuestreo basados en costes presentan una fuerte similitud con los métodos de submuestreo o sobremuestreo (Weiss y otros, 2007). En este sentido, alterar el ratio entre clases positiva y negativa equivale a asignar un ratio de coste por error (Elkan, 2001). Por el contrario, Ciraco y otros (2005) mencionan que cambiar la distribución de los datos no es equivalente a alterar el ratio de costes.

Teniendo en cuenta estos dos anteriores puntos de vista opuestos sobre los ratios de desbalance y de costes, Liu y Zhou (2006) proponen que los costes de errores de una clase deberían ser normalizados por el número de muestras en dicha clase.

3.3.1.2.2. Algoritmos de clasificación con sesgo hacia la clase minoritaria. Los algoritmos de clasificación de este tipo son adaptados internamente en su proceso de discriminación hacia la clase menos representada (Pazzani y otros, 1994, Zhang y otros, 2004), ya sea mediante el empleo de un umbral (Kubat y otros, 1998) o por las relaciones existentes entre ciertos atributos (Ezawa y otros, 1996).

Algunas propuestas están dirigidas a incorporar los diferentes pesos de error de clases positiva y negativa en el algoritmo de entrenamiento. Como ejemplo de esto último se encuentra el trabajo de Cohen y otros (2006) en el área de las SVM. La incorporación de ambos pesos de error induce al clasificador a establecer una mejor frontera de decisión. Imam y otros (2006), al contrario de la anterior propuesta, incorpora un peso asociado a cada vector soporte de la clase positiva. Podemos encontrar otras propuestas similares en Zhuang y Dai (2006), Quang y otros (2002) y Eitrich y Lang (2005).

En lo que respecta a la regla del vecino más cercano, Tan (2005) propone ponderar los k vecinos más cercanos asignando un mayor peso a aquellos que pertenecen a la clase minoritaria. Cabe mencionar que este factor de ponderación no influye al momento de determinar los vecinos más cercanos.

En la misma línea de emplear un factor de ponderación, se encuentra el algoritmo IG-CBL (*Information–Gain–weighted Case–Based Learning algorithm*) (Cardie y Nowe, 1997), el cual asigna a cada característica un peso calculado por *ganancia de información*.

3.3.1.2.3. Clasificadores de una clase. En contraste con los problemas de clasificación estándar, donde el objetivo es *discriminar* entre dos (o más) clases de objetos, la clasificación de una clase intenta *describir* una única clase de objetos (la clase objetivo) distinguiéndolos de cualquier otro posible objeto. Los clasificadores de una clase (*OCC, One-Class Classifier*) se entrenan para rechazar un objeto y

etiquetarlo como salida falsa o valor atípico (*outlier*), siempre que no pueda asignarse a la clase objetivo.

Los OCC se han propuesto como una solución alternativa al problema del desbalance (Japkowicz y otros, 1995, Juszczak y Duin, 2003, Tax, 2001). Raskutti y Kowalczyk (2004) demuestran que, en entornos con un alto ratio de desbalance y una alta dimensionalidad con ruido, los OCC son una técnica de clasificación robusta. Lee y Cho (2006) obtienen similares conclusiones empleando una máquina de vectores soporte. Adicionalmente, mencionan que los OCC presentan un mejor comportamiento en problemas donde el objetivo es discriminar una clase del resto (problemas no-simétricos).

Generalmente, el proceso de aprendizaje de los OCC requiere que el conjunto de entrenamiento contenga únicamente ejemplos de la clase objetivo. Sin embargo, éstos pueden ser modificados para aceptar información de la otra clase (Lee y Cho, 2006).

3.3.1.3. Combinación de estrategias

Varios estudios sobre conjuntos desbalanceados presentan diversos puntos de vista acerca de la mejor estrategia de solución para el problema del sesgo en los datos (Chawla, 2003, Drummond y Holte, 2003).

Japkowicz y Stephen (2002) evalúan tres estrategias para tratar el problema del desbalance: sobremuestreo, submuestreo y aprendizaje por costes de error. Los resultados sobre conjuntos artificiales muestran que la tercera estrategia es más efectiva que ambos métodos de remuestro en conjuntos de datos que presentan poca complejidad y una talla media en el conjunto de entrenamiento. Cabe mencionar que los algoritmos de sobremuestreo considerados en este estudio no crean ningún nuevo ejemplo. Un estudio en esta misma línea sobre conjuntos reales lo presentan McCarthy y otros (2005), con la diferencia de que ambas estrategias de remuestreo son aleatorias. En un sentido que pudiera ser contradictorio a los trabajos de Japkowicz y Stephen, McCarthy y otros mencionan que el aprendizaje basado en costes presenta un mejor comportamiento en conjuntos de datos con más de $10,000$ ejemplos.

Por otra parte, Barandela y otros (2004) y Prati y otros (2004a) realizan experimentos similares con el objetivo de comparar las diversas técnicas de sobremuestreo y submuestreo. En ambos casos, las conclusiones coinciden al indicar que reducir la talla de la clase mayoritaria resulta efectivo para aquellos conjuntos de datos que no presentan un ratio de desbalance severo. Por el contrario, si el conjunto de datos está fuertemente desbalanceado, los algoritmos de sobremuestreo constituyen una solución más apropiada.

A pesar de la existencia de diversos estudios con múltiples estrategias, no es concluyente todavía bajo qué situaciones una solución presenta un mejor comportamiento que otras. Asimismo, las conclusiones actuales pudieran estar influenciadas por otros factores, tales como el clasificador, la complejidad de los datos, la medida de evaluación, etc. En relación a este punto, los trabajos experimentales de Hulse

y otros (2007) muestran que la interpretación de los resultados presenta una fuerte dependencia de la medida de efectividad a emplear. En consecuencia, la decisión acerca de la mejor técnica depende, en gran manera, de la naturaleza del clasificador y de la medida de efectividad.

Como ya se ha comentado, la mayor crítica a las estrategias a nivel de datos se refiere al cambio que realizan sobre la distribución original del conjunto de entrenamiento. Por lo que respecta a las estrategias a nivel de algoritmos, su mayor obstáculo radica en la difícil adaptación o modificación de dicha estrategia a otros algoritmos o aplicaciones. Como alternativa a estas dos opciones, algunos investigadores proponen emplear una combinación de estrategias de ambos niveles para tratar el problema del desbalance (Ling y Li, 1998). El objetivo es aprovechar las ventajas que cada uno de los métodos ofrece, logrando así métodos más robustos, aunque con un incremento en términos de coste computacional.

La combinación de estrategias no sólo se lleva a cabo empleando las técnicas de desbalance, sino que algunas propuestas buscan complementar tales técnicas con metodologías ya establecidas, como suelen ser los sistemas múltiples de clasificación (Estabrooks y otros, 2004).

3.3.1.3.1. Sistemas de múltiple clasificación. Los sistemas múltiples de clasificación (ensembles) consisten en un conjunto de clasificadores entrenados individualmente cuyas decisiones posteriormente son combinadas. Estos sistemas surgen con el principal propósito de mejorar la precisión en la clasificación de nuevas muestras (Hansen y Salamon, 1990).

Experimentos empleando esta técnica muestran que la precisión de las predicciones colectivas suelen ser mejores que la precisión de los clasificadores individuales. Esta ventaja que presentan los ensembles hace atractiva su aplicación a problemas desbalanceados.

En general, la mayoría de las propuestas existentes para tratar el desbalance mediante ensembles combinan dicha metodología con otras aproximaciones. Un planteamiento de este tipo puede verse como una estructura formada por tres niveles jerárquicos:

1. Nivel de entrada.
2. Nivel de expertos o ensemble.
3. Nivel de salida.

En el primer nivel, el conjunto de entrenamiento original puede ser preprocesado por alguna técnica de reducción o filtrado. A nivel de expertos, los clasificadores individuales son entrenados empleando los conjuntos provenientes del primer nivel, donde cada uno de estos clasificadores produce como única salida una etiqueta de clase. En el nivel de salida, se combinan las decisiones individuales del ensemble mediante algún esquema de toma de decisiones para determinar la clase de la muestra de entrada.

Estabrooks y Japkowicz (2001) diseñaron un esquema de combinación de clasificadores C4.5. En el nivel de entrada, el conjunto de entrenamiento es remuestreado aleatoriamente a diferentes porcentajes. En particular, el nivel de expertos cuenta con 20 clasificadores, 10 de los cuales son entrenados con datos sobremuestreados, mientras que los 10 restantes son entrenados con conjuntos submuestreados. La decisión final es tomada de acuerdo al esquema más simple de combinación: la votación por mayoría.

En la misma línea de emplear conjuntos balanceados por submuestreo, se encuentran los trabajos de Chan y Stolfo (1998) y de Barandela y otros (2003b). Este último se caracteriza por establecer el número de clasificadores individuales por el ratio de desbalance, es decir, la cantidad de clasificadores individuales (en este caso, 1−NN) se fija en función de la diferencia entre la talla de la clase mayoritaria y la talla de la clase minoritaria.

Los trabajos anteriores se caracterizan por emplear métodos aleatorios para remuestrear el conjunto de datos. Con la finalidad de evitar las desventajas que presentan los métodos no dirigidos, He y otros (2005) proponen un ensemble de clasificadores C4.5 entrenados con conjuntos de datos sobremuestreados por el algoritmo SMOTE. En cuanto a la toma de decisiones, el esquema utiliza una votación ponderada, donde aquellos clasificadores con una mejor tasa de exactitud basada en la medida f cuentan con un peso mayor.

Como un paso previo al remuestreo, Kotsiantis y Pintelas (2003) proponen reducir la dimensionalidad del conjunto de datos por selección de variables. Por otra parte, es importante mencionar que este esquema se caracteriza por emplear clasificadores individuales de distinta naturaleza: C4.5, 5−NN y naïve Bayes.

A nivel de clasificadores, se emplean otros modelos complementarios, como el *boosting* y *los clasificadores en cascada*, para tratar el desbalance. El boosting es un algoritmo iterativo donde, en cada paso, se entrena el modelo de clasificación sobre un conjunto diferente de muestras ponderadas. Después de cada iteración, se actualizan los pesos de los ejemplos. El objetivo es incrementar los pesos de las muestras erróneamente clasificadas, lo que ocasiona que cada nuevo clasificador se enfoque más hacia las muestras clasificadas erróneamente en la iteración anterior. Algunas propuestas basadas en este modelo son las siguientes:

- Chawla y otros (2003b) diseñan el algoritmo *SMOTEBoost* basado en un combinación de SMOTE y boosting. El procedimiento de crear muestras sintéticas asigna de manera indirecta una ponderación hacia la clase minoritaria.

- Guo y Viktor (2004) combinan la generación de datos y boosting en un algoritmo llamado *Databoost–IM*. Este algoritmo identifica por boosting las muestras de ambas clases difíciles de clasificar. A partir de la identificación, se generan nuevas muestras (de ambas clases) para ser incorporadas al conjunto original.

- Joshi y otros (2001) exploran cuatro algoritmos de boosting, incluyendo una versión basada en costes sobre conjuntos artificiales desbalanceados. Los cuatro

algoritmos presentan variaciones en su forma de ponderación. El clasificador elegido fue el C4.5.

- Liu y otros (2006b) usan un ensemble de SVM entrenados sobre conjuntos preprocesados por SMOTE y edición de Wilson.

- Liu y Zhou (2006) diseñan el algoritmo *EasyEnsemble*, que consta de un grupo de clasificadores entrenados por Adaboost cuyos subconjuntos de entrenamiento son aleatoriamente balanceados.

- Molinara y otros (2007) exploran dos formas de crear subconjuntos balanceados: aleatorio y basado en clustering. Este último método de partición permite ubicar un clasificador en una región del espacio de representación en particular.

- Sun y otros (2007) llevan a cabo una extensa experimentación empleando diversos métodos de boosting basados en costes. Los resultados permiten concluir que las aproximaciones de ensembles basados en costes incrementan la tasa de reconocimiento sobre la clase minoritaria.

Viola y Jones (2001) construyen un método de *clasificadores en cascada* para la detección de caras en secuencias de vídeo. La arquitectura puede verse como un árbol de decisión, donde cada estado representa un clasificador que es entrenado para detectar muchos de los objetos de interés mientras que se rechaza una cierta fracción de muestras de no interés. Tomando como base esta estructura, Liu y otros (2006a) diseñan el algoritmo *BalanceCascade* que, a diferencia del método clásico, cada clasificador es entrenado con subconjuntos submuestreados, a la vez que se eliminan del conjunto original las muestras de la clase negativa que son erróneamente clasificadas. En general, los resultados muestran que este método de ensemble presenta mejores resultados que los métodos clásicos de remuestreo, principalmente en bases de datos fuertemente desbalanceadas.

Finalmente, a nivel de salida, las decisiones finales son tomadas por votación mayoritaria, ponderación o media (Molinara y otros, 2007). Chawla y Sylvester (2007) proponen ponderar cada clasificador del ensemble mediante el uso de algoritmos genéticos.

3.3.1.3.2. Combinación de técnicas de remuestreo. Barandela y otros (2004) indican que, para el caso de conjuntos con un alto ratio de desbalance, las técnicas de sobremuestreo son la mejor opción para tratar el problema. Adicionalmente, se eliminan casos de ambas clases mediante la edición de Wilson. Experimentos sobre seis bases de datos reales muestran que la combinación de ambos métodos reduce la talla del conjunto, a la vez que incrementa la tasa de reconocimiento.

En la misma línea, Batista y otros (2004) combinan SMOTE con dos algoritmos de submuestreo, links de Tomek y edición de Wilson, llegando a conclusiones similares a las obtenidas por Barandela y otros.

3.3.2. Evaluación de la efectividad de los clasificadores

Las métricas de evaluación más empleadas para medir la efectividad de los clasificadores son la *tasa de errores* y la *tasa de aciertos* (Demšar, 2006). Éstas, para un problema de dos clases, pueden obtenerse fácilmente a partir de una *matriz de confusión*[1] como la mostrada en la Tabla 3.1.

Tabla 3.1: *Matriz de confusión para un problema de dos clases*

	Predicción Positiva	Predicción Negativa
Clase Positiva	Verdadero Positivo (vp)	Falso Negativo (fn)
Clase Negativa	Falso Positivo (fp)	Verdadero Negativo (vn)

Estas tasas pueden calcularse como

$$Acc = \frac{vp + vn}{vp + fn + vn + fp} \tag{3.1}$$

y

$$Err = \frac{fp + fn}{vp + fn + vn + fp} \tag{3.2}$$

Sin embargo, diversos experimentos han demostrado que, cuando las probabilidades *a priori* no se encuentran equilibradas, ambas medidas no resultan apropiadas debido a que no consideran distintos tipos de errores. En consecuencia, estas métricas se muestran fuertemente sesgadas en favor de la clase mayoritaria (Daskalaki y otros, 2006, Provost y Fawcett, 1997). Por ejemplo, considérese un problema binario cuya clase positiva contiene un 1 % de objetos sobre el conjunto total; en tal situación, una simple estrategia de asignar todas las muestras a la clase negativa ofrecería una tasa de aciertos del 99 %. Sin embargo, tal clasificador carecería de valor alguno.

Las desventajas de estos evaluadores han motivado la búsqueda de medidas alternativas. Algunos ejemplos sencillos son:

- Tasa de Verdaderos Positivos (*recall* o *sensitivity*), es el porcentaje de ejemplos positivos que son correctamente clasificados.

$$tvp = \frac{vp}{vp + fn} \tag{3.3}$$

- Tasa de Verdaderos Negativos (*specificity*), es el porcentaje de ejemplos negativos que son clasificados como positivos.

$$tvn = \frac{vn}{vn + fp} \tag{3.4}$$

[1] Cada columna de la matriz representa el número de predicciones de cada clase, mientras que la fila representa a las instancias en la clase real.

- Tasa de Falsos Positivos, es el porcentaje de ejemplos negativos que son erróneamente clasificados.

$$tfp = \frac{fp}{vn + fp} \tag{3.5}$$

- Tasa de Falsos Negativos, es el porcentaje de ejemplos positivos que son clasificados como negativos.

$$tfn = \frac{fn}{vp + fn} \tag{3.6}$$

- *Precision*, conocida también como *pureza*, se define como el porcentaje de ejemplos que fueron etiquetados correctamente como positivos, con respecto a todas las muestras que fueron etiquetadas como tal.

$$precision = \frac{vp}{vp + fp} \tag{3.7}$$

Recientemente, se han propuesto algunas métricas tomadas de otros dominios para la evaluación de clasificadores en entornos desbalanceados. Algunos ejemplos ampliamente conocidos son las curvas *ROC* (*Receiver Operating Characteristic*), el área bajo la curva *ROC* (*AUC, Area Under Curve*), la media geométrica (*Gm*) y la medida f. Otra métrica, aunque menos conocida, es la precisión optimizada (*OP, Optimized Precision*) (Ranawana y Palade, 2006).

3.3.2.1. Curvas *ROC*

La curva *ROC* es una metodología de análisis desarrollada por ingenieros eléctricos y de radar durante la Segunda Guerra Mundial, con la finalidad de resolver problemas prácticos en la detección de señales. La ventaja que proporciona este análisis ha motivado su uso en el area de la medicina para evaluar y visualizar el comportamiento de los sistemas de diagnóstico (Swets, 1988).

En los últimos años, la utilización de las curvas *ROC* se ha extendido a otras áreas muy diversas, como el aprendizaje automático y la minería de datos, en respuesta a tareas de clasificación con distribuciones de clases variantes y costes de errores no uniformes (Spackman, 1989, Fawcett, 2006), situaciones en las que la simple tasa de aciertos falla a la hora de evaluar la efectividad del clasificador (Provost y Fawcett, 1997).

El espacio de una curva *ROC* es un gráfico bidimensional que permite visualizar, organizar y seleccionar clasificadores basados en su efectividad. Mediante esta representación es posible conocer la relación entre los beneficios (verdaderos positivos) y los costes (falsos positivos). La Figura 3.1 muestra un espacio *ROC*, cuyo eje Y representa la tasa de verdaderos positivos y el eje X la tasa de falsos positivos (tfp). En esta misma figura, se encuentran cinco clasificadores etiquetados de las letras A a la E.

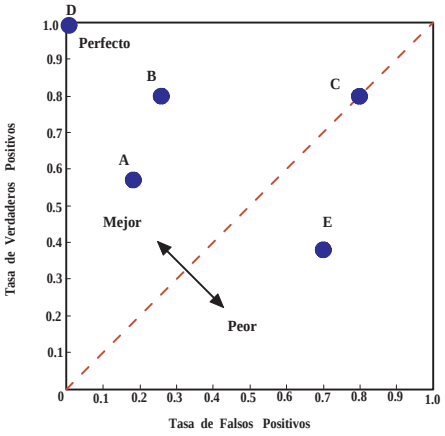


Figura 3.1: *Espacio de una curva ROC*

Un clasificador discreto es aquel cuyas salidas son etiquetas de clase. Cada clasificador discreto produce un par (tfp, tvp) que corresponde a un punto en el espacio ROC. Los clasificadores en la Figura 3.1 son todos discretos.

En el espacio ROC, el punto $(0, 0)$ indica que todos los ejemplos han sido etiquetados como de la clase negativa. El caso contrario sucede en el punto $(1, 1)$ al etiquetar todas los ejemplos como positivos. La clasificación perfecta ocurre en el punto $(0, 1)$, mientras que el caso de un clasificador con decisiones aleatorias viene representado por la diagonal X = Y.

Informalmente, podemos decir que un punto en el espacio ROC es mejor que otro si el primero está más al noroeste del segundo (tvp más alto, tfp más bajo, o ambos). Los clasificadores que aparecen en el lado izquierdo de la figura, cerca del eje X, podrían considerarse como *conservadores*, ya que su decisión de clasificar un ejemplo como positivo ocurre si existe una fuerte evidencia, lo que puede provocar una tfp muy baja. Por lo que respecta a los clasificadores que se ubican en el lado derecho del espacio, éstos son considerados como *liberales*. En este caso, la tendencia es a etiquetar como positivo un ejemplo teniendo una débil evidencia. Por supuesto, lo anterior implica un aumento en el índice de falsos positivos. En la Figura 3.1, por ejemplo, el clasificador A es más conservador que el B.

3.3.2.1.1. Curvas *ROC* con *precision* y *posfrac*. A pesar de las múltiples ventajas que presentan las curvas *ROC*, algunas investigaciones muestran que, en situaciones donde el sesgo cambia, esta forma de evaluar no permite expresar dicho cambio. Landgrebe y otros (2006) proponen emplear en el eje X *precision* o pureza en lugar de tfp. Esta medida indica la fracción de ejemplos clasificados correctamente como positivos. Además, la medida *precision* es dependiente del grado de sesgo presente en el conjunto de datos. Las curvas de *precision-tvp* (*P-ROC, Precision-Recall Characteristic*) son conocidas también como curvas ROC bayesianas (*B-ROC, Bayesian ROC*) (Cárdenas y Baras, 2006).

Una segunda variante de curvas *ROC* incorpora una medida de efectividad llamada *Posfrac*, la cual se emplea en aquellas aplicaciones que requieren una segunda fase de tratamiento manual de las salidas positivas del clasificador (Landgrebe y otros, 2004) (por ejemplo, en los tests de exploraciones médicas).

Recientemente, algunos trabajos optan por otros métodos para visualizar el comportamiento del clasificador. Uno de éstos se refiere a las curvas de costes (Drummond y Holte, 2006), que definen un método gráfico en un espacio bidimensional

3.3.2.2. Área bajo la curva *ROC*

A partir de las curvas *ROC*, es posible obtener un sencillo valor numérico que represente el comportamiento de un clasificador en el espacio *ROC*. Dicho valor se obtiene al calcular la fracción del área total que se encuentra bajo la curva *ROC*, llamada comúnmente *AUC* (ver Figura 3.2).

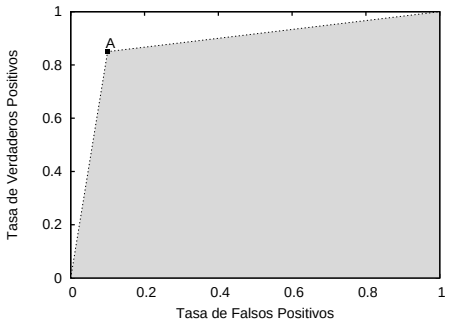


Figura 3.2: *Área bajo la curva ROC de un clasificador discreto*

Algunos trabajos de Bradley (1997) evaluando varios clasificadores muestran las ventajas de emplear *AUC* sobre la clásica medida de efectividad basada en aciertos. Algunas de éstas son:

- Es invariante a cambios en las probabilidades *a priori* de las clases.
- Es un buen indicador del trabajo bien hecho por un esquema de clasificación. Para clasificadores aleatorios o de una sola clase, sus valores son bajos.
- Su escala se encuentra entre los valores 0 y 1.

Uno de los métodos más simples para calcular el *AUC* es por la sucesiva adición de areas trapezoidales. Sin embargo, para el caso de un clasificador discreto con una matriz de confusión (una única ejecución), es fácil demostrar que el *AUC* puede calcularse de la siguiente manera (Sokolova y otros, 2006):

$$AUC = \frac{tvp + tvn}{2} \tag{3.8}$$

Sin embargo, esta forma sencilla de evaluar que no requiere un análisis ROC, puede producir el mismo resultado para diversas (y, posiblemente, antagónicas) combinaciones de tvp y tvn (Ferri y otros, 2003).

3.3.2.2.1. AUCPREC y Weighted-AUC. En la Sección 3.3.2.1, se comentó que es posible sustituir la tfp de la curva ROC por $precision$ con la finalidad de poder estimar con más exactitud los cambios en el sesgo de los datos. En este tipo de curvas, también es posible calcular el área que se encuentra bajo ellas ($AUCPREC$). Los valores del $AUCPREC$, al igual que AUC, pueden calcularse empleando una aproximación trapezoidal. Adicionalmente, Landgrebe y otros (2006) proponen incorporar al cálculo del $AUCPREC$ los valores de sesgo del conjunto de datos, en una medida llamada $IAUCPREC$.

Por otra parte, Weng y Poon (2008) proponen calcular de la clásica curva ROC un área bajo la curva ponderada ($weighted\text{-}AUC$). Dicha medida se caracteriza por realizar sucesivas sumas ponderadas de fracciones de áreas bajo la curva. La idea es otorgar más importancia a aquellos clasificadores que están cerca de la parte superior del espacio ROC.

3.3.2.3. Media geométrica

En el problema del desbalance, una de las métricas más empleadas es la media geométrica (Gm) de tasas de aciertos por clase (Kubat y Matwin, 1997):

$$Gm = \sqrt{tvp \cdot tvn} \tag{3.9}$$

La media geométrica representa un punto en el espacio ROC y su objetivo es maximizar la efectividad sobre ambas clases mientras se mantienen equilibradas.

Una propiedad importante de la media geométrica es su independencia de la distribución de los datos entre las clases. Además, esta medida es no lineal.

3.3.2.4. Medida f

Otra métrica de la comunidad de la recuperación de información empleada en la evaluación de clasificadores en entornos desbalanceados es la medida f,

$$f = \frac{(1+\beta^2) \cdot tvp \cdot precision}{\beta^2 \cdot tvp + precision} \tag{3.10}$$

la cual, se define a partir de $precision$ y tvp. En principio, la medida f representa una media armónica entre estas dos métricas.

El factor β es un parámetro que permite controlar de forma separada la influencia de tvp y $precision$. En la mayoría de los casos, β es puesto a 1. Por tanto, la medida puede ser reescrita como sigue:

$$f_1 = \frac{2 \cdot tvp \cdot precision}{tvp + precision} \tag{3.11}$$

Un valor alto de f_1 indica que ambas, *tvp* y *precision*, son razonablemente altas.

A diferencias de las anteriores métricas, la medida f no considera, dentro de sus cálculos, índices sobre la clase negativa. En este sentido, diversos investigadores no la consideran idónea para evaluar clasificadores en entornos desbalanceados.

3.3.2.5. Media geométrica de *tvp* y *precision*

Trabajos de Daskalaki y otros (2006) muestran una propuesta para combinar *precision* y *tvp* por medio de una media geométrica de ambas métricas,

$$Gpr = \sqrt{precision \cdot tvp} \tag{3.12}$$

Los resultados experimentales muestran que esta métrica presenta un comportamiento similar al de la medida f.

3.3.2.6. Precisión optimizada

Recientemente, Ranawana y Palade (2006) han propuesto una nueva medida en el área del desbalance llamada precisión optimizada (*OP*, *Optimized Precision*), que se formula de la siguiente manera:

$$OP = Acc - \frac{|tvn - tvp|}{tvn + tvp} \tag{3.13}$$

Esta métrica refleja la diferencia entre la tasa de precisión global (*Acc*) y un segundo término llamado *índice de relación* (*RI*, *Relation Index*). Esté segundo término indica qué tan balanceadas están las tasas de aciertos sobre las dos clases. Valores altos de *OP* requieren altas tasas globales de acierto y tasas balanceadas de aciertos por clase. No obstante, ha de señalarse que esta medida puede verse fuertemente influenciada por la tasa de exactitud global.

3.3.3. Relación con otras complejidades

La tercera y última línea de investigación en el tema del desbalance se centra en estudiar si el sesgo en los datos es el principal causante de la baja efectividad de un clasificador. En diversos dominios de aplicación, los algoritmos de clasificación son capaces de obtener altas tasas de reconocimiento en ambas clases, aún cuando el conjunto de datos presente un alto ratio de desbalance (6.50 % de ejemplos pertenecen a la clase minoritaria, (Asuncion y Newman, 2007)).

Japkowicz y Stephen (2002) presentan uno de los primeros trabajos dirigidos a analizar el efecto del desbalance en diversos clasificadores. Sus experimentos sobre datos sintéticos generados con una serie de parámetros que controlan la complejidad, la talla y el grado de desbalance, les permiten concluir que, en conjuntos separables pocos complejos, el clasificador no se ve afectado por el desbalance presente en los datos. Sin embargo, en conjuntos con un alto grado de complejidad, el sistema se hace más sensible al sesgo de los datos. Esto sugiere que existe una fuerte dependencia conjunta de otras fuentes de complejidad.

Otras líneas de investigación buscan establecer el efecto conjunto del desbalance y el solapamiento entre clases en algoritmos como los árboles de decisión (Prati y otros, 2004a) y los clasificadores difusos (Visa y Ralescu, 2003). Ambos concluyen que, a bajos niveles de solapamiento, cualquier ratio de desbalance puede conducir a altas tasas de efectividad. En caso contrario, cuando existe un alto nivel de solapamiento entre clases, la efectividad del clasificador decae a pesar de que el ratio de desbalance sea mínimo. Asimismo, los resultados sugieren que en situaciones con bajos niveles de desbalance y de solapamiento, este último tiene una mayor influencia en el comportamiento del clasificador.

Por último, otros trabajos exploran la relación entre desbalance y *casos raros*. Los *casos excepcionales* o *raros* son definidos como aquellos ejemplos que suceden con poca frecuencia dentro de un dominio en áreas muy particulares del espacio de características (Weiss, 1995). Trabajos de Jo y Japkowicz (2004) muestran que, además del problema del desbalance entre clases, la presencia de casos raros (en una clase o en ambas) afecta de manera negativa al comportamiento del clasificador. Similares conclusiones se presentan en otro trabajo de Prati y otros (2004b).

Capítulo 4

Índice de precisión balanceada

Las medidas que evalúan la efectividad de un clasificador en entornos desbalanceados, a pesar de considerar tasas de aciertos por clase, no suelen explicar la importancia de cada una de ellas sobre el índice global de rendimiento. En el presente capítulo, se define y fundamenta una nueva función que combina una métrica global y un índice que mide cuán dominante es la clase con la mayor tasa de aciertos.

4.1. Introducción

Las medidas o funciones de efectividad evalúan el comportamiento de un clasificador en un determinado ámbito, lo cual, realizado para un conjunto de modelos candidatos, permite compararlos con el objetivo de elegir el mejor. Esta comparación empírica se realiza frecuentemente aplicando pruebas de significancia estadística o intervalos de confianza a los resultados de varios clasificadores sobre una o más bases de datos (Demšar, 2006, Ferri y otros, 2009, Sokolova, 2006). Adicionalmente, se emplean estas medidas como criterio de optimización en procesos de aprendizaje de sistemas de clasificación (Ranawana y Palade, 2006).

Una de las medidas que más se emplean es la *tasa global de aciertos* del clasificador (en inglés, *accuracy*, a la que también nos referiremos como *exactitud global*)[1]. Esta función es el cociente entre el número de muestras correctamente clasificadas y la cantidad total de muestras del conjunto de datos. Sin embargo, al tratarse de una medida de evaluación global, dicha función no distingue entre el número de aciertos de las diferentes clases, asumiendo así una distribución uniforme de muestras por clases y la no existencia de diferencias entre los distintos tipos de error. Debido a esta particularidad, la exactitud global resulta inadecuada como medida de evaluación en entornos con clases no balanceadas pues, a mayor nivel de desbalance, es más probable que oculte resultados de clasificación deficientes sobre la clase minoritaria (o positiva).

Para ilustrar este comportamiento, se ha definido un experimento simple a partir de una serie de cuatro conjuntos sintéticos de dos clases con solapamiento constante y desbalance creciente. Una de las clases está compuesta por una cantidad fija de 400 instancias (clase mayoritaria), mientras que la otra clase (la minoritaria) parte del caso balanceado 1:1 hasta alcanzar la situación desbalanceada con una relación de 1:8. En todos los casos, las muestras fueron generadas aleatoriamente siguiendo

[1] Demšar (2006) indica que el 70 % de los artículos presentados en la International Conference on Machine Learning de 2003 emplearon este índice o, complementariamente, la *tasa global de errores*

distribuciones uniformes en un espacio de 2 dimensiones de tamaño 100×100. Por último, se generaron 50 pares de conjuntos distintos de entrenamiento y evaluación, por cada uno de los niveles de desbalance. La Figura 4.1(a) representa el caso del conjunto de datos desbalanceado 1:8.

En la Figura 4.1(b), se pueden observar los resultados medios de clasificación obtenidos con la regla 1−NN. El eje X representa el ratio mayoritaria/minoritaria de desbalance, mientras que el eje Y agrupa la tasa global de aciertos (*Acc*) y la tasa de aciertos sobre la clase positiva (o tasa de verdaderos positivos, *tvp*). Los resultados reflejan que, conforme aumenta el desbalance, crece el valor de *Acc*, manteniéndose por encima del 0.90 a pesar de que *tvp* se reduce significativamente. Esta situación es aún más notoria en el caso extremo del desbalance 1:8, donde $tvp \approx 0.76$ y $Acc \approx 0.98$.

Este comportamiento se puede explicar por el funcionamiento sesgado del modelo 1−NN, que favorece a la clase cada vez más representada, y por la naturaleza de la tasa global de aciertos que no penaliza el sesgo del clasificador al considerar todos los aciertos por igual (Daskalaki y otros, 2006, Drummond y otros, 2006, Folleco y otros, 2008, Landgrebe y otros, 2006, Liu y otros, 2007, Provost y Fawcett, 1997).

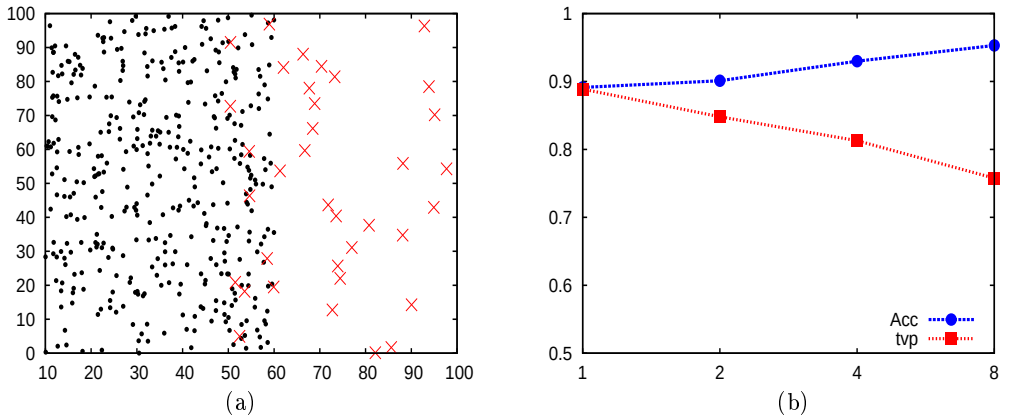


Figura 4.1: *(a) Conjunto de datos con un desbalance 1:8 y (b) Resultados de clasificación.*

Debido a este comportamiento de la métrica *Acc* en entornos desbalanceados, diversos investigadores han propuesto medidas de efectividad alternativas para evaluar clasificadores en dichos contextos. Algunas de estas medidas ampliamente conocidas son la curva ROC (Fawcett y Provost, 1997, Hernández-Orallo y otros, 2004, Provost y Fawcett, 2001, Fawcett, 2006), el área bajo la curva ROC (AUC) (Bradley, 1997, Jin y C.X.Ling, 2005, Hulse y otros, 2007, Wong y otros, 2006), la media geométrica de tasas de acierto por clase (Kubat y Matwin, 1997, Kubat y otros, 1998) y la medida f (Daskalaki y otros, 2006, Tan, 2005, Zhang y Mani, 2003). Otras medidas han sido propuestas teniendo como base las anteriores, por lo cual son combinaciones de tasas de error, de aciertos y *precision*. Entre éstas, podemos mencionar las curvas P-ROC (Landgrebe y otros, 2006), las curvas B-ROC (Cárdenas y Baras, 2006), las

curvas basadas en costes (Drummond y Holte, 2006), AUC con ponderación (Weng y Poon, 2008) y *precisión optimizada* (Ranawana y Palade, 2006). Muchas de estas funciones de evaluación son combinaciones aritméticas de tasas de errores y aciertos evaluadas separadamente sobre cada clase, y que fueron formuladas para aliviar los resultados sesgados de clasificación. Sin embargo, a pesar de esto, algunas de ellas presentan las siguientes desventajas:

1. No consideran la magnitud, ni el sentido de la diferencia entre las tasas de aciertos por clase, es decir, no reflejan el sentido del sesgo del clasificador.

2. Las medidas basadas en *precision* ignoran por completo la información de la tasa de aciertos de la clase negativa.

3. Producen el mismo resultado para infinitas y diversas combinaciones de tasas de aciertos en las clases negativa y positiva. Por ejemplo, dados dos resultados hipotéticos de clasificación, $[tvp = 0.5, tvn = 1]$ y $[tvp = 0.71, tvn = 0.71]$, la media geométrica (ver Ecuación 3.9) produce el mismo resultado, mientras que AUC (ver Ecuación 3.8) produce un valor más alto para la primera configuración, la cual podría admitirse como poco útil. Una situación más acentuada ocurre cuando, para un par cualquiera $[tvp, tvn]$, sus componentes son intercambiadas, manteniéndose invariantes los resultados de AUC y Gm; por ejemplo, en casos tan opuestos como $[tvp = 0.5, tvn = 1]$ y $[tvp = 1, tvn = 0.5]$.

La siguiente sección propone una nueva medida de evaluación de la efectividad de la clasificación en problemas de 2 clases, que resuelve las desventajas anteriores y que se define a partir de la relación entre una medida global no sesgada y un índice simple que evalúa la diferencia entre las tasas de aciertos sobre cada una de las dos clases.

4.2. Una nueva medida de efectividad

Esta sección introduce una nueva medida que evalúa la efectividad de un clasificador en problemas binarios no balanceados. La idea que subyace detrás de esta propuesta es la de considerar, además de las tasas individuales por clases, la diferencia con signo entre ambas tasas. El uso de esta información contribuirá a penalizar situaciones en las que la tasa de éxito sobre la clase mayoritaria es mucho mayor que en la clase minoritaria, así como a favorecer resultados con alta efectividad sobre la clase de interés.

La nueva medida es una función que calcula el área de una región rectangular en un espacio bidimensional que se ha denominado *Gráfico de Precisión Balanceada* (BAG, *Balanced Accuracy Graph*). Este espacio se define en función de dos medidas: *i)* el cuadrado de la media geométrica de las tasas de acierto por clases, la cual es una estimación no sesgada de la precisión global del clasificador, y *ii)* una expresión simple, *Índice de dominio*, que mide cuán dominante es la clase más favorecida por el clasificador.

4.2.1. Índice de dominio

Con la finalidad de medir la magnitud y el sentido de la diferencia entre *tvp* y *tvn*, se propone una sencilla función llamada *índice de dominio* (*dominance*), cuyo objetivo es informar acerca de cuál es la clase prevalente y cuán significativa es esta relación de dominio. Este índice se formula de la siguiente manera:

$$dominance = tvp - tvn \tag{4.1}$$

Debido a que *tvp* y *tvn* se encuentran en el rango $[0, +1]$, *dominance* puede tomar cualquier valor entre -1 y $+1$. Un valor de dominance de $+1$ representa una clasificación perfecta sobre la clase positiva, pero fallando en todos los casos negativos. Complementariamente, un valor de -1 corresponde a la situación opuesta. Por otra parte, valores de *dominance* cercanos a 0 indican que las tasas de aciertos entre clases están significativamente balanceadas. Sin embargo, cualquier valor de *dominance* podría representar diferentes escenarios. Así, $dominance = 0$ puede corresponder al caso óptimo $tvp = tvn = 1$, pero también a la dramática situación $tvp = tvn = 0$. En la práctica, dominance puede entenderse como un indicador del balance entre *tvp* y *tvn*.

Esta circunstancia sugiere la necesidad de considerar la información que aporta *dominance* conjuntamente con otra medida capaz de evaluar, sin sesgo, la efectividad global del clasificador.

4.2.2. Gráfico de precisión balanceada

En esta sección, se presenta el gráfico de precisión balanceada (BAG), el cual es una herramienta que permite visualizar y evaluar el comportamiento de cualquier clasificador desde la perspectiva conjunta de la media geométrica y el *dominance*. Con la finalidad de proveer sencillez al modelo, se emplea la media geométrica al cuadrado (Gm^2), en lugar de la versión original con raíz cuadrada.

La Figura 4.2 muestra el gráfico BAG, el cual, para un clasificador cualquiera, permite expresar en un sistema de coordenadas bidimensional la relación entre *dominance* en el eje X y Gm^2 en el eje Y. A partir de este espacio, es posible representar el comportamiento de un clasificador como un par (d, g) que corresponde a un punto en este gráfico, siendo d y g los valores de *dominance* y Gm^2, respectivamente.

Dentro del espacio BAG, es posible destacar diversos puntos importantes. El punto superior central $(0, +1)$ representa la situación de un clasificador perfecto ($tvp = tvn = 1$), mientras que los puntos $(-1, 0)$ y $(+1, 0)$ corresponden a los casos extremos de $tvn = 0$, $tvn = 1$ y $tvp = 1, tvn = 0$, respectivamente.

Los puntos $(-1, +1)$ y $(+1, +1)$ representan situaciones no factibles pues, al ser *dominance* igual $+1$ o -1, algunas de las dos tasas de aciertos es 0, por lo cual es imposible que Gm produzca valores mayores a 0. En realidad, existe un número infinito de puntos en BAG que representan casos imposibles. La Figura 4.3 muestra en color gris la región que contiene las combinaciones posibles de *dominance* y Gm^2.

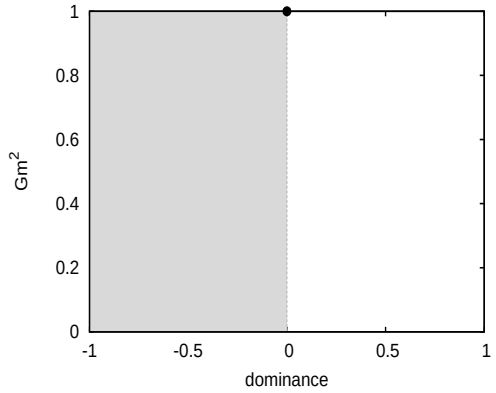


Figura 4.2: *Gráfico de precisión balanceada: el punto óptimo (dominance* $= 0$ *y* $Gm^2 = 1$*) que da lugar al rectángulo con área máxima.*

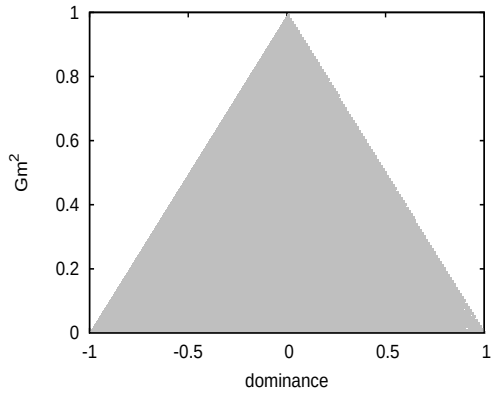


Figura 4.3: *Región que contiene las combinaciones posibles de dominance y* Gm^2*.*

4.2.3. Índice de precisión balanceada

A partir de un punto (d, g), es posible obtener un valor numérico que represente la relación entre *dominance* y Gm^2. Para este fin, se propone emplear el área del rectángulo formado por los vértices $(-1, 0)$,$(-1, g)$ y $(d, 0)$ (ver la Figura 4.2 para el caso de $d = 0$ y $g = 1$). El cálculo de dicha área rectangular, a la que denominaremos *índice de precisión balanceada* (*IBA, Index of Balanced Accuracy*), se expresa por medio de la siguiente ecuación:

$$IBA = (1 + dominance) \cdot Gm^2 \tag{4.2}$$

Al sustituir, en la ecuación anterior, *dominance* y Gm^2 por sus expresiones en términos de *tvp* y *tvn*, la función resultante provee detalles útiles para una mejor comprensión acerca de cómo IBA establece la relación entre ambas medidas.

$$IBA = (1 + (tvp - tvn)) \cdot tvp \cdot tvn \tag{4.3}$$

La propia naturaleza de IBA beneficia a aquellos clasificadores que presenten mejores resultados sobre la clase positiva. Con la finalidad de moderar la importancia de dominance sobre los resultados globales de clasificación, se añade un parámetro $0 \leq \alpha \leq 1$ tal y como muestra la siguiente expresión,

$$IBA_\alpha = (1 + \alpha \cdot (tvp - tvn)) \cdot tvp \cdot tvn \tag{4.4}$$

Cambios significativos en el comportamiento de IBA se obtienen para $\alpha \leq 0.5$. En este sentido, si $\alpha = 0$, la función IBA se reduce a Gm^2. Por otra parte, obsérvese que el primer factor de la Ecuación 4.3 se encuentra dentro del rango $[1 - \alpha, 1 + \alpha]$, por lo que dicha función puede ser vista como una ponderación de Gm^2.

IBA puede tomar valores entre 0 y +1, representando este último el valor del área del rectángulo más grande posible, el cual se corresponde con el punto $dominance = 0$ y $Gm^2 = 1$ y que resulta de la clasificación óptima. En la Figura 4.4, se muestra el valor de IBA como una función de tvn y tvp, mostrando que el máximo +1 ocurre para $tvp = tvn = 1$. Esto también podría demostrarse analíticamente mediante la optimización de la expresión matemática de IBA (ver Ecuación 4.4)

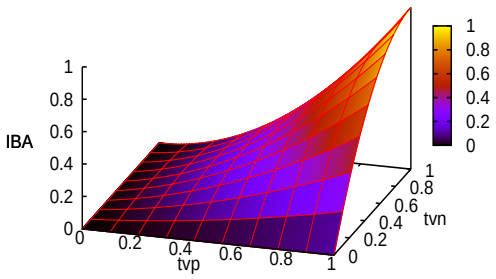
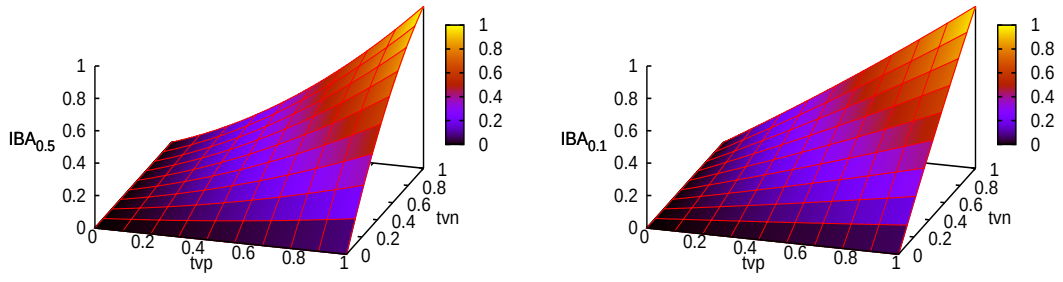


Figura 4.4: *Función IBA.*

Finalmente, es importante mencionar que la incorporación del parámetro α no afecta al rango de resultados de IBA, tal y como se puede observar en las Figuras 4.5(a) y 4.5(b).

4.2.4. Un ejemplo teórico

Con la finalidad de mostrar las diferencias de IBA_α con respecto a otras medidas de evaluación del clasificador, en esta sección se desarrolla un ejemplo artificial a partir de cuatro posibles resultados de clasificación. A pesar de que en el problema del desbalance se han empleado con éxito diversas métricas basadas en *precision/pureza*,



(a) $\text{IBA}_{0.5}$ (b) $\text{IBA}_{0.1}$

Figura 4.5: *Función* IBA_α.

en nuestro caso, éstas se omitirán debido a su naturaleza enfocada a evaluar únicamente el comportamiento del clasificador sobre la clase positiva.

Sea $f(\theta)$ un clasificador que depende de un conjunto de parámetros θ, de cuya optimización dependerá la capacidad del clasificador para diferenciar las dos clases de un problema desbalanceado con un ratio de 1:10. A partir de un conjunto T de entrenamiento, se obtienen cuatro posibles configuraciones (θ_1, θ_2, θ_3, θ_4) con sus correspondientes clasificadores $f(\theta_i)$, los cuales son evaluados sobre un conjunto V de validación con distintas métricas. La Tabla 4.1 muestra valores hipotéticos de *tvp* y *tvn* para cada configuración.

Como se puede observar, las configuraciones θ_1 y θ_4 corresponden a situaciones con comportamientos fuertemente sesgados, mientras que θ_2 y θ_3 muestran una menor diferencia entre los valores de *tvp* y *tvn*. El objetivo de este ejemplo es juzgar la elección de cada medida a partir de la valoración empírica de la utilidad de la configuración seleccionada.

De las cuatro configuraciones, *AUC* y *Acc* sugieren que las asociadas a los casos más sesgados θ_1 y θ_4 representan la mejor opción sobre el resto. Este comportamiento es aún más acentuado para *Acc*, la cual depende fuertemente de la tasa de aciertos sobre la clase mayoritaria. Por lo que respecta a *Gm*, esta medida propone alguna de las configuraciones moderadas θ_2 y θ_3, pero no es capaz de distinguir entre ellas e ignora que la clase minoritaria es usualmente la de mayor importancia. Asimismo, mientras que *Gm* no distingue entre ambas configuraciones, OP se inclina por θ_2 antes que por θ_3, como resultado de la influencia que ejerce *Acc* en el cálculo de dicha medida. Con respecto a IBA, la versión con $\alpha = 0.1$ sugiere que θ_3 es la mejor opción, la cual corresponde al caso moderado con el valor más alto de *tvp*.

Por otra parte, los resultados de IBA_1 e $IBA_{0.5}$ muestran la tendencia de IBA_α a favorecer altas tasas de *tvp*. En consecuencia, como se comentó en la Sección 4.2.3, la fuerte influencia de *dominance* justifica la necesidad de añadir el parámetro α para ponderar o disminuir la importancia de este término. Los ejemplos sintéticos

Tabla 4.1: *Ejemplo teórico en resultados de clasificación (tvp, tvn) hipotéticos, y valores de medidas de efectividad (los mejores resultados han sido resaltados).*

	tvp	*tvn*	*Acc*	*Gm*	*AUC*	*OP*	IBA_1	$IBA_{0.5}$	$IBA_{0.1}$
θ_1	0.550	0.950	**0.914**	0.723	**0.750**	0.647	0.314	0.418	0.502
θ_2	0.680	0.810	0.798	**0.742**	0.745	**0.711**	0.479	0.515	0.544
θ_3	0.810	0.680	0.692	**0.742**	0.745	0.605	0.622	0.587	**0.558**
θ_4	0.950	0.550	0.586	0.723	**0.750**	0.320	**0.732**	**0.627**	0.543

anteriores sugieren que valores pequeños de α (por ejemplo, $\alpha = 0.1$) consiguen corregir este efecto negativo. Por tal motivo, en las siguientes secciones, se empleará IBA_α con $\alpha = 0.1$

4.3. Análisis teórico de IBA

En la sección anterior, mediante un ejemplo sencillo, se mostró el comportamiento y las ventajas que presenta IBA_α al ser empleada para evaluar la efectividad de los clasificadores en entornos desbalanceados. En esta sección, se presentan dos estudios teóricos que aportan análisis complementarios sobre IBA_α y que corroboran su utilidad con respecto a otras métricas en conjuntos de datos desbalanceados de dos clases.

El primer estudio consiste en el cálculo de los coeficientes de correlación de Pearson entre cada par de medidas, a partir de diversas colecciones de datos sintéticos que simulan la salida de un clasificador. Estos datos han sido generados aleatoriamente considerando diversos niveles de desbalance. El objetivo es conocer cuán correlacionada está IBA_α con otras medidas que pueden ser aceptadas como buenas o malas opciones en problemas desbalanceados, así como las valoraciones que producen distintas medidas de efectividad.

El segundo estudio está enfocado a analizar la estabilidad de una métrica de evaluación ante diversos cambios sobre la matriz de confusión. En principio, una métrica robusta debe ser capaz de detectar cualquier cambio en la matriz de confusión, pues éste define un nuevo escenario de clasificación. Sin embargo, el no registrar dicho cambio puede implicar una ventaja o desventaja dependiendo del propósito del aprendizaje. Con relación a esto, se presenta una comparación entre métricas de efectividad en lo que se refiere a cuán sensibles son en presencia de cambios en la matriz.

4.3.1. Análisis por correlación de Pearson

Diversas investigaciones sobre medidas de efectividad sostiene que la similitud entre un par cualquiera puede ser evaluada a través de la correlación de Pearson (Ferri y otros, 2009, Tan y otros, 2002). En el presente estudio, se calcula esta correlación entre IBA y diversas medidas a partir de cinco colecciones que representan las salidas artificiales de un clasificador, las cuales fueron generadas de acuerdo al

método propuesto por Huang y Ling (2007). El objetivo es valorar la similitud de la nueva métrica de efectividad con respecto al resto de las métricas habitualmente empleadas en el ámbito de las distribuciones desbalanceadas.

En este trabajo se denomina salida correcta de un clasificador a una lista de talla n con valores entre 0 y 1, los cuales se definen para n muestras como la probabilidad $p_i = \frac{i}{n}$ de pertenecer a la clase positiva. A partir de esta lista *original* o *patrón*, se generó una serie de listas modificadas variando aleatoriamente cada probabilidad original p_i en el rango $[max(0, p_i - \varepsilon), min(1, p_i + \varepsilon)]$. Sin embargo, a diferencia del trabajo de Huang y Ling, la lista *patrón* se definió considerando un nivel particular de desbalance en la asignación de las probabilidades originales. Así, para un ratio de desbalance de 1:3, al primer 75 % de muestras de la lista se le asignaron probabilidades en el rango $[0, 0.5]$ (clase negativa), mientras que el resto (25 %) fueron asociadas a probabilidades en el rango $(0.5, 1]$ (clase positiva). Por lo que respecta a las listas modificadas, éstas fueron generadas haciendo fluctuar aleatoriamente las probabilidades originales p de las muestras de la clase negativa en el rango $[max(0, p - \varepsilon_n), min(1, p + \varepsilon_n)]$, mientras que las probabilidades p de las muestras positivas fueron generadas en el rango $[max(0, p - \varepsilon_p), min(1, p + \varepsilon_p)]$.

El uso dos términos de distorsión, ε_n y ε_p para la clase negativa y positiva, respectivamente, tiene por objetivo simular diferentes escenarios de un clasificador con aprendizaje sesgado. Es decir, para $\varepsilon_n > \varepsilon_p$, una mayor proporción de muestras de la clase negativa será erróneamente clasificada, mientras que cuando $\varepsilon_n < \varepsilon_p$ la clase más afectada corresponderá a la positiva. En la Tabla 4.2, se muestra un ejemplo de una lista *patrón* para 12 muestras y un ratio de desbalance de 1:3, así como dos listas *modificadas* obtenidas de $(\varepsilon_n = 0.3, \varepsilon_p = 0)$ y $(\varepsilon_n = 0, \varepsilon_p = 0.3)$, respectivamente. En el caso de la primera lista modificada, P1, las salidas simulan un clasificador cuyo aprendizaje está enfocado a la clase positiva, mientras que en la segunda lista, P2, los resultados representan un clasificador sesgado que favorece a la clase negativa.

Tabla 4.2: *Un ejemplo de una lista original (T) y dos listas alteradas (P1 y P2) para un desbalance de 1:3. Los elementos en negrita representan las muestras etiquetadas erróneamente.*

T	0.06	0.11	0.17	0.22	0.28	0.33	0.39	0.44	0.5	0.67	0.83	1.0
	−	−	−	−	−	−	−	−	−	+	+	+
P1	0.0	0.0	0.23	**0.52**	0.26	0.49	**0.55**	0.24	**0.61**	0.67	0.83	1.0
	−	−	−	+	−	−	+	−	+	+	+	+
P2	0.06	0.11	0.17	0.22	0.28	0.33	0.39	0.44	0.5	**0.45**	0.57	1.0
	−	−	−	−	−	−	−	−	−	−	+	+

Se generaron cinco colecciones de salidas de clasificación considerando cinco niveles diferentes de desbalance, los cuales son expresados en términos del porcentaje de muestras de la clase positiva: 5 %, 10 %, 15 %, 20 % y 25 %. Cada una las colecciones la conforman un total de 130 listas: 10 listas por cada una de las 13 combinaciones

de términos de distorsión desde $(\varepsilon_n = 0.6, \varepsilon_p = 0)$ hasta $(\varepsilon_n = 0, \varepsilon_p = 0.6)$ con saltos de $(-0.05, 0.05)$ y cumpliendo $\varepsilon_n + \varepsilon_p = 0.6$.

A partir de cada una de las colecciones de datos, se construyó una matriz independiente de correlación entre todos los pares de métricas de efectividad empleadas en este análisis: *tvp*, *tvn*, *Acc*, *Gm*, *Op* e $IBA_{0.1}$. Con la finalidad de facilitar la comprensión de los resultados, los coeficientes de correlación han sido representados en diagramas como los de la Figura 4.6. Cada diagrama está vinculado a una medida específica y muestra los coeficientes de correlación (eje X) y los diferentes porcentajes de muestras de la clase positiva (eje Y).

Los diagramas de la Figura 4.6 permiten obtener diversas conclusiones interesantes relacionadas con la medida $IBA_{0.1}$. Partiendo de la Figura 4.6(a), $IBA_{0.1}$ muestra la correlación más pequeña con respecto a *Acc*, lo cual parece conveniente dado el comportamiento sesgado de esta medida al evaluar clasificadores en entornos desbalanceados. En términos absolutos, $IBA_{0.1}$ presenta coeficientes de correlación ligeramente más bajos que *AUC* y *Gm*, los cuales inclusive son positivos. Sin embargo, a pesar de esta diferencia, $IBA_{0.1}$ guarda una relación directa con ambas métricas al tener altas correlaciones positivas próximas a 1 (ver Figuras 4.6(b) y 4.6(c)). Esto sugiere que la nueva métrica puede resultar idónea para evaluar la efectividad de los clasificadores en datos con distribuciones no balanceadas.

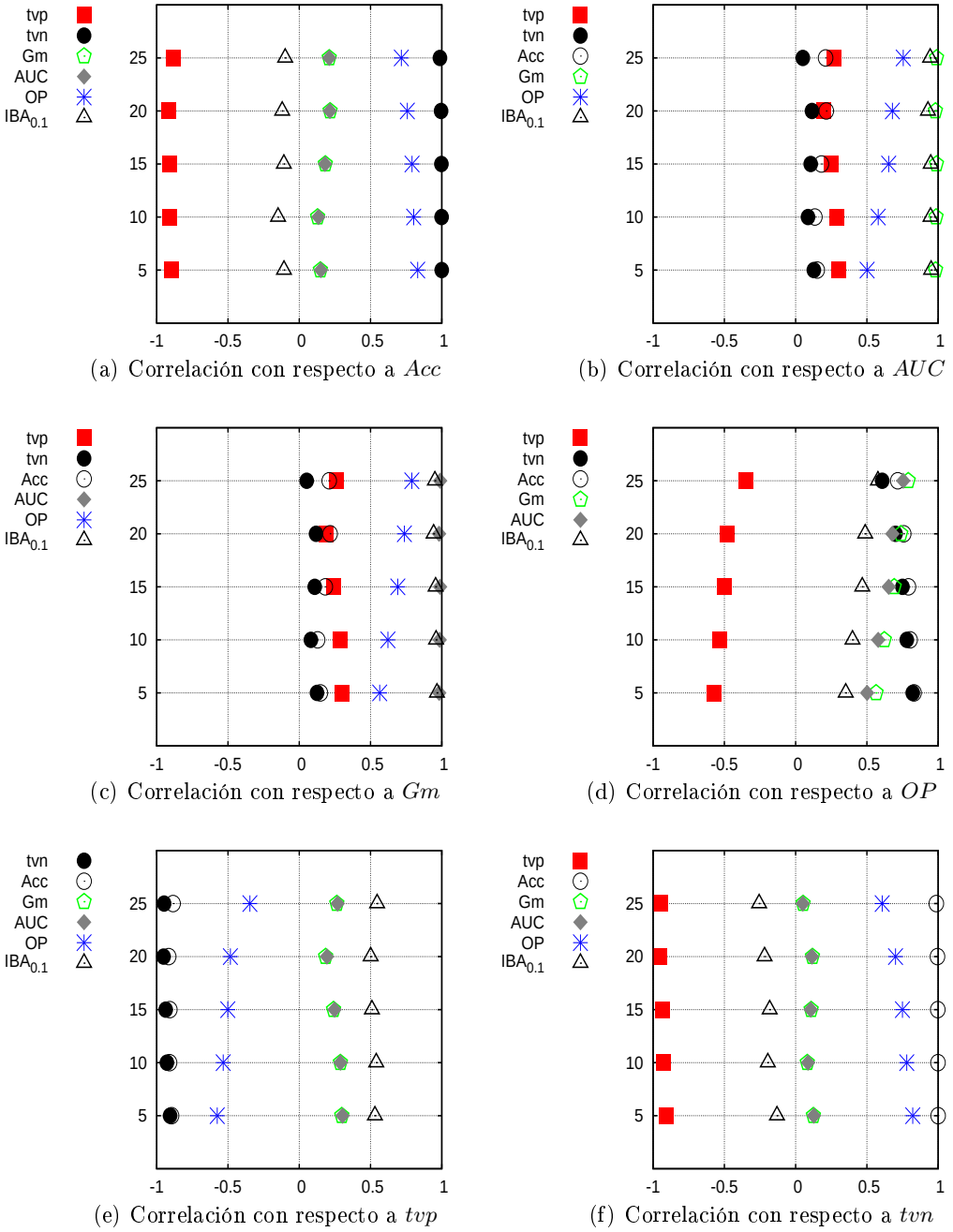
Con respecto a las métricas básicas *tvp* y *tvn*, en las Figuras 4.6(d) y 4.6(e), respectivamente, $IBA_{0.1}$ aparece más positivamente correlacionada con *tvp*. Esta comparación parece favorable a $IBA_{0.1}$, pues *tvp* representa la efectividad sobre la clase positiva, que es generalmente la clase más importante. En sentido opuesto, $IBA_{0.1}$ presenta correlaciones negativas y bajas con *tvn*, que es la tasa de exactitud sobre la clase mayoritaria y que, usualmente, conduce a resultados sesgados de clasificación. Pese a que *AUC* y *Gmean* muestran bajas correlaciones con respecto a *tvn*, sus coeficientes son aún positivos.

A pesar de que *OP* se ha propuesto como una métrica especialmente apropiada para entornos desbalanceados, estos análisis indican que presenta una fuerte correlación con *Acc* y *tvn*. Este hecho se confirma al analizar la Ecuación 3.13, en la que se aprecia de forma clara la dependencia directa de *OP* del valor de *Acc*.

4.3.2. Análisis de las propiedades invariantes de las medidas

Este segundo análisis está enfocado a estudiar el comportamiento de una medida en base a sus *propiedades invariantes* con respecto a cuatro cambios básicos en la matriz de confusión: *i)* inversión de la matriz, *ii)* modificación en el número de verdaderos negativos, *iii)* modificación en el número de falsos positivos y, *iv)* escalado de columnas (Tan y otros, 2002, Sokolova y otros, 2006, Sokolova, 2006, Sokolova y Lapalme, 2007, 2009).

Una vez que se ha producido un cambio en la matriz de confusión, la invariancia de una medida informa que ésta no es capaz de distinguir la nueva configuración de la anterior. En general, en esta tesis, se asume que una métrica de efectividad robusta debe ser capaz de detectar cada una de las transformaciones sobre la matriz



(a) Correlación con respecto a *Acc*

(b) Correlación con respecto a *AUC*

(c) Correlación con respecto a *Gm*

(d) Correlación con respecto a *OP*

(e) Correlación con respecto a *tvp*

(f) Correlación con respecto a *tvn*

Figura 4.6: *Diagramas de coeficientes de correlación de Pearson entre las medidas de efectividad consideradas.*

de confusión.

En esta sección, se emplean las cuatro propiedades invariantes propuestas por Sokolova y otros (2006) para demostrar la sensibilidad de $IBA_{0.1}$ en comparación

al resto de las métricas. Para la matriz de confusión mostrada en la Tabla 3.1, estas cuatro propiedades pueden ser definidas de la siguiente manera:

p1 Invariante a intercambios entre las cantidades de verdaderos positivos (vp) con verdaderos negativos (vn) y de falsos negativos (fn) con falsos positivos (fp).

p2 Invariante a cambios en la cantidad de vn, mientras que el resto de entradas de la matriz se mantienen constantes.

p3 Invariante a cambios en el número de fp, mientras que el resto de las entradas de la matriz se mantienen constantes.

p4 Invariante a escalados en los índices de efectividad por clases: $vp \to k_1 vp$, $vn \to k_2 vn$, $fp \to k_1 fp$, $fn \to k_2 fn$, para $k_1 k_2 > 0$.

La Tabla 4.3 muestra las medidas consideradas en este estudio y el cumplimiento o no de las propiedades invariantes antes enumeradas. Como se puede observar, la tasa de exactitud global (Acc) no es capaz de distinguir vp de vn y fn de fp (p1), de modo que no detecta la asimetría en los resultados de clasificación, mientras que Gm, AUC y OP presentan esta misma propiedad invariante. Por lo que respecta a *precision* y la medida f_1, ambas no son capaces de detectar cambios en vn (p2), ya que están enfocadas a evaluar la efectividad sobre la clase positiva. Finalmente, IBA_α es capaz de detectar los cuatro tipos de transformaciones analizadas y, por tanto, se puede concluir que esta métrica es más sensible que el resto de las medidas con respecto a las cuatro propiedades definidas.

Tabla 4.3: *Propiedades invariantes. '+' y '–' indican la invariancia y variancia, respectivamente.*

Métrica \ Propiedades	p1	p2	p3	p4
Acc	+	–	–	–
tvp	–	+	+	–
tvn	–	–	–	–
$Precision$	–	+	–	+
f_1	–	+	–	–
Gm	+	–	–	–
AUC^*	+	–	–	–
OP	+	–	–	–
IBA_α	–	–	–	–

* Valido para la situación de un único resultado de clasificación.

4.4. Experimentos en bases de datos reales

Con el objetivo de validar la utilidad y la consistencia de IBA_α, se llevó a cabo un experimento de clasificación para comparar $IBA_{0.1}$ con otras métricas, tales

Tabla 4.4: *Conjuntos de datos usados en los experimentos.*

	Muestras Positivas	Muestras Negativas	Clases	Clase Mayoritaria
Breast	81	196	2	1
Ecoli	35	301	8	1,2,3,5,6,7,8
German	300	700	2	1
Glass	17	197	9	1,2,4,5,6,7,8,9
Haberman	81	225	2	1
Laryngeal_2	53	639	2	1
Letter_a	789	19211	26	$2,\ldots 26$
Phoneme	1586	3818	2	1
Optidigts	554	5066	10	1,2,3,4,5,6,7,8,10
Pendigits	1055	9937	10	1,2,3,4,5,7,8,9,10
Pima	268	500	2	1
Satimage	626	5809	7	1,2,3,5,6,7
Scrapie	531	2582	2	1
Segmentation	330	1980	6	1,2,3,4,6
Spambase	1813	2788	2	1
Vehicle	212	634	4	2,3,4
Yeast	429	1055	10	1,3,4,5,6,7,8,9,10

como *Acc*, *Gm*, *AUC* y *OP*. Para ello, se empleó un total de 17 bases de datos reales, de las que la mayoría fue obtenida del UCI Machine Learning Database Repository (Asuncion y Newman, 2007) y el resto de una biblioteca privada (`http://www.vision.uji.es/~sanchez/Databases/`). La lista la conforman 8 bases de datos de dos clases y 9 de naturaleza multiclase, siendo estas últimas convertidas a conjuntos de dos clases manteniendo una clase original (minoritaria) y uniendo el resto para formar la clase mayoritaria. Por ejemplo, en el conjunto de datos Vehicle, las muestras de las clases 2, 3 y 4 fueron unidas para formar la clase mayoritaria (negativa), mientras que la clase 1 fue elegida como la clase minoritaria (positiva). La Tabla 4.4 ofrece detalles adicionales para cada una de las bases de datos, como la cantidad de muestras de las clases y la clase elegida como minoritaria.

De los conjuntos de datos indicados en la Tabla 4.4, se han incluido para este análisis los resultados de cinco casos que representan diversas situaciones de desbalance. El ratio minoritaria/mayoritara para cada una de estas bases de datos es: Breast (0.41), Vehicle (0.33), Glass (0.09), Satimage (0.11) y Laryngeal_2 (0.08). El resto de resultados se pueden encontrar en el Anexo A de esta tesis.

Para cada uno de estos conjuntos de datos, las métricas de efectividad arriba mencionadas se calcularon a partir de 10 particiones de validación cruzada con 5 repeticiones, donde los conjuntos resultantes fueron usados para entrenar/evaluar diversos clasificadores: la regla del vecino más cercano (1−NN) (Aha y otros, 1991), una red neuronal del tipo perceptron multicapa (MLP), una máquina de vectores soporte (SVM) (Platt, 1998), un árbol de decisión (J48) (Quinlan, 1993), el algoritmo naïve Bayes (NBC) (Duda y otros, 2001) y una red de funciones de base radial (RBF). Todos estos algoritmos fueron tomados del kit de herramientas Weka (Witten

y Frank, 2005), empleando para todos ellos la correspondiente configuración por defecto.

4.4.1. Análisis de resultados experimentales

La Tabla 4.5 muestra los resultados de los seis clasificadores en términos de las cinco métricas de efectividad antes referidas, así como de las tasas de exactitud por clase (*tvp* y *tvn*).

El estudio presentado en esta sección tiene por objetivo analizar la elección del mejor clasificador que se deriva de cada una de las medidas y, en particular, la decisión resultante de $IBA_{0.1}$.

Obsérvese que, como se esperaba, la tasa de exactitud global (*Acc*) resulta inadecuada para evaluar la efectividad del clasificador en entornos desbalanceados, mientras que $IBA_{0.1}$, en todas las bases de datos, establece como mejor modelo el clasificador que presenta una alta tasa de aciertos sobre la clase positiva (*tvp*). El resto de las métricas, por lo general, suelen verse afectadas por las altas tasas de aciertos en la clase negativa (*tvn*), minimizando con esto la relevancia de *tvp*.

Centrando el análisis sobre cada una de las bases de datos, se puede observar que, en el caso de Breast, los clasificadores 1−NN y NBC obtienen resultados muy similares en términos de *tvp* y *tvn*. No obstante, OP e $IBA_{0.1}$ sugieren como mejor clasificador el 1−NN debido a que sus tasas de efectividad por clase están más balanceadas. Por lo que respecta al resto de las métricas, la tendencia hacia NBC está marcada por su menor error global, producto de la fuerte influencia que ejerce la clase mayoritaria.

Un comportamiento similar se puede observar en el conjunto de datos Vehicle, donde $IBA_{0.1}$ y OP destacan a SVM como el mejor modelo de clasificación, mientras que el resto de métricas sugieren a MLP. Una vez más, esta diferencia se puede explicar al calcular los valores de *dominance*, en donde, para el caso de SVM, indican que las tasas de acierto entre ambas clases están más balanceadas que las obtenidas por MLP. No obstante, $IBA_{0.1}$ obtiene para MLP prácticamente el mismo valor que para SVM.

Los resultados sobre el conjunto Glass revelan que la clase positiva se ve fuertemente afectada por el desbalance, lo cual se ve reflejado en los resultados de la mayoría de los clasificadores. $IBA_{0.1}$ opta por NBC, el cual presenta el valor más alto de *tvp* y, además, éste es mayor que *tvn*. Por el contrario, Gm, AUC y OP se inclinan por J48, caso en el que la tasa de aciertos sobre la clase positiva es muy baja.

Satimage representa un caso evidente, en el que todas las métricas (excepto *Acc*) dan como mejor clasificador a NBC. Puede observarse que ambas tasas de acierto por clase, *tvp* y *tvn*, son suficientemente altas y se encuentran bastante balanceadas. Finalmente, en $Laryngeal_2$, Gm, AUC e $IBA_{0.1}$ prefieren a NBC, el cual claramente presenta el mejor rendimiento. Por lo que se refiere a OP, esta medida se inclina por MLP debido a que se ve afectada por la tasa de aciertos global, la cual coincide en esta decisión.

Tabla 4.5: *Resultados experimentales (en negrita mejores valores para cada métrica).*

Breast							
	tvp	*tvn*	*Acc*	*Gm*	*AUC*	*OP*	$IBA_{0.1}$
1−NN	0.454	0.761	0.671	0.588	0.608	**0.419**	**0.335**
MLP	0.368	0.845	0.705	0.558	0.606	0.312	0.296
SVM	0.328	0.864	0.708	0.533	0.596	0.258	0.269
NBC	0.402	0.865	**0.730**	**0.590**	**0.634**	0.365	0.332
J48	0.321	0.891	0.724	0.535	0.606	0.254	0.270
RBF	0.390	0.867	0.728	0.582	0.629	0.348	0.322
Vehicle							
	tvp	*tvn*	*Acc*	*Gm*	*AUC*	*OP*	$IBA_{0.1}$
1−NN	0.458	0.844	0.747	0.622	0.651	0.451	0.372
MLP	0.617	0.909	**0.835**	**0.749**	**0.763**	0.644	0.544
SVM	0.645	0.864	0.809	0.747	0.755	**0.664**	**0.545**
NBC	0.614	0.718	0.692	0.664	0.666	0.614	0.436
J48	0.509	0.862	0.774	0.663	0.686	0.517	0.424
RBF	0.397	0.870	0.752	0.588	0.634	0.378	0.329
Satimage							
	tvp	*tvn*	*Acc*	*Gm*	*AUC*	*OP*	$IBA_{0.1}$
1−NN	0.705	0.960	**0.935**	0.823	0.833	0.782	0.660
MLP	0.637	0.966	0.934	0.785	0.802	0.729	0.596
SVM	0.000	1.000	0.903	0.000	0.500	-0.097	0.000
NBC	0.870	0.815	0.820	**0.842**	**0.843**	**0.788**	**0.713**
J48	0.550	0.959	0.919	0.726	0.755	0.648	0.506
RBF	0.000	1.000	0.903	0.000	0.500	-0.097	0.000
Glass							
	tvp	*tvn*	*Acc*	*Gm*	*AUC*	*OP*	$IBA_{0.1}$
1−NN	0.318	0.948	0.898	0.549	0.633	0.400	0.282
MLP	0.141	0.969	0.903	0.370	0.555	0.157	0.125
SVM	0.000	0.991	0.912	0.000	0.495	-0.088	0.000
NBC	0.753	0.444	0.468	0.578	0.598	0.210	**0.344**
J48	0.353	0.963	0.915	**0.583**	**0.658**	**0.451**	0.319
RBF	0.000	0.999	**0.920**	0.000	0.499	-0.080	0.000
$Laryngeal_2$							
	tvp	*tvn*	*Acc*	*Gm*	*AUC*	*OP*	$IBA_{0.1}$
1−NN	0.694	0.970	0.949	0.821	0.832	0.783	0.655
MLP	0.770	0.979	**0.963**	0.868	0.875	**0.844**	0.738
SVM	0.672	0.942	0.922	0.796	0.807	0.754	0.616
NBC	0.943	0.844	0.851	**0.892**	**0.894**	0.796	**0.804**
J48	0.638	0.980	0.954	0.791	0.809	0.742	0.604
RBF	0.558	0.985	0.952	0.742	0.772	0.676	0.526

4.5. Conclusiones

La mayoría de las métricas de evaluación de la efectividad de un clasificador empleadas en entornos desbalanceados son combinaciones de tasas de aciertos (errores) por clase. Sin embargo, ninguna de estas métricas, a excepción de OP, tiene en cuenta la información acerca del grado de balance que existe entre dichas tasas. Por tanto, estas métricas no reflejan cuán dominante es la tasa de aciertos de la clase más beneficiada por el clasificador.

Con la finalidad de poder solventar esta limitación, se ha propuesto un nuevo método de evaluación que permite medir la efectividad de un sistema de clasificación en problemas desbalanceados de dos clases. Esta métrica está definida como la relación entre una métrica global no sesgada (media geométrica) y un índice simple que refleja cuán balanceadas están las tasas de acierto por clase (*dominance*). Valores altos de la nueva métrica IBA_α se obtienen cuando las tasas de aciertos de ambas clases son altas y están significativamente balanceadas. A la vez, esta métrica intenta favorecer a aquellos clasificadores que presentan mejores resultados sobre la clase positiva (generalmente, la más importante). En los casos evidentes, IBA_α coincide con el resto de medidas en cuanto a la selección del mejor clasificador.

Los dos estudios teóricos aquí presentados han mostrado los beneficios de la nueva métrica con respecto a otras medidas tradicionales. En este sentido, los análisis de correlación han demostrado que IBA_α tiene una fuerte correlación con AUC y Gm, dos de las métricas que son aceptadas, por lo general, como buenas funciones de evaluación para problemas desbalanceados. Adicionalmente, dicha medida está muy correlacionada con la tasa de aciertos de la clase positiva y poco correlacionada con la tasa de precisión global (accuracy).

Por otra parte, el estudio de propiedades invariantes ha revelado que IBA_α es más sensible a cambios en las entradas de la matriz de confusión que el resto de medidas consideradas en esta tesis.

Por último, los experimentos empleando conjuntos de datos reales han permitido constatar la robustez y la consistencia de IBA_α, evidenciando además ciertos beneficios con respecto al resto de métricas empleadas en escenarios desbalanceados de dos clases.

Capítulo 5

Técnicas para tratar el problema del desbalance

El presente capítulo se centra en el estudio de diversas metodologías para tratar el problema del desbalance. La primera parte del capítulo introduce nuevos esquemas de remuestreo y analiza su comportamiento mediante un extenso banco de experimentos, comparándolos con los principales algoritmos presentes en la literatura y, al mismo tiempo, utilizando clasificadores de naturaleza diversa. Una segunda parte del capítulo compara las estrategias de submuestreo y sobremuestreo, tratando de analizar la efectividad de cada una de ellas en diferentes contextos de clasificación según el ratio de desbalance. Por último, se presenta también una solución a nivel de algoritmos mediante la definición de una distancia ponderada que intenta potenciar la influencia de la clase minoritaria en el proceso de toma de decisiones.

5.1. Introducción

La búsqueda/propuesta de técnicas para tratar el problema del desbalance en el marco de la clasificación automática es una de las líneas de investigación más ampliamente estudiada dentro de este ámbito[1]. De acuerdo a su naturaleza, éstas pueden ser catalogadas como *soluciones a nivel de datos* y *soluciones a nivel de algoritmos*.

Las soluciones a nivel de datos tienen por objetivo remuestrear el conjunto de datos original hasta lograr un cierto balance o equilibrio entre clases, ya sea por la eliminación/selección de muestras de la clase mayoritaria (submuestreo) o por la adición de (nuevas) muestras en la clase minoritaria (sobremuestreo). Por lo que respecta a las soluciones a nivel de algoritmos, las propuestas dentro de esta categoría tratan de adaptar el propio sistema de clasificación, ya sea mediante una modificación del mismo o por el empleo de una matriz de costes.

Los puntos de vista acerca de los beneficios y las desventajas que presenta cada estrategia son muy diversos, por lo que no existe un claro consenso acerca de qué solución puede ser la más apropiada. Los algoritmos basados en remuestreo son criticados principalmente por la alteración que producen sobre la distribución original de los datos. En el caso de las soluciones a nivel de algoritmos, su posible desventaja radica en la difícil adaptación de dichas soluciones a otros problemas desbalanceados.

[1] Otras líneas de investigación están orientadas a evaluar la efectividad del clasificador y el estudio de la complejidad de los datos.

Dentro de las soluciones a nivel de datos, los algoritmos de sobremuestreo resultan fuertemente criticados no solo por la modificación artificial del conjunto de datos, sino también por el incremento en el coste computacional derivado del aumento de la talla de la clase minoritaria. Por su parte, la reducción de la clase mayoritaria (submuestreo) pudiera eliminar información útil para la etapa de aprendizaje del clasificador. Con respecto a esto, algunas propuestas mencionan que la desventaja de los algoritmos de submuestreo podría solventarse al emplear algoritmos *dirigidos*, cuyo objetivo fuese eliminar información "no útil".

En el presente capítulo, se propone emplear dos algoritmos de edición y condensado, ampliamente utilizados para reducir la talla del conjunto de datos, como algoritmos de submuestreo *dirigidos* que permiten la eliminación de muestras erróneamente etiquetadas, muestras cercanas a la frontera de decisión y muestras redundantes.

Por último, se incorpora una adaptación a la distancia Euclídea en la regla del vecino más cercano con la finalidad de adecuar el algoritmo de clasificación al problema del desbalance, sin que ello implique una alteración sobre la distribución original de los datos.

5.2. Soluciones a nivel de datos: algoritmos de submuestreo

En el Capítulo 3, se han presentado algunos de los algoritmos de submuestreo más populares, entre los que se encuentran el submuestreo aleatorio y el algoritmo OSS de Kubat y Matwin (1997). Debido a las desventajas que presentan los algoritmos inspirados en el esquema general del OSS, aquí se propone una nueva metodología de submuestreo que emplea dos algoritmos basados en la regla del vecino más cercano: el conocido método de edición de Wilson (1972) y una variante del algoritmo de condensado de Ritter y otros (1975), llamado selectivo modificado (Barandela y otros, 2005a). El primer algoritmo reduce el conjunto de entrenamiento original por la eliminación de muestras atípicas o mal etiquetadas, mientras que el segundo, basado en la idea de un subconjunto consistente, minimiza el número de objetos almacenados garantizando una aproximación satisfactoria a las fronteras de decisión.

5.2.1. Edición de Wilson

Este algoritmo corresponde a la primera propuesta para editar o filtrar las muestras de entrenamiento en la regla del vecino más cercano (NN). De forma sencilla, el algoritmo de Wilson consiste en emplear el clasificador $k-$NN para estimar la etiqueta de clase de todas las muestras del conjunto de entrenamiento original, y descartar aquellas cuya etiqueta de clase no coincida con la estimada por la mayoría de sus k vecinos. El Algoritmo 3 muestra un esquema algorítmico de este método.

El Algoritmo 4 corresponde a una versión modificada de la edición de Wilson para editar únicamente la clase mayoritaria. Como se puede observar, dicho cambio

no implica una modificación radical del algoritmo de Wilson, sino que simplemente proporciona como salida un subconjunto del original que contiene todas las muestras positivas y un (posiblemente) reducido número de muestras negativas.

Algoritmo 3 Edición de Wilson (EW)

Entrada:
T // Conjunto de entrenamiento original
k // Número de vecinos más cercanos
Salida:
S // Conjunto editado
Inicio:
$S \leftarrow T$
Para $\mathbf{x}_i$ en T **hacer** // $\mathbf{x}_i$ Muestra de entrenamiento
$P \leftarrow T - \{x_i\}$
Encontrar los $k-$NN de $\mathbf{x}_i$ en P
Descartar $\mathbf{x}_i$ de S si la mayoría de los vecinos más cercanos son de clase diferente
Fin Para
Fin

Algoritmo 4 Edición de Wilson sobre clase mayoritaria (EW$^-$)

Entrada:
T // Conjunto de entrenamiento original
k // Número de vecinos más cercanos
Salida:
S // Conjunto editado
Inicio:
$S \leftarrow T$
Para $\mathbf{x}n_i$ en T **hacer** // $\mathbf{x}n_i$ muestra de entrenamiento de clase mayoritaria
$P \leftarrow T - \{\mathbf{x}n_i\}$
Encontrar los $k-$NN de $\mathbf{x}n_i$ en P
Descartar $\mathbf{x}n_i$ de S si la mayoría de los k vecinos más cercanos son de clase diferente a la de $\mathbf{x}n_i$
Fin Para
Fin

5.2.2. Subconjunto selectivo modificado

Como resultado del algoritmo de condensando de Hart (1968), cada muestra del conjunto original de entrenamiento (T) estará más próxima a una muestra del subconjunto condensado de su misma clase que a cualquier otra de clase distinta del propio subconjunto. Esta idea es extendida por Ritter y otros (1975) a través de la introducción de una condición de consistencia más fuerte, mediante la cual cada muestra de T deberá estar más próxima a otra de su misma clase en el subconjun-

to condensado (en (Ritter y otros, 1975), Subconjunto Selectivo) que a cualquier instancia de clase diferente en T.

Barandela y otros (2001, 2005a) proponen una modificación al esquema de Ritter y otros en el algoritmo denominado *Subconjunto Selectivo Modificado* (SSM). Éste intenta obtener una mejor aproximación a las fronteras de decisión, a través de una función que da preferencia a las muestras más próximas a las fronteras definidas por la regla del vecino más cercano. El criterio empleado para medir la proximidad de una muestra a la frontera de decisión es la distancia más cercana a una muestra con etiqueta de clase distinta.

Al igual que la propuesta de Ritter y otros, el método SSM puede seleccionar instancias de forma independiente por clase. El Algoritmo 5 muestra el esquema de solución para un problema de dos clases, que permite obtener un subconjunto selectivo para una clase determinada (para este ejemplo, la clase 1). Esta característica del método admite su uso como metodología de submuestreo, ya que da la posibilidad de seleccionar la clase a reducir. En nuestro caso, el algoritmo SSM sólo se aplicará sobre la clase mayoritaria (SSM^{-}).

Algoritmo 5 Subconjunto Selectivo Modificado (SSM)

Entrada:
T // Conjunto de entrenamiento original
Salida:
SSM // Subconjunto selectivo de T
Inicio:
$S \leftarrow T$
$C \leftarrow T$
$D_i \leftarrow min\ d(\mathbf{x}_i, \mathbf{x}_j),\ clase(\mathbf{x}_i) \neq clase(\mathbf{x}_j),\ \forall \mathbf{x}_i \mathbf{x}_j \in T$
$\mathbf{y}_i \leftarrow argmin\ d(\mathbf{x}_i, \mathbf{y}_j),\ clase(\mathbf{x}_i) \neq clase(\mathbf{y}_j),\ \forall \mathbf{y}_j \in T$
$\mathbf{v}_i \leftarrow \{\mathbf{x}_j \in T | clase(\mathbf{x}_i) = clase(\mathbf{x}_j), d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_i, \mathbf{y}_i)\}$
mientras $C \neq \emptyset$ **hacer**
$\mathbf{x}_k \leftarrow argmin\ D_i$
$C \leftarrow C - \{\mathbf{x}_k\}$
$S_k \leftarrow \{\mathbf{x}_i \in S | \mathbf{x}_k \in v_i\}$
Si $S_k \cap S \neq \emptyset$ **entonces**
$SSM \leftarrow SSM \cup \{\mathbf{x}_k\}$
$S \leftarrow S - S_k$
Fin Si
Fin mientras
Fin

5.2.3. Propuesta de un algoritmo de submuestreo

Teniendo como base la idea de Kubat y Matwin, se propone submuestrear el conjunto de datos original aplicando el algoritmo de edición de Wilson y, posteriormente, el algoritmo SSM (ver Algoritmo 6).

A diferencia del algoritmo OSS, el primer paso en la metodología propuesta consiste en editar el conjunto de entrenamiento original (ver Figura 5.1(a)) para eliminar muestras de la clase mayoritaria erróneamente etiquetadas o atípicas. Posteriormente, a partir del conjunto resultante (ver Figura 5.1(b)), se obtiene un subconjunto condensado por medio de SSM$^-$. El empleo de ambos algoritmos, en este orden, tiene por objetivo evitar que en el proceso final puedan existir muestras poco confiables (mal etiquetadas o en zonas de confusión). Idealmente, el subconjunto reducido estará formado por todas las muestras de la clase minoritaria y un selecto grupo de muestras de la clase mayoritaria próximas a la frontera de decisión.

Algoritmo 6 Submuestreo por edición y condensado

Entrada:
T // Conjunto de entrenamiento original
k // Número de vecinos más cercanos
Salida:
S // Conjunto de entrenamiento editado-condensado
Inicio:
$E \leftarrow EW^-(T, k)$
$S \leftarrow SSM^-(E)$
Fin

Por otra parte, Tomek (1976a) demuestra que la edición del conjunto de datos de forma repetitiva implica una mejora en las tasas de efectividad del clasificador. Por ello, esta idea se incorpora al proceso de submuestreo, donde se edita repetidamente el conjunto de datos hasta que no se eliminen más muestras de la clase mayoritaria. Posteriormente, cada conjunto resultante (editado) será condensado mediante el algoritmo SSM$^-$.

5.2.4. Experimentos

El algoritmo de submuestreo aquí propuesto ha sido evaluado empleando los 17 conjuntos de datos desbalanceados descritos en el Capítulo 4. Asimismo, los resultados mostrados corresponden a los promedios obtenidos de repetir 5 veces una validación cruzada de 10 particiones.

La metodología propuesta se compara con diversas técnicas de remuestreo ampliamente utilizadas en el problema del desbalance. Una de éstas hace referencia al algoritmo OSS de Kubat y Matwin, sobre el que resulta necesario mencionar dos cuestiones que posteriormente serán tratadas con más detalle en el Anexo B.

1. Kubat y Matwin mencionan que el procedimiento de submuestreo denominado OSS se deriva de la aplicación del algoritmo de condensado de Hart seguido de la eliminación de muestras de la clase negativa que participan en los links de Tomek. Sin embargo, diversos artículos que emplean este esquema indican que consiste en la aplicación de los links de Tomek seguido del algoritmo

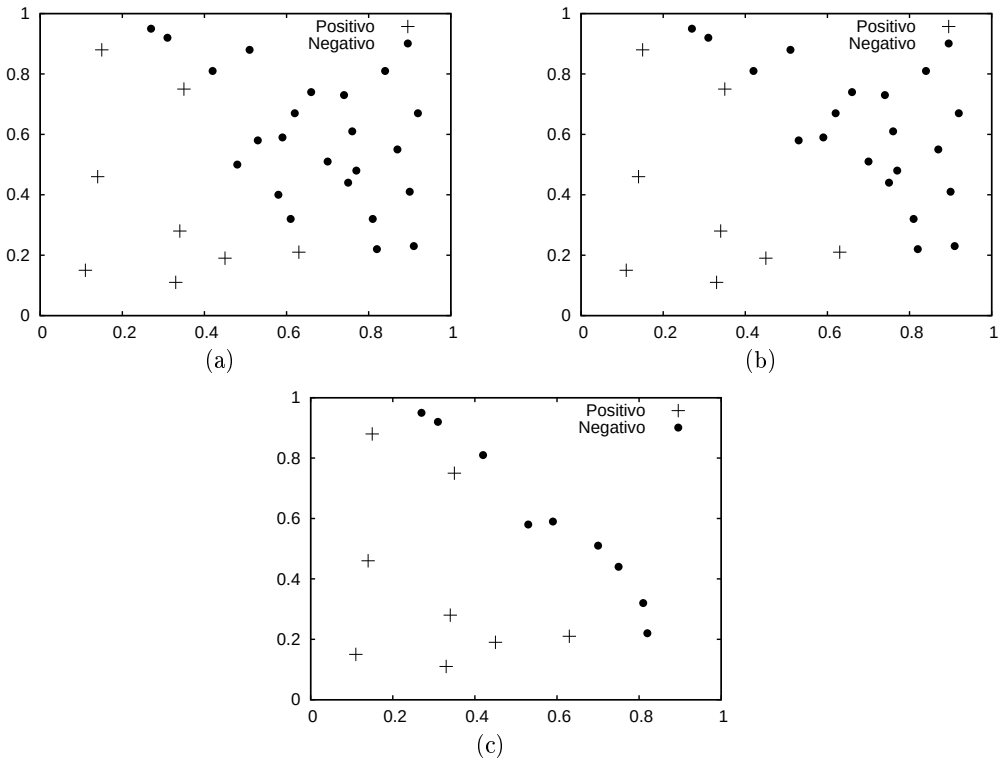


Figura 5.1: *Submuestreo por EW⁻ + SSM⁻: (a) Conjunto de entrenamiento original, (b) Conjunto de datos resultado de EW⁻, (c) Conjunto de datos después de EW⁻ + SSM⁻.*

de condensado de Hart (Batista y otros, 2004, García y Herrera, 2009, He y Garcia, 2009, Hulse y otros, 2007, Liao, 2008).

2. Aún cuando el artículo menciona que el algoritmo OSS emplea los links de Tomek para eliminar muestras que se encuentran en la frontera de decisión, la descripción dada por Kubat y Matwin no se corresponde con la que se proporciona en el trabajo original de Tomek (1976b).

En concreto, las técnicas de remuestreo que se han utilizado en los experimentos corresponden a 11 esquemas para reducir la talla de la clase mayoritaria:

1. Subconjunto selectivo modificado (SSM$^-$).
2. Edición de Wilson iterativo (EW$^-$).
3. Edición de Wilson en combinación con subconjunto selectivo modificado (EW$^-$ + SSM$^-$).
4. Submuestreo aleatorio (RUS).
5. Condensado de Hart (CNN$^-$).

6. Links de Tomek según la versión original (TMK_1).

7. Links de Tomek según la versión de Kubat y Matwin (TMK_2).

8. Condensado de Hart seguido de la versión original de los links de Tomek (CNN^{-}+TMK_1).

9. Condensado de Hart seguido de los links de Tomek según la versión de Kubat y Matwin (CNN^{-}+TMK_2).

10. Links de Tomek en su versión original seguido del condensado de Hart (TMK_1 + CNN^{-}).

11. Links de Tomek (versión de Kubat y Matwin) seguido del condensado de Hart (TMK_2 + CNN^{-}).

Los resultados de clasificación para cada uno de los conjuntos originales y submuestreados corresponden a las salidas obtenidas por 8 clasificadores de distinta naturaleza: $1, 7, 13-$NN, un árbol de decisión (J48), el clasificador naïve Bayes (NBC), la red neuronal perceptron multicapa (MLP), la red neuronal de funciones de base radial (RBF) y la máquina de vectores soporte (SVM). Todos ellos fueron tomados del kit de herramientas Weka (Witten y Frank, 2005), empleando los parámetros de su configuración por defecto.

5.2.4.1. Análisis de resultados

El diseño de los experimentos puede resumirse como sigue. Para cada uno de los 17 conjuntos de datos, se llevaron a cabo 5 repeticiones de una validación cruzada de 10 particiones. Cada una de las 11 técnicas de submuestreo se aplicó sobre los conjuntos de entrenamiento; posteriormente, se construyeron los 8 clasificadores antes mencionados a partir de los conjuntos de entrenamiento originales y submuestreados. Cada uno de los clasificadores se evaluó empleando los conjuntos de test.

Cada una de las salidas de los clasificadores fue evaluada empleando diversas métricas de efectividad apropiadas para problemas desbalanceados, así como la conocida tasa de precisión global (Acc). Con la finalidad de verificar la existencia de diferencias significativas en los resultados, se llevó a cabo entre cada par de ellos un análisis estadístico mediante la prueba t (paired t-test) (Demšar, 2006) para un nivel de confianza del 95 %. La Tabla 5.1 muestra un ejemplo del análisis estadístico al ser clasificadas por la regla $1-$NN y evaluadas con las métricas $IBA_{0.1}$ y la tasa de precisión global (Acc). En cada una de las entradas, se indica cuántas veces el algoritmo asociado a esa fila ha sido significativamente mejor/igual/peor que el algoritmo que determina la correspondiente columna. El resto de las tablas se pueden consultar en el Anexo C.

A partir de los resultados de la prueba $t-$student, se calculó un índice de efectividad que puede expresarse como la diferencia entre el número de veces que cada algoritmo fue significativamente mejor y el número de veces que resultó peor (ver

Tabla 5.1: *Prueba t−student sobre los resultados promedios de clasificación por 1−NN en las 17 bases de datos originales y submuestreadas: métricas de efectividad $IBA_{0.1}$ y Acc.*

	$IBA_{0.1}$										
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) Original	2/4/11	3/3/11	2/5/10	3/4/10	2/7/8	7/3/7	3/3/11	13/3/1	2/5/10	11/2/4	2/4/11
(2) SSM^-		3/6/8	2/5/10	3/4/10	3/12/2	9/2/6	7/8/2	15/2/0	4/7/6	12/0/5	1/11/5
(3) EW^-			4/6/7	6/5/6	7/8/2	9/2/6	10/6/1	15/2/0	6/6/5	12/1/4	5/9/3
(4) EW^-+SSM^-				6/8/3	8/6/3	10/2/5	9/5/3	16/1/0	7/8/2	11/3/3	6/8/3
(5) RUS					10/4/3	9/3/5	8/6/3	16/0/1	4/11/2	11/3/3	6/8/3
(6) CNN^-						8/4/5	4/9/4	14/3/0	5/4/8	12/0/5	3/8/6
(7) TMK_1							7/1/9	17/0/0	4/4/9	12/4/1	4/3/10
(8) TMK_2								14/3/0	3/7/7	12/1/4	2/8/7
(9) CNN^-+TMK_1									0/0/17	0/0/17	0/1/16
(10) CNN^-+TMK_2										11/1/5	3/8/6
(11) TMK_1+CNN^-											4/1/12
(12) TMK_2+CNN^-											

	Acc										
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original	16/1/0	12/2/3	16/0/1	17/0/0	17/0/0	17/0/0	9/6/2	17/0/0	17/0/0	17/0/0	17/0/0
SSM^-		3/1/13	10/5/2	12/3/2	6/10/1	13/2/2	1/3/13	17/0/0	15/1/1	17/0/0	11/5/1
EW^-			16/1/0	16/1/0	13/1/3	16/0/1	3/5/9	17/0/0	15/1/1	16/0/1	14/1/2
EW^-+SSM^-				9/3/5	4/6/7	12/1/4	1/0/16	17/0/0	9/7/1	16/0/1	4/7/6
RUS					2/4/11	9/0/8	0/0/17	16/0/1	6/4/7	14/1/2	2/5/10
CNN^-						13/2/2	2/1/14	17/0/0	16/1/0	17/0/0	10/7/0
TMK_1							0/0/17	17/0/0	4/3/10	16/1/0	3/2/12
TMK_2								17/0/0	17/0/0	17/0/0	15/2/0
CNN^-+TMK_1									0/0/17	0/0/17	0/0/17
CNN^-+TMK_2										15/1/1	0/1/16
TMK_1+CNN^-											0/0/17
TMK_2+CNN^-											

Tabla 5.2). Con la finalidad de facilitar el posterior análisis de los resultados, cada índice de efectividad puede representarse por medio de un rango que determina la posición que ocupa cada uno de los algoritmos de submuestreo: al algoritmo con el índice más alto se le asignará el rango 1, al siguiente el rango 2 y así sucesivamente hasta que cada algoritmo tenga asignado un rango. En caso de empate, se asigna un rango promedio a cada uno de los algoritmos implicados.

La Figura 5.2 representa los rangos correspondientes al índice de efectividad para seis de las métricas utilizadas en los experimentos sobre cada uno de los clasificadores.

- Las Figuras 5.2(a) y 5.2(b) muestran los rankings sobre las tasas de verdaderos positivos y verdaderos negativos, respectivamente. Como se puede observar, los algoritmos de submuestreo conllevan un deterioro de la tasa de aciertos sobre la clase mayoritaria y, al mismo tiempo, un incremento de la correspondiente tasa de aciertos sobre la clase minoritaria. Se puede decir que ambas figuras muestran comportamientos inversamente proporcionales. En este sentido, en el caso de *tvp*, los algoritmos mejor posicionados son aquellos con las mayores tasas de reducción sobre la clase mayoritaria. Por el contrario, en el caso de *tvn*, los algoritmos con un mejor rango se refieren a los que no han sido capaces de eliminar el desbalance (ver Tabla 5.3).

- Como puede deducirse de lo comentado en el punto anterior, la Figura 5.2(c) muestra que los mejores resultados utilizando la métrica *Acc* se obtienen con

los conjuntos originales. Se observa que, en general, el rango de los algoritmos de submestreo de acuerdo a Acc viene determinado por el ratio de desbalance (Tabla 5.3): cuanto más alta sea la tasa de reducción sobre la clase mayoritaria, peor será el ranking del correspondiente algoritmo de submuestreo. Esto demuestra una vez más la influencia que ejerce la clase mayoritaria al emplear esta métrica global.

- Si nos centramos en algunas de las medidas más apropiadas para problemas desbalanceados, como f_1, $IBA_{0.1}$ y Gm (ver Figuras 5.2(d), 5.2(e) y 5.2(f)), se puede observar que las primeras posiciones están ocupadas ya por los resultados de algunos de los algoritmos de submuestreo. En concreto, tanto para $IBA_{0.1}$ como para Gm, los resultados sobre el conjunto original se encuentran entre las tres últimas posiciones, independientemente del clasificador utilizado. La única excepción se produce en el caso de NBC, donde los resultados con el conjunto original dan lugar a unos valores de ranking intermedios.

Fijándonos en los rangos para los diferentes algoritmos de submuestreo al utilizar las métricas $IBA_{0.1}$ y Gm (ver Figuras 5.2(e) y 5.2(f)), es posible destacar los siguientes resultados:

- Sorprendentemente, el submuestreo aleatorio se sitúa como una de las mejores estrategias para la mayoría de los clasificadores. El único caso en el que no se da esta circunstancia corresponde al clasificador 1−NN, cuyo modelo se caracteriza por presentar un aprendizaje más local.
- Los algoritmos de submuestreo con una moderada reducción de la clase mayoritaria obtienen mejores resultados en el caso de los clasificadores con un aprendizaje más global. Esto se puede observar especialmente a través de los rangos correspondientes a los algoritmos de submuestreo aleatorio y por edición de Wilson.
- Por el contrario, las peores estrategias de submuestreo son aquellas cuyo proceso de reducción invierte el sentido del desbalance (la clase mayoritaria pasa a ser minoritaria, ver Tabla 5.3). Los casos más evidentes corresponden a los diferentes algoritmos que combinan los links de Tomek con el condensado de Hart, posicionándose incluso por detrás de los rangos obtenidos para el conjunto original.

5.3. Soluciones a nivel de datos: algoritmos de sobremuestreo

La estrategia opuesta a la reducción de la clase mayoritaria viene representada por los algoritmos de sobremuestreo, los cuales logran el equilibrio en la base de datos mediante el incremento de la talla de la clase minoritaria.

Tabla 5.2: *Índice de efectividad empleando $IBA_{0.1}$ sobre los algoritmos de submuestreo.*

	Mejor	Peor	Índice		Mejor	Peor	Índice
	1−NN				7−NN		
EW^-+SSM^-	100	30	70	RUS	138	15	123
RUS	93	38	55	EW^-	109	48	61
EW^-	93	40	53	EW^-+SSM^-	102	41	61
TMK_2+CNN^-	82	36	46	TMK_2+CNN^-	93	53	40
CNN^-+TMK_2	80	46	34	SSM^-	88	57	31
SSM^-	70	56	14	TMK_1	85	75	10
CNN^-	64	58	6	CNN^-+TMK_2	76	67	9
TMK_2	64	66	-2	CNN^-	75	74	1
TMK_1	78	81	-3	TMK_2	76	79	-3
Original	50	94	-44	Original	55	116	-61
TMK_1+CNN^-	55	116	-61	TMK_1+CNN^-	32	137	-105
CNN^-+TMK_1	2	170	-168	CNN^-+TMK_1	9	176	-167
	13−NN				J48		
RUS	135	14	121	RUS	128	7	121
EW^-	107	50	57	EW^-	93	26	67
EW^-+SSM^-	101	48	53	TMK_2+CNN^-	75	40	35
TMK_2+CNN^-	97	52	45	TMK_1	90	58	32
SSM^-	96	56	40	TMK_2	71	46	25
CNN^-+TMK_2	82	74	8	CNN^-+TMK_2	60	52	8
TMK_1	79	75	4	EW^-+SSM^-	64	60	4
CNN^-	78	76	2	CNN^-	55	60	-5
TMK_2	81	81	0	SSM^-	53	71	-18
Original	58	108	-50	Original	46	93	-47
TMK_1+CNN^-	29	143	-114	TMK_1+CNN^-	50	112	-62
CNN^-+TMK_1	5	171	-166	CNN^-+TMK_1	9	169	-160
	MLP				NBC		
RUS	131	5	126	RUS	107	23	84
EW^-	86	37	49	EW^-	107	24	83
CNN^-+TMK_2	66	33	33	TMK_2	89	50	39
TMK_2+CNN^-	65	33	32	TMK_2+CNN^-	83	52	31
EW^-+SSM^-	68	47	21	CNN^-+TMK_2	78	58	20
TMK_1	87	72	15	EW^-+SSM^-	76	63	13
CNN^-	54	52	2	Original	71	67	4
SSM^-	55	58	-3	TMK_1	74	83	-9
TMK_2	58	70	-12	TMK_1+CNN^-	66	94	-28
Original	45	88	-43	CNN^-	57	92	-35
TMK_1+CNN^-	45	104	-59	SSM^-	50	102	-52
CNN^-+TMK_1	8	169	-161	CNN^-+TMK_1	11	161	-150
	RBF				SVM		
RUS	139	10	129	RUS	156	6	150
TMK_2+CNN^-	102	33	69	EW^-+SSM^-	108	44	64
EW^-+SSM^-	101	41	60	TMK_2+CNN^-	106	46	60
CNN^-+TMK_2	89	48	41	SSM^-	96	50	46
SSM^-	81	65	16	CNN^-	87	69	18
CNN^-	71	58	13	EW^-	84	70	14
TMK_1	90	80	10	CNN^-+TMK_2	83	72	11
EW^-	71	72	-1	TMK_1	78	85	-7
TMK_1+CNN^-	69	88	-19	TMK_2	59	102	-43
TMK_2	45	107	-62	Original	40	117	-77
Original	19	141	-122	TMK_1+CNN^-	34	132	-98
CNN^-+TMK_1	18	152	-134	CNN^-+TMK_1	13	151	-138

Como ya se ha comentado en el Capítulo 3, el algoritmo SMOTE y algunas de

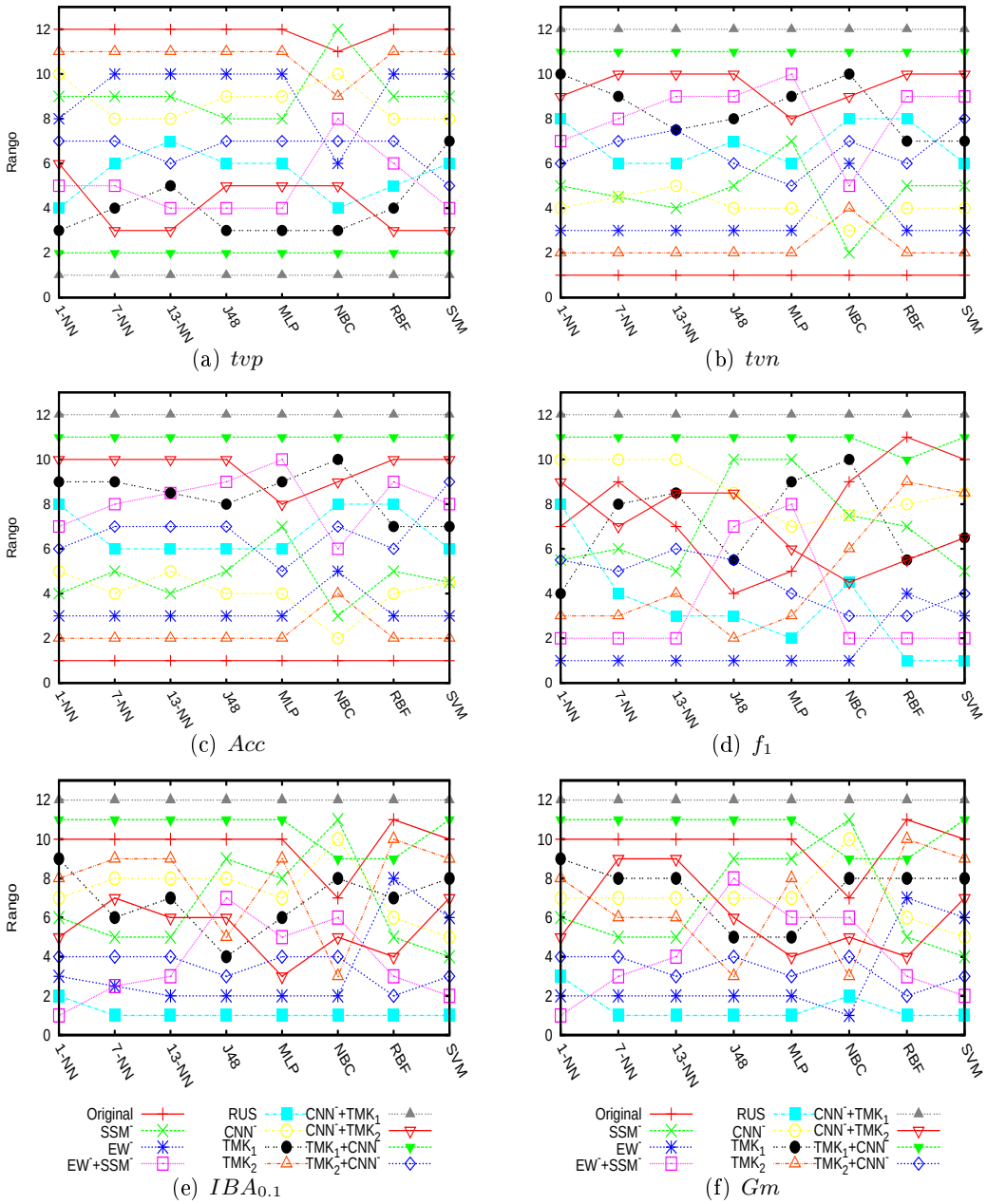


Figura 5.2: *Ranking de los algoritmos de submuestreo empleando los clasificadores* $1-NN$, $7-NN$, $13-NN$, *J48, MLP, NBC y SVM.*

sus extensiones (por ejemplo, el Borderline–SMOTE) han demostrado, en muchos casos, ser una solución adecuada para tratar el desbalance a partir de la generación de muestras sintéticas en la clase minoritaria. Siguiendo con el esquema general del SMOTE, en la siguiente sección, se presentan tres variantes basadas en el concepto de vecindad envolvente (Sánchez y otros, 1997a).

Tabla 5.3: *Ratios de desbalance en conjuntos submuestreados*

	Ratios de desbalance								
	Breast	Ecoli	German	Glass	Haberman	Laryngeal_2	Letter_a	Optdigits	Pendigits
Original	0.41	0.12	0.43	0.09	0.36	0.08	0.04	0.11	0.11
SSM^-	0.84	0.81	0.95	0.41	0.90	1.07	1.52	1.61	2.76
EW^-	0.67	0.12	0.57	0.09	0.47	0.09	0.04	0.11	0.11
EW^-+SSM^-	1.83	1.05	1.50	0.47	1.45	1.34	1.52	1.61	2.78
RUS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CNN^-	0.71	0.88	0.85	0.45	0.76	0.90	2.30	2.81	6.80
TMK_1	1.49	0.15	15.90	0.14	0.75	0.10	0.06	7.86	0.13
TMK_2	0.54	0.12	0.49	0.09	0.44	0.08	0.04	0.11	0.11
$\text{CNN}^-+\text{TMK}_1$	13.25	11.41	355.26	4.53	7.25	4.67	263.00	0.00	600.95
$\text{CNN}^-+\text{TMK}_2$	1.30	1.12	1.34	0.56	1.41	1.14	2.32	2.89	7.07
$\text{TMK}_1+\text{CNN}^-$	3.71	2.59	28.60	1.03	2.37	1.62	4.90	40.34	13.19
$\text{TMK}_2+\text{CNN}^-$	1.07	0.98	0.49	0.49	1.08	0.99	2.30	2.81	6.84

	Ratios de desbalance							
	Phoneme	Pima	Satimage	Scrapie	Segmentation	Spambase	Vehicle	Yeast
Original	0.42	0.54	0.11	0.21	0.17	0.65	0.33	0.41
SSM^-	2.11	1.24	0.79	1.40	1.49	2.13	1.00	1.00
EW^-	0.45	0.78	0.11	0.56	0.17	0.81	0.41	0.51
EW^-+SSM^-	2.80	2.49	1.11	3.46	1.73	3.59	1.55	1.50
RUS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CNN^-	2.21	1.13	0.88	0.26	1.86	1.91	0.90	0.98
TMK_1	0.75	2.74	0.25	0.23	0.21	2.41	1.64	2.51
TMK_2	0.43	0.60	0.11	0.61	0.17	0.69	0.37	0.45
$\text{CNN}^-+\text{TMK}_1$	86.72	74.44	130.42	0.30	22.47	102.88	109.66	311.37
$\text{CNN}^-+\text{TMK}_2$	3.00	1.66	0.99	1.85	2.08	2.52	1.41	1.39
$\text{TMK}_1+\text{CNN}^-$	7.93	8.02	10.07	0.29	3.65	10.74	8.80	13.88
$\text{TMK}_2+\text{CNN}^-$	2.48	1.38	0.91	1.61	1.98	2.17	1.13	0.45

5.3.1. SMOTE basado en vecinos envolventes

Idealmente, el concepto de vecindad implicaría que los vecinos de un objeto deberían no únicamente estar tan próximos a él como fuera posible, sino también que éstos se encontraran distribuidos homogéneamente alrededor de dicho objeto. Esta segunda condición es una consecuencia de la primera en el caso asintótico. Sin embargo, en algunas situaciones prácticas, la ubicación geométrica puede ser mucho más importante que las propias distancias reales. En el caso de la regla del vecino más cercano, ésta solamente tiene en cuenta la primera propiedad, por lo que pudiera ocurrir que los vecinos no se encontraran situados simétricamente alrededor del objeto.

En la literatura, se han propuesto algunos conceptos alternativos de vecindad con el fin de solventar el mencionado problema que presenta la vecindad tradicionalmente utilizada en la regla $k-$NN. Estas propuestas definen la vecindad de un objeto en términos de proximidad y simetría, razón por la cual han recibido el nombre genérico de *vecindad envolvente* (Sánchez y otros, 1997a). Algunos de los ejemplos más representativos dentro de esta categoría de vecindad vienen dados por la *vecindad del centroide más cercano* (*NCN*, *Nearest Centroid Neighbourhood*) (Chaudhuri, 1996) y los *grafos de proximidad*. De este último grupo, dos de las aproximaciones más conocidas son los *grafos de Gabriel* (*GG*, *Gabriel Graph*) y los *grafos de vecindad relativa* (*RNG*, *Relative Neighbourhood Graph*) (Jaromczyk y Toussaint, 1992). Diversos trabajos han demostrado empíricamente que estas aproximaciones, aplicadas a problemas de clasificación, presentan un mejor comportamiento que la regla del vecino más cercano convencional (Sánchez y otros, 1997a,b, Zhang y otros, 1997).

Basado en el concepto de vecindad envolvente, se propone emplear las tres apro-

ximaciones antes mencionadas para sobremuestrear la clase minoritaria siguiendo la estrategia general del algoritmo SMOTE. Así, en lugar de emplear el vecino más cercano para generar muestras artificiales, el algoritmo selecciona vecinos envolventes de la clase positiva. El objetivo detrás de esta modificación se centra en extender la región de las nuevas muestras sintéticas.

El Algoritmo 7 muestra un ejemplo del SMOTE basado en vecinos envolventes mediante la aproximación NCN (ncn-SMOTE). Obsérvese que la incorporación de estos enfoques no implica una modificación sustancial sobre la versión original del algoritmo SMOTE. Cabe mencionar que, en el caso de GG y RNG, el parámetro k no es necesario, ya que cada muestra de entrenamiento de la clase positiva puede tener un diferente número de vecinos. A pesar de esto, el procedimiento de SMOTE por grafos de vecindad podría expresarse en los mismos términos que el algoritmo ncn-SMOTE.

Algoritmo 7 ncn–SMOTE

Entrada:
$P = \{p_1, \ldots, p_{min}\}$ // Muestras de la clase minoritaria
min // Talla de la clase minoritaria
n_s // Número de muestras artificiales a generar por cada muestra positiva
k // Número de vecinos más cercanos
Salida:
$Synthetic$ // Conjunto de muestras artificiales
Inicio:
Para $i = 1$ hasta min **hacer**
Encontrar los k vecinos envolvente de p_i
Para $j = 1$ hasta n_s **hacer**
Elegir aleatoriamente uno de los k vecinos de p_i, sea $n(p_i)$
$diff = p_i - n(p_i)$
$gap =$ Número aleatorio entre 0 y 1
$NewSample = p_i + gap * diff$
$Synthetic \longleftarrow NewSample$
Fin Para
Fin Para
Fin

5.3.2. Experimentos y análisis de resultados

En esta sección, se comparan diversos algoritmos de sobremuestreo cuyo procedimiento se basa en el esquema general del SMOTE. En concreto, se analizan cinco técnicas que incrementan la talla de la clase minoritaria por la creación de muestras sintéticas de la clase positiva. Tres de estas aproximaciones tienen en cuenta la proximidad y la distribución espacial de las muestras (vecinos envolventes): ncn-SMOTE, gg-SMOTE y rng-SMOTE. Las dos restantes son el algoritmo SMOTE original y el Borderline-SMOTE (b-SMOTE).

Para la obtención y el análisis de los resultados con los algoritmos de sobremuestreo, se ha mantenido la misma configuración de los experimentos (clasificadores, bases de datos, métricas de efectividad, etc.) empleada en la Sección 5.2.

Por otra parte, para todas las versiones de SMOTE se incrementó la talla de la clase minoritaria hasta alcanzar un equilibrio perfecto entre clases, exceptuando el caso del algoritmo b-SMOTE, en donde el número de muestras sintéticas generadas por cada ejemplo de la clase positiva fue igual a tres, por lo que es posible que no se llegue a obtener un balance absoluto entre las clases.

La Figura 5.3 muestra los resultados de los diversos algoritmos de sobremuestreo en términos de rangos para seis métricas de efectividad, los cuales se obtuvieron a partir de la diferencia entre los números de veces que cada algoritmo fue significativamente mejor y peor.

A partir de las gráficas, merece la pena destacar las siguientes observaciones:

- En un comportamiento similar a los obtenidos por los algoritmos de submuestreo, la Figura 5.3(a) muestra que, en general, los esquemas de sobremuestreo son superiores al conjunto original en términos de la tasa de aciertos sobre la clase positiva. Sin embargo, esto implica simultáneamente un deterioro de la tasa de aciertos sobre la clase negativa, tal y como se puede observar en la Figura 5.3(b), donde la estrategia de emplear el conjunto original sin ningún tipo de preprocesado obtiene el mejor rango.

- De lo observado anteriormente, resulta evidente que las tasas de precisión global (Acc) obtenidas por los algoritmos de sobremuestreo son inferiores (ver Figura 5.3(c)), donde los correspondientes rangos son peores que el del conjunto original).

- Considerando las métricas de evaluación más apropiadas para entornos desbalanceados, Figuras 5.3(d), 5.3(e) y 5.3(f), éstas coinciden en señalar que las peores estrategias corresponden al algoritmo b-SMOTE y la opción de utilizar directamente el conjunto original sin ningún tipo de sobremuestreo.

Analizando con más detalle las Figuras 5.3(d), 5.3(e) y 5.3(f), que corresponden a las métricas de efectividad adecuadas para entornos desbalanceados, cabe reseñar las siguientes conclusiones:

- Las tres métricas coinciden en indicar que, en el caso del clasificador 1−NN, la mejor técnica de sobremuestreo viene dada por el algoritmo ncn-SMOTE.

- Los resultados, en general, no son concluyentes acerca de qué estrategia de sobremuestreo es la mejor. Únicamente en el caso de $IBA_{0.1}$, el esquema original de SMOTE se sitúa como la mejor solución en la mayoría de los clasificadores.

- A pesar de las diferencias existentes, es posible observar que la técnica gg-SMOTE presenta un comportamiento más estable que el resto de los algoritmos de sobremuestreo basados en vecindades envolventes.

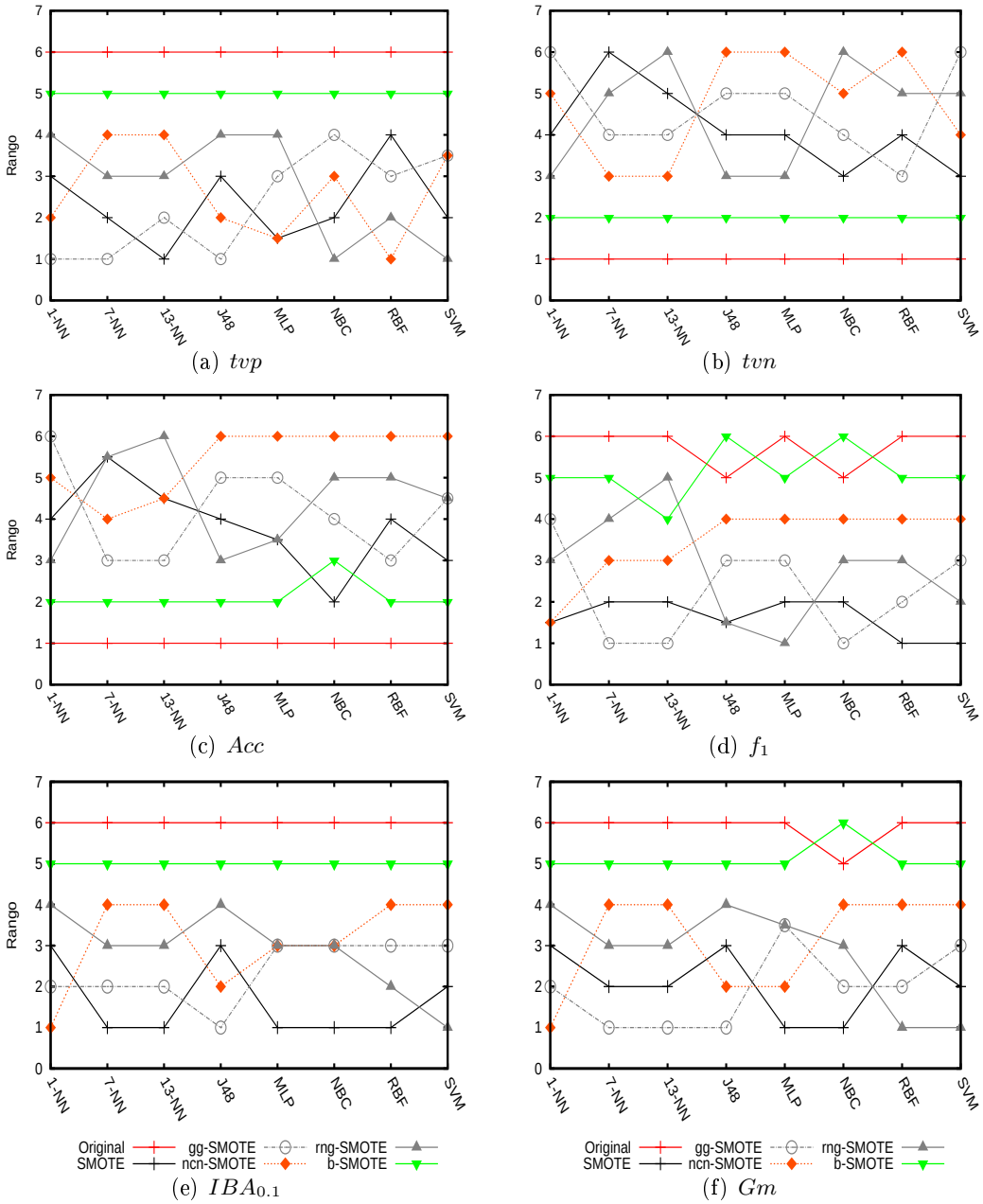


Figura 5.3: *Ranking de los algoritmos de sobremuestreo empleando los clasificadores* $1-NN$, $7-NN$, $13-NN$, *J48, MLP, NBC y SVM.*

5.4. Submuestreo vs. sobremuestreo

Como ya se comentó previamente, dentro de las soluciones a nivel de datos, existe un fuerte debate acerca de qué método representa la mejor propuesta para tratar el problema del desbalance (Batista y otros, 2004, Barandela y otros, 2004, Chawla, 2003, Drummond y Holte, 2003, Japkowicz y Stephen, 2002, Hulse y otros, 2007,

Orriols-Puig y Bernadó-Mansilla, 2009). En este sentido, la gran diversidad de opiniones acerca de la efectividad de los algoritmos de remuestreo giran en torno a dos factores: *i)* el ratio de desbalance y *ii)* la naturaleza del clasificador.

Respecto al primer punto, es posible mencionar los trabajos de Barandela y otros (2004), Batista y otros (2004) y Orriols-Puig y Bernadó-Mansilla (2009) quienes sugieren que, en general, los algoritmos de submuestreo resultan más apropiados para problemas con un débil desbalance entre clases, así como para bases de datos desbalanceados con un elevado número de muestras por atributo. Por el contrario, los algoritmos de sobremuestreo aparecen como una mejor solución en situaciones con un fuerte desbalance. En cuanto al segundo punto, Batista y otros (2004), Chawla (2003), Japkowicz y Stephen (2002) y Orriols-Puig y Bernadó-Mansilla (2009) coinciden en señalar que los algoritmos de sobremuestreo son más efectivos que las técnicas de submuestreo cuando se utilizan clasificadores como árboles de decisión (C4.5) o máquinas de vectores soporte (SVM). Por otra parte, además de estos dos factores, Hulse y otros (2007) destacan que la efectividad de los métodos de remuestreo depende, no sólo del modelo de clasificación y de la naturaleza de la base de datos, sino también de la métrica de efectividad a emplear.

Con la finalidad de comprobar si las conclusiones antes mencionadas pueden extenderse a otros clasificadores y a otras métricas de efectividad, en la siguiente sección se comparan cuatro de los algoritmos de submuestreo más representativos con dos algoritmos de sobremuestreo: el SMOTE original (Chawla y otros, 2002) y una de las variantes basadas en vecinos envolventes que aquí se han propuesto (gg-SMOTE).

Además, con el objetivo de analizar el rendimiento de estos algoritmos de remuestreo en función del ratio de desbalance, se agruparon las bases de datos en dos categorías. En un primer grupo, se encuentran aquellas bases de datos que podrían considerarse como fuertemente desbalanceadas (un ratio minoratia/mayoritaria menor a 0.27): Ecoli, Glass, Laryngeal$_2$, Letter$_a$, Optdigits, Pendigits, Satimage, Scrapie y Segmentation. En el segundo grupo, se incluyeron las bases de datos débilmente desbalanceadas (un ratio mayor o igual a 0.27): Breast, German, Haberman, Phoneme, Pima, Spambase, Vehicle y Yeast.

5.4.1. Experimentos y análisis de resultados

A diferencia de los experimentos de las secciones anteriores, los resultados incluidos en esta sección están enfocados a analizar la efectividad de los métodos de remuestreo considerando las dos categorías de desbalance antes mencionadas.

Por lo que respecta a los clasificadores, las métricas de evaluación de la efectividad, el número de bases de datos y la técnica de validación cruzada, éstas son idénticas a las empleadas en las Secciones 5.2 y 5.3.

Las Figuras 5.4 y 5.5 muestran los rangos asignados a las diferentes técnicas de sobremuestreo y submuestreo de acuerdo al nivel de desbalance que presentan las bases de datos: un alto nivel de desbalance y un bajo nivel de desbalance, respectivamente.

Al analizar los resultados que se obtuvieron sobre las bases de datos con una fuerte diferencia en la distribución de las clases (ver Figura 5.4), es posible comentar diversas situaciones interesantes. Las dos primeras métricas de efectividad, Figuras 5.4(a) (*tvp*) y 5.4(b) (*tvn*), confirman lo observado en las Secciones 5.2 y 5.3: todos los algoritmos de remuestreo benefician a la clase minoritaria, a la vez que producen un efecto negativo sobre la tasa de aciertos de la clase mayoritaria. Sin embargo, podemos apreciar algunas diferencias destacables en cuanto a la intensidad de este efecto. Así, mientras que los algoritmos de sobremuestreo no causan un deterioro tan drástico sobre la clase mayoritaria al mejorar la tasa de aciertos de la clase minoritaria, los esquemas de submuestreo benefician claramente a una de las clases al tiempo que perjudican a la otra en la misma medida. Por supuesto, este comportamiento de las estrategias de submuestreo dirigido obedece a la elevada tasa de reducción de la clase mayoritaria (ver Tabla 5.3).

Considerando los resultados sobre la métrica f_1 (Figura 5.4(d)), se puede ver que en cuatro clasificadores la mejor estrategia corresponde al submuestreo por edición de Wilson, mientras que para el resto, los algoritmos de sobremuestreo son los que obtienen los mejores resultados. En este sentido, es importante mencionar que dicha métrica podría llegar a favorecer a clasificadores con bajas tasas de aciertos en la clase positiva, debido a que la medida *precision* no tiene en cuenta los errores sobre las muestras positivas.

Para las métricas $IBA_{0.1}$ y Gm (ver Figura 5.4(e) y 5.4(f)), en la mayoría de los clasificadores, el algoritmo original de SMOTE obtiene el mejor rango. La excepción en ambos casos ocurre con el clasificador NBC, donde los mejores rangos corresponden al submuestreo aleatorio y a la versión basada en edición de Wilson. En el caso de $IBA_{0.1}$, se observa también este mismo comportamiento en el árbol de decisión J48.

En lo que respecta a los resultados sobre bases de datos con un bajo nivel de desbalance, existe una gran diversidad de resultados. Partiendo de la métricas básicas, Figura 5.5(a) y Figura 5.5(b), obsérvese que la combinación de edición de Wilson con selectivo modificado obtiene el mejor rango en *tvp* y el peor en *tvn*. Este comportamiento opuesto a lo observado en el caso de las bases de datos fuertemente desbalanceadas nos sugiere que dicha técnica acentúa de forma muy significativa el desbalance inverso, lo cual puede confirmarse al analizar los ratios de la Tabla 5.3.

Al centrarnos en la medida $IBA_{0.1}$ (ver Figura 5.5(e)), se observa que los clasificadores 1−NN, NBC y RBF obtienen los mejores resultados con tres algoritmos distintos de submuestreo, EW^-+SSM^-, EW^- y RUS, respectivamente. Por el contrario, el resto de clasificadores coinciden en señalar al algoritmo de sobremuestreo gg-SMOTE como la mejor solución.

En cuanto a Gm (ver Figura 5.5(f)), el clasificador NBC presenta el mejor comportamiento utilizando el submuestreo por edición de Wilson, mientras que para RBF corresponde al submuestreo aleatorio. Para el resto de clasificadores, los resultados indican que, en general, el sobremuestreo con gg-SMOTE se convierte en la mejor estrategia.

A pesar de que en trabajos anteriores se ha demostrado la efectividad de los

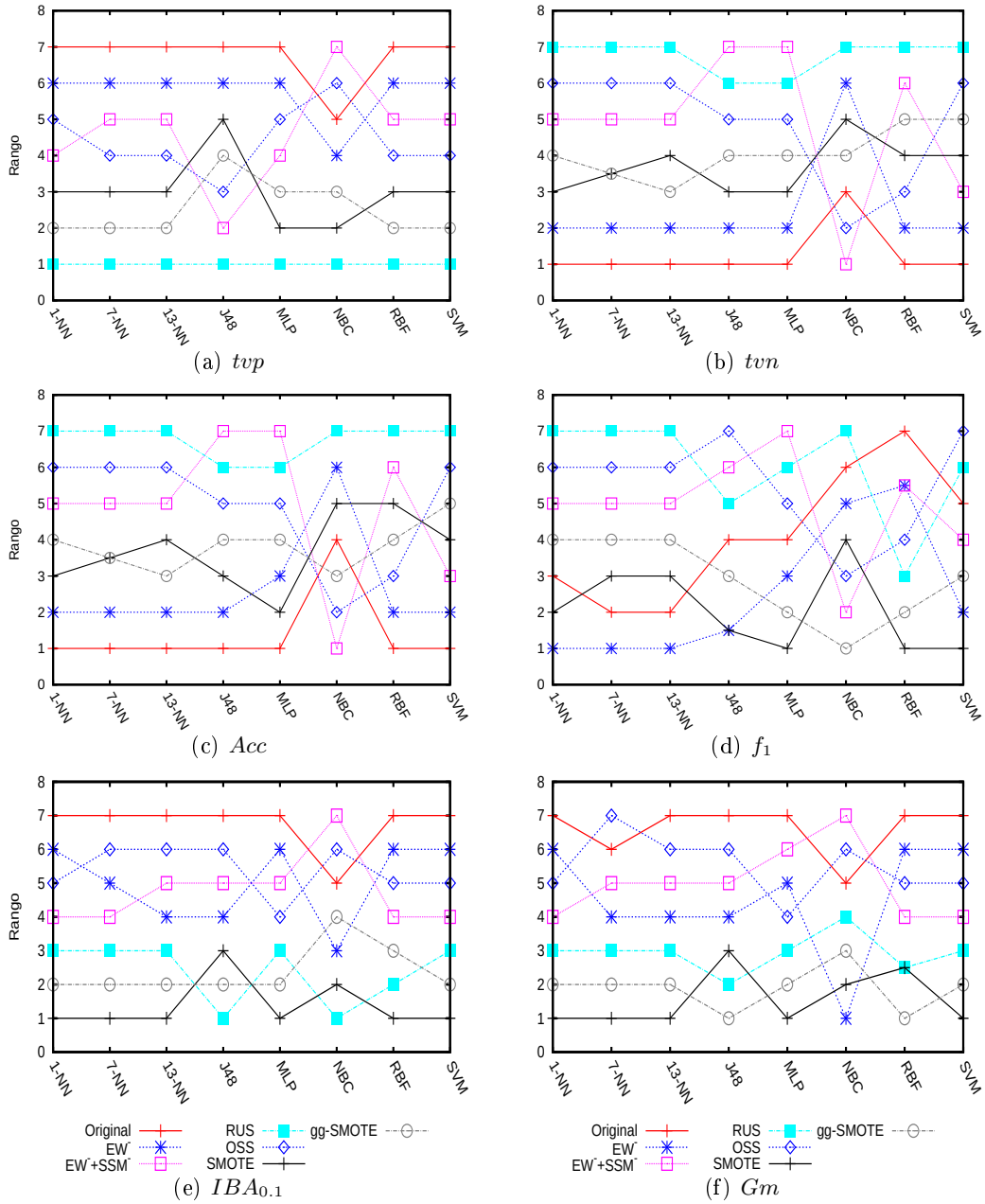
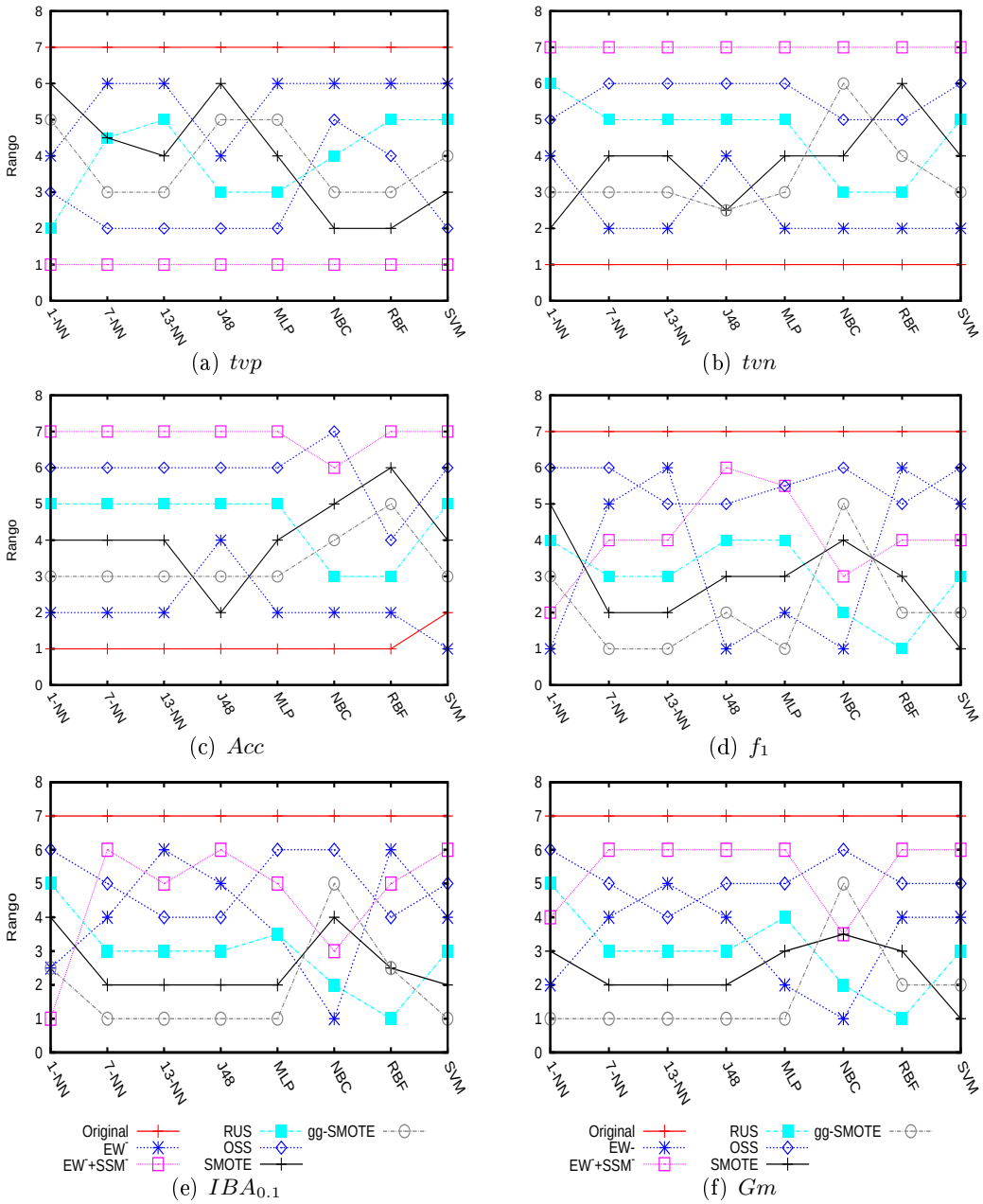


Figura 5.4: *Ranking de los algoritmos de submuestreo y sobremuestreo sobre bases de datos con un alto grado de desbalance empleando los clasificadores* $1-NN$*,* $7-NN$*,* $13-NN$*, J48, MLP, NBC y SVM.*

algoritmos de submuestreo dirigidos, los resultados aquí mostrados sugieren que factores como las características del conjunto de datos, el ratio de desbalance y la naturaleza del clasificador afectan al rendimiento de dichas estrategias. Por ejemplo, en bases de datos con un débil desbalance, estos algoritmos perjudican en gran manera a la clase mayoritaria, llegando a invertir el sentido del desbalance (la clase

mayoritaria pasa a ser minoritaria). Este efecto da lugar a muy malos resultados en el caso de clasificadores con un aprendizaje global, mientras que aquellos que se basan en un aprendizaje local son capaces de resistir mejor dicha situación.



(a) tvp (b) tvn (c) Acc (d) f_1 (e) $IBA_{0.1}$ (f) Gm

Figura 5.5: *Ranking de los algoritmos de submuestreo y sobremuestreo sobre bases de datos con un bajo nivel de desbalance empleando los clasificadores* $1-NN$, $7-NN$, $13-NN$, *J48, MLP, NBC y SVM.*

5.5. Soluciones a nivel de algoritmos

Los experimentos de la sección anterior han probado la efectividad de los algoritmos de remuestreo para tratar el problema del desbalance. Sin embargo, su dependencia de la complejidad y características de los datos provoca que, en determinadas ocasiones, no se logre un absoluto equilibrio entre clases y, en ciertos casos, la reducción puede incluso dar lugar a que el sentido del desbalance se invierta. Asimismo, algunos investigadores mencionan que remuestrear las bases de datos implica una alteración artificial de la distribución original de los datos (McCarthy y otros, 2005, Seiffert y otros, 2008), lo que podría añadir nuevos problemas (Orriols-Puig y Bernadó-Mansilla, 2009).

Una línea de investigación alternativa a las técnicas que preprocesan las bases de datos se refiere a las soluciones a nivel de algoritmos, las cuales tienen por objetivo adaptar los propios clasificadores al problema del desbalance. Sin embargo, las críticas a estas estrategias establecen que, al tratarse de soluciones específicas para cada modelo, se hace difícil su extensión a otros problemas.

Siguiendo la línea general de las soluciones a nivel de algoritmos, en la presente sección, se introduce una distancia ponderada para la regla del vecino más cercano, cuyo objetivo será predisponer al clasificador hacia la clase menos representativa.

5.5.1. Distancia ponderada

Con el objetivo de establecer una tendencia positiva del clasificador 1−NN hacia la clase menos representada, se sugiere aquí una modificación sobre la distancia Euclídea. Esta adaptación puede verse como una función de distancia ponderada, la cual puede expresarse de la siguiente forma:

$$d_w(\mathbf{y}, \mathbf{x}) = \left(\frac{N_i}{N}\right)^{\frac{1}{n}} \times d_E(\mathbf{y}, \mathbf{x}) \tag{5.1}$$

donde $d_E(.)$ representa la distancia Euclídea, $\mathbf{y}$ la muestra a ser etiquetada y $\mathbf{x}$ la muestra de entrenamiento. La primera parte de la fórmula constituye un coeficiente de ponderación que se obtiene del cociente entre la talla de la clase correspondiente a la muestra de entrenamiento N_i y el número total de muestras del conjunto de datos N, todo ello elevado al inverso del número de atributos n. El objetivo de esta potenciación es moderar las diferencias entre los coeficientes de ponderación.

La idea básica detrás de esta distancia ponderada (DP) es compensar el desbalance presente en el conjunto de entrenamiento original sin alterar la distribución entre clases, puesto que los pesos son asignados a las respectivas clases y no a los muestras individuales.

El factor de ponderación será más alto para la clase mayoritaria que para la clase minoritaria, por lo que la distancia (final) de las muestras positivas tenderá a ser más pequeña que la distancia a las muestras de la clase negativa. Esto produce una predisposición del clasificador a buscar el vecino más cercano de las nuevas muestras entre los objetos de la clase minoritaria.

5.5.2. Experimentos y análisis de resultados

La configuración de los experimentos de este sección consiste en la clasificación de las bases de datos originales por la regla de los k vecinos más cercanos para $k = 1, 7, 13$ y la versión $1-$NN con la distancia ponderada. Las salidas de los clasificadores fueron evaluados por las seis métricas de efectividad que se han venido empleando a lo largo de este capítulo. Asimismo, a través de los datos obtenidos por una prueba $t-$student, se calculó un rango de efectividad de los clasificadores para cada una de las seis métricas (ver Figura 5.6).

Se puede observar que la versión del clasificador que emplea la distancia ponderada presenta el mejor rango en términos de tvp, aunque se sitúa como el peor algoritmo al reconocer la clase mayoritaria (tvn). Cabe recordar que este mismo comportamiento se pudo advertir también en las soluciones a nivel de datos. Sin embargo, a pesar de ello, las tres métricas de efectividad apropiadas para entornos no balanceados (f_1, $IBA_{0.1}$, Gm) lo ubican como el mejor clasificador, indicando que a pesar de que se reduce la tasa de aciertos sobre la clase negativa, este descenso no llega a ser excesivamente fuerte (ver Anexo A).

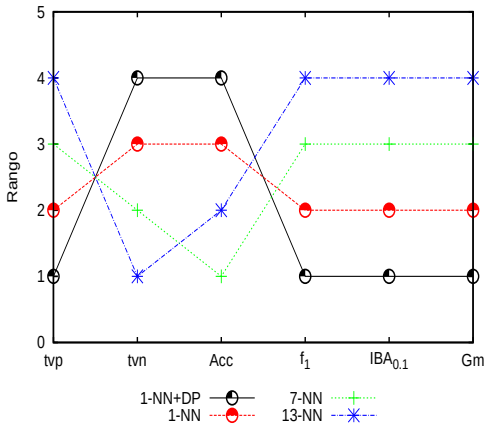


Figura 5.6: *Ranking de los clasificadores $1-NN$, $7-NN$, $13-NN$ y $1-NN+DP$ sobre los conjuntos originales empleando seis métricas de efectividad.*

5.6. Combinación de estrategias

Una de las propuestas para tratar el problema del desbalance sugiere que la mejor técnica es aquella que combina soluciones a nivel de datos con otras a nivel de algoritmos. El objetivo es aprovechar las ventajas que cada una de ellas ofrece y, por consiguiente, superar las debilidades que presentan por separado.

En la misma línea, proponemos dos estrategias que combinan diversas técnicas que se han introducido previamente. La primera propuesta tendrá por objetivo aumentar el poder de reducción de la edición de Wilson sobre la clase mayoritaria. Por su parte, la segunda estrategia empleará el clasificador con distancia ponderada

(a nivel de algoritmos) sobre los conjuntos submuestreados (a nivel de datos), los cuales aún presentan un cierto desbalance entre clases (Tabla 5.3).

5.6.1. Edición de Wilson con distancia ponderada

En la Sección 5.2, se pudo comprobar que el algoritmo de Wilson produce una baja tasa de reducción sobre la clase mayoritaria (ver Tabla 5.3), motivo por el cual no se consigue aliviar de forma significativa el desbalance existente.

La edición de Wilson utiliza la regla $k-$NN para estimar, para cada muestra del conjunto de entrenamiento, su etiqueta de clase y, posteriormente, eliminar aquellas cuya etiqueta verdadera no coincida con la asignada por la mayoría de sus k vecinos. Puesto que el clasificador $k-$NN se ve afectado por el problema del desbalance, el hecho de emplearlo en el proceso de edición podría resultar en una muy escasa eliminación de muestras negativas, dado que la mayor parte de los k vecinos seleccionados podrían pertenecer a la clase mayoritaria.

Para tratar el problema que acabamos de exponer, se propone emplear la distancia ponderada no sólo en la fase de clasificación, sino también al editar la clase mayoritaria.

5.6.1.1. Experimentos y análisis de resultados

El primer experimento consiste en submuestrear las 17 bases de datos por edición de Wilson con distancia ponderada (EWW$^-$) y, posteriormente, condensar cada uno de los conjuntos resultantes mediante el algoritmo selectivo modificado (EWW$^-$+SSM$^-$). Cada conjunto preprocesado fue clasificado por los ocho modelos que se han venido empleando y, seguidamente, fueron evaluados por seis métricas de efectividad. Los resultados se comparan con el submuestreo aleatorio (RUS), la edición de Wilson con distancia Euclídea (EW$^-$) y la combinación de ésta con selectivo modificado (EW$^-$+SSM$^-$).

La Figura 5.7 muestra los rangos que se obtienen al emplear los conjuntos originales y las cinco estrategias de submuestreo (RUS, EW$^-$, EWW$^-$, EW$^-$+SSM$^-$ y EWW$^-$+SSM$^-$), en términos de las seis métricas de efectividad.

Los resultados en términos de *tvp*, *tvn* y *Acc* son, conceptualmente, similares a los de la Sección 5.2: la mejor propuesta según la tasa de precisión global y *tvn* corresponde a los algoritmos originales, mientras que para la *tvp* viene dada por la técnica que reduce de manera significativa la clase mayoritaria (EWW$^-$+SSM$^-$). Con las tres medidas apropiadas para el desbalance (f_1, $IBA_{0.1}$, Gm), las dos técnicas que, en general, se sitúan como las mejores corresponden a la versión de edición de Wilson con distancia ponderada (EWW$^-$) y al submuestreo aleatorio (RUS).

Comparando las dos técnicas que incluyen selectivo modificado, resulta interesante destacar que, en la mayoría de las situaciones, la combinación con edición de Wilson mediante distancia ponderada es mejor a pesar de incrementar la tasa de reducción, tal como se puede observar en la Tabla 5.4.

En general, como cabía esperar, el empleo de la distancia ponderada aumenta el

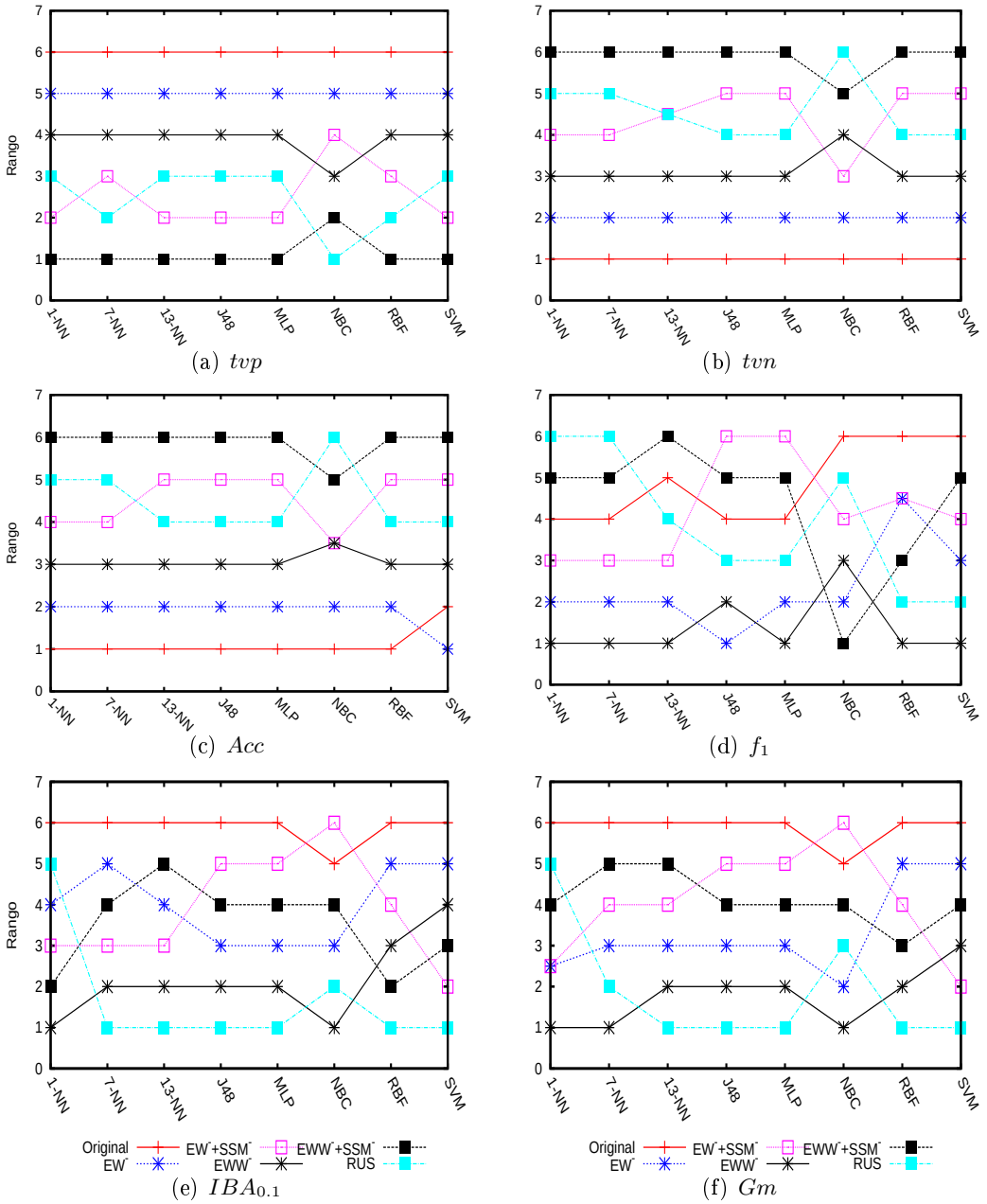


Figura 5.7: *Ranking de las dos versiones de submuestreo empleando los clasificadores 1−NN, 7−NN, 13−NN, J48, MLP, NBC y SVM.*

poder de reducción sobre la clase mayoritaria. Sin embargo, existen situaciones en las que se mantiene una tasa de reducción similar a la que se obtiene de aplicar el algoritmo clásico de Wilson. Resulta significativo que es precisamente en las bases de datos con altas tasas de acierto sobre ambas clases (Letter, Optdigits, Pendigits, Segmentation) donde se produce esta circunstancia, probablemente porque no existe un fuerte solapamiento entre clases. Por tanto, en estas bases de datos, pese a existir

Tabla 5.4: *Ratios de desbalance y porcentajes de reducción empleando edición de Wilson con distancia ponderada (EWW^-).*

	Ratios de desbalance								
	Breast	Ecoli	German	Glass	Haberman	$Laryngeal_2$	$Letter_a$	Optdigits	Pendigits
Original	0.41	0.12	0.43	0.09	0.36	0.08	0.04	0.11	0.11
EW^-	0.67	0.12	0.57	0.09	0.47	0.09	0.04	0.11	0.11
EW^-+SSM^-	1.83	1.05	1.50	0.47	1.45	1.34	1.52	1.61	2.78
EWW^-	0.68	0.13	0.61	0.11	0.55	0.09	0.04	0.11	0.11
EWW^-+SSM^-	1.87	1.76	1.66	0.68	1.96	1.55	1.53	1.62	2.79
	Ratios de desbalance							
	Phoneme	Pima	Satimage	Scrapie	Segmentation	Spambase	Vehicle	Yeast	
Original	0.42	0.54	0.11	0.21	0.17	0.65	0.33	0.41	
EW^-	0.45	0.78	0.11	0.56	0.17	0.81	0.41	0.51	
EW^-+SSM^-	2.80	2.49	1.11	3.46	1.73	3.59	1.55	1.50	
EWW^-	0.47	0.84	0.12	0.56	0.17	0.82	0.44	0.59	
EWW^-+SSM^-	3.24	2.79	1.11	3.46	1.79	3.65	1.83	1.91	

un importante desequilibrio en las distribuciones de clases, la ausencia de cualquier otra complejidad permite obtener buenos resultados sin apenas reducir la talla de la clase mayoritaria.

5.6.2. Solución a nivel de datos más solución a nivel de algoritmos

La segunda propuesta de combinación consiste en emplear la clasificación con distancia ponderada sobre los conjuntos submuestreados, donde el primero representa la solución a nivel de algoritmos y el segundo la solución a nivel de datos. El objetivo es solventar el desbalance (aún) presente en los conjuntos submuestreados, lo que en teoría incrementaría la tasa de aciertos sobre la clase positiva.

5.6.2.1. Experimentos y análisis de resultados

La configuración de estos experimentos se reduce a emplear la regla 1−NN con distancia ponderada sobre los conjuntos originales y los conjuntos submuestreados mediante cuatro diferentes esquemas: las dos versiones de edición de Wilson (EW^- y EWW^-) y sus correspondientes combinaciones con el condensado por selectivo modificado (EW^-+SSM^- y EWW^-+SSM^-).

Al analizar la Figura 5.8, podemos ver que las únicas estrategias que proporcionan buenos resultados corresponden a la combinación de clasificar con 1−NN+DP los conjuntos submuestrados mediante la edición de Wilson (tanto EW^- como EWW^-). La razón por la que las combinaciones con selectivo modificado no producen buenas tasas de efectividad debemos buscarla en el hecho de que el sentido del desbalance se ha invertido, tal como ya se comentó en experimentos anteriores.

5.7. Conclusiones preliminares

Las principales propuestas para tratar el problema del desbalance pueden agruparse en dos grandes categorías: soluciones a nivel de datos y soluciones a nivel de algorit-

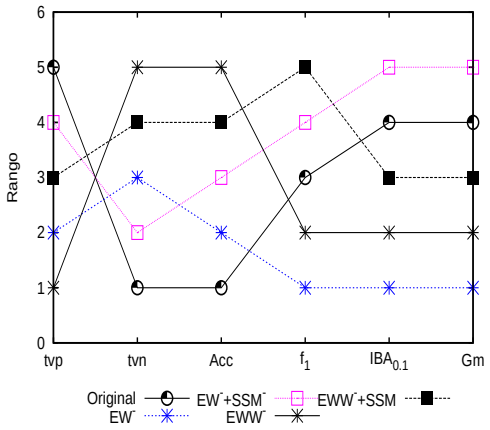


Figura 5.8: *Ranking de la combinación de* $1-NN+DP$ *con los conjuntos submuestreados empleando seis métricas de efectividad.*

mos. La diferencia entre ambas radica fundamentalmente en el procedimiento que emplean para beneficiar a la clase minoritaria. Las soluciones pertenecientes al primer grupo intentan favorecer a la clase positiva a través de la alteración artificial de la distribución de clases en el conjunto de entrenamiento con el objetivo de reducir el desbalance. Por el contrario, las soluciones a nivel de algoritmos suelen basarse en la ponderación de las muestras o en la inclusión de una matriz de costes.

Actualmente, es posible encontrar diversos estudios acerca de las ventajas e inconvenientes que cada solución aporta y bajo qué condiciones resulta mejor aplicar las diferentes propuestas.

En el presente capítulo, se han analizado diversas técnicas de submuestreo y sobremuestreo, es decir, soluciones a nivel de datos, así como una propuesta para la regla del vecino más cercano que puede considerarse como una solución a nivel de algoritmos. El estudio se ha llevado a cabo teniendo en cuenta tres factores que podríamos considerar críticos en el proceso de clasificación: el ratio de desbalance, la naturaleza del clasificador y la métrica de efectividad utilizada.

Al analizar las diferentes técnicas de remuestreo, los resultados revelan que la utilidad de éstas depende fuertemente de la naturaleza del clasificador. Así, en general, es posible afirmar que para clasificadores con un aprendizaje global, independientemente del ratio de desbalance que presente el conjunto de datos, la mejor estrategia consiste en sobremuestrear la clase minoritaria. Sin embargo, en el caso de modelos más locales (por ejemplo, el clasificador $1-$NN), si el conjunto de datos presenta un alto desbalance, los resultados experimentales sugieren que la mejor opción viene determinada por los algoritmos de sobremuestreo; por el contrario, cuando existe un nivel de desbalance moderado o bajo, el submuestreo aparece como la solución más adecuada. Sin embargo, aún así, debemos tener en cuenta que estas conclusiones dependerán también de la métrica de efectividad que vayamos a emplear.

Por otra parte, en diversas publicaciones, se ha mencionado que el remuestreo

aleatorio constituye una estrategia bastante competitiva frente a las técnicas dirigidas. En este sentido, se ha podido observar que una de las principales desventajas de los algoritmos dirigidos es que, en la mayoría de las ocasiones, no se logra un balance absoluto entre las clases, llegando en algunos casos a mantener un ratio de desbalance igual al conjunto original y, en el peor de los casos, se llega a invertir el sentido del desbalance, siendo esto especialmente contraproducente en los clasificadores de naturaleza global. El hecho de obtener una variabilidad en las tasas de reducción sugiere que estos algoritmos dirigidos dependen fuertemente de la complejidad de los datos. Asimismo, fue posible observar que no siempre la mejor distribución es aquella que presenta un equilibrio perfecto entre las clases.

Finalmente, la incorporación de un factor de ponderación en la regla 1−NN ha exhibido un buen comportamiento, especialmente al combinarse con alguna técnica de submuestreo. Sin embargo, al igual que ocurría con los esquemas anteriores, el beneficio sobre la clase minoritaria parece conllevar paralelamente una leve degeneración de la clase mayoritaria.

Capítulo 6

Análisis del comportamiento de algoritmos de clasificación en conjuntos de datos con solapamiento y desbalance

Generalmente, se asume que la diferencia significativa en el número de muestras por clase es una de las causas más importantes de la baja tasa de efectividad de un clasificador sobre la clase menos representada. Sin embargo, existen otros factores que, en combinación con el desbalance, causan dicha degradación, siendo uno de éstos el solapamiento. En el capítulo actual, se desarrolla una serie de experimentos para evaluar la influencia que ejercen el desbalance y el solapamiento en la efectividad de la regla del vecino más cercano. Este estudio se extiende a diversos clasificadores de distinta naturaleza, con la finalidad de comprobar en qué medida la influencia que ejercen ambas complejidades sobre los resultados de clasificación depende del tipo de clasificador.

6.1. Introducción

En el problema de los conjuntos desbalanceados, una de las líneas de investigación se centra en el estudio de la influencia que ejerce dicha complejidad sobre los resultados de clasificación. Diversos estudios han demostrado que, en entornos desbalanceados, los clasificadores presentan una pérdida significativa de efectividad debido al sesgo que existe en las distribuciones de las clases (Al-Haddad y otros, 2000, Fernández y otros, 2009, Japkowicz y Stephen, 2002, Kubat y otros, 1997, Mazurowski y otros, 2008, Weiss y Provost, 2003, Zhang y Mani, 2003). No obstante, múltiples investigaciones sugieren que existen otros muchos factores que contribuyen a tal degradación de la efectividad, como pueden ser la talla global del conjunto de datos, el nivel de desbalance, la densidad, el solapamiento entre clases, entre otros (Al-Haddad y otros, 2000, Batista y otros, 2005, Hulse y Khoshgoftaar, 2009, Japkowicz y Stephen, 2002, Jo y Japkowicz, 2004, Murphey y otros, 2004, Prati y otros, 2004a, Visa y Ralescu, 2003, Weiss, 2003, Weiss y otros, 2007).

Estudios llevados a cabo empleando clasificadores C4.5 y difusos (Fuzzy Classifier) permiten concluir que, si bien es cierto que un gran porcentaje de errores ocurren en la zona de confusión, existe un punto en el cual el desbalance entre

las clases predomina más que el solapamiento (Japkowicz y Stephen, 2002, Visa y Ralescu, 2003). Con respecto a esto, podría ser interesante identificar el grado de influencia que ejerce cada una de estas complejidades (y sus interdependencias) en la operación de cada clasificador. En este sentido, podría intentarse relacionar la naturaleza local o global de un clasificador con el grado y tipo de afectación en sus resultados. En el presente estudio, se han considerado los clasificadores naïve Bayes (NBC), la red neuronal perceptrón multicapa (MLP), la red neuronal de funciones de base radial (RBF), el árbol de decisión (J48) y el clasificador del vecino más cercano con $k = 1, 3, 7, 9, 13$, poniéndose especial énfasis en este último.

6.2. $k-$NN: un escenario infrecuente de desbalance y solapamiento

La regla $k-$NN ha sido históricamente uno de los algoritmos más estudiados y analizados en la literatura. Diversos estudios teóricos y empíricos han permitido concluir que este clasificador es extremadamente sensible a ciertos tipos de complejidades de los datos, principalmente a las imperfecciones en los conjuntos de datos y a la densidad de las muestras por clases (Dasarathy, 1991, Devijver y Kittler, 1982). Entre otras características, este método requiere de altas densidades para poder ofrecer estimaciones confiables de las clases, lo cual suele estar directamente relacionado con la talla de las mismas. Por tanto, el algoritmo k-NN es susceptible al nivel de desbalance, pues cuanto más representada (o densa) sea una clase, mejores serán los resultados sobre dicha clase (Hand y Vinciotti, 2003, Kubat y otros, 1997, Zhang y Mani, 2003). En realidad, esta dependencia es especialmente crítica en las regiones de solapamiento, donde las decisiones entre dos (o más) clases depende de la relación entre las densidades locales.

En un problema desbalanceado de dos clases, donde la clase mayoritaria está generalmente más representada en la región de solapamiento, los resultados de clasificación del k-NN son normalmente mejores en esta clase que en la minoritaria. Así, conforme el valor de k y/o el desbalance se incrementan, este efecto se acentúa con una reducción progresiva de la influencia de la clase minoritaria. La situación anterior puede representar el comportamiento esperado pero, ¿es posible encontrar situaciones donde el clasificador k-NN beneficie a la clase minoritaria?. En este sentido, ¿es probable que, debido a la naturaleza local del algoritmo, el desbalance en la región de solapamiento influya más en la efectividad del k-NN que el propio volumen (o talla) de la región de solapamiento? y, finalmente, considerando que la operación del k-NN depende de sus densidades locales, ¿podría decirse que el desbalance global es mucho menos importante para el k-NN que para otros clasificadores con una naturaleza menos local?

Estas tres cuestiones que acabamos de plantear establecen una línea de investigación a través del análisis de la efectividad del algoritmo k-NN, empleando dos series de datos sintéticos con complejidades opuestas en las regiones de solapamiento:

Serie 1 El ratio de desbalance en la región de solapamiento es igual al ratio de

desbalance global (caso general).

Serie 2 El tipo de desbalance en la región de solapamiento es inverso al ratio global, es decir, la clase minoritaria es más densa que la clase mayoritaria en la zona de solapamiento (podría verse como un caso atípico).

Para tal propósito, se han planteado dos estudios sobre conjuntos de datos sintéticos de dos clases con un nivel fijo de desbalance global, con el objetivo de que los resultados no dependiesen de dicho parámetro.

El primer experimento consiste en una serie de datos en la que ambos ratios de desbalance, el de la región de solapamiento y el global, son idénticos, mientras que el tamaño de la región de solapamiento cambia. Esta situación, que puede considerarse típica, proporcionará criterios y bases para analizar los resultados del segundo caso.

El segundo experimento se lleva a cabo sobre una serie de datos caracterizada por una mayor densidad local de la clase minoritaria, con respecto a la mayoritaria, en la zona de solapamiento. Para esta situación, se llevó a cabo otro estudio considerando una relación de desbalance de 1:50, con la finalidad de verificar los resultados experimentales en un escenario con un desbalance muy acentuado.

Finalmente, con el objetivo de responder a la tercera pregunta, se realizaron experimentos complementarios empleando otros cuatro algoritmos de clasificación de diferente naturaleza. Esto permitió comparar los resultados obtenidos para la regla k-NN en los experimentos antes descritos con los producidos por el esto de clasificadores, así como analizar los efectos del desbalance sobre los resultados de clasificación en relación a la naturaleza local o global del propio clasificador.

6.3. Experimentos y análisis de resultados

Los dos escenarios descritos en la sección anterior se desarrollaron a través de dos series de conjuntos de datos sintéticos. Para cada conjunto, se generaron 500 muestras pseudoaleatorias con una distribución uniforme en un espacio bidimensional de 100 $\times$ 100. De ellas, 400 son de la clase mayoritaria y 100 de la clase minoritaria, manteniendo así un relación igual a 1:4. Es importante mencionar que, aún cuando sólo una dimensión aparece como discriminante, la inclusión de dos dimensiones facilita la distribución de las muestras y la interpretación de los resultados, principalmente en las regiones de solapamiento en situaciones de grandes tallas o volúmenes.

El primer escenario constituye un típico problema de desbalance con solapamiento, de forma que el desequilibrio afecta de igual manera a todo el espacio de representación. El segundo escenario se refiere a una situación poco común, donde el desbalance en la región de solapamiento es inverso al desbalance global, es decir, la clase mayoritaria y minoritaria intercambian sus roles. Con esta situación singular es posible explicar la dependencia del desbalance global y local sobre la regla k-NN, así como del volumen de la región de solapamiento. Se incluyó un tercer grupo de experimentos con datos sintéticos con una relación de desbalance de 1:50 con el objetivo de corroborar las conclusiones obtenidas en el segundo escenario.

Los resultados mostrados son promedios de 10 repeticiones de una validación cruzada de 10 particiones, es decir, cada promedio es obtenido a partir de 100 resultados. Los experimentos consisten en el cálculo de siete métricas de evaluación de la efectividad de un clasificador apropiadas para entornos desbalanceados: *i)* la tasa de verdaderos positivos (*tvp*), *ii)* la tasa de verdaderos negativos (*tvn*), *iii)* *precision*, *iv)* la media geométrica (*Gm*) de tasas de aciertos por clase, *v)* la media geométrica de *precision* y de la tasa de aciertos de la clase positiva (*Gpr*), *vi)* el área bajo la curva ROC (*AUC*) y *vii)* el índice de precisión balanceada ($IBA_{0.1}$). Asimismo, dentro de este análisis, también se incluyó la tasa de precisión global (*Acc*), que se caracteriza por presentar un comportamiento sesgado hacia la clase mayoritaria. Por otra parte, cabe recordar que los clasificadores empleados fueron tomados del kit de herramientas Weka (Witten y Frank, 2005).

6.3.1. Desbalance homogéneo

El primer experimento consta de una colección de seis conjuntos de datos con un solapamiento que se incrementa gradualmente. En todos los casos, las muestras positivas (100) están definidas en el eje X en el rango [50–100], mientras que aquellas que pertenecen a la clase mayoritaria (400) son generadas dentro de [0–50] para 0 % de solapamiento, [10–60] para 20 %, [20–70] para 40 %, [30–80] para 60 %, [40–90] para 80 % y [50–100] para el caso de un solapamiento total, 100 %. Como se puede observar, en todos los casos, el ratio de desbalance global es igual al ratio de desbalance en la región de solapamiento, por lo que podría verse como un caso típico que se caracteriza por presentar un desbalance homogéneo. La Figura 6.1 muestra dos ejemplos de estos conjuntos de datos.

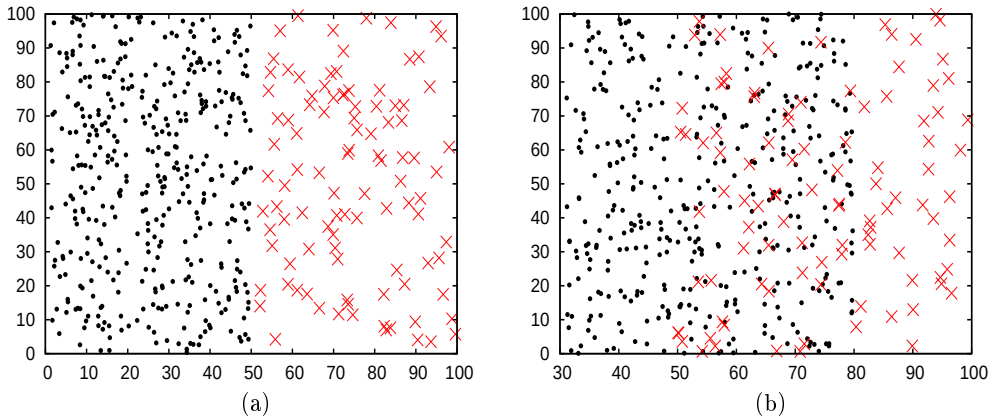


Figura 6.1: *Dos casos diferentes de desbalance homogéneo: a) 0 % de solapamiento y b) 60 % de solapamiento*

6.3.1.1. Comportamiento del algoritmo k-NN

El algoritmo k-NN aquí empleado utiliza la siguiente configuración: distancia Euclídea para valores de k entre 1 y 15. Por simplicidad y claridad, en las figuras sólo se representan resultados para valores de $k = 1, 3, 7, 9, 13$.

La Figura 6.2 exhibe resultados de ocho medidas de efectividad para todas las versiones de k-NN a medida que el solapamiento se incrementa. A partir de las dos primeras gráficas correspondientes a *tvp* y *tvn*, es posible observar dos situaciones importantes. Primero, conforme el solapamiento se va incrementando, las tasas de reconocimiento sobre las muestras positivas decrece de forma rápida y significativa. Una situación distinta ocurre en el caso de la tasa de verdaderos negativos (para cualquier valor fijo de k), la cual decrece aunque muy moderadamente. Segundo, el incremento en el valor de k da lugar a que el algoritmo sea cada vez menos local, de modo que, para cualquier porcentaje fijo de solapamiento, disminuye gradualmente la tasa de verdaderos positivos mientras que *tvn* se incrementa hasta alcanzar valores cercanos al 100 %.

Con respecto a la primera cuestión, el incremento en los niveles de solapamiento implica también un aumento en la cantidad de ejemplos de ambas clases en la región de confusión, en la que se mantiene la relación global entre las densidades (la clase negativa es más densa que la clase positiva). Esto ocasiona que aumente significativamente la probabilidad de que el algoritmo k-NN clasifique erróneamente las muestras positivas. Por lo que respecta al segundo punto, con el incremento de los valores de k y del volumen local de decisión, se eleva la probabilidad de encontrar una mayoría de la clase negativa y, por tanto, clasificar erróneamente muestras de la clase positiva.

Como conclusión preliminar, es posible decir que el incremento del nivel de solapamiento o del valor de k en una situación de desbalance homogéneo o típico afecta de forma significativa a la clase minoritaria. Por otra parte, es importante recordar que los resultados fueron obtenidos considerando un ratio fijo de desbalance, lo cual sugiere que el desbalance global por sí mismo no afecta necesariamente el rendimiento del algoritmo k-NN.

Finalmente, a partir de las métricas globales de efectividad, se observan dos situaciones particulares. Primero, *precision*, *Gpr*, *AUC* y *Acc* (ver Figuras 6.2(c), 6.2(e), 6.2(f), 6.2(h)) muestran una tendencia a favorecer a aquellos clasificadores que presentan altas tasas de aciertos sobre la clase negativa, es decir, a las versiones de k-NN con mejores valores de k. Segundo, las métricas como *Gm* e $IBA_{0.1}$ de naturaleza no sesgada (ver Figuras 6.2(d), 6.2(g)), a diferencia de las anteriores, muestran como esquema más discriminante al algoritmo 1-NN para la mayoría de los conjuntos de datos con diversos grados de solapamiento. La razón de este comportamiento se debe a que dicho clasificador presenta un mejor *tvp* que el resto de los de k-NN con $k > 1$.

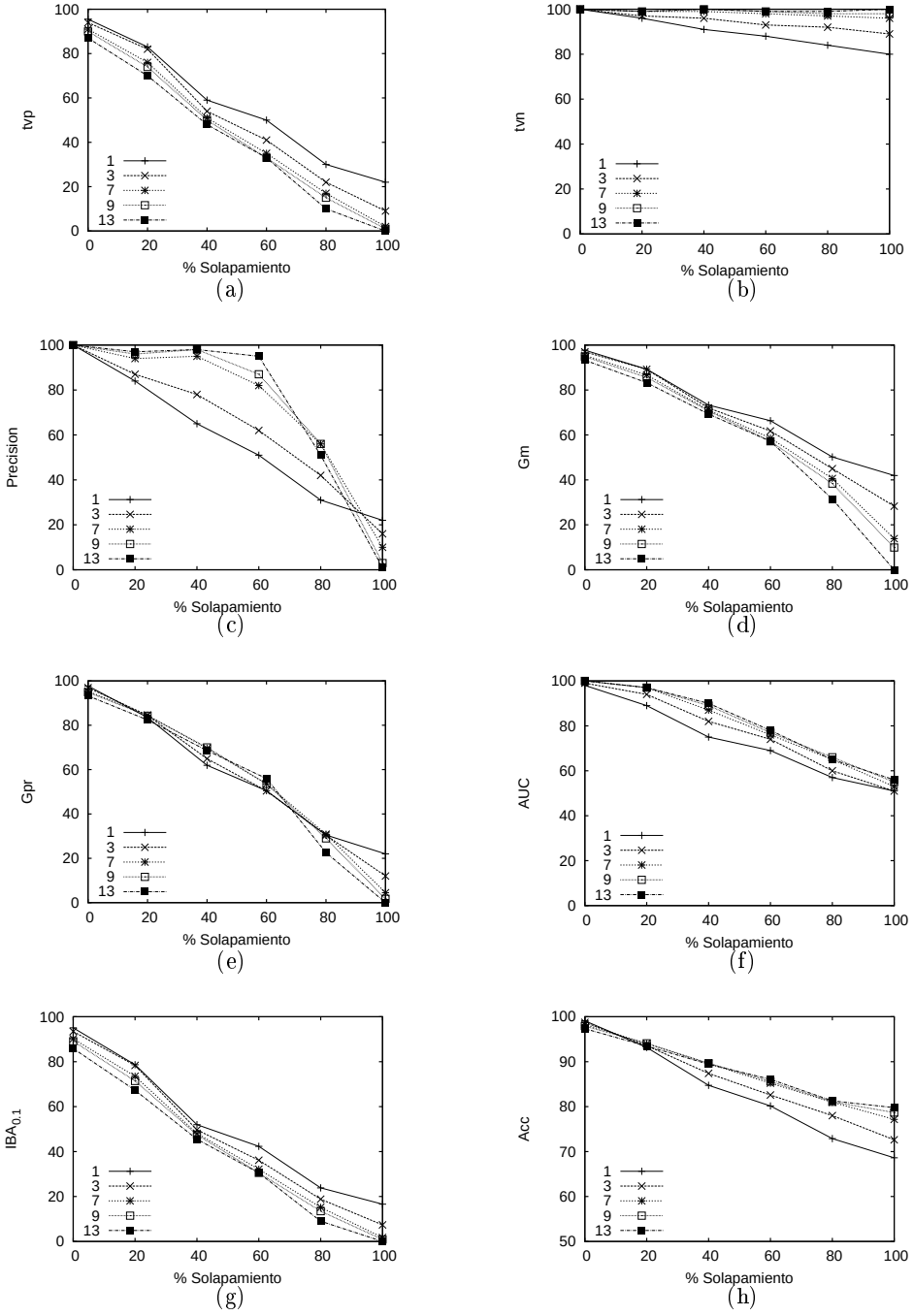


Figura 6.2: *Desbalance homogéneo: evaluación de la efectividad de los clasificadores* $1, 3, 7, 9$ *y* 13*-NN según las métricas (a) tvp, (b) tvn, (c) precision, (d) Gm, (e) Gpr, (f) AUC, (g)* $IBA_{0.1}$ *y (h) Acc*

6.3.1.2. k-NN frente a otros modelos de clasificación

En esta sección, el algoritmo k-NN se compara con otros cuatro clasificadores de naturaleza diversa, empleando para ello la misma serie de datos desbalanceados del apartado anterior. Estos cuatro métodos son la red MLP, el algoritmo naïve Bayes NBC, la red neuronal RBF y el árbol de decisiones J48. En el caso de k-NN, sólo se usaron valores de $k = 1, 3, 9$, ya que se considera que son suficientes para demostrar las tendencias locales de este esquema.

Los resultados son conceptualmente similares a los de la sección anterior: los esquemas más locales reconocen mejor la clase minoritaria (especialmente para grandes niveles de solapamiento), mientras que los modelos basados en un aprendizaje global son más robustos al clasificar a las muestras de la clase mayoritaria (ver Figuras 6.3(a), 6.3(b)). En particular, la regla 1-NN produce las mejores tasas de reconocimiento sobre las muestras de la clase positiva. Por el contrario, el clasificador MLP, cuyo aprendizaje es global, se muestra como el mejor modelo de clasificación para la clase mayoritaria. En este sentido, el aumento de la precisión global del algoritmo k-NN sugiere que dicho clasificador va adoptando un comportamiento menos local a medida que se incrementa el valor de k.

Puesto que los valores de tvn son significativamente más altos que los de tvp en todos los clasificadores, los métodos que mejor reconocen la clase negativa son también los que generalmente alcanzan las más altas tasas de efectividad global (ver Figuras 6.3(c), 6.3(e), 6.3(f), 6.3(h)). En este sentido, el clasificador MLP resulta ser el más consistente, mientras que el algoritmo 1-NN es uno de los que presentan el peor comportamiento para estas cuatro medidas de efectividad global. La excepción vuelve a presentarse con la media geométrica y el índice de precisión balanceada (ver Figuras 6.3(d), 6.3(g)), donde el clasificador 1-NN supera al resto de los clasificadores en los escenarios con un alto grado de solapamiento. Cabe mencionar que estos resultados se han obtenido con las métricas más empleadas en problemas desbalanceados. Aún así, el uso de la conocida medida de precisión global (Acc, ver Figura 6.3(h)) conduce a resultados similares (clasificadores globales son mejores que los locales), pero con diferencias más acentuadas.

Una vez más los resultados confirman el hecho de que la clase mayoritaria tiende a ser mejor discriminada por aquellos clasificadores que se basan en un aprendizaje global. En el caso de la clase minoritaria, los clasificadores con un aprendizaje local presentan un mejor comportamiento sobre dicha clase. Al igual que en la sección anterior, esta conclusión es válida para las situaciones con un desbalance homogéneo (el desbalance global es similar al desbalance en la región de solapamiento).

6.3.2. Desbalance heterogéneo

El segundo experimento se llevó a cabo sobre una colección de cinco conjuntos de datos sintéticos desbalanceados, donde la clase minoritaria se convierte en mayoritaria en la región de solapamiento. Para este fin, las 400 muestras negativas que conforman la clase mayoritaria se distribuyeron sobre el eje X en el rango [0–100]

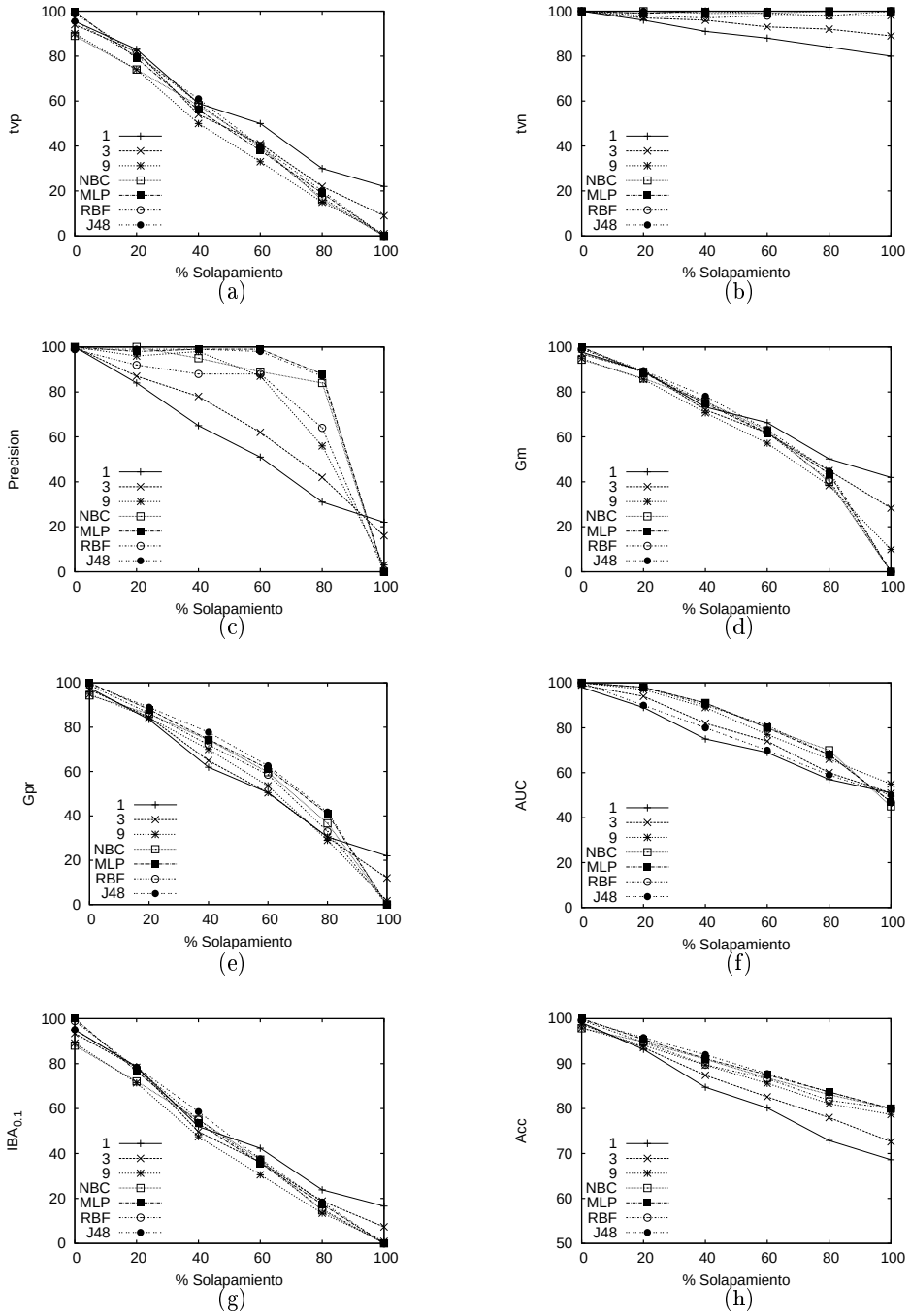


Figura 6.3: *Desbalance homogéneo: evaluación de la efectividad de los clasificadores* $1, 3, 9$-*NN, NBC, MLP, RBF y J48 según las métricas (a) tvp, (b) tvn, (c) precision, (d) Gm, (e) Gpr, (f) AUC, (g)* $IBA_{0.1}$ *y (h) Acc*

para todos los conjuntos de datos, mientras que las 100 muestras de la clase mi-

noritaria se generaron en un rango distinto para cada conjunto de datos: [75–100], [80–100], [85–100], [90–100] y [95–100]. Nótese que la región que ocupa la clase minoritaria coincide, en todos los casos, con la región de solapamiento. Con esto, se logra que el número de elementos en esta región varíe de una situación balanceada en la primera serie de datos, en la cual ambas clases presentan los mismos valores esperados en número de muestras y densidad, hasta una situación crítica de desbalance como ocurre en la última colección de datos, en la que las 100 muestras de la clase minoritaria aparecen como la clase dominante en la región de solapamiento, donde la clase mayoritaria debería contar, aproximadamente, con 20 muestras. En la Figura 6.4, se muestran dos ejemplos de estos conjuntos de datos. En resumen, la clase minoritaria (positiva) se convierte en la clase más representada en la región de solapamiento. A situaciones como ésta, que se caracteriza porque el desbalance global difiere del desbalance en las regiones locales, nos referiremos como desbalance heterogéneo.

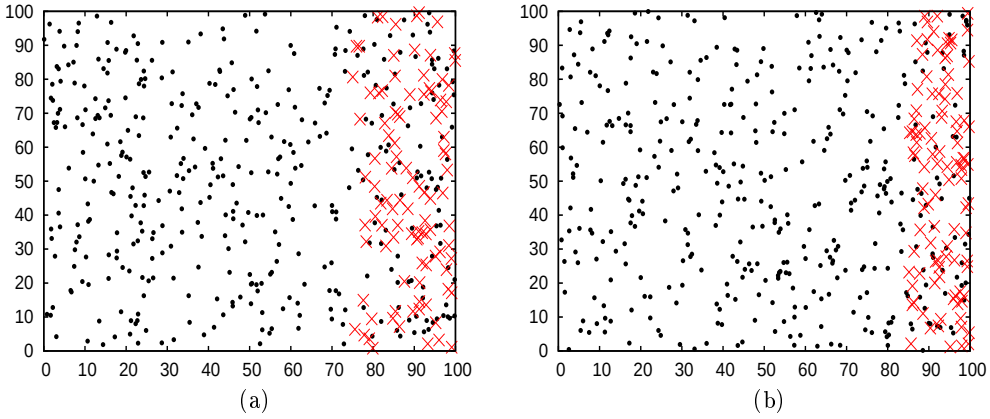


Figura 6.4: *Dos casos diferentes de desbalance heterogéneo: [75–100] y [80–100]. El primer caso representa una situación balanceada en la región de solapamiento. Para el último caso, en dicha región, la clase mayoritaria aparece menos representada que la clase minoritaria*

El objetivo de este experimento es analizar la importancia del ratio de desbalance global, la talla de la región de solapamiento y el ratio de desbalance local en dicha región, dentro del comportamiento de diversos algoritmos de clasificación.

6.3.2.1. Comportamiento del algoritmo k-NN

En el análisis del comportamiento del clasificador k-NN, se emplearon los mismos valores de k y la misma distancia que en los experimentos de la Sección 6.3.1.1.

La Figura 6.5 muestra los valores de las ocho métricas de efectividad para todas las versiones de k-NN en función del solapamiento (de mayor a menor volumen de solapamiento). Analizando las dos primeras medidas, tvp y tvn, es posible remarcar dos puntos. En primera instancia, conforme el valor de k se incrementa y, al contrario del caso típico analizado en la Sección 6.3.1.1, la tasa de reconocimiento de la

clase positiva aumenta en los tres casos en los que la diferencia entre las densidades en la región de solapamiento es más significativa, mientras que la tasa de reconocimiento de las muestras negativas cae ligeramente para la mayoría de las situaciones. En segundo lugar, conforme el volumen en la zona de solapamiento se reduce, la densidad de la clase positiva aumenta mientras que se reduce el número de muestras negativas en dicha región. Esto ocasiona que *tvp* se incremente al tiempo que *tvn* se mantiene casi constante.

Con respecto al primer punto, a diferencia de los primeros experimentos con un desbalance homogéneo, el aumento en los valores de k beneficia claramente a la clase minoritaria. Sin embargo, al igual que en los experimentos del primer escenario, esta clase es la más representada en la zona de solapamiento. Esto confirma que el algoritmo k-NN depende, en realidad, del desbalance local en la zona de solapamiento y no tanto del desbalance global.

En el caso del segundo punto, la reducción del volumen de la zona de solapamiento afecta a ambas clases de diferente forma. La densidad de la clase positiva se incrementa debido a que la región de solapamiento engloba a todas las muestras positivas, mientras que para la clase negativa, la desventaja creciente en términos de densidad local se compensa con una cantidad cada vez menor de muestras en la región de solapamiento. Como resultado, el algoritmo k–NN (para todos los valores de k) mejora la tasa de reconocimiento de la clase positiva (toda la clase minoritaria), mientras que se producen curvas estables de tasas de aciertos sobre la clase negativa (la mayoritaria). La explicación de este resultado sobre la clase positiva se debe al incremento de la densidad de ésta con respecto a la clase negativa, por lo que se incrementa la probabilidad de que el clasificador mejore la tasa de reconocimiento sobre las muestras positivas. En el caso de los resultados de la clase negativa, cada nueva reducción en el volumen de la zona de solapamiento deja nuevas muestras negativas fuera de dicha zona, lo cual reduce la cantidad de las que quedan coexistiendo con la clase positiva. Por consiguiente, muchas de las primeras pasarán a ser bien clasificadas, mientras que la mayoría de estas últimas tendrán una mayor probabilidad de ser etiquetadas erróneamente. Ambos efectos se compensan entre sí, produciendo curvas de precisión global más estables, como ocurre en los casos centrales de solapamiento (85, 90 y 95–100), en cuyas zonas de confusión la clase minoritaria es más densa que la mayoritaria, siendo cada vez menos representada.

Este resultado puede también obtenerse para una región fija de solapamiento, la cual contenga todas las muestras positivas mientras se reduce el número de muestras negativas en dicha zona (manteniéndose ambas clases uniformemente distribuidas). En otras palabras, este comportamiento del k-NN no depende tanto del grado de solapamiento, sino de la relación entre las densidades locales de ambas clases en la región de solapamiento. El ejemplo muestra que, cuando los datos solapados no están balanceados, el ratio de desbalance en dicha zona puede ser más importante para la efectividad del k-NN que la propia talla del solapamiento.

Por último, las cinco medidas globales de rendimiento, Gm, Gpr, AUC, $IBA_{0.1}$ y Acc (ver Figuras 6.5(d), 6.5(e), 6.5(f), 6.5(g), 6.5(h)) muestran como tendencia general en muchos de los casos de solapamiento que, a un mayor valor de k, mejor

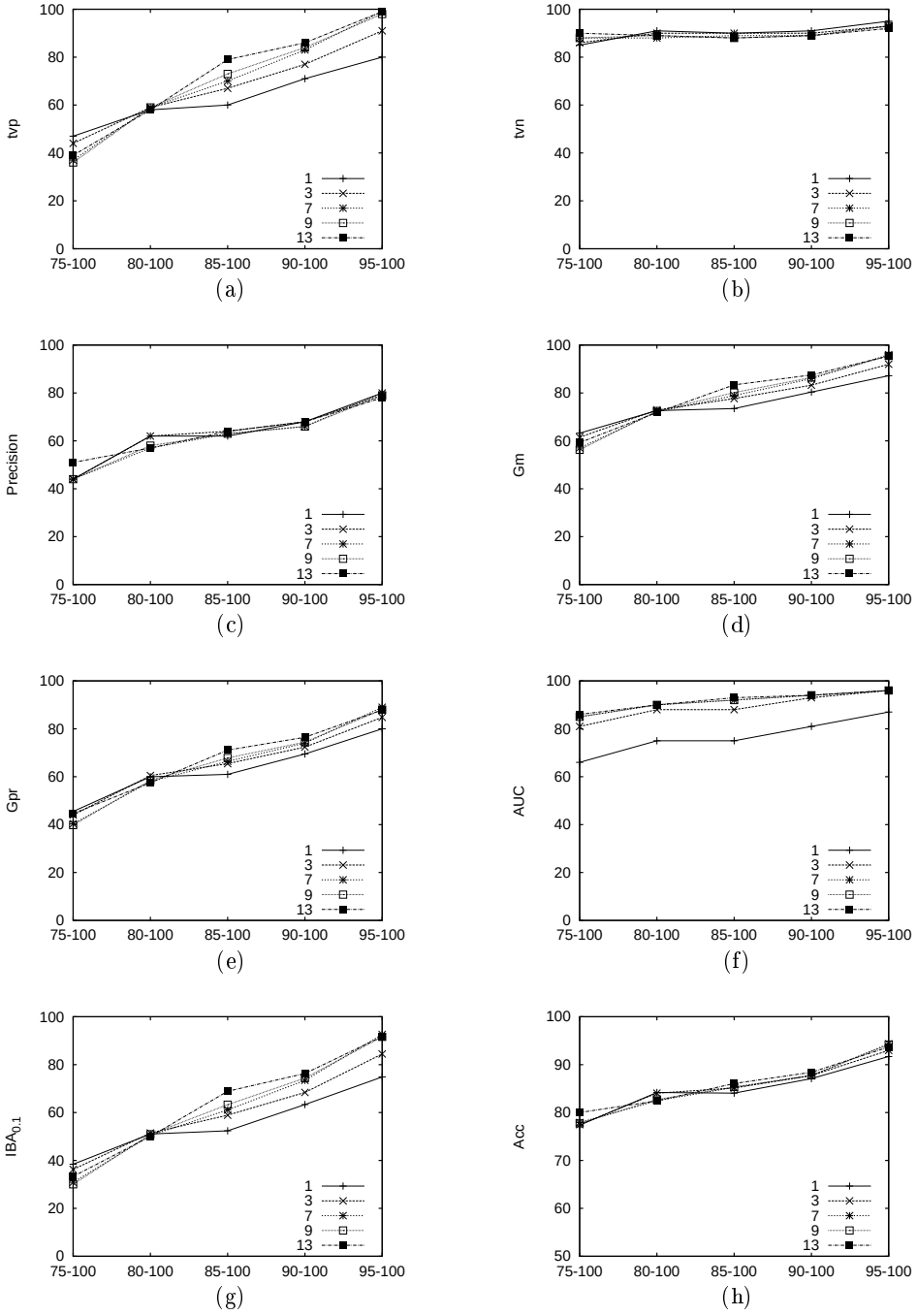


Figura 6.5: *Desbalance heterogéneo: evaluación de la efectividad de los clasificadores* $1, 3, 7, 9$ *y* 13*-NN según las métricas (a) tvp, (b) tvn, (c) precision, (d) Gm, (e) Gpr, (f) AUC, (g)* $IBA_{0.1}$ *y (h) Acc*

es la efectividad. Una revisión de la discusión anterior y una vistazo a los valores de

tvp y *tvn* (ver Figuras 6.5(a), 6.5(b)) permiten explicar este comportamiento.

6.3.2.2. k-NN frente a otras aproximaciones

Al igual que la Sección 6.3.1.2, los clasificadores 1, 3, 9–NN se comparan aquí con los clasificadores MLP, NBC, RBF y el árbol de decisión J48, aunque, en este caso, la comparación se realiza sobre conjuntos de datos con un desbalance heterogéneo.

En los experimentos anteriores, el clasificador 1-NN, que corresponde a la versión más local del k-NN, mostró ser el más apropiado para reconocer la clase menos representada en la zona de solapamiento. En la Sección 6.3.1, esta clase fue la minoritaria, mientras que en la Sección 6.3.2.1 ocurrió para la clase mayoritaria. Por lo que respecta a los clasificadores con un aprendizaje global, éstos presentan un mejor comportamiento sobre la clase que está fuertemente representada en la región de confusión, en particular, la clase mayoritaria en la Sección 6.3.1 y la clase minoritaria en la Sección 6.3.2.

Los nuevos resultados de esta sección coinciden con las deducciones obtenidas en la sección anterior. Los clasificadores con un aprendizaje más global, como el MLP, el NBC y el J48, logran altos valores de *tvp* (ver Figura 6.6(a)), mientras que los modelos como k-NN y RBF, los cuales están basados en un aprendizaje más local, obtienen mejores tasas de acierto sobre la clase negativa (ver Figura 6.6(b)). Estos resultados son más visibles para los casos centrales de solapamiento (85, 90 y 95–100), los cuales excluyen los dos casos de desbalance más extremos (uno moderado y otro intenso).

Debido a que la mayor diferencia entre el aprendizaje local y el aprendizaje global se encuentra en las tasas de verdaderos positivos con respecto a las tasas de verdaderos negativos, aquellos clasificadores con altos valores de *tvp* son los que presentan también valores más altos en las medidas de efectividad global. En particular, MLP, NBC y J48 son notablemente mejores que la familia de clasificadores k-NN, considerando las medidas Gm, Gpr e $IBA_{0.1}$ (ver Figura 6.6(d), 6.6(e), 6.6(g)). Sin embargo, desde el punto de vista de *precision*, AUC y Acc (ver Figura 6.6(c), 6.6(f), 6.6(h)), la mayoría de los clasificadores presentan valores muy similares, por lo que sus curvas de efectividad se encuentran muy próximas.

Los experimentos de la Sección 6.3.1 sobre conjuntos de datos con un desbalance homogéneo sugerían que la naturaleza del aprendizaje local/global está interrelacionada con la condición minoritaria/mayoritaria de las clases. Sin embargo, el presente experimento demuestra que tal conclusión no es válida en el caso de un desbalance heterogéneo. Una conclusión más general podría relacionar la naturaleza del aprendizaje con la condición local minoritaria/mayoritaria de las clases en las regiones de solapamiento. En este sentido, se puede establecer que la clase más representada en las regiones de solapamiento tiende a ser mejor clasificada por los métodos que se basan en un aprendizaje global, mientras que la clase menos representada en tales regiones resulta mejor clasificada por medio de los métodos basados en un aprendizaje local

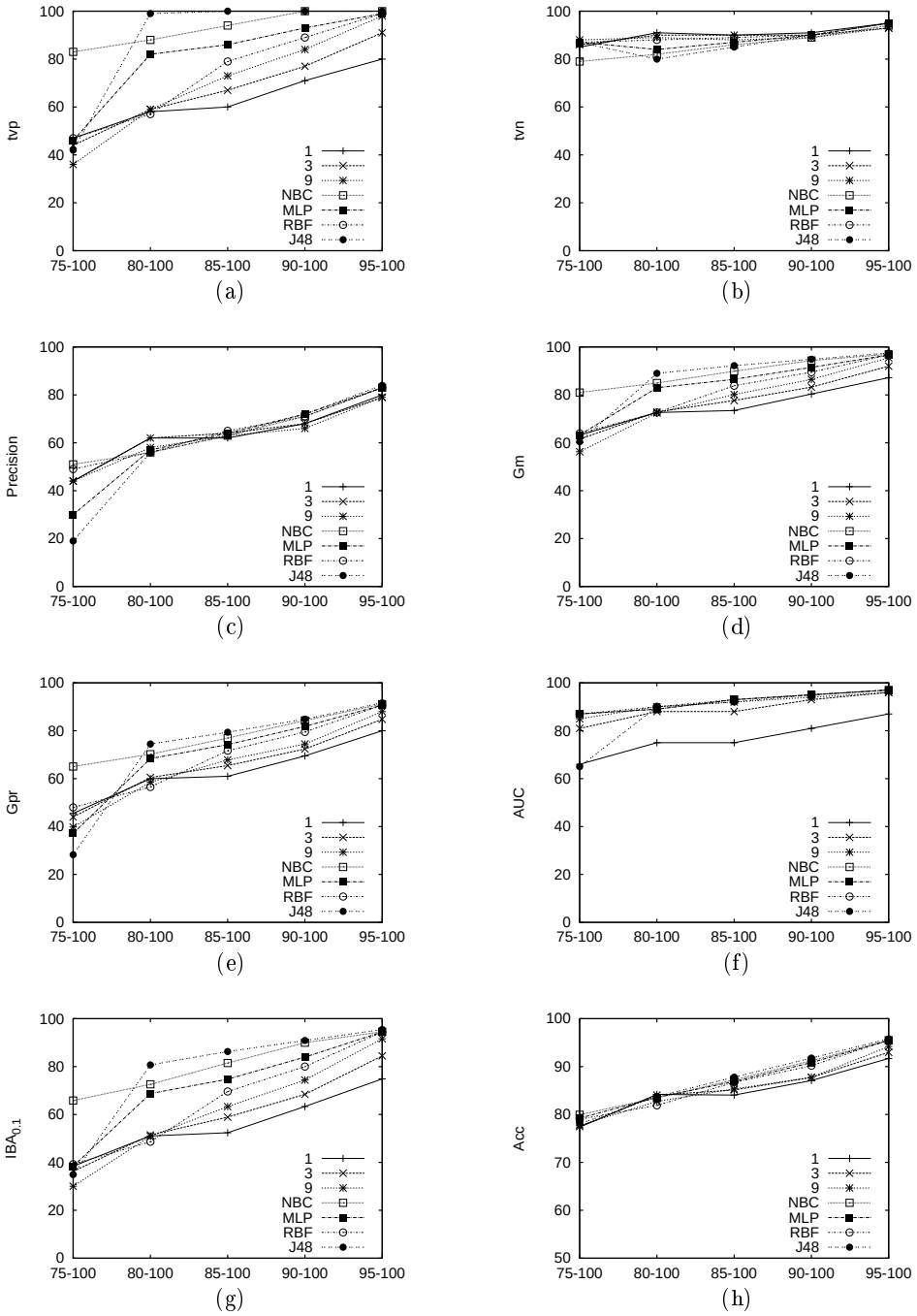


Figura 6.6: *Desbalance heterogéneo: evaluación de la efectividad de los clasificadores* $1, 3, 9$-*NN, NBC, MLP, RBF y J48 según las métricas (a) tvp, (b) tvn, (c) precision, (d) Gm, (e) Gpr, (f) AUC, (g)* $IBA_{0.1}$ *y (h) Acc*

6.3.3. Experimentos en bases de datos con un enorme ratio de desbalance

El tercer experimento se ha llevado a cabo sobre una colección de cinco conjuntos de datos sintéticos con un desbalance muy acentuado, teniendo como base la idea de los experimentos de la Sección 6.3.2. Los conjuntos de datos cuentan con dos clases concéntricas en el punto $(0, 0)$ y cuyas muestras se generaron a partir de funciones paramétricas variando aleatoriamente el radio y el ángulo entre $[0, 2\pi]$.

La clase mayoritaria consta de $5,000$ muestras situadas dentro de una región circular con radio igual a 50 (para todos los conjuntos de datos). Para definir la clase minoritaria, con el objetivo de producir un desbalance heterogéneo (en la región de solapamiento el desbalance difiere del desbalance global), se calculó de forma empírica los radios mínimos en los que es posible encontrar aproximadamente 20, 40, 60, 80 y 100 muestras de la clase mayoritaria. Los radios resultantes fueron 0.20, 0.40, 0.60, 0.80 y 1.00, y se obtuvieron aproximando los valores medios de los radios individuales estimados para 10 generaciones independientes de la clase mayoritaria. Posteriormente, para cada uno de ellos, se generaron las 100 muestras pertenecientes a la clase positiva, dando a lugar a los 5 problemas relacionados. En el caso del radio igual a 1.00, las cantidades esperadas de muestras de ambas clases dentro de la región de solapamiento son iguales, mientras que para el resto de los casos se espera, en dicha región, más muestras positivas que negativas.

El objetivo de este experimento es validar los resultados que se obtuvieron en la Sección 6.3.2, pero en una situación con un alto ratio de desbalance. La Figura 6.7 muestra los resultados de los siete algoritmos de aprendizaje evaluados en términos de las mismas métricas de efectividad que se emplearon en secciones anteriores.

Analizando la familia de los clasificadores k-NN, los resultados de *tvp* y *tvn* confirman que el desbalance local influye más que el desbalance global. El comportamiento sobre la clase positiva mejora conforme el valor de k se incrementa, principalmente en aquellos conjuntos de datos donde la clase minoritaria es más densa que la clase mayoritaria en la región de solapamiento (radio igual a 0.20, 0.40 y 0.60).

Por lo que respecta al resto de los clasificadores, éstos presentan resultados similares a los de la Sección 6.3.2. Sin embargo, se observan ciertas diferencias, como en el caso del clasificador MLP, donde para todas las series de datos, presenta una muy baja tasa de reconocimiento sobre la clase positiva. Por consiguiente, esta pérdida se ve reflejada de igual manera en las métricas de efectividad como *precision*, Gm, Gpr, AUC e $IBA_{0.1}$ (ver Figuras 6.7(c), 6.7(d), 6.7(e), 6.7(f), 6.7(g)). Este comportamiento nos sugiere que la naturaleza de su aprendizaje global en combinación con un alto ratio de desbalance global ocasiona una reducción significativa en la tasa de reconocimiento sobre la clase positiva. En el caso de los clasificadores NBC y J48, generalmente, presentan mejores tasas de efectividad que la familia de clasificadores k-NN y RBF.

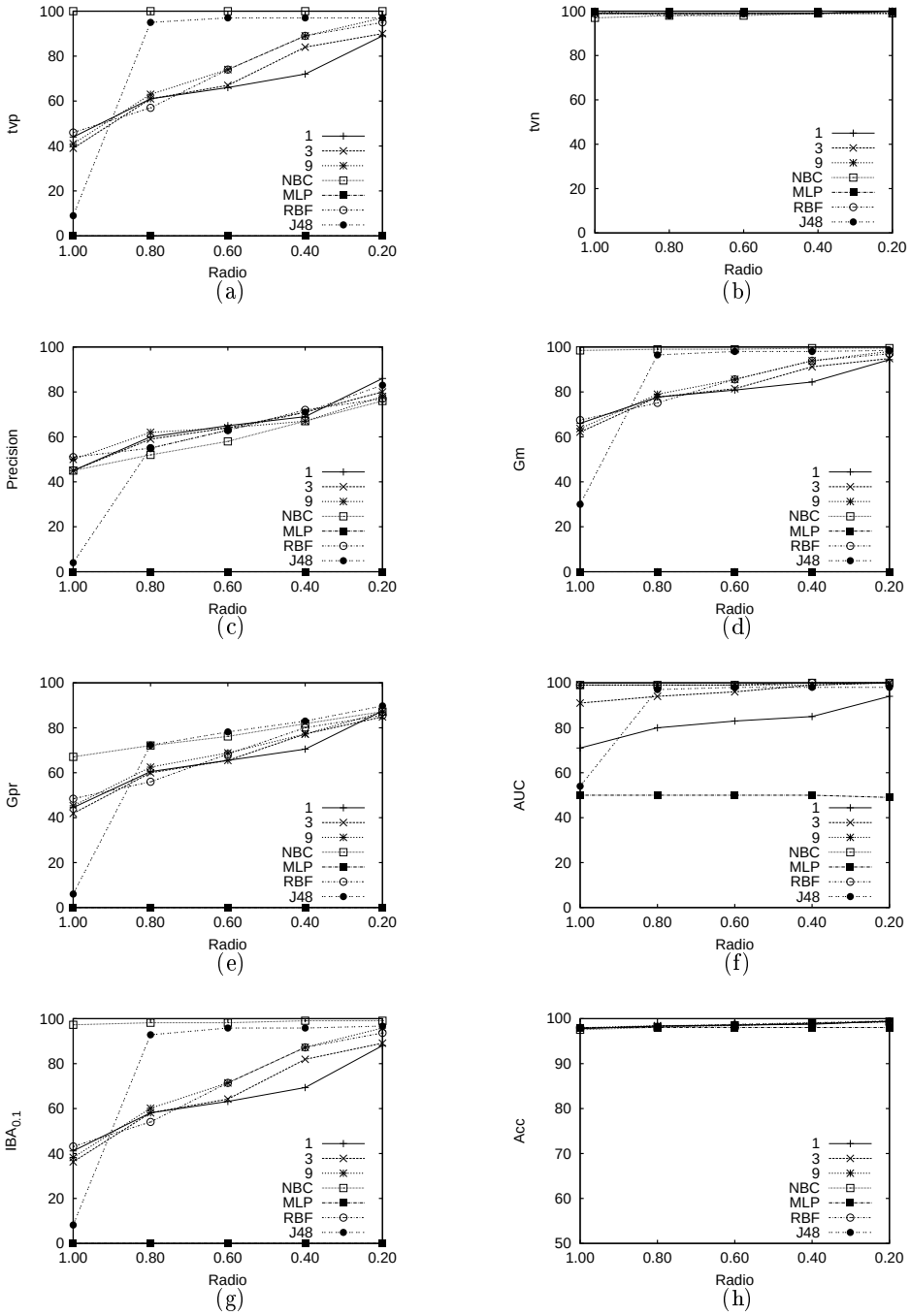


Figura 6.7: *Alto nivel de desbalance homogéneo: evaluación de la efectividad de los clasificadores* $1, 3, 9$*-NN, NBC, MLP, RBF y J48 según las métricas (a) tvp, (b) tvn, (c) precision, (d) Gm, (e) Gpr, (f) AUC, (g)* $IBA_{0.1}$ *y (h) Acc*

6.4. Conclusiones preliminares

El problema de las clases desbalanceadas y su relación con otras complejidades, como el solapamiento, se ha estudiado fundamentalmente desde el punto de vista de un desbalance homogéneo, es decir, asumiendo un ratio de desbalance uniforme en todo el espacio de representación de los datos.

En el presente capítulo, se ha diseñado un problema de dos clases, cuya complejidad, poco común, se caracteriza porque el ratio de desbalance en la zona de solapamiento es inverso al ratio de desbalance global. En otras palabras, la clase mayoritaria se convierte en minoritaria en la región de solapamiento.

El análisis se ha enfocado principalmente en analizar el comportamiento de los clasificadores k-NN considerando tres tipos de complejidades: el ratio de desbalance global, la talla de la zona de solapamiento y su ratio de desbalance local en dicha zona. Los resultados han mostrado que el algoritmo k-NN presenta una fuerte dependencia de los cambios en el ratio de desbalance local en la zona de solapamiento, mucho más significativa que de los cambios en la talla de dicha región. Los experimentos también han confirmado que el ratio de desbalance local y la talla de la región son más importantes que el ratio de desbalance global. Un comportamiento similar se ha podido observar en otro ejemplo con el mismo tipo de complejidad, aunque con un muy alto grado de desbalance.

El uso de cuatro clasificadores de distinta naturaleza, además de la regla k-NN, ha permitido estudiar la relación entre la efectividad en la clasificación de las clases minoritaria/mayoritaria y la naturaleza local/global de los clasificadores. Los resultados han indicado que las clases más representadas en la zona de solapamiento y, por tanto, localmente mayoritarias, tienden a ser mejor reconocidas por los métodos basados en un aprendizaje global, mientras que la clase menos representada en dicha región es mejor clasificada por los métodos locales. En este sentido, a un incremento en los valores de k, el comportamiento de la regla k-NN tiende a aproximarse progresivamente a los métodos globales.

Capítulo 7

Conclusiones y trabajos futuros

El desbalance en las distribuciones de clases se ha identificado como uno de los principales problemas que afectan de forma negativa al comportamiento de un sistema de clasificación. Se dice que un conjunto de dos clases está desbalanceado si una de ellas (clase minoritaria) está formada por un número de muestras muy inferior al de la otra clase (mayoritaria). Este problema aparece principalmente en aplicaciones donde la clase de mayor importancia corresponde a la menos representada, implicando además un elevado coste por los errores cometidos en la clasificación de las muestras de esta clase minoritaria.

Ante tal problemática, la gran mayoría de las investigaciones se han centrado en la búsqueda de soluciones que permitan tratar con el desequilibrio en las distribuciones por clases. Sin embargo, en los últimos años, han surgido otras líneas de investigación, como pueden ser la definición de nuevas métricas de efectividad y el análisis de la complejidad de los datos. La primera de ellas tiene por objetivo buscar medidas alternativas a la conocida tasa de precisión global, que permitan evaluar y representar de forma más efectiva los resultados de clasificación bajo condiciones de desbalance entre clases. La segunda cuestión está dirigida a determinar el grado de influencia que ejerce el sesgo presente en los datos sobre los resultados de clasificación, así como su relación con otras complejidades que pudieran presentar los datos.

La presente tesis se ha planteado analizar y revisar la problemática del desbalance, aunque no solamente enfocada a la búsqueda de soluciones, sino también a la evaluación de la efectividad y al estudio de situaciones en las que aparecen simultáneamente distintas complejidades. Este capítulo final resume las contribuciones principales y discute posibles direcciones futuras para extender la investigación dentro del problema del desbalance.

7.1. Contribuciones

El objetivo principal de esta tesis ha sido la de proporcionar una visión estructurada del problema del desbalance abarcando tres líneas de investigación, así como el desarrollo de nuevas herramientas capaces de tratar el problema del sesgo entre los datos.

7.1.1. Las métricas de efectividad

Una de las métricas más empleadas para evaluar la efectividad del clasificador es la tasa de precisión global. Sin embargo, en entornos desbalanceados, resulta poco

útil debido a que se ve fuertemente influenciada por la clase de mayor representatividad, así como también a que considera iguales costes de error para ambas clases. En este sentido, algunas de las actuales medidas de evaluación constituyen buenas alternativas al emplear ciertas combinaciones de métricas más puras. Sin embargo, a pesar de que han demostrado su utilidad, no consideran dentro de sus resultados finales la información que pudiera aportar el desbalance, lo cual puede conducir en algunos casos a decisiones poco confiables.

Con la finalidad de solventar el inconveniente que acabamos de comentar, en el Capítulo 4 se ha desarrollado una nueva métrica de evaluación que resulta de la combinación de un índice básico que mide el equilibrio entre las tasas de acierto por clase y una métrica global no sesgada. Esta nueva propuesta favorece a aquellos resultados que presenten altas tasas de aciertos sobre la clase positiva.

Estudios experimentales llevados a cabo evaluando propiedades invariantes a partir de cambios en la matriz de confusión y coeficientes de correlación de Pearson han mostrado que el *índice de precisión balanceada* (IBA) se presenta como una métrica adecuada para evaluar clasificadores sobre bases de datos desbalanceadas, exhibiendo comportamientos similares a los de las métricas de evaluación aceptadas en contextos de desbalance, mientras mantiene una alta correlación con la clase positiva. Asimismo, esta nueva métrica se ha mostrado más sensible a cambios realizados sobre la matriz de confusión y, por tanto, más capaz para distinguir escenarios diferentes.

7.1.2. Soluciones para tratar el desbalance

En general, las soluciones en el problema del desbalance están dirigidas a remuestrear el conjunto de datos (soluciones a nivel de datos) o a la adaptación de los algoritmos de aprendizaje al problema del desbalance mediante el empleo de alguna matriz de costes (soluciones a nivel de algoritmos).

En el Capítulo 5, se han presentado diversas propuestas dirigidas a tratar el problema del desbalance. Así, dentro de las soluciones basadas en submuestreo, se ha formulado una técnica que reduce de forma inteligente la clase mayoritaria a través de la eliminación de muestras redundantes, atípicas o con etiqueta ruidosa, empleando para ello algoritmos conocidos de edición y condensado. En el caso de los algoritmos de sobremuestreo, se han incorporado tres modificaciones sobre el esquema original del SMOTE que permiten generar nuevas muestras considerando la vecindad y la distribución espacial. Por lo que respecta a nivel de datos, se ha propuesto una modificación a la distancia Euclídea en la regla 1-NN con el objetivo de lograr una tendencia del clasificador hacia la clase menos representada.

Estas tres propuestas han sido analizadas considerando tres factores: la naturaleza del clasificador, la métrica de evaluación de la efectividad y el nivel de desbalance. Los resultados obtenidos sugieren que, en general, las técnicas de remuestreo dependen fuertemente de la característica que presentan los datos, así como de la naturaleza del clasificador. Así, para clasificadores de naturaleza global, parece más idóneo sobremuestrear la clase mayoritaria, independientemente del nivel de desbalance.

Por otra parte, el submuestreo dirigido presenta la clara desventaja de depender de la naturaleza del conjunto de datos, por lo que es posible obtener situaciones adversas y opuestas: una moderada o excesiva tasa de reducción que puede conducir a cambiar el sentido del desbalance.

Por otra parte, la combinación de estrategias ha demostrado ser una técnica conveniente a la hora de tratar con problemas desbalanceados, principalmente para solventar las desventajas que presenta cada una de las soluciones aplicadas de manera individual.

7.1.3. Complejidad de los datos

En el Capítulo 6, se ha analizado la relación entre el desequilibrio que presentan las distribuciones por clases y otras complejidades (por ejemplo, el solapamiento), así como el grado de influencia sobre el comportamiento de los clasificadores.

En este proceso, se han llevado a cabo varios experimentos sobre bases de datos sintéticas con complejidades diversas en la zona de confusión, con el objetivo de evaluar la influencia que presenta el desbalance a nivel local y a nivel global sobre clasificadores de distinta naturaleza.

Los resultados experimentales han permitido concluir que:

- Los clasificadores con aprendizaje local suelen presentar un mejor comportamiento sobre la clase menos representada en la zona de solapamiento que los clasificadores globales.
- El comportamiento de cualquier clasificador podría depender más del desbalance local que se presenta en la zona de confusión que del desbalance a nivel global.

7.2. Direcciones para trabajos futuros

Las investigaciones y resultados descritos en esta memoria de tesis pueden extenderse en diferentes direcciones. Una de éstas puede consistir en llevar a cabo este mismo estudio sobre conjuntos de datos multiclase, lo que implicaría adaptar las soluciones, las métricas de efectividad y el estudio de la complejidad de los datos a estos entornos. Asimismo, es posible incluir en estos análisis el aprendizaje sensible al coste.

Por otra parte, todo el estudio desarrollado en esta tesis ha sido en un contexto de aprendizaje supervisado. Sin embargo, es posible que esta problemática se pueda extender a otros modelos, tales como el aprendizaje semi-supervisado y una derivación de éste conocido como aprendizaje on-line.

Por último, una extensión natural del trabajo realizado se refiere a la aplicación de estos algoritmos en algunos de los problemas reales en los que viene investigando nuestro grupo. En concreto, hemos previsto utilizar algunas de las propuestas que aquí se han presentado para el análisis y procesamiento de imágenes hiperespectrales

en aplicaciones de percepción remota. Otra línea en la que ya se ha empezado a trabajar se centra en la clasificación de pistas musicales en ficheros MIDI.

7.3. Publicaciones resultantes

Los diferentes resultados de la investigación realizada en esta tesis han dado lugar tanto a publicaciones en revistas internacionales del área del reconocimiento de formas y aprendizaje automático, como a comunicaciones en diversos congresos nacionales e internacionales.

7.3.1. Publicaciones en revistas internacionales

1. García, V.; Sánchez, J.S. y Mollineda, R.A.: Surrounding–SMOTE Algorithms for Learning from Imbalanced Data Sets. *WSEAS Transactions on Computers*, 2009, (en revisión).

2. García, V.; Mollineda, R.A. y Sánchez, J.S.: On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, 2008, **11**(**3**), pp. 269-280.

3. Barandela, R.; Sánchez, J.S.; García, V. y Rangel, E.: Strategies for learning in class imbalance problems. *Pattern Recognition*, 2003, **36**(**3**), pp. 849–851.

7.3.2. Publicaciones en congresos internacionales y nacionales

1. García, V.; Mollineda, R.A. y Sánchez, J.S.: Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. En: LNCS, 5524, pp. 441-448, 2009.

2. Cleofas, L.; Valdovinos, R.M.; García, V. y Alejo, R.: Use of Ensemble Based on GA for Imbalance Problem. En: LNCS, 5552, pp. 547-554, 2009.

3. Martín, R.; Mollineda, R.A. y García, V.: Melodic Track Identification in MIDI Files Considering the Imbalanced Context. En: LNCS, 5524, pp. 489-496, 2009.

4. García, V.; Mollineda, R.A. y Sánchez, J.S.: A new performance evaluation method for two-class imbalanced problems. En: LNCS, 5342, pp. 917-925, 2008.

5. García, V.; Sánchez, J.S. y Mollineda, R.A.: On the Use of Surrounding Neighbors for Synthetic Over-Sampling of the Minority Class. En: WSEAS SMO, pp. 389-394, 2008.

6. Alejo,R.; García, V.; García, F. y de la Rosa, M.: Redes Neuronales Artificiales y Distribuciones no Balanceadas. En: CICOS, pp. 77–88, 2008.

7. García, V.; Mollineda, R.A.; Sánchez, J.S., Alejo, R. y Sotoca, J.M.: When overlapping unexpectedly alters the class imbalance effects. En: LNCS, 4478, pp. 499-506, 2007.

8. Alejo, R.; García,V.; Sotoca, J.M.; Mollineda, R.A. y Sánchez, J.S.: Improving the performance of the RBF neural network trained with imbalanced samples. En: LNCS, 4507, pp. 162–169, 2007.

9. García, V.; Sánchez, J.S. y Mollineda, R.A.: An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. En: LNCS, 4756, pp. 397-406, 2007.

10. García, V.; Mollineda, R.A.; Sánchez, J.S.; Alejo, R. y Sotoca, J.M.: The class imbalance problem in pattern classification and learning. En: CEDI, pp. 283–291, 2007.

11. Alejo, R.; García,V.; Sotoca, J.M.; Mollineda, R.A. y Sánchez, J.S.: Improving the classification accuracy of RBF and MLP neural networks trained with imbalanced samples. En: LNCS, 4224, pp. 464–471, 2006.

12. García, V.; Alejo, R.; Sánchez, J.S.; Sotoca, J.M. y Mollineda, R.A.: Combined effects of class imbalance and class overlap on instance-based classification. En: LNCS, 4224, pp. 371–378, 2006.

13. Barandela, R.; Sánchez, J.S.; García, V. y Ferri, F.J.: Learning from imbalanced sets through resampling and weighting. En: LNCS, 2652, pp. 80–88, 2003.

14. Barandela, R.; Sánchez, J.S.; García, V. y Rangel, E.: Fusion of techniques for handling the imbalanced training sample problem. En: 6th Ibero-American Symposium on Pattern Recognition, pp. 34-40, 2001.

Anexo A

Clasificación sobre conjuntos originales

Tabla A.1: *Resultados de clasificación sobre el conjunto de datos Breast.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.454	0.761	0.671	0.440	0.447	0.447	0.588	0.608	**0.419**	**0.240**	**0.293**	**0.335**
J48	0.321	0.891	0.724	0.549	0.420	0.405	0.535	0.606	0.254	0.123	0.204	0.270
MLP	0.368	0.845	0.705	0.495	0.427	0.422	0.558	0.606	0.312	0.163	0.237	0.296
NBC	0.402	0.865	0.730	0.553	0.472	0.466	0.590	0.634	0.365	0.187	0.268	0.332
RBF	0.390	0.867	0.728	0.549	0.463	0.456	0.582	0.629	0.348	0.177	0.258	0.322
SVM	0.328	0.864	0.708	0.500	0.405	0.396	0.533	0.596	0.258	0.132	0.208	0.269
1−NN + DP	0.558	0.629	0.608	0.383	0.462	0.454	0.592	0.593	0.548	0.326	0.338	0.348

Tabla A.2: *Resultados de clasificación sobre el conjunto de datos Ecoli.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.514	0.942	0.897	0.506	0.510	0.510	0.696	0.728	0.604	0.277	0.381	0.464
J48	0.531	0.969	0.923	0.664	0.594	0.590	0.718	0.750	0.632	0.290	0.402	0.492
MLP	0.714	0.958	0.933	0.665	0.689	0.689	0.827	0.836	0.787	0.517	0.601	0.668
NBC	0.851	0.893	0.889	0.481	0.640	0.614	0.872	0.872	0.865	0.729	0.745	0.757
RBF	0.554	0.956	0.914	0.595	0.574	0.574	0.728	0.755	0.648	0.317	0.423	0.509
SVM	0.006	0.999	0.896	0.500	0.053	0.011	0.076	0.503	-0.093	0.000	0.003	0.005
1−NN + DP	0.846	0.883	0.879	0.457	0.622	0.593	0.864	0.864	0.858	0.719	0.733	0.744

Tabla A.3: *Resultados de clasificación sobre el conjunto de datos German.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.451	0.760	0.667	0.447	0.449	0.449	0.586	0.606	0.413	0.237	0.290	0.333
J48	0.472	0.835	0.726	0.551	0.510	0.508	0.628	0.653	0.448	0.251	0.323	0.380
MLP	0.504	0.810	0.718	0.533	0.518	0.518	0.639	0.657	0.485	0.283	0.346	0.396
NBC	0.505	0.859	0.753	0.605	0.553	0.551	0.659	0.682	0.494	0.281	0.357	0.419
RBF	0.442	0.841	0.721	0.544	0.491	0.488	0.610	0.642	0.410	0.223	0.298	0.357
SVM	0.493	0.803	0.710	0.519	0.506	0.506	0.630	0.648	0.471	0.273	0.335	0.384
1−NN + DP	0.477	0.725	0.650	0.427	0.451	0.450	0.588	0.601	0.444	0.260	0.303	0.337

Tabla A.4: *Resultados de clasificación sobre el conjunto de datos Glass.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.318	0.948	0.898	0.346	0.332	0.331	0.549	0.633	0.400	0.111	0.206	0.282
J48	0.353	0.963	0.915	0.455	0.401	0.397	0.583	0.658	0.451	0.132	0.236	0.319
MLP	0.141	0.969	0.903	0.279	0.198	0.188	0.370	0.555	0.157	0.024	0.080	0.125
NBC	0.753	0.444	0.468	0.105	0.281	0.184	0.578	0.598	0.210	0.437	0.386	0.344
RBF	0.000	0.999	0.920	0.000	0.000	0.000	0.000	0.499	-0.080	0.000	0.000	0.000
SVM	0.000	0.991	0.912	0.000	0.000	0.000	0.000	0.495	-0.088	0.000	0.000	0.000
1−NN + DP	0.482	0.827	0.800	0.194	0.306	0.277	0.632	0.655	0.537	0.261	0.330	0.385

Tabla A.5: *Resultados de clasificación sobre el conjunto de datos Haberman.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.328	0.771	0.654	0.340	0.503	0.334	0.503	0.550	0.251	**0.141**	0.197	0.242
J48	0.338	0.852	0.716	0.452	0.537	0.387	0.537	0.595	0.285	0.140	0.214	0.274
MLP	0.259	0.912	0.739	0.515	0.486	0.345	0.486	0.586	0.182	0.082	0.159	0.221
NBC	0.202	0.943	0.747	0.562	0.437	0.298	0.437	0.573	0.101	0.050	0.120	0.177
RBF	0.170	0.949	0.743	0.548	0.402	0.260	0.402	0.560	0.047	0.036	0.099	0.149
SVM	0.148	0.952	0.739	0.526	0.376	0.231	0.376	0.550	0.009	0.028	0.084	0.130
1−NN + DP	0.560	0.643	0.621	0.361	0.600	0.439	0.600	0.602	0.553	0.331	0.345	0.357

Tabla A.6: *Resultados de clasificación sobre el conjunto de datos* $Laryngeal_2$.

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.694	0.970	0.949	0.657	0.675	0.675	0.821	0.832	0.783	0.488	0.581	0.655
J48	0.638	0.980	0.954	0.728	0.682	0.680	0.791	0.809	0.742	0.411	0.518	0.604
MLP	0.770	0.979	0.963	0.756	0.763	0.763	0.868	0.875	0.844	0.596	0.675	0.738
NBC	0.943	0.844	0.851	0.334	0.561	0.493	0.892	0.894	0.796	0.875	0.836	0.804
RBF	0.558	0.985	0.952	0.751	0.648	0.641	0.742	0.772	0.676	0.316	0.433	0.526
SVM	0.672	0.942	0.922	0.492	0.575	0.568	0.796	0.807	0.754	0.462	0.547	0.616
1−NN + DP	0.804	0.958	0.946	0.612	0.701	0.695	0.877	0.881	0.859	0.651	0.711	0.758

Tabla A.7: *Resultados de clasificación sobre el conjunto de datos* $Letter_2$.

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.995	1.000	1.000	0.995	0.995	0.995	0.998	0.998	0.997	0.991	0.993	0.995
J48	0.943	0.999	0.996	0.966	0.955	0.955	0.971	0.971	0.968	0.890	0.916	0.937
MLP	0.886	1.000	0.995	0.998	0.940	0.938	0.941	0.943	0.935	0.784	0.835	0.875
NBC	0.858	0.992	0.986	0.811	0.834	0.833	0.922	0.925	0.914	0.737	0.794	0.839
RBF	0.830	0.996	0.990	0.899	0.864	0.863	0.909	0.913	0.898	0.689	0.758	0.813
SVM	0.850	0.997	0.992	0.931	0.890	0.889	0.921	0.924	0.912	0.722	0.785	0.835
1−NN+ DP	0.999	0.999	0.999	0.971	0.985	0.985	0.999	0.999	0.999	0.997	0.997	0.998

Tabla A.8: *Resultados de clasificación sobre el conjunto de datos Optdigits.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.966	0.999	0.996	0.993	0.979	0.979	0.982	0.982	0.979	0.933	0.949	0.962
J48	0.828	0.988	0.972	0.879	0.853	0.853	0.904	0.908	0.884	0.687	0.753	0.805
MLP	0.965	0.998	0.995	0.982	0.973	0.973	0.981	0.981	0.978	0.931	0.947	0.960
NBC	0.899	0.860	0.864	0.413	0.609	0.566	0.879	0.880	0.842	0.804	0.789	0.777
RBF	0.000	1.000	0.901	0.000	0.000	0.000	0.000	0.500	-0.099	0.000	0.000	0.000
SVM	0.785	0.991	0.971	0.909	0.845	0.842	0.882	0.888	0.855	0.617	0.698	0.762
1−NN + DP	0.977	0.999	0.996	0.987	0.982	0.982	0.988	0.988	0.986	0.955	0.965	0.974

Tabla A.9: *Resultados de clasificación sobre el conjunto de datos Pendigits.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.992	0.999	0.999	0.993	0.993	0.993	0.995	0.995	0.995	0.983	0.987	0.990
J48	0.947	0.997	0.992	0.971	0.959	0.959	0.972	0.972	0.967	0.898	0.921	0.940
MLP	0.593	1.000	0.961	0.997	0.769	0.744	0.770	0.796	0.705	0.352	0.472	0.569
NBC	0.583	0.972	0.935	0.690	0.634	0.632	0.753	0.777	0.684	0.346	0.456	0.544
RBF	0.725	0.993	0.967	0.917	0.815	0.810	0.849	0.859	0.811	0.527	0.624	0.701
SVM	0.711	0.991	0.964	0.892	0.797	0.791	0.839	0.851	0.800	0.507	0.606	0.685
1−NN + DP	0.996	0.999	0.998	0.986	0.991	0.991	0.997	0.997	0.997	0.991	0.993	0.994

Tabla A.10: *Resultados de clasificación sobre el conjunto de datos Phoneme.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.804	0.943	0.903	0.855	0.829	0.829	0.871	0.874	0.823	0.653	0.706	0.748
J48	0.787	0.898	0.865	0.761	0.774	0.774	0.841	0.842	0.800	0.629	0.668	0.699
MLP	0.718	0.849	0.811	0.664	0.691	0.690	0.781	0.784	0.727	0.530	0.570	0.602
NBC	0.730	0.773	0.760	0.572	0.646	0.641	0.751	0.751	0.731	0.539	0.552	0.561
RBF	0.564	0.866	0.777	0.636	0.599	0.598	0.699	0.715	0.566	0.341	0.414	0.473
SVM	0.573	0.857	0.773	0.624	0.598	0.000	0.701	0.715	0.575	0.351	0.421	0.477
1−NN + DP	0.876	0.916	0.904	0.812	0.843	0.843	0.896	0.896	0.882	0.771	0.786	0.799

Tabla A.11: *Resultados de clasificación sobre el conjunto de datos Pima.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.525	0.792	0.699	0.575	0.549	0.549	0.645	0.658	0.496	0.304	0.360	0.404
J48	0.586	0.821	0.739	0.637	0.611	0.610	0.694	0.704	0.572	0.368	0.424	0.470
MLP	0.547	0.854	0.747	0.667	0.604	0.601	0.683	0.700	0.528	0.324	0.395	0.453
NBC	0.600	0.838	0.755	0.666	0.632	0.631	0.709	0.719	0.589	0.383	0.443	0.491
RBF	0.522	0.850	0.735	0.651	0.583	0.579	0.666	0.686	0.497	0.299	0.371	0.429
SVM	0.531	0.875	0.755	0.694	0.607	0.000	0.681	0.703	0.510	0.304	0.384	0.448
1−NN + DP	0.613	0.681	0.657	0.507	0.558	0.555	0.646	0.647	0.605	0.389	0.404	0.415

Tabla A.12: *Resultados de clasificación sobre el conjunto de datos Satimage.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.705	0.960	0.935	0.656	0.680	0.679	0.823	0.833	0.782	0.504	0.591	0.660
J48	0.550	0.959	0.919	0.589	0.569	0.569	0.726	0.755	0.648	0.312	0.420	0.506
MLP	0.637	0.966	0.934	0.669	0.653	0.653	0.785	0.802	0.729	0.413	0.515	0.596
NBC	0.870	0.815	0.820	0.337	0.541	0.485	0.842	0.843	0.788	0.749	0.729	0.713
SVM	0.000	1.000	0.903	0.000	0.000	0.000	0.000	0.500	-0.097	0.000	0.000	0.000
RBF	0.000	1.000	0.903	0.000	0.000	0.000	0.000	0.500	-0.097	0.000	0.000	0.000
1−NN + DP	0.837	0.933	0.924	0.574	0.693	0.681	0.884	0.885	0.869	0.706	0.743	0.773

Tabla A.13: *Resultados de clasificación sobre el conjunto de datos Scrapie.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.805	0.274	0.365	0.186	0.387	0.302	0.470	0.540	-0.127	0.338	0.280	0.233
J48	0.116	0.974	0.828	0.482	0.236	0.187	0.336	0.545	0.040	0.016	0.064	0.103
MLP	0.168	0.956	0.822	0.441	0.272	0.243	0.400	0.562	0.120	0.034	0.097	0.148
NBC	0.030	0.992	0.828	0.437	0.115	0.056	0.173	0.511	-0.113	0.001	0.016	0.027
RBF	0.009	0.996	0.828	0.311	0.052	0.017	0.093	0.502	-0.155	0.000	0.004	0.008
SVM	0.000	1.000	0.829	0.000	0.000	0.000	0.000	0.500	-0.171	0.000	0.000	0.000
1−NN + DP	0.805	0.274	0.365	0.186	0.387	0.302	0.470	0.540	-0.127	0.338	0.280	0.233

Tabla A.14: *Resultados de clasificación sobre el conjunto de datos Segmentation.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.903	0.987	0.975	0.922	0.912	0.912	0.944	0.945	0.931	0.816	0.854	0.884
J48	0.869	0.977	0.962	0.865	0.867	0.867	0.922	0.923	0.903	0.757	0.803	0.840
MLP	0.914	0.984	0.974	0.905	0.909	0.909	0.948	0.949	0.937	0.836	0.868	0.893
NBC	0.922	0.686	0.720	0.329	0.551	0.485	0.796	0.804	0.573	0.783	0.708	0.648
RBF	0.276	0.972	0.873	0.623	0.415	0.382	0.518	0.624	0.315	0.081	0.175	0.249
SVM	0.896	0.981	0.969	0.889	0.893	0.893	0.938	0.939	0.924	0.805	0.842	0.872
1−NN + DP	0.945	0.976	0.972	0.867	0.905	0.905	0.961	0.961	0.956	0.895	0.909	0.920

Tabla A.15: *Resultados de clasificación sobre el conjunto de datos Spambase.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.881	0.925	0.908	0.885	0.883	0.883	0.903	0.903	0.883	0.779	0.797	0.811
J48	0.904	0.942	0.927	0.910	0.907	0.907	0.923	0.923	0.906	0.819	0.835	0.848
MLP	0.905	0.930	0.920	0.894	0.899	0.899	0.917	0.917	0.906	0.820	0.831	0.839
NBC	0.953	0.692	0.795	0.668	0.798	0.786	0.812	0.823	0.636	0.832	0.746	0.677
RBF	0.849	0.782	0.808	0.717	0.780	0.777	0.815	0.815	0.767	0.709	0.686	0.668
SVM	0.832	0.952	0.905	0.918	0.874	0.873	0.890	0.892	0.838	0.698	0.745	0.783
1−NN + DP	0.781	0.850	0.823	0.772	0.777	0.777	0.815	0.816	0.781	0.619	0.641	0.659

Tabla A.16: *Resultados de clasificación sobre el conjunto de datos Vehicle.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.458	0.844	0.747	0.495	0.477	0.476	0.622	0.651	0.451	0.238	0.312	0.372
J48	0.509	0.862	0.774	0.553	0.531	0.530	0.663	0.686	0.517	0.284	0.362	0.424
MLP	0.617	0.909	0.835	0.693	0.654	0.653	0.749	0.763	0.644	0.397	0.479	0.544
NBC	0.614	0.718	0.692	0.421	0.509	0.500	0.664	0.666	0.614	0.395	0.418	0.436
RBF	0.397	0.870	0.752	0.506	0.448	0.445	0.588	0.634	0.378	0.182	0.264	0.329
SVM	0.645	0.864	0.809	0.613	0.629	0.629	0.747	0.755	0.664	0.436	0.497	0.545
1−NN + DP	0.458	0.790	0.706	0.421	0.439	0.439	0.601	0.624	0.440	0.241	0.301	0.349

Tabla A.17: *Resultados de clasificación sobre el conjunto de datos Yeast.*

	tvp	*tvn*	*Acc*	*precision*	*Gpr*	f_1	*Gm*	*AUC*	*OP*	*IBA*	$IBA_{0.5}$	$IBA_{0.1}$
1−NN	0.478	0.818	0.720	0.516	0.497	0.496	0.625	0.648	0.457	0.258	0.324	0.378
J48	0.462	0.874	0.755	0.598	0.525	0.521	0.635	0.668	0.446	0.237	0.320	0.387
MLP	0.515	0.866	0.765	0.610	0.561	0.559	0.668	0.691	0.511	0.290	0.368	0.431
NBC	0.486	0.865	0.756	0.595	0.538	0.535	0.648	0.676	0.475	0.261	0.341	0.404
RBF	0.312	0.916	0.741	0.602	0.433	0.411	0.535	0.614	0.249	0.113	0.199	0.268
SVM	0.175	0.971	0.741	0.708	0.352	0.281	0.412	0.573	0.047	0.035	0.102	0.157
1−NN + DP	0.606	0.733	0.696	0.480	0.539	0.535	0.666	0.669	0.601	0.387	0.416	0.438

Anexo B

Interpretación de los links de Tomek

La definición de los links que aparece en el artículo original de Tomek (1976b) puede describirse de la siguiente manera. Para un conjunto de puntos D, un par $(\mathbf{x}, \mathbf{y})$ será un link de Tomek si no existe ningún elemento de D que caiga dentro de una hiperesfera con centro en el punto medio entre $\mathbf{x}$ e $\mathbf{y}$ y diámetro igual a la distancia entre ellos. Esta descripción puede verse con más claridad en la Figura B.1, donde la región de color azul representa la circunferencia vacía definida por los puntos $\mathbf{x}$ e $\mathbf{y}$ y cuyo centro corresponde al punto Z_1.

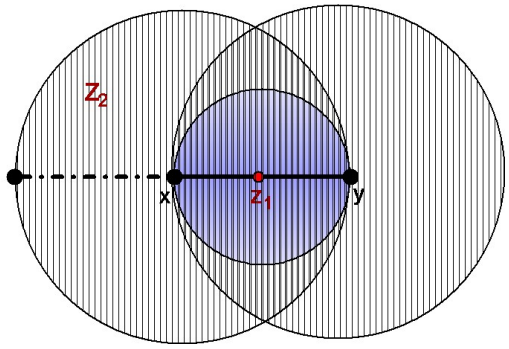


Figura B.1: *Ejemplo para comparar diferentes versiones de los links de Tomek.*

Cabe mencionar que la descripción de los links de Tomek coincide con la idea de los grafos de Gabriel (Gabriel y Sokal, 1969). La única diferencia entre estos dos conceptos se refiere a que los links de Tomek se definen para pares de puntos de clase diferente (Toussaint, 1994).

Al analizar la definición que ofrecen Kubat y Matwin (1997) sobre lo que ellos denominan links de Tomek, podemos observar que dicha definición presenta una importante variación que modifica en realidad el propio concepto. Para ellos, un par de puntos $(\mathbf{x}, \mathbf{y})$ forman un link de Tomek si no existe ningún ejemplo $\mathbf{z}$ tal que $d(\mathbf{x}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$ o $d(\mathbf{y}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$. De acuerdo a esta definición, la distancia entre $\mathbf{x}$ e $\mathbf{y}$ determina el radio de la hiperesfera, dando lugar a una región mucho más grande que la producida por los verdaderos links de Tomek. El resultado de esta importante diferencia está en que habrá muchos menos pares de puntos que puedan cumplir la condición impuesta para formar los links, es decir, el número de links producidos mediante el procedimiento de Kubat y Matwin (1997) será (muy)

inferior al número de los links de Tomek.

La diferencia que acabamos de comentar puede observarse en la Figura B.1, donde se han representado las circunferencias resultantes de aplicar las dos versiones de links de Tomek. Así, dado un punto Z_2, éste quedaría fuera de la circunferencia que se forma de acuerdo a la versión original del algoritmo de Tomek (región azul) y, por tanto, el par $(\mathbf{x}, \mathbf{y})$ constituiría un link. Sin embargo, de acuerdo a la versión modificada, este par no representaría un link porque el punto Z_2 queda dentro de la región que forman las dos circunferencias.

Anexo C

Tablas de pruebas $t-$student

Tabla C.1: *Prueba $t-$student sobre resultados de clasificación por la regla del vecino más cercano con distancia Euclídea y distancia ponderada según las métricas tvp, tvn, Acc, f_1, $IBA_{0.1}$ y Gm*

		tvp			tvn			Acc	
	(2)	(3)	(4)	(2)	(3)	(4)	(2)	(3)	(4)
(1) 1-NN+DP	13/3/1	16/0/1	16/0/1	0/1/16	0/1/16	1/3/13	1/4/12	3/2/12	5/0/12
(2) 1-NN		12/4/1	13/3/1		3/3/11	3/2/12		4/2/11	6/0/11
(3) 7-NN			10/6/1			2/6/9			5/9/3
(4) 13-NN									

		f_1			$IBA_{0.1}$			Gm	
	(2)	(3)	(4)	(2)	(3)	(4)	(2)	(3)	(4)
(1) 1-NN+DP	5/8/4	7/8/2	11/4/2	10/6/1	13/2/2	15/1/1	11/5/1	13/2/2	15/0/2
(2) 1-NN		8/6/3	11/3/3		10/4/3	13/2/2		10/5/2	13/2/2
(3) 7-NN			8/8/1			9/7/1			9/7/1
(4) 13-NN									

Tabla C.2: *Prueba $t-$student del clasificador $1-NN$ en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/3/14	0/2/15	0/1/16	0/0/17	0/1/16	0/0/17	0/4/13	0/0/17	0/0/17	0/0/17	0/0/17	0/1/16	0/1/16	0/1/16	0/0/17	0/1/16	0/1/16	2/10/5
(2) SSM$^-$		4/6/7	0/2/15	1/1/15	8/6/3	0/0/17	12/4/1	0/0/17	0/2/15	0/0/17	0/5/12	3/4/10	0/1/16	5/6/6	3/4/10	4/5/8	6/6/5	12/4/1
(3) EW$^-$			0/1/16	1/4/12	8/4/5	1/0/16	14/3/0	1/0/16	2/3/12	1/0/16	4/4/9	0/7/10	0/1/16	5/5/7	5/4/8	5/4/8	5/7/5	11/6/0
(4) EW$^-$+SSM$^-$				7/4/6	13/2/2	1/4/12	16/1/0	1/0/16	7/8/2	1/0/16	11/4/2	11/5/1	0/7/10	10/3/4	9/4/4	9/4/4	12/2/3	15/2/0
(5) RUS					15/1/1	2/5/10	17/0/0	1/1/15	8/5/4	1/4/12	11/5/1	10/4/3	4/5/8	13/4/0	12/5/0	12/5/0	14/3/0	16/1/0
(6) CNN$^-$						0/2/15	9/6/2	0/0/17	0/2/15	0/0/17	0/4/13	5/3/9	1/2/14	3/7/7	1/6/10	1/9/7	4/7/6	14/1/2
(7) TMK$_1$							17/0/0	0/2/15	13/3/1	0/3/14	15/2/0	14/2/1	11/4/2	15/2/0	14/3/0	13/4/0	15/2/0	17/0/0
(8) TMK$_2$								0/0/17	0/0/17	0/0/17	0/0/17	0/1/16	0/1/16	2/0/15	2/0/15	2/0/15	2/2/13	10/5/2
(9) CNN$^-$+TMK$_1$									17/0/0	15/2/0	17/0/0	16/0/1	16/0/1	17/0/0	16/1/0	16/1/0	17/0/0	17/0/0
(10) CNN$^-$+TMK$_2$										1/0/16	13/4/0	7/6/4	1/5/11	10/5/2	9/5/3	9/5/3	11/5/1	16/1/0
(11) TMK$_1$+CNN$^-$											17/0/0	15/1/1	15/1/1	17/0/0	16/1/0	16/1/0	17/0/0	17/0/0
(12) TMK$_2$+CNN$^-$												5/4/8	1/4/12	9/5/3	6/7/4	7/6/4	10/4/3	14/2/1
(13) EWW$^-$													0/2/15	8/4/5	8/4/5	8/3/6	8/5/4	15/2/0
(14) EWW$^-$+SSM$^-$														11/3/3	10/4/3	10/4/3	13/2/2	16/1/0
(15) SMOTE															0/11/6	0/13/4	5/12/0	16/1/0
(16) gg-SMOTE																3/14/0	10/7/0	16/1/0
(17) ncn-SMOTE																	8/9/0	16/1/0
(18) rng-SMOTE																		15/2/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	17/0/0	15/2/0	17/0/0	17/0/0	17/0/0	17/0/0	14/3/0	17/0/0	17/0/0	17/0/0	17/0/0	16/1/0	17/0/0	16/1/0	17/0/0	17/0/0	15/2/0	6/9/2
(2) SSM$^-$		3/2/12	13/4/0	14/1/2	5/6/6	14/2/1	2/1/14	17/0/0	16/1/0	17/0/0	14/2/1	6/3/8	15/2/0	2/4/11	4/4/9	2/5/10	1/2/14	0/2/15
(3) EW$^-$			17/0/0	15/2/0	11/3/3	16/0/1	1/3/13	17/0/0	16/0/1	16/0/1	14/1/2	12/5/0	17/0/0	8/3/6	9/4/4	8/5/4	8/3/6	1/4/12
(4) EW$^-$+SSM$^-$				8/3/6	3/2/12	12/2/3	0/1/16	17/0/0	8/5/4	16/0/1	4/2/11	1/0/16	13/4/0	2/0/15	2/0/15	2/0/15	1/1/15	0/1/16
(5) RUS					1/1/15	9/1/7	0/0/17	16/0/1	8/2/7	14/1/2	2/3/12	1/2/14	8/1/8	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(6) CNN$^-$						14/2/1	2/1/14	17/0/0	17/0/0	17/0/0	14/3/0	6/2/9	14/0/3	2/4/11	3/6/8	2/5/10	1/3/13	0/2/15
(7) TMK$_1$							0/0/17	17/0/0	5/1/11	17/0/0	1/3/13	1/0/16	5/0/12	0/1/16	0/1/16	0/1/16	0/1/16	0/0/17
(8) TMK$_2$								17/0/0	17/0/0	17/0/0	17/0/0	16/1/0	17/0/0	14/1/2	15/0/2	15/0/2	11/4/2	2/4/11
(9) CNN$^-$+TMK$_1$									0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(10) CNN$^-$+TMK$_2$										16/0/1	0/0/17	1/1/15	11/3/3	0/2/15	1/1/15	0/2/15	0/0/17	0/0/17
(11) TMK$_1$+CNN$^-$											0/0/17	1/0/16	1/0/16	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(12) TMK$_2$+CNN$^-$												4/1/12	14/0/3	1/1/15	1/1/15	1/2/14	1/1/15	0/0/17
(13) EWW$^-$													17/0/0	4/6/7	5/4/8	6/4/7	4/4/9	0/2/15
(14) EWW$^-$+SSM$^-$														0/1/16	1/0/16	0/0/17	0/0/17	0/0/17
(15) SMOTE															6/11/0	7/9/1	1/12/4	0/1/16
(16) gg-SMOTE																2/10/5	0/5/12	0/1/16
(17) ncn-SMOTE																	1/5/11	0/2/15
(18) rng-SMOTE																		1/2/14
(19) b-SMOTE																		

Tabla C.3: *Prueba $t-$student del clasificador $1-NN$ en términos de Acc y f_1*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/1/0	12/2/3	16/0/1	17/0/0	17/0/0	17/0/0	9/6/2	17/0/0	17/0/0	17/0/0	17/0/0	11/4/2	16/1/0	12/4/1	11/6/0	10/7/0	10/6/1	2/13/2
(2) SSM$^-$		3/1/13	10/5/2	12/3/2	6/10/1	13/2/2	1/3/13	17/0/0	15/1/1	17/0/0	11/5/1	4/3/10	11/6/0	1/5/11	2/4/11	1/4/12	1/2/14	0/3/14
(3) EW$^-$			16/1/0	16/1/0	13/1/3	16/0/1	3/5/9	17/0/0	15/1/1	16/0/1	14/1/2	9/8/0	17/0/0	9/4/4	9/3/5	7/6/4	8/5/4	1/7/9
(4) EW$^-$+SSM$^-$				9/3/5	4/6/7	12/1/4	1/0/16	17/0/0	9/7/1	16/0/1	4/7/6	1/1/15	12/5/0	2/0/15	2/0/15	2/0/15	2/0/15	1/0/16
(5) RUS					2/4/11	9/0/8	0/0/17	16/0/1	6/4/7	14/1/2	2/5/10	1/1/15	7/2/8	0/2/15	0/2/15	0/1/16	0/0/17	0/0/17
(6) CNN$^-$						13/2/2	2/1/14	17/0/0	16/1/0	17/0/0	10/7/0	3/2/12	10/4/3	1/4/12	2/3/12	1/4/12	1/3/13	0/2/15
(7) TMK$_1$							0/0/17	17/0/0	4/3/10	16/1/0	3/2/12	1/0/16	5/0/12	0/1/16	0/1/16	0/2/15	0/1/16	0/0/17
(8) TMK$_2$								17/0/0	17/0/0	17/0/0	15/2/0	10/6/1	16/1/0	9/4/4	9/5/3	10/4/3	9/4/4	3/7/7
(9) CNN$^-$+TMK$_1$									0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(10) CNN$^-$+TMK$_2$										15/1/1	0/1/16	1/2/14	4/9/4	0/0/17	1/0/16	0/0/17	0/0/17	0/0/17
(11) TMK$_1$+CNN$^-$											0/0/17	1/0/16	1/2/14	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(12) TMK$_2$+CNN$^-$												2/2/13	10/4/3	1/1/15	1/1/15	1/1/15	0/1/16	0/0/17
(13) EWW$^-$													16/1/0	5/6/6	6/6/5	4/9/4	5/7/5	1/7/9
(14) EWW$^-$+SSM$^-$														0/2/15	1/2/14	0/1/16	0/1/16	0/0/17
(15) SMOTE															3/13/1	2/14/1	1/15/1	1/5/11
(16) gg-SMOTE																1/12/4	0/13/4	0/8/9
(17) ncn-SMOTE																	0/14/3	0/7/10
(18) rng-SMOTE																		1/7/9
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	6/5/6	5/4/8	7/2/8	8/4/5	8/5/4	7/2/8	2/6/9	11/5/1	7/4/6	8/2/7	7/3/7	4/3/10	7/2/8	2/6/9	2/7/8	2/7/8	3/6/8	1/12/4
(2) SSM$^-$		1/4/12	4/4/9	6/7/4	7/9/1	6/2/9	2/7/8	13/3/1	6/9/2	9/5/3	5/10/2	1/2/14	5/5/7	2/5/10	1/4/12	2/4/11	1/7/9	4/5/8
(3) EW$^-$			9/5/3	12/3/2	13/3/1	9/7/1	8/5/4	13/3/1	12/4/1	12/4/1	12/5/0	4/9/4	9/4/4	6/8/3	5/8/4	5/9/3	5/10/2	5/8/4
(4) EW$^-$+SSM$^-$				9/7/1	10/5/2	10/4/3	7/2/8	16/0/1	9/8/0	13/3/1	9/5/3	3/5/9	6/10/1	7/2/8	7/3/7	6/3/8	6/4/7	7/3/7
(5) RUS					4/9/4	5/5/7	4/5/8	15/1/1	4/9/4	9/6/2	2/10/5	1/3/13	3/4/10	2/6/9	2/5/10	2/5/10	2/6/9	6/3/8
(6) CNN$^-$						6/3/8	0/8/9	13/3/1	6/6/5	8/4/5	3/10/4	1/2/14	4/3/10	0/6/11	1/3/13	0/4/13	0/4/13	2/5/10
(7) TMK$_1$							5/5/7	16/1/0	7/4/6	13/4/0	6/5/6	1/5/11	2/4/11	6/3/8	2/8/7	4/5/8	4/6/7	6/4/7
(8) TMK$_2$								13/3/1	9/6/2	8/6/3	8/6/3	4/4/9	8/2/7	4/8/5	3/10/4	4/9/4	5/7/5	7/5/5
(9) CNN$^-$+TMK$_1$									1/2/14	0/1/16	1/0/16	0/3/14	0/1/16	1/3/13	1/3/13	1/3/13	1/3/13	1/3/13
(10) CNN$^-$+TMK$_2$										9/5/3	0/9/8	1/1/15	3/5/9	1/6/10	1/7/9	1/6/10	1/7/9	5/4/8
(11) TMK$_1$+CNN$^-$											2/6/9	0/4/13	2/2/13	3/4/10	2/6/9	3/5/9	3/6/8	6/3/8
(12) TMK$_2$+CNN$^-$												1/2/14	4/4/9	2/6/9	2/7/8	3/5/9	3/5/9	6/3/8
(13) EWW$^-$													10/5/2	7/6/4	7/7/3	7/8/2	6/8/3	8/5/4
(14) EWW$^-$+SSM$^-$														8/0/9	7/2/8	7/1/9	7/1/9	7/3/7
(15) SMOTE															2/13/2	1/15/1	1/16/0	4/11/2
(16) gg-SMOTE																0/15/2	1/14/2	5/10/2
(17) ncn-SMOTE																	0/17/0	4/11/2
(18) rng-SMOTE																		5/9/3
(19) b-SMOTE																		

Tabla C.4: *Prueba $t-$student del clasificador $1-NN$ en términos de $IBA_{0.1}$ y Gm*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/4/11	3/3/11	2/5/10	3/4/10	2/7/8	7/3/7	3/3/11	13/3/1	2/5/10	11/2/4	2/4/11	2/2/13	2/4/11	0/1/16	1/0/16	0/1/16	0/2/15	1/10/6
(2) SSM$^-$		3/6/8	2/5/10	3/4/10	3/12/2	9/2/6	7/8/2	15/2/0	4/7/6	12/0/5	1/11/5	2/5/10	3/5/9	0/7/10	1/3/13	0/7/10	1/8/8	8/6/3
(3) EW$^-$			4/6/7	6/5/6	7/8/2	9/2/6	10/6/1	15/2/0	6/6/5	12/1/4	5/9/3	2/8/7	4/8/5	2/6/9	2/4/11	2/6/9	2/7/8	10/5/2
(4) EW$^-$+SSM$^-$				6/8/3	8/6/3	10/2/5	9/5/3	16/1/0	7/8/2	11/3/3	6/8/3	2/9/6	3/10/4	3/6/8	3/6/8	3/7/7	4/7/6	9/6/2
(5) RUS					10/4/3	9/3/5	8/6/3	16/0/1	4/11/2	11/3/3	6/8/3	1/10/6	3/9/5	1/8/8	2/7/8	2/7/8	3/7/7	10/4/3
(6) CNN$^-$						8/4/5	4/9/4	14/3/0	5/4/8	12/0/5	3/8/6	3/5/9	4/5/8	0/6/11	1/2/14	0/3/14	0/6/11	5/7/5
(7) TMK$_1$							7/1/9	17/0/0	4/4/9	12/4/1	4/3/10	3/4/10	3/4/10	3/4/10	1/6/10	2/5/10	3/5/9	6/4/7
(8) TMK$_2$								14/3/0	3/7/7	12/1/4	2/8/7	2/4/11	3/4/10	1/4/12	0/4/13	0/4/13	1/4/12	7/6/4
(9) CNN$^-$+TMK$_1$									0/0/17	0/0/17	0/1/16	0/1/16	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	1/1/15
(10) CNN$^-$+TMK$_2$										11/1/5	3/8/6	2/7/8	2/7/8	0/8/9	0/8/9	0/7/10	0/8/9	6/9/2
(11) TMK$_1$+CNN$^-$											4/1/12	2/3/12	2/3/12	1/3/13	1/3/13	1/3/13	2/2/13	4/3/10
(12) TMK$_2$+CNN$^-$												3/6/8	3/8/6	1/6/10	0/8/9	0/7/10	2/7/8	9/5/3
(13) EWW$^-$													6/8/3	5/5/7	4/4/9	5/4/8	5/6/6	10/5/2
(14) EWW$^-$+SSM$^-$														3/6/8	4/6/7	4/6/7	4/7/6	9/6/2
(15) SMOTE															1/14/2	0/14/3	5/12/0	13/3/1
(16) gg-SMOTE																0/15/2	7/9/1	14/2/1
(17) ncn-SMOTE																	7/10/0	13/3/1
(18) rng-SMOTE																		12/4/1
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/4/10	3/2/12	2/4/11	3/4/10	2/7/8	8/3/6	3/4/10	14/2/1	4/4/9	12/1/4	2/3/12	2/3/12	2/6/9	0/1/16	1/0/16	0/2/15	0/2/15	1/11/5
(2) SSM$^-$		3/5/9	4/4/9	4/4/9	2/13/2	10/2/5	4/10/3	15/2/0	4/7/6	12/0/5	2/11/4	2/5/10	4/4/9	0/7/10	1/3/13	0/5/12	0/9/8	6/5/6
(3) EW$^-$			4/8/5	8/3/6	7/9/1	10/1/6	10/5/2	15/2/0	7/6/4	12/1/4	5/10/2	3/7/7	5/7/5	2/6/9	2/4/11	2/5/10	2/7/8	6/9/2
(4) EW$^-$+SSM$^-$				5/8/4	8/7/2	10/2/5	10/4/3	15/2/0	7/8/2	11/3/3	5/9/3	1/10/6	4/9/4	2/6/9	1/7/9	3/6/8	4/6/7	8/7/2
(5) RUS					10/3/4	10/2/5	7/6/4	16/0/1	5/10/2	12/3/2	6/7/4	1/8/8	4/7/6	1/8/8	2/7/8	2/7/8	4/5/8	8/6/3
(6) CNN$^-$						10/3/4	3/10/4	15/2/0	5/5/7	12/1/4	3/8/6	2/6/9	3/7/7	0/5/12	1/3/13	0/3/14	0/4/13	4/8/5
(7) TMK$_1$							5/3/9	16/1/0	4/3/10	13/3/1	4/3/10	3/4/10	3/4/10	2/4/11	1/5/11	2/4/11	2/5/10	5/4/8
(8) TMK$_2$								14/3/0	4/7/6	13/0/4	2/9/6	3/3/11	4/5/8	0/6/11	0/6/11	0/4/13	1/5/11	7/6/4
(9) CNN$^-$+TMK$_1$									0/1/16	0/0/17	0/2/15	0/2/15	0/1/16	0/0/17	0/1/16	0/1/16	0/1/16	0/2/15
(10) CNN$^-$+TMK$_2$										11/2/4	3/8/6	1/7/9	2/8/7	0/7/10	0/5/12	0/6/11	0/7/10	5/8/4
(11) TMK$_1$+CNN$^-$											4/1/12	3/2/12	2/3/12	2/2/13	2/2/13	2/2/13	2/2/13	3/3/11
(12) TMK$_2$+CNN$^-$												2/7/8	4/8/5	0/6/11	0/7/10	0/6/11	0/7/10	7/7/3
(13) EWW$^-$													6/9/2	5/4/8	4/4/9	5/4/8	4/7/6	9/6/2
(14) EWW$^-$+SSM$^-$														3/7/7	3/6/8	3/6/8	3/8/6	8/7/2
(15) SMOTE															1/14/2	0/15/2	5/12/0	13/3/1
(16) gg-SMOTE																0/15/2	7/9/1	14/2/1
(17) ncn-SMOTE																	5/12/0	13/3/1
(18) rng-SMOTE																		12/4/1
(19) b-SMOTE																		

Tabla C.5: *Prueba t−student del clasificador 7−NN en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/16	0/3/14	0/1/16	0/0/17	0/0/17	0/0/17	0/4/13	0/0/17	0/0/17	0/0/17	0/0/17	0/2/15	0/1/16	0/0/17	0/0/17	0/0/17	0/0/17	0/6/11
(2) SSM$^-$		12/2/3	0/4/13	3/0/14	4/7/6	2/2/13	15/2/0	1/0/16	0/1/16	1/0/16	0/1/16	7/4/6	0/3/14	3/2/12	3/1/13	3/2/12	3/1/13	13/2/2
(3) EW$^-$			0/1/16	0/3/14	4/2/11	1/0/16	12/4/1	1/0/16	0/0/17	1/0/16	0/1/16	0/5/12	0/1/16	2/1/14	2/1/14	2/1/14	2/1/14	11/4/2
(4) EW$^-$+SSM$^-$				9/1/7	14/0/3	4/3/10	16/1/0	1/0/16	4/6/7	1/0/16	10/4/3	13/3/1	0/7/10	9/2/6	8/3/6	9/2/6	9/2/6	14/3/0
(5) RUS					13/1/3	7/1/9	16/0/1	1/1/15	6/1/10	1/4/12	7/5/5	14/2/1	7/1/9	6/9/2	5/10/2	11/4/2	7/9/1	16/1/0
(6) CNN$^-$						2/3/12	14/2/1	0/0/17	0/2/15	0/0/17	0/3/14	9/1/7	3/0/14	4/1/12	3/1/13	3/3/11	3/2/12	13/2/2
(7) TMK$_1$							16/0/1	0/2/15	8/3/6	0/2/15	10/3/4	16/0/1	8/4/5	10/0/7	10/0/7	10/0/7	10/0/7	15/2/0
(8) TMK$_2$								1/0/16	0/0/17	1/0/16	0/0/17	1/2/14	0/1/16	1/0/16	1/0/16	1/0/16	1/0/16	9/5/3
(9) CNN$^-$+TMK$_1$									15/1/1	11/6/0	16/0/1	16/0/1	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	17/0/0
(10) CNN$^-$+TMK$_2$										1/1/15	14/3/0	16/0/1	4/3/10	11/2/4	11/2/4	11/2/4	11/3/3	16/1/0
(11) TMK$_1$+CNN$^-$											16/0/1	16/0/1	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	17/0/0
(12) TMK$_2$+CNN$^-$												13/3/1	3/0/14	8/3/6	7/3/7	8/3/6	7/4/6	14/3/0
(13) EWW$^-$													0/1/16	3/1/13	3/1/13	2/2/13	3/1/13	13/3/1
(14) EWW$^-$+SSM$^-$														11/2/4	9/4/4	10/2/5	10/3/4	16/1/0
(15) SMOTE															1/15/1	3/14/0	1/16/0	17/0/0
(16) gg-SMOTE																6/10/1	2/14/1	17/0/0
(17) ncn-SMOTE																	1/12/4	17/0/0
(18) rng-SMOTE																		16/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	17/0/0	15/2/0	17/0/0	17/0/0	17/0/0	17/0/0	13/4/0	17/0/0	17/0/0	17/0/0	17/0/0	16/1/0	17/0/0	17/0/0	17/0/0	17/0/0	17/0/0	9/8/0
(2) SSM$^-$		1/2/14	14/3/0	12/2/3	7/4/6	12/2/3	0/0/17	16/0/1	16/1/0	16/0/1	15/2/0	4/2/11	16/1/0	8/2/7	7/2/8	7/2/8	8/1/8	1/2/14
(3) EW$^-$			17/0/0	16/0/1	14/1/2	16/0/1	1/3/13	16/0/1	17/0/0	16/0/1	17/0/0	14/3/0	17/0/0	15/1/1	13/2/2	14/2/1	14/2/1	3/1/13
(4) EW$^-$+SSM$^-$				7/1/9	3/2/12	9/3/5	0/0/17	16/0/1	11/3/3	16/0/1	4/3/10	1/0/16	14/3/0	4/0/13	4/0/13	4/0/13	4/0/13	0/1/16
(5) RUS					5/2/10	9/0/8	1/0/16	16/0/1	13/0/4	15/1/1	10/3/4	1/1/15	9/1/7	0/2/15	0/2/15	0/2/15	0/1/16	0/0/17
(6) CNN$^-$						13/1/3	2/0/15	17/0/0	17/0/0	17/0/0	15/2/0	4/3/10	13/1/3	9/2/6	10/1/6	9/2/6	10/1/6	1/4/12
(7) TMK$_1$							1/0/16	17/0/0	7/2/8	17/0/0	5/1/11	1/0/16	9/1/7	4/1/12	5/0/12	5/1/11	5/1/11	0/2/15
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	17/0/0	15/1/1	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	4/3/10
(9) CNN$^-$+TMK$_1$									1/0/16	0/1/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										16/0/1	0/0/17	0/1/16	9/3/5	3/1/13	4/0/13	4/0/13	3/1/13	0/0/17
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(12) TMK$_2$+CNN$^-$												0/2/15	13/1/3	4/1/12	4/1/12	4/0/13	4/1/12	0/1/16
(13) EWW$^-$													17/0/0	13/3/1	13/2/2	13/3/1	13/2/2	1/2/14
(14) EWW$^-$+SSM$^-$														2/1/14	2/1/14	2/1/14	2/1/14	0/0/17
(15) SMOTE															3/7/7	3/9/5	3/11/3	0/0/17
(16) gg-SMOTE																4/7/6	5/9/3	0/0/17
(17) ncn-SMOTE																	3/13/1	0/0/17
(18) rng-SMOTE																		0/0/17
(19) b-SMOTE																		

Tabla C.6: *Prueba $t-$student del clasificador $7-NN$ en términos de Acc y f_1*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	15/2/0	8/8/1	16/1/0	17/0/0	15/2/0	17/0/0	4/12/1	17/0/0	17/0/0	17/0/0	16/1/0	11/6/0	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	4/12/1
(2) SSM$^-$		1/2/14	13/3/1	10/2/5	7/5/5	11/3/3	0/1/16	16/0/1	16/1/0	16/0/1	12/5/0	3/2/12	16/1/0	5/3/9	5/3/9	5/2/10	5/3/9	1/3/13
(3) EW$^-$			16/1/0	16/0/1	14/1/2	16/0/1	2/8/7	16/0/1	17/0/0	16/0/1	16/1/0	8/9/0	17/0/0	13/2/2	12/1/4	13/1/3	13/2/2	3/6/8
(4) EW$^-$+SSM$^-$				7/2/8	4/1/12	9/3/5	0/0/17	16/0/1	10/4/3	16/0/1	5/4/8	0/2/15	14/3/0	4/0/13	4/0/13	4/0/13	4/0/13	0/1/16
(5) RUS					5/3/9	9/0/8	1/0/16	16/0/1	13/0/4	15/1/1	9/4/4	1/0/16	9/1/7	0/2/15	0/1/16	0/1/16	0/1/16	0/0/17
(6) CNN$^-$						10/4/3	2/1/14	17/0/0	17/0/0	17/0/0	13/4/0	2/3/12	13/1/3	8/3/6	8/2/7	8/2/7	8/2/7	0/6/11
(7) TMK$_1$							1/0/16	17/0/0	7/3/7	17/0/0	5/2/10	1/0/16	8/2/7	5/0/12	5/0/12	5/0/12	5/0/12	0/2/15
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	16/1/0	10/6/1	17/0/0	15/1/1	14/1/2	15/1/1	15/1/1	2/12/3
(9) CNN$^-$+TMK$_1$									1/0/16	0/1/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										16/0/1	0/0/17	0/1/16	5/7/5	3/1/13	3/1/13	3/1/13	2/2/13	0/0/17
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(12) TMK$_2$+CNN$^-$												0/2/15	13/1/3	4/1/12	4/0/13	4/0/13	4/0/13	0/2/15
(13) EWW$^-$													17/0/0	13/1/3	12/1/4	13/1/3	12/2/3	1/6/10
(14) EWW$^-$+SSM$^-$														2/1/14	2/1/14	2/1/14	2/1/14	0/0/17
(15) SMOTE															3/7/7	3/11/3	2/13/2	1/0/16
(16) gg-SMOTE																5/8/4	6/8/3	1/1/15
(17) ncn-SMOTE																	3/12/2	1/0/16
(18) rng-SMOTE																		1/0/16
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	6/2/9	2/6/9	7/1/9	7/2/8	7/2/8	7/2/8	0/7/10	12/1/4	7/2/8	10/2/5	7/2/8	4/4/9	7/2/8	6/1/10	5/2/10	6/1/10	6/1/10	0/13/4
(2) SSM$^-$		2/4/11	5/4/8	5/3/9	8/8/1	7/6/4	5/3/9	14/1/2	7/6/4	14/1/2	6/8/3	2/4/11	7/3/7	2/3/12	2/3/12	2/3/12	2/2/13	5/5/7
(3) EW$^-$			9/3/5	10/2/5	12/3/2	11/5/1	7/5/5	15/0/2	10/5/2	15/0/2	11/2/4	4/8/5	10/2/5	7/2/8	5/5/7	8/1/8	8/3/6	8/5/4
(4) EW$^-$+SSM$^-$				7/7/3	10/4/3	9/5/3	6/3/8	15/2/0	9/6/2	15/2/0	5/10/2	1/7/9	9/7/1	4/5/8	3/7/7	3/5/9	3/6/8	6/4/7
(5) RUS					13/2/2	8/3/6	7/1/9	16/0/1	8/6/3	14/2/1	8/6/3	2/4/11	3/7/7	0/6/11	1/5/11	1/5/11	1/5/11	5/3/9
(6) CNN$^-$						6/5/6	3/5/9	13/2/2	8/4/5	13/1/3	5/5/7	1/3/13	4/4/9	2/3/12	2/2/13	2/3/12	2/2/13	3/5/9
(7) TMK$_1$							3/6/8	15/1/1	4/5/8	14/2/1	5/3/9	1/2/14	6/3/8	3/3/11	2/3/12	2/3/12	2/3/12	3/7/7
(8) TMK$_2$								13/2/2	9/3/5	13/2/2	8/2/7	5/5/7	9/2/6	6/4/7	5/3/9	7/2/8	6/2/9	4/11/2
(9) CNN$^-$+TMK$_1$									1/0/16	0/3/14	1/1/15	0/2/15	0/1/16	0/2/15	0/1/16	0/1/16	0/1/16	1/4/12
(10) CNN$^-$+TMK$_2$										16/0/1	3/3/11	1/5/11	3/5/9	3/3/11	2/5/10	2/5/10	2/5/10	7/3/7
(11) TMK$_1$+CNN$^-$											1/1/15	0/2/15	0/2/15	0/2/15	0/1/16	0/2/15	0/2/15	1/5/11
(12) TMK$_2$+CNN$^-$												2/4/11	4/8/5	3/4/10	2/5/10	2/5/10	2/5/10	6/3/8
(13) EWW$^-$													10/6/1	8/4/5	6/6/5	7/5/5	7/6/4	6/7/4
(14) EWW$^-$+SSM$^-$														3/6/8	2/7/8	2/7/8	2/7/8	6/4/7
(15) SMOTE															3/8/6	3/13/1	2/15/0	8/3/6
(16) gg-SMOTE																3/10/4	6/9/2	9/3/5
(17) ncn-SMOTE																	1/15/1	8/3/6
(18) rng-SMOTE																		8/3/6
(19) b-SMOTE																		

Tabla C.7: *Prueba t−student del clasificador 7−NN en términos de $IBA_{0.1}$ y Gm*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/1/12	1/2/14	4/1/12	2/1/14	4/1/12	5/1/11	0/4/13	15/0/2	4/3/10	12/0/5	4/2/11	1/1/15	4/2/11	0/1/16	0/0/17	0/0/17	0/1/16	0/7/10
(2) SSM$^-$		5/3/9	3/6/8	2/2/13	6/10/1	9/3/5	8/4/5	17/0/0	6/7/4	14/2/1	6/4/7	2/4/11	3/6/8	0/2/15	0/2/15	0/2/15	0/2/15	7/4/6
(3) EW$^-$			6/4/7	3/4/10	11/1/5	9/3/5	11/5/1	16/0/1	8/4/5	14/1/2	8/3/6	1/5/11	7/3/7	3/1/13	1/2/14	2/1/14	3/1/13	11/3/3
(4) EW$^-$+SSM$^-$				2/7/8	11/4/2	8/5/4	9/4/4	16/0/1	8/6/3	14/1/2	7/6/4	5/5/7	6/8/3	0/5/12	0/5/12	0/6/11	0/7/10	7/6/4
(5) RUS					15/2/0	10/3/4	13/2/2	17/0/0	11/6/0	15/2/0	12/5/0	8/5/4	7/8/2	1/7/9	1/6/10	1/7/9	1/7/9	12/2/3
(6) CNN$^-$						9/2/6	8/4/5	16/0/1	7/4/6	12/3/2	3/7/7	1/4/12	4/3/10	0/3/14	0/1/16	0/3/14	0/2/15	7/4/6
(7) TMK$_1$							8/2/7	16/0/1	7/2/8	14/2/1	5/4/8	3/4/10	5/3/9	1/1/15	1/2/14	1/2/14	1/3/13	7/5/5
(8) TMK$_2$								16/0/1	5/4/8	13/2/2	5/1/11	1/3/13	4/5/8	1/2/14	1/0/16	1/1/15	1/1/15	8/6/3
(9) CNN$^-$+TMK$_1$									1/0/16	1/1/15	0/1/16	0/0/17	1/0/16	0/0/17	0/0/17	0/0/17	0/0/17	1/0/16
(10) CNN$^-$+TMK$_2$										14/2/1	2/6/9	3/4/10	2/6/9	0/3/14	0/3/14	0/6/11	0/3/14	7/4/6
(11) TMK$_1$+CNN$^-$											1/2/14	2/0/15	1/2/14	0/1/16	0/1/16	0/1/16	0/1/16	3/1/13
(12) TMK$_2$+CNN$^-$												3/5/9	6/2/9	0/5/12	0/4/13	0/5/12	0/5/12	8/4/5
(13) EWW$^-$													7/6/4	3/3/11	2/3/12	3/3/11	3/3/11	13/3/1
(14) EWW$^-$+SSM$^-$														0/5/12	0/4/13	1/4/12	0/6/11	9/4/4
(15) SMOTE															2/14/1	5/12/0	2/15/0	15/2/0
(16) gg-SMOTE																5/10/2	3/13/1	16/1/0
(17) ncn-SMOTE																	1/14/2	16/1/0
(18) rng-SMOTE																		15/2/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/1/12	1/2/14	4/5/8	2/3/12	5/0/12	6/2/9	0/4/13	15/0/2	7/1/9	12/1/4	5/2/10	1/1/15	5/4/8	0/1/16	0/0/17	0/0/17	0/1/16	0/7/10
(2) SSM$^-$		2/6/9	3/7/7	2/3/12	5/10/2	9/3/5	8/3/6	15/2/0	7/6/4	13/3/1	6/5/6	1/5/11	5/6/6	0/2/15	0/2/15	1/1/15	0/2/15	6/5/6
(3) EW$^-$			7/4/6	5/3/9	10/3/4	11/1/5	11/5/1	16/0/1	8/5/4	14/1/2	9/2/6	1/6/10	7/4/6	2/2/13	1/2/14	2/1/14	2/3/12	11/3/3
(4) EW$^-$+SSM$^-$				2/7/8	10/3/4	8/5/4	8/4/5	16/0/1	8/6/3	14/1/2	6/7/4	2/6/9	7/7/3	0/5/12	0/5/12	0/5/12	0/5/12	5/8/4
(5) RUS					15/2/0	11/2/4	11/4/2	17/0/0	11/6/0	15/2/0	12/5/0	7/4/6	7/8/2	1/7/9	1/7/9	1/7/9	1/8/8	10/4/3
(6) CNN$^-$						9/2/6	7/4/6	15/1/1	8/5/4	12/3/2	4/7/6	1/5/11	4/3/10	0/3/14	0/2/15	0/3/14	0/2/15	7/3/7
(7) TMK$_1$							6/3/8	16/0/1	7/2/8	14/1/2	6/1/10	3/3/11	6/2/9	1/3/13	1/2/14	1/2/14	1/3/13	5/6/6
(8) TMK$_2$								16/0/1	7/2/8	13/2/2	6/2/9	1/3/13	6/3/8	0/4/13	1/0/16	0/2/15	1/1/15	7/7/3
(9) CNN$^-$+TMK$_1$									1/0/16	1/1/15	1/0/16	0/0/17	1/0/16	0/0/17	0/0/17	0/0/17	0/0/17	1/0/16
(10) CNN$^-$+TMK$_2$										14/2/1	2/5/10	3/2/12	2/7/8	0/2/15	0/2/15	0/3/14	0/3/14	6/3/8
(11) TMK$_1$+CNN$^-$											2/1/14	2/0/15	1/2/14	0/2/15	0/2/15	0/2/15	0/2/15	3/0/14
(12) TMK$_2$+CNN$^-$												3/4/10	6/3/8	0/4/13	0/3/14	0/5/12	0/5/12	7/4/6
(13) EWW$^-$													9/5/3	3/4/10	2/3/12	3/3/11	3/4/10	13/3/1
(14) EWW$^-$+SSM$^-$														0/5/12	0/3/14	1/3/13	0/5/12	7/4/6
(15) SMOTE															2/12/3	4/13/0	2/14/1	15/2/0
(16) gg-SMOTE																4/11/2	2/14/1	16/1/0
(17) ncn-SMOTE																	1/14/2	16/1/0
(18) rng-SMOTE																		15/2/0
(19) b-SMOTE																		

Tabla C.8: *Prueba $t-$student del clasificador $13-NN$ en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/16	0/3/14	0/0/17	0/0/17	0/1/16	0/1/16	0/4/13	0/0/17	0/0/17	0/0/17	0/0/17	0/3/14	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	0/6/11
(2) SSM$^-$		13/3/1	0/2/15	3/2/12	6/3/8	2/4/11	14/3/0	1/0/16	0/1/16	1/0/16	0/1/16	10/2/5	0/2/15	3/1/13	3/1/13	3/2/12	3/2/12	14/1/2
(3) EW$^-$			0/0/17	0/2/15	3/4/10	1/1/15	13/3/1	1/0/16	0/0/17	1/0/16	0/0/17	0/6/11	0/0/17	1/1/15	1/0/16	1/2/14	1/2/14	12/2/3
(4) EW$^-$+SSM$^-$				9/2/6	14/0/3	4/4/9	17/0/0	1/0/16	3/5/9	1/0/16	9/5/3	16/1/0	0/7/10	9/3/5	8/4/5	9/3/5	9/3/5	15/2/0
(5) RUS					12/1/4	8/0/9	16/0/1	1/1/15	3/3/11	1/2/14	6/5/6	15/1/1	6/2/9	5/9/3	4/10/3	8/9/0	6/8/3	17/0/0
(6) CNN$^-$						3/4/10	14/1/2	0/2/15	0/3/14	0/2/15	0/3/14	9/2/6	3/0/14	4/1/12	4/0/13	4/1/12	4/1/12	12/3/2
(7) TMK$_1$							15/1/1	0/2/15	7/2/8	0/2/15	8/3/6	15/1/1	8/3/6	9/1/7	9/1/7	10/0/7	9/2/6	14/2/1
(8) TMK$_2$								1/0/16	0/0/17	1/0/16	0/0/17	1/3/13	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	8/6/3
(9) CNN$^-$+TMK$_1$									14/2/1	9/8/0	14/2/1	16/0/1	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	16/1/0
(10) CNN$^-$+TMK$_2$										1/2/14	14/3/0	16/1/0	4/5/8	11/3/3	11/3/3	11/3/3	11/3/3	17/0/0
(11) TMK$_1$+CNN$^-$											14/2/1	16/0/1	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	16/1/0
(12) TMK$_2$+CNN$^-$												16/1/0	3/1/13	7/5/5	7/2/8	8/5/4	8/3/6	15/2/0
(13) EWW$^-$													0/0/17	1/3/13	1/2/14	2/2/13	1/3/13	13/2/2
(14) EWW$^-$+SSM$^-$														10/2/5	10/2/5	10/2/5	10/2/5	17/0/0
(15) SMOTE															4/10/3	8/9/0	2/15/0	17/0/0
(16) gg-SMOTE																6/10/1	3/12/2	17/0/0
(17) ncn-SMOTE																	2/11/4	17/0/0
(18) rng-SMOTE																		16/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	17/0/0	13/4/0	17/0/0	17/0/0	17/0/0	17/0/0	13/4/0	17/0/0	17/0/0	17/0/0	17/0/0	15/2/0	17/0/0	17/0/0	17/0/0	17/0/0	17/0/0	9/8/0
(2) SSM$^-$		1/1/15	16/1/0	12/1/4	9/2/6	12/1/4	0/0/17	16/0/1	17/0/0	16/0/1	17/0/0	4/1/12	17/0/0	9/1/7	9/1/7	9/1/7	9/1/7	1/2/14
(3) EW$^-$			17/0/0	16/0/1	15/0/2	16/0/1	1/4/12	16/0/1	17/0/0	16/0/1	17/0/0	11/6/0	17/0/0	16/0/1	15/1/1	16/0/1	16/0/1	3/3/11
(4) EW$^-$+SSM$^-$				7/0/10	3/1/13	8/2/7	0/0/17	16/0/1	13/2/2	16/0/1	4/5/8	0/1/16	15/2/0	4/0/13	4/0/13	4/0/13	4/0/13	0/0/17
(5) RUS					6/1/10	9/0/8	1/0/16	16/0/1	13/0/4	15/1/1	11/2/4	1/1/15	11/0/6	2/2/13	0/3/14	1/2/14	0/4/13	0/0/17
(6) CNN$^-$						12/2/3	2/0/15	17/0/0	17/0/0	17/0/0	15/2/0	4/2/11	13/1/3	11/0/6	11/0/6	10/1/6	11/0/6	2/3/12
(7) TMK$_1$							1/0/16	16/1/0	8/2/7	16/1/0	5/4/8	1/0/16	9/1/7	5/1/11	5/1/11	5/1/11	5/1/11	0/2/15
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	17/0/0	14/2/1	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	3/5/9
(9) CNN$^-$+TMK$_1$									1/0/16	0/2/15	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16
(10) CNN$^-$+TMK$_2$										16/0/1	0/0/17	0/0/17	8/2/7	2/2/13	3/1/13	3/1/13	3/1/13	0/0/17
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16
(12) TMK$_2$+CNN$^-$												0/1/16	13/1/3	4/2/11	4/1/12	4/1/12	4/1/12	0/0/17
(13) EWW$^-$													17/0/0	16/0/1	14/2/1	15/1/1	15/1/1	2/1/14
(14) EWW$^-$+SSM$^-$														2/0/15	2/0/15	2/0/15	2/0/15	0/0/17
(15) SMOTE															2/8/7	4/8/5	4/10/3	0/0/17
(16) gg-SMOTE																4/7/6	6/8/3	0/0/17
(17) ncn-SMOTE																	6/10/1	0/0/17
(18) rng-SMOTE																		0/0/17
(19) b-SMOTE																		

Tabla C.9: *Prueba t−student del clasificador* 13*−NN en términos de Acc y* f_1

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/1/0	7/10/0	17/0/0	17/0/0	15/2/0	16/1/0	5/11/1	17/0/0	17/0/0	17/0/0	17/0/0	9/8/0	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	4/13/0
(2) SSM$^-$		0/3/14	15/2/0	11/1/5	9/4/4	12/1/4	0/2/15	16/0/1	17/0/0	16/0/1	16/1/0	2/3/12	17/0/0	7/3/7	6/3/8	7/1/9	7/2/8	0/3/14
(3) EW$^-$			16/1/0	16/0/1	14/1/2	16/0/1	1/10/6	16/0/1	17/0/0	16/0/1	16/1/0	8/8/1	17/0/0	15/0/2	13/2/2	14/1/2	15/0/2	2/9/6
(4) EW$^-$+SSM$^-$				6/1/10	3/3/11	8/3/6	0/0/17	16/0/1	14/1/2	16/0/1	4/5/8	0/1/16	15/2/0	4/1/12	4/0/13	4/0/13	4/0/13	0/0/17
(5) RUS					6/1/10	9/0/8	1/0/16	16/0/1	13/0/4	15/1/1	10/3/4	1/1/15	11/1/5	1/3/13	1/3/13	1/4/12	1/4/12	0/0/17
(6) CNN$^-$						12/2/3	1/1/15	17/0/0	17/0/0	17/0/0	15/2/0	2/5/10	13/1/3	8/3/6	8/3/6	8/3/6	9/2/6	1/4/12
(7) TMK$_1$							1/0/16	16/1/0	8/2/7	15/2/0	5/4/8	1/0/16	8/2/7	5/0/12	5/0/12	5/0/12	5/0/12	0/2/15
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	17/0/0	9/7/1	17/0/0	15/0/2	15/0/2	15/0/2	15/0/2	2/12/3
(9) CNN$^-$+TMK$_1$									1/0/16	0/2/15	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16
(10) CNN$^-$+TMK$_2$										16/0/1	0/0/17	0/0/17	6/3/8	2/2/13	3/1/13	3/1/13	3/1/13	0/0/17
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16
(12) TMK$_2$+CNN$^-$												0/1/16	13/1/3	4/2/11	4/1/12	4/1/12	4/1/12	0/0/17
(13) EWW$^-$													17/0/0	14/0/3	12/2/3	13/1/3	14/0/3	1/8/8
(14) EWW$^-$+SSM$^-$														2/0/15	2/0/15	2/0/15	2/0/15	0/0/17
(15) SMOTE															2/8/7	5/9/3	4/11/2	1/0/16
(16) gg-SMOTE																4/10/3	5/10/2	1/0/16
(17) ncn-SMOTE																	4/11/2	1/0/16
(18) rng-SMOTE																		1/0/16
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/2/8	1/7/9	7/2/8	8/1/8	7/3/7	7/2/8	0/7/10	11/2/4	7/2/8	10/2/5	7/2/8	3/5/9	8/1/8	6/1/10	6/1/10	7/0/10	6/1/10	0/10/7
(2) SSM$^-$		2/6/9	6/6/5	5/4/8	11/6/0	8/4/5	5/4/8	14/2/1	9/4/4	14/2/1	8/3/6	1/5/11	9/3/5	2/2/13	2/3/12	2/3/12	2/4/11	6/3/8
(3) EW$^-$			8/4/5	10/1/6	11/6/0	12/4/1	7/7/3	15/1/1	9/6/2	14/1/2	9/3/5	2/10/5	9/4/4	7/3/7	7/4/6	8/3/6	8/3/6	7/8/2
(4) EW$^-$+SSM$^-$				5/8/4	10/4/3	11/2/4	7/2/8	16/1/0	12/5/0	15/2/0	5/11/1	3/6/8	11/6/0	4/3/10	4/5/8	4/5/8	4/5/8	7/3/7
(5) RUS					13/1/3	9/3/5	8/1/8	15/2/0	10/4/3	14/3/0	10/4/3	3/4/10	6/7/4	2/3/12	2/4/11	1/6/10	2/4/11	7/2/8
(6) CNN$^-$						6/5/6	5/4/8	13/2/2	9/2/6	13/1/3	5/4/8	0/6/11	4/4/9	3/1/13	3/1/13	3/0/14	3/1/13	4/4/9
(7) TMK$_1$							3/6/8	13/2/2	7/1/9	13/2/2	4/5/8	1/2/14	5/2/10	2/2/13	2/2/13	2/3/12	3/2/12	3/6/8
(8) TMK$_2$								13/2/2	10/1/6	13/2/2	9/2/6	5/5/7	9/1/7	7/2/8	7/1/9	7/1/9	8/1/8	4/11/2
(9) CNN$^-$+TMK$_1$									0/2/15	0/3/14	1/1/15	1/1/15	0/1/16	0/1/16	0/1/16	0/1/16	0/1/16	1/4/12
(10) CNN$^-$+TMK$_2$										15/2/0	2/5/10	1/6/10	2/4/11	1/6/10	2/5/10	2/5/10	3/4/10	7/1/9
(11) TMK$_1$+CNN$^-$											1/2/14	1/2/14	0/2/15	1/1/15	1/0/16	0/2/15	1/1/15	1/5/11
(12) TMK$_2$+CNN$^-$												1/9/7	4/8/5	3/4/10	3/4/10	3/5/9	3/5/9	6/3/8
(13) EWW$^-$													9/6/2	7/4/6	7/4/6	7/6/4	8/3/6	7/5/5
(14) EWW$^-$+SSM$^-$														3/3/11	3/3/11	3/4/10	3/5/9	8/1/8
(15) SMOTE															3/9/5	4/12/1	6/10/1	9/1/7
(16) gg-SMOTE																5/10/2	5/11/1	9/1/7
(17) ncn-SMOTE																	3/13/1	9/1/7
(18) rng-SMOTE																		9/0/8
(19) b-SMOTE																		

Tabla C.10: *Prueba $t-$student del clasificador $13-NN$ en términos de $IBA_{0.1}$ y Gm*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/2/11	0/3/14	6/0/11	1/3/13	5/1/11	4/3/10	0/5/12	15/1/1	5/2/10	13/1/3	5/0/12	1/2/14	6/1/10	0/0/17	0/0/17	0/0/17	0/0/17	0/8/9
(2) SSM$^-$		6/3/8	5/6/6	2/2/13	9/6/2	9/4/4	8/3/6	16/1/0	9/3/5	15/1/1	6/4/7	3/4/10	7/4/6	0/2/15	0/3/14	0/3/14	0/3/14	8/3/6
(3) EW$^-$			7/2/8	3/4/10	8/4/5	8/3/6	12/5/0	16/1/0	9/2/6	15/0/2	7/3/7	0/7/10	9/3/5	1/2/14	1/2/14	2/1/14	2/1/14	11/3/3
(4) EW$^-$+SSM$^-$				2/7/8	9/5/3	8/5/4	9/2/6	16/0/1	11/2/4	15/1/1	6/8/3	6/4/7	9/5/3	1/4/12	1/3/13	1/5/11	2/3/12	6/5/6
(5) RUS					14/3/0	10/3/4	13/3/1	17/0/0	11/5/1	15/2/0	11/6/0	9/4/4	10/4/3	1/6/10	1/7/9	1/8/8	2/6/9	11/5/1
(6) CNN$^-$						8/3/6	9/2/6	15/1/1	7/4/6	14/1/2	4/3/10	2/4/11	6/2/9	0/1/16	0/1/16	0/2/15	0/2/15	7/3/7
(7) TMK$_1$							8/2/7	14/2/1	6/2/9	13/2/2	4/4/9	4/5/8	7/1/9	1/2/14	1/1/15	1/3/13	1/3/13	7/4/6
(8) TMK$_2$								16/1/0	6/2/9	15/0/2	6/0/11	1/3/13	6/4/7	1/0/16	1/0/16	1/0/16	1/0/16	6/8/3
(9) CNN$^-$+TMK$_1$									1/0/16	0/3/14	0/1/16	0/1/16	1/0/16	0/0/17	0/0/17	0/0/17	0/0/17	0/1/16
(10) CNN$^-$+TMK$_2$										14/2/1	2/7/8	3/4/10	3/7/7	0/5/12	0/2/15	0/4/13	0/4/13	7/3/7
(11) TMK$_1$+CNN$^-$											1/2/14	2/0/15	1/0/16	0/1/16	0/1/16	0/1/16	0/1/16	2/1/14
(12) TMK$_2$+CNN$^-$												6/3/8	7/5/5	0/6/11	0/4/13	0/7/10	1/5/11	8/4/5
(13) EWW$^-$													9/5/3	1/5/11	1/4/12	3/1/13	3/2/12	13/1/3
(14) EWW$^-$+SSM$^-$														1/2/14	0/1/16	1/2/14	1/1/15	8/1/8
(15) SMOTE															2/12/3	5/12/0	4/12/1	15/2/0
(16) gg-SMOTE																5/10/2	3/13/1	15/2/0
(17) ncn-SMOTE																	1/15/1	15/2/0
(18) rng-SMOTE																		15/2/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	5/2/10	0/3/14	6/2/9	3/1/13	5/1/11	6/1/10	0/5/12	15/1/1	6/1/10	13/1/3	5/2/10	1/2/14	7/2/8	0/0/17	0/0/17	0/0/17	0/0/17	0/8/9
(2) SSM$^-$		4/5/8	6/5/6	2/3/12	10/5/2	10/2/5	7/4/6	15/2/0	9/4/4	14/2/1	6/5/6	2/5/10	9/3/5	0/2/15	0/3/14	1/2/14	0/3/14	7/3/7
(3) EW$^-$			8/2/7	4/4/9	8/6/3	9/2/6	11/6/0	16/1/0	10/3/4	15/0/2	8/2/7	0/7/10	9/3/5	1/2/14	1/2/14	2/2/13	2/2/13	11/3/3
(4) EW$^-$+SSM$^-$				2/6/9	9/4/4	9/4/4	8/2/7	16/0/1	11/2/4	15/0/2	4/9/4	5/4/8	10/4/3	1/4/12	1/2/14	1/5/11	2/3/12	5/6/6
(5) RUS					14/3/0	10/3/4	12/2/3	17/0/0	12/4/1	15/2/0	10/6/1	7/6/4	11/4/2	2/5/10	2/6/9	1/8/8	2/6/9	11/3/3
(6) CNN$^-$						8/3/6	7/4/6	15/2/0	8/5/4	14/1/2	4/5/8	2/3/12	6/4/7	0/2/15	0/1/16	0/3/14	0/3/14	7/3/7
(7) TMK$_1$							7/3/7	15/1/1	7/1/9	13/2/2	3/6/8	5/2/10	7/1/9	1/2/14	1/2/14	1/3/13	2/2/13	6/5/6
(8) TMK$_2$								16/1/0	7/3/7	15/0/2	6/2/9	1/3/13	7/4/6	0/1/16	1/0/16	1/0/16	1/0/16	6/8/3
(9) CNN$^-$+TMK$_1$									1/0/16	0/3/14	0/1/16	0/1/16	1/0/16	0/0/17	0/0/17	0/0/17	0/0/17	0/1/16
(10) CNN$^-$+TMK$_2$										14/2/1	2/4/11	3/4/10	3/7/7	0/2/15	0/2/15	0/2/15	0/3/14	7/2/8
(11) TMK$_1$+CNN$^-$											1/2/14	2/0/15	1/0/16	0/1/16	0/1/16	0/1/16	0/1/16	2/1/14
(12) TMK$_2$+CNN$^-$												4/5/8	7/5/5	0/6/11	1/2/14	0/7/10	1/4/12	7/5/5
(13) EWW$^-$													11/3/3	0/6/11	1/4/12	2/4/11	3/2/12	12/2/3
(14) EWW$^-$+SSM$^-$														1/2/14	1/0/16	1/1/15	1/1/15	6/2/9
(15) SMOTE															2/11/4	5/12/0	4/12/1	15/2/0
(16) gg-SMOTE																6/9/2	3/13/1	15/2/0
(17) ncn-SMOTE																	1/15/1	15/2/0
(18) rng-SMOTE																		15/1/1
(19) b-SMOTE																		

Tabla C.11: *Prueba t−student del clasificador J48 en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/15	0/4/13	0/1/16	0/0/17	1/2/14	0/0/17	0/8/9	0/0/17	0/2/15	0/0/17	1/1/15	0/4/13	0/1/16	0/2/15	0/1/16	0/1/16	0/2/15	1/10/6
(2) SSM$^-$		9/4/4	0/2/15	3/3/11	3/11/3	2/2/13	11/3/3	1/0/16	0/4/13	1/0/16	0/8/9	7/4/6	0/3/14	6/7/4	5/6/6	6/5/6	6/9/2	14/1/2
(3) EW$^-$			1/1/15	1/3/13	5/6/6	1/0/16	11/5/1	1/0/16	1/2/14	1/0/16	3/2/12	0/8/9	0/1/16	5/4/8	5/2/10	5/2/10	5/6/6	12/4/1
(4) EW$^-$+SSM$^-$				8/5/4	12/3/2	3/3/11	15/1/1	1/0/16	6/7/4	1/0/16	10/5/2	15/1/1	0/9/8	12/3/2	11/4/2	11/4/2	13/2/2	15/2/0
(5) RUS					12/2/3	5/2/10	15/1/1	1/1/15	6/1/10	1/3/13	9/2/6	12/3/2	4/4/9	13/3/1	13/3/1	11/6/0	14/3/0	16/1/0
(6) CNN$^-$						1/2/14	11/2/4	0/0/17	0/6/11	0/0/17	0/9/8	6/5/6	2/2/13	5/5/7	5/4/8	5/6/6	5/7/5	12/4/1
(7) TMK$_1$							16/0/1	0/0/17	11/3/3	0/2/15	12/3/2	16/0/1	9/6/2	11/5/1	13/3/1	12/3/2	13/3/1	17/0/0
(8) TMK$_2$								1/0/16	0/3/14	1/0/16	1/2/14	2/4/11	0/1/16	1/6/10	2/2/13	2/2/13	2/3/12	7/10/0
(9) CNN$^-$+TMK$_1$									16/0/1	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	17/0/0	16/0/1	16/0/1	17/0/0
(10) CNN$^-$+TMK$_2$										1/0/16	11/6/0	13/3/1	2/6/9	12/4/1	12/4/1	12/3/2	13/3/1	16/1/0
(11) TMK$_1$+CNN$^-$											16/0/1	16/0/1	14/2/1	16/0/1	17/0/0	16/0/1	16/0/1	17/0/0
(12) TMK$_2$+CNN$^-$												9/5/3	2/3/12	11/4/2	9/6/2	10/4/3	12/3/2	15/2/0
(13) EWW$^-$													0/0/17	7/4/6	5/5/7	4/5/8	7/5/5	12/5/0
(14) EWW$^-$+SSM$^-$														15/2/0	14/3/0	13/4/0	15/2/0	17/0/0
(15) SMOTE															1/11/5	1/13/3	2/15/0	15/2/0
(16) gg-SMOTE																0/16/1	4/12/1	16/1/0
(17) ncn-SMOTE																	3/14/0	15/2/0
(18) rng-SMOTE																		15/2/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/1/0	13/4/0	17/0/0	17/0/0	17/0/0	17/0/0	10/7/0	17/0/0	17/0/0	17/0/0	17/0/0	14/3/0	17/0/0	16/1/0	16/1/0	16/1/0	16/1/0	10/7/0
(2) SSM$^-$		1/3/13	13/4/0	6/6/5	2/8/7	11/0/6	1/2/14	16/0/1	15/1/1	16/0/1	6/8/3	2/4/11	16/1/0	1/3/13	1/3/13	1/3/13	1/4/12	0/4/13
(3) EW$^-$			17/0/0	14/1/2	11/4/2	16/0/1	2/4/11	16/0/1	17/0/0	16/0/1	13/4/0	10/7/0	17/0/0	9/4/4	10/3/4	9/3/5	8/5/4	2/5/10
(4) EW$^-$+SSM$^-$				4/1/12	2/2/13	8/3/6	0/0/17	16/0/1	3/9/5	16/0/1	1/5/11	0/1/16	7/10/0	0/2/15	0/2/15	0/2/15	0/1/16	0/1/16
(5) RUS					3/4/10	10/0/7	1/1/15	16/0/1	12/0/5	13/2/2	6/5/6	2/3/12	13/0/4	1/0/16	1/0/16	0/1/16	1/0/16	0/1/16
(6) CNN$^-$						12/1/4	2/2/13	17/0/0	15/2/0	17/0/0	10/7/0	3/6/8	15/0/2	3/4/10	2/6/9	3/3/11	3/4/10	0/5/12
(7) TMK$_1$							1/0/16	17/0/0	7/2/8	17/0/0	6/1/10	1/0/16	8/1/8	1/2/14	0/1/16	1/1/15	1/0/16	0/0/17
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	17/0/0	12/4/1	17/0/0	14/1/2	13/2/2	14/2/1	13/2/2	1/12/4
(9) CNN$^-$+TMK$_1$									1/0/16	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										16/0/1	0/4/13	0/1/16	7/9/1	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	2/0/15	1/0/16	0/0/17	1/0/16	1/0/16	0/0/17
(12) TMK$_2$+CNN$^-$												0/5/12	13/3/1	0/3/14	0/3/14	0/2/15	0/3/14	0/2/15
(13) EWW$^-$													17/0/0	4/5/8	7/3/7	6/4/7	4/4/9	0/4/13
(14) EWW$^-$+SSM$^-$														0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(15) SMOTE															5/10/2	5/10/2	0/16/1	0/4/13
(16) gg-SMOTE																3/13/1	2/8/7	0/2/15
(17) ncn-SMOTE																	0/10/7	0/2/15
(18) rng-SMOTE																		0/5/12
(19) b-SMOTE																		

Tabla C.12: *Prueba $t-$student del clasificador J48 en términos de Acc y f_1*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	15/2/0	8/9/0	17/0/0	17/0/0	15/2/0	17/0/0	3/14/0	17/0/0	17/0/0	17/0/0	16/1/0	12/5/0	17/0/0	13/3/1	13/3/1	13/3/1	10/5/2	7/10/0
(2) SSM$^-$		2/2/13	11/6/0	6/4/7	2/9/6	9/2/6	0/3/14	16/0/1	12/4/1	16/0/1	5/9/3	2/4/11	13/4/0	1/4/12	1/4/12	1/4/12	1/2/14	0/5/12
(3) EW$^-$			17/0/0	14/2/1	12/2/3	15/1/1	2/10/5	16/0/1	16/1/0	16/0/1	13/4/0	7/10/0	17/0/0	6/6/5	6/6/5	8/4/5	5/7/5	2/11/4
(4) EW$^-$+SSM$^-$				4/2/11	2/2/13	8/2/7	0/0/17	16/0/1	3/9/5	15/1/1	2/4/11	0/1/16	6/10/1	0/1/16	0/1/16	0/1/16	0/1/16	0/1/16
(5) RUS					3/5/9	10/0/7	1/2/14	16/0/1	11/2/4	13/2/2	7/5/5	1/4/12	13/0/4	1/0/16	1/1/15	0/1/16	1/0/16	0/1/16
(6) CNN$^-$						10/2/5	2/2/13	17/0/0	15/2/0	17/0/0	8/9/0	3/5/9	14/1/2	2/4/11	1/5/11	2/5/10	2/5/10	0/5/12
(7) TMK$_1$							1/0/16	17/0/0	7/2/8	17/0/0	6/3/8	1/2/14	8/1/8	1/2/14	0/3/14	1/2/14	1/2/14	0/1/16
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	16/1/0	11/5/1	17/0/0	9/6/2	11/2/4	8/7/2	6/8/3	4/12/1
(9) CNN$^-$+TMK$_1$									1/0/16	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										14/2/1	0/6/11	0/2/15	6/10/1	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	2/0/15	1/0/16	0/0/17	1/0/16	1/0/16	0/0/17
(12) TMK$_2$+CNN$^-$												0/5/12	11/5/1	0/4/13	0/4/13	0/2/15	0/2/15	0/2/15
(13) EWW$^-$													16/1/0	4/7/6	4/7/6	5/4/8	2/7/8	0/7/10
(14) EWW$^-$+SSM$^-$														0/0/17	0/0/17	0/0/17	0/0/17	0/0/17
(15) SMOTE															3/13/1	3/12/2	0/16/1	1/10/6
(16) gg-SMOTE																4/12/1	2/10/5	1/11/5
(17) ncn-SMOTE																	1/10/6	1/9/7
(18) rng-SMOTE																		2/11/4
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/4/4	1/7/9	9/2/6	8/2/7	7/8/2	8/6/3	0/11/6	14/0/3	10/2/5	10/4/3	8/4/5	2/7/8	9/1/7	1/7/9	2/7/8	2/8/7	1/6/10	1/15/1
(2) SSM$^-$		1/3/13	3/8/6	1/5/11	4/9/4	5/4/8	0/7/10	15/1/1	5/7/5	11/4/2	2/8/7	1/2/14	4/7/6	0/4/13	1/5/11	0/4/13	0/4/13	2/6/9
(3) EW$^-$			12/4/1	8/7/2	12/4/1	11/5/1	1/15/1	15/2/0	10/7/0	14/3/0	11/5/1	2/14/1	11/6/0	3/11/3	6/9/2	5/9/3	3/10/4	11/5/1
(4) EW$^-$+SSM$^-$				2/6/9	4/10/3	6/4/7	0/8/9	15/1/1	4/9/4	12/3/2	3/10/4	0/5/12	3/11/3	1/4/12	2/4/11	2/1/14	2/2/13	4/4/9
(5) RUS					8/6/3	9/2/6	1/8/8	15/2/0	9/7/1	12/4/1	7/8/2	1/7/9	9/5/3	2/5/10	2/4/11	2/5/10	3/4/10	5/4/8
(6) CNN$^-$						5/2/10	1/5/11	13/1/3	6/8/3	11/3/3	2/12/3	1/2/14	5/7/5	0/4/13	0/4/13	0/5/12	0/4/13	2/9/6
(7) TMK$_1$							1/6/10	15/2/0	8/3/6	12/5/0	7/3/7	1/4/12	7/2/8	0/6/11	0/7/10	1/4/12	1/4/12	4/6/7
(8) TMK$_2$								15/2/0	11/6/0	13/4/0	9/7/1	2/12/3	9/8/0	1/12/4	3/11/3	3/10/4	1/11/5	7/10/0
(9) CNN$^-$+TMK$_1$									0/1/16	1/1/15	0/2/15	0/2/15	0/2/15	0/1/16	1/1/15	1/0/16	0/2/15	2/1/14
(10) CNN$^-$+TMK$_2$										12/3/2	2/9/6	0/6/11	2/10/5	1/3/13	3/1/13	2/1/14	1/2/14	4/4/9
(11) TMK$_1$+CNN$^-$											1/4/12	1/2/14	1/4/12	0/3/14	1/3/13	1/2/14	1/2/14	3/5/9
(12) TMK$_2$+CNN$^-$												1/5/11	4/9/4	1/7/9	2/4/11	1/5/11	0/5/12	5/5/7
(13) EWW$^-$													11/6/0	3/9/5	6/8/3	5/7/5	5/7/5	8/8/1
(14) EWW$^-$+SSM$^-$														1/6/10	2/5/10	2/3/12	2/4/11	5/5/7
(15) SMOTE															3/13/1	3/13/1	1/16/0	10/6/1
(16) gg-SMOTE																3/12/2	0/13/4	8/7/2
(17) ncn-SMOTE																	1/14/2	7/8/2
(18) rng-SMOTE																		10/6/1
(19) b-SMOTE																		

Tabla C.13: *Prueba t−student del clasificador J48 en términos de $IBA_{0.1}$ y Gm*

	$IBA_{0.1}$																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/5/8	0/5/12	5/3/9	0/4/13	3/6/8	5/3/9	0/9/8	14/1/2	3/6/8	9/2/6	3/4/10	0/5/12	4/3/10	0/2/15	0/1/16	0/4/13	0/2/15	1/14/2
(2) SSM$^-$		1/6/10	3/8/6	1/1/15	2/11/4	6/3/8	3/7/7	15/1/1	2/10/5	12/1/4	0/10/7	1/4/12	5/7/5	1/3/13	2/3/12	1/3/13	0/5/12	6/6/5
(3) EW$^-$			8/6/3	2/7/8	8/8/1	8/2/7	7/10/0	15/2/0	6/10/1	11/2/4	6/10/1	1/11/5	7/7/3	3/7/7	4/3/10	2/7/8	2/9/6	12/5/0
(4) EW$^-$+SSM$^-$				0/5/12	5/9/3	6/4/7	5/6/6	15/0/2	3/9/5	10/5/2	2/8/7	1/6/10	1/12/4	2/3/12	2/4/11	2/3/12	2/3/12	8/5/4
(5) RUS					12/4/1	8/8/1	9/8/0	17/0/0	11/6/0	13/3/1	10/6/1	5/10/2	11/6/0	5/9/3	3/10/4	5/8/4	5/9/3	12/5/0
(6) CNN$^-$						5/3/9	4/7/6	14/1/2	3/11/3	10/2/5	2/10/5	1/6/10	4/6/7	2/2/13	2/0/15	2/3/12	1/4/12	6/8/3
(7) TMK$_1$							6/5/6	16/0/1	8/2/7	12/5/0	7/4/6	5/3/9	6/4/7	2/6/9	1/7/9	3/4/10	5/2/10	8/4/5
(8) TMK$_2$								16/1/0	6/8/3	11/2/4	5/7/5	1/6/10	7/5/5	0/8/9	1/4/12	1/5/11	1/7/9	7/10/0
(9) CNN$^-$+TMK$_1$									0/1/16	1/0/16	0/2/15	0/2/15	1/0/16	0/1/16	1/1/15	0/0/17	0/1/16	2/1/14
(10) CNN$^-$+TMK$_2$										11/2/4	1/10/6	1/6/10	3/11/3	1/5/11	2/4/11	2/4/11	3/5/9	9/5/3
(11) TMK$_1$+CNN$^-$											4/1/12	2/3/12	1/5/11	2/2/13	1/4/12	1/3/13	2/3/12	5/1/11
(12) TMK$_2$+CNN$^-$												1/8/8	6/6/5	1/8/8	2/5/10	2/5/10	2/8/7	10/4/3
(13) EWW$^-$													7/7/3	4/7/6	3/8/6	2/9/6	5/6/6	11/6/0
(14) EWW$^-$+SSM$^-$														3/3/11	4/4/9	4/3/10	2/5/10	10/3/4
(15) SMOTE															1/12/4	1/13/3	2/15/0	14/3/0
(16) gg-SMOTE																0/16/1	4/12/1	15/2/0
(17) ncn-SMOTE																	4/13/0	13/4/0
(18) rng-SMOTE																		13/4/0
(19) b-SMOTE																		

	Gm																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	5/4/8	0/6/11	6/3/8	0/5/12	4/7/6	6/3/8	0/10/7	14/1/2	3/7/7	9/3/5	4/4/9	0/5/12	5/5/7	0/3/14	0/1/16	0/4/13	0/2/15	1/14/2
(2) SSM$^-$		1/4/12	5/7/5	1/2/14	3/10/4	6/3/8	3/5/9	15/1/1	5/7/5	12/1/4	0/10/7	1/3/13	6/6/5	0/4/13	1/4/12	0/4/13	0/4/13	7/4/6
(3) EW$^-$			11/4/2	3/7/7	9/7/1	9/2/6	6/11/0	15/2/0	8/9/0	12/2/3	7/9/1	0/13/4	8/7/2	1/8/8	4/3/10	2/8/7	2/10/5	12/5/0
(4) EW$^-$+SSM$^-$				0/3/14	5/9/3	6/4/7	3/6/8	15/0/2	3/9/5	12/2/3	2/9/6	0/5/12	1/13/3	2/3/12	0/5/12	0/4/13	1/4/12	7/5/5
(5) RUS					11/5/1	10/4/3	9/8/0	17/0/0	13/4/0	14/2/1	10/6/1	4/10/3	13/4/0	4/9/4	3/8/6	4/9/4	5/8/4	10/7/0
(6) CNN$^-$						5/3/9	3/6/8	14/1/2	4/10/3	11/2/4	2/11/4	1/5/11	5/6/6	2/2/13	0/3/14	0/5/12	0/5/12	6/7/4
(7) TMK$_1$							6/4/7	16/0/1	9/0/8	13/3/1	7/4/6	5/3/9	5/5/7	2/4/11	1/7/9	3/4/10	4/3/10	7/4/6
(8) TMK$_2$								15/2/0	6/10/1	12/1/4	5/9/3	1/7/9	8/6/3	0/9/8	1/5/11	0/6/11	0/8/9	7/10/0
(9) CNN$^-$+TMK$_1$									0/3/14	1/0/16	1/1/15	0/2/15	1/1/15	0/1/16	1/0/16	0/1/16	0/2/15	3/0/14
(10) CNN$^-$+TMK$_2$										11/3/3	1/9/7	1/6/10	3/11/3	1/4/12	2/2/13	1/3/13	0/5/12	9/2/6
(11) TMK$_1$+CNN$^-$											3/1/13	2/2/13	2/4/11	1/3/13	2/2/13	1/2/14	1/3/13	5/1/11
(12) TMK$_2$+CNN$^-$												1/8/8	6/6/5	1/8/8	2/5/10	2/4/11	0/8/9	8/5/4
(13) EWW$^-$													9/7/1	3/8/6	2/9/6	2/9/6	4/7/6	11/6/0
(14) EWW$^-$+SSM$^-$														2/5/10	2/5/10	2/5/10	2/6/9	7/4/6
(15) SMOTE															1/12/4	1/13/3	2/15/0	13/4/0
(16) gg-SMOTE																1/15/1	4/12/1	14/3/0
(17) ncn-SMOTE																	3/13/1	12/5/0
(18) rng-SMOTE																		13/4/0
(19) b-SMOTE																		

Tabla C.14: *Prueba $t-$student del clasificador J48 en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/3/14	0/7/10	0/1/16	0/1/16	0/6/11	0/0/17	1/10/6	0/0/17	0/1/16	0/0/17	0/2/15	0/4/13	0/1/16	0/0/17	0/0/17	0/0/17	0/0/17	2/9/6
(2) SSM$^-$		7/8/2	0/6/11	3/3/11	1/14/2	2/2/13	12/4/1	1/0/16	0/6/11	1/0/16	0/9/8	7/4/6	0/4/13	2/8/7	2/8/7	2/8/7	2/9/6	11/5/1
(3) EW$^-$			0/3/14	3/2/12	4/8/5	1/1/15	9/7/1	1/0/16	0/4/13	1/0/16	2/5/10	0/10/7	0/1/16	3/1/13	2/2/13	3/2/12	2/3/12	9/5/3
(4) EW$^-$+SSM$^-$				7/4/6	10/6/1	2/3/12	16/1/0	1/0/16	5/8/4	1/0/16	8/7/2	12/4/1	0/8/9	7/7/3	9/5/3	8/6/3	9/7/1	14/3/0
(5) RUS					11/4/2	2/6/9	15/1/1	1/0/16	6/5/6	1/3/13	6/7/4	10/4/3	4/4/9	8/7/2	9/6/2	9/6/2	10/5/2	16/1/0
(6) CNN$^-$						1/4/12	11/4/2	0/0/17	1/5/11	0/1/16	1/10/6	5/4/8	1/5/11	2/7/8	2/6/9	2/8/7	3/5/9	10/4/3
(7) TMK$_1$							16/0/1	0/0/17	11/3/3	0/4/13	11/4/2	15/1/1	9/4/4	10/5/2	10/5/2	11/4/2	12/3/2	17/0/0
(8) TMK$_2$								1/0/16	0/1/16	1/0/16	0/0/17	1/3/13	0/1/16	2/0/15	1/1/15	1/1/15	2/1/14	1/13/3
(9) CNN$^-$+TMK$_1$									16/0/1	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	16/0/1	16/0/1	16/0/1	17/0/0
(10) CNN$^-$+TMK$_2$										1/1/15	5/12/0	12/4/1	1/6/10	8/5/4	8/7/2	9/6/2	8/8/1	16/0/1
(11) TMK$_1$+CNN$^-$											16/0/1	15/1/1	12/4/1	16/0/1	16/0/1	15/1/1	16/0/1	17/0/0
(12) TMK$_2$+CNN$^-$												10/3/4	1/6/10	6/7/4	6/6/5	7/6/4	7/8/2	13/2/2
(13) EWW$^-$													0/2/15	4/3/10	4/2/11	5/5/7	5/2/10	13/4/0
(14) EWW$^-$+SSM$^-$														10/6/1	10/6/1	10/6/1	12/4/1	16/1/0
(15) SMOTE															2/14/1	1/15/1	2/14/1	15/2/0
(16) gg-SMOTE																0/16/1	3/13/1	15/2/0
(17) ncn-SMOTE																	2/15/0	14/3/0
(18) rng-SMOTE																		14/3/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	17/0/0	10/7/0	17/0/0	16/1/0	16/1/0	17/0/0	6/10/1	17/0/0	17/0/0	17/0/0	17/0/0	14/3/0	17/0/0	16/1/0	17/0/0	17/0/0	16/1/0	8/8/1
(2) SSM$^-$		2/1/14	12/5/0	5/6/6	1/8/8	10/1/6	0/1/16	16/0/1	8/6/3	14/2/1	5/8/4	3/3/11	11/5/1	3/3/11	3/2/12	2/5/10	3/3/11	0/3/14
(3) EW$^-$			17/0/0	13/2/2	10/4/3	16/0/1	1/6/10	16/0/1	16/1/0	16/0/1	15/2/0	10/7/0	17/0/0	12/3/2	13/1/3	12/2/3	12/3/2	3/5/9
(4) EW$^-$+SSM$^-$				3/2/12	0/4/13	8/2/7	0/0/17	16/0/1	1/7/9	13/2/2	0/5/12	0/1/16	6/10/1	0/1/16	0/0/17	0/1/16	0/0/17	0/0/17
(5) RUS					3/4/10	10/0/7	1/1/15	16/0/1	9/4/4	15/0/2	5/7/5	2/2/13	13/1/3	2/4/11	1/3/13	1/4/12	1/4/12	0/1/16
(6) CNN$^-$						12/2/3	2/0/15	17/0/0	9/8/0	16/0/1	7/10/0	6/1/10	14/3/0	5/4/8	4/7/6	3/8/6	3/7/7	1/3/13
(7) TMK$_1$							1/0/16	17/0/0	5/2/10	17/0/0	4/2/11	1/0/16	8/1/8	2/2/13	3/1/13	2/3/12	2/2/13	0/0/17
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	17/0/0	12/4/1	17/0/0	14/2/1	15/1/1	15/1/1	14/2/1	4/12/1
(9) CNN$^-$+TMK$_1$									1/0/16	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										14/2/1	0/8/9	0/2/15	12/5/0	0/4/13	0/3/14	0/3/14	0/1/16	0/0/17
(11) TMK$_1$+CNN$^-$											2/0/15	1/0/16	1/4/12	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(12) TMK$_2$+CNN$^-$												2/3/12	13/4/0	1/4/12	2/4/11	2/5/10	1/4/12	0/2/15
(13) EWW$^-$													17/0/0	7/5/5	9/3/5	10/2/5	7/5/5	0/6/11
(14) EWW$^-$+SSM$^-$														0/1/16	0/0/17	0/0/17	0/0/17	0/0/17
(15) SMOTE															6/9/2	6/9/2	1/16/0	0/3/14
(16) gg-SMOTE																3/12/2	2/9/6	0/2/15
(17) ncn-SMOTE																	1/9/7	0/1/16
(18) rng-SMOTE																		0/4/13
(19) b-SMOTE																		

Tabla C.15: *Prueba t−student del clasificador MLP en términos de Acc y* f_1

									Acc									
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/0/1	8/8/1	16/0/1	15/1/1	15/1/1	16/1/0	4/13/0	17/0/0	16/0/1	16/0/1	16/0/1	11/5/1	16/0/1	13/2/2	15/1/1	14/2/1	13/2/2	8/9/0
(2) SSM$^-$		2/2/13	12/4/1	5/5/7	1/9/7	10/2/5	1/1/15	16/0/1	6/8/3	14/1/2	2/11/4	2/5/10	11/5/1	0/6/11	0/5/12	1/5/11	0/6/11	1/4/12
(3) EW$^-$			16/0/1	14/1/2	11/3/3	16/0/1	2/9/6	16/0/1	15/1/1	15/0/2	14/2/1	9/8/0	16/0/1	7/6/4	10/4/3	9/5/3	8/6/3	2/9/6
(4) EW$^-$+SSM$^-$				3/3/11	1/3/13	9/2/6	1/0/16	16/0/1	1/9/7	13/1/3	0/5/12	1/1/15	4/12/1	0/1/16	0/0/17	0/2/15	0/0/17	1/0/16
(5) RUS					3/6/8	11/0/6	2/1/14	16/0/1	9/4/4	15/0/2	6/6/5	3/2/12	12/2/3	0/5/12	0/4/13	1/4/12	0/6/11	1/2/14
(6) CNN$^-$						13/2/2	2/2/13	17/0/0	8/9/0	16/0/1	4/13/0	5/4/8	13/4/0	2/7/8	2/8/7	2/7/8	3/6/8	2/4/11
(7) TMK$_1$							1/1/15	17/0/0	4/3/10	16/0/1	3/3/11	1/0/16	8/0/9	2/1/14	3/0/14	1/3/13	1/2/14	0/2/15
(8) TMK$_2$								16/0/1	16/0/1	15/0/2	16/0/1	11/4/2	16/0/1	10/4/3	12/3/2	12/3/2	11/3/3	2/13/2
(9) CNN$^-$+TMK$_1$									1/0/16	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										14/2/1	0/10/7	1/2/14	10/7/0	0/3/14	0/2/15	0/3/14	0/2/15	1/0/16
(11) TMK$_1$+CNN$^-$											2/0/15	2/0/15	2/3/12	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16
(12) TMK$_2$+CNN$^-$												2/4/11	13/4/0	1/4/12	1/3/13	1/4/12	1/5/11	1/3/13
(13) EWW$^-$													16/0/1	4/9/4	5/7/5	7/4/6	5/6/6	1/6/10
(14) EWW$^-$+SSM$^-$														0/0/17	0/0/17	0/0/17	0/0/17	1/0/16
(15) SMOTE															4/11/2	6/11/0	1/16/0	2/5/10
(16) gg-SMOTE																3/13/1	1/11/5	1/7/9
(17) ncn-SMOTE																	0/11/6	1/6/10
(18) rng-SMOTE																		2/5/10
(19) b-SMOTE																		

									f_1									
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/3/5	0/10/7	9/2/6	6/4/7	8/5/4	8/4/5	1/13/3	13/1/3	8/3/6	11/1/5	7/5/5	1/8/8	9/2/6	2/7/8	2/7/8	3/7/7	3/5/9	1/13/3
(2) SSM$^-$		2/4/11	2/9/6	2/6/9	1/12/4	9/2/6	3/5/9	14/1/2	0/12/5	11/3/3	0/13/4	1/4/12	2/9/6	0/3/14	0/5/12	0/4/13	0/5/12	3/5/9
(3) EW$^-$			10/5/2	9/3/5	11/4/2	12/2/3	6/11/0	13/4/0	12/3/2	12/3/2	9/6/2	2/12/3	10/5/2	3/8/6	3/10/4	6/7/4	4/7/6	8/8/1
(4) EW$^-$+SSM$^-$				1/7/9	4/7/6	9/3/5	5/3/9	14/3/0	2/10/5	11/4/2	3/7/7	1/6/10	0/16/1	1/4/12	1/5/11	1/5/11	1/4/12	4/5/8
(5) RUS					5/8/4	11/0/6	6/5/6	16/0/1	5/10/2	14/1/2	6/7/4	4/4/9	8/8/1	0/8/9	1/6/10	0/9/8	0/7/10	6/5/6
(6) CNN$^-$						8/3/6	4/3/10	14/1/2	1/13/3	12/2/3	0/15/2	1/4/12	5/9/3	0/7/10	0/6/11	0/7/10	0/7/10	2/7/8
(7) TMK$_1$							4/4/9	16/0/1	4/3/10	13/3/1	4/3/10	0/6/11	7/1/9	2/3/12	3/1/13	2/3/12	1/4/12	4/5/8
(8) TMK$_2$								13/2/2	9/5/3	12/1/4	7/6/4	1/10/6	9/3/5	1/9/7	1/10/6	4/7/6	2/8/7	1/13/3
(9) CNN$^-$+TMK$_1$									0/3/14	0/1/16	0/3/14	0/1/16	0/1/16	1/0/16	0/2/15	0/2/15	1/0/16	1/2/14
(10) CNN$^-$+TMK$_2$										12/4/1	0/16/1	1/5/11	5/9/3	1/5/11	0/6/11	0/7/10	1/4/12	4/5/8
(11) TMK$_1$+CNN$^-$											3/1/13	1/4/12	1/4/12	1/1/15	1/1/15	1/2/14	1/2/14	3/2/12
(12) TMK$_2$+CNN$^-$												1/7/9	7/7/3	0/9/8	0/7/10	1/7/9	0/8/9	5/5/7
(13) EWW$^-$													9/7/1	3/8/6	3/12/2	4/9/4	3/9/5	7/8/2
(14) EWW$^-$+SSM$^-$														1/3/13	1/5/11	1/4/12	1/3/13	6/2/9
(15) SMOTE															3/14/0	6/11/0	1/14/2	8/9/0
(16) gg-SMOTE																3/12/2	0/13/4	7/8/2
(17) ncn-SMOTE																	0/12/5	6/8/3
(18) rng-SMOTE																		8/8/1
(19) b-SMOTE																		

Tabla C.16: *Prueba $t-$student del clasificador MLP en términos de $IBA_{0.1}$ y Gm*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/5/8	0/8/9	3/4/10	0/4/13	4/6/7	5/3/9	1/11/5	14/1/2	3/4/10	8/3/6	3/5/9	0/4/13	3/3/11	0/2/15	0/2/15	1/1/15	0/3/14	2/11/4
(2) SSM$^-$		3/6/8	2/11/4	0/4/13	2/11/4	8/2/7	5/8/4	16/0/1	1/10/6	9/6/2	1/11/5	2/3/12	1/12/4	0/4/13	0/4/13	0/4/13	0/4/13	6/7/4
(3) EW$^-$			6/7/4	2/5/10	7/7/3	8/2/7	9/8/0	15/2/0	6/8/3	10/3/4	6/8/3	0/12/5	6/6/5	1/5/11	2/5/10	2/4/11	2/5/10	10/5/2
(4) EW$^-$+SSM$^-$				1/6/10	6/8/3	8/3/6	7/6/4	15/1/1	1/12/4	10/4/3	2/10/5	2/5/10	0/14/3	1/5/11	0/5/12	0/6/11	1/5/11	6/7/4
(5) RUS					12/5/0	10/5/2	12/5/0	17/0/0	11/6/0	13/4/0	10/7/0	7/7/3	9/7/1	1/12/4	1/11/5	0/14/3	1/11/5	11/6/0
(6) CNN$^-$						7/3/7	6/6/5	15/1/1	0/15/2	9/5/3	0/14/3	3/3/11	2/9/6	0/5/12	0/5/12	0/4/13	0/5/12	5/9/3
(7) TMK$_1$							9/1/7	16/0/1	5/4/8	14/1/2	5/4/8	5/4/8	6/3/8	1/6/10	3/3/11	3/3/11	2/4/11	9/2/6
(8) TMK$_2$								15/1/1	4/5/8	10/1/6	4/7/6	0/5/12	4/5/8	1/3/13	0/4/13	0/4/13	1/2/14	1/13/3
(9) CNN$^-$+TMK$_1$									0/2/15	1/0/16	0/2/15	0/1/16	0/2/15	0/0/17	0/1/16	0/0/17	0/1/16	1/1/15
(10) CNN$^-$+TMK$_2$										10/6/1	0/16/1	3/5/9	3/12/2	0/5/12	0/6/11	1/4/12	2/4/11	8/4/5
(11) TMK$_1$+CNN$^-$											2/5/10	3/2/12	3/3/11	1/2/14	0/4/13	1/3/13	1/3/13	5/4/8
(12) TMK$_2$+CNN$^-$												3/6/8	4/11/2	0/7/10	0/5/12	1/6/10	0/7/10	8/5/4
(13) EWW$^-$													9/5/3	2/8/7	3/7/7	3/7/7	2/8/7	12/5/0
(14) EWW$^-$+SSM$^-$														1/5/11	0/4/13	0/5/12	1/3/13	5/9/3
(15) SMOTE															2/14/1	2/13/2	2/13/2	12/5/0
(16) gg-SMOTE																1/15/1	1/15/1	12/5/0
(17) ncn-SMOTE																	1/16/0	11/6/0
(18) rng-SMOTE																		13/4/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/5/8	0/8/9	4/5/8	1/3/13	4/6/7	6/2/9	1/12/4	14/1/2	3/6/8	9/2/6	3/5/9	0/5/12	3/6/8	1/1/15	0/2/15	1/1/15	1/3/13	2/11/4
(2) SSM$^-$		3/5/9	2/11/4	0/4/13	2/11/4	8/2/7	5/8/4	15/2/0	1/10/6	9/5/3	1/10/6	2/4/11	1/12/4	0/5/12	0/4/13	0/3/14	0/4/13	6/5/6
(3) EW$^-$			9/5/3	2/7/8	9/5/3	8/3/6	7/10/0	15/1/1	9/5/3	11/3/3	6/8/3	0/12/5	9/3/5	1/6/10	2/6/9	2/5/10	2/5/10	10/5/2
(4) EW$^-$+SSM$^-$				1/6/10	5/8/4	8/4/5	6/7/4	14/1/2	1/11/5	11/2/4	1/9/7	2/5/10	0/14/3	0/5/12	0/4/13	0/4/13	1/3/13	5/7/5
(5) RUS					12/5/0	10/5/2	11/6/0	17/0/0	11/6/0	14/3/0	10/7/0	7/6/4	12/4/1	1/11/5	1/11/5	1/12/4	1/11/5	10/6/1
(6) CNN$^-$						7/3/7	6/6/5	15/1/1	1/14/2	9/5/3	0/14/3	3/5/9	2/9/6	0/6/11	0/5/12	0/3/14	0/5/12	5/8/4
(7) TMK$_1$							9/1/7	16/0/1	5/5/7	14/1/2	4/5/8	4/4/9	5/3/9	1/6/10	3/3/11	3/3/11	2/3/12	9/2/6
(8) TMK$_2$								14/2/1	5/7/5	10/2/5	4/7/6	0/6/11	4/7/6	1/4/12	0/5/12	0/4/13	1/4/12	1/12/4
(9) CNN$^-$+TMK$_1$									1/1/15	0/1/16	0/2/15	0/2/15	1/1/15	0/0/17	0/1/16	0/1/16	0/0/17	1/2/14
(10) CNN$^-$+TMK$_2$										10/6/1	0/16/1	3/4/10	4/10/3	0/5/12	0/4/13	1/4/12	0/5/12	6/6/5
(11) TMK$_1$+CNN$^-$											1/7/9	3/2/12	4/2/11	1/2/14	0/3/14	1/3/13	1/3/13	5/2/10
(12) TMK$_2$+CNN$^-$												3/6/8	7/8/2	0/7/10	0/5/12	1/5/11	0/8/9	6/7/4
(13) EWW$^-$													11/4/2	2/8/7	3/7/7	3/7/7	2/7/8	10/7/0
(14) EWW$^-$+SSM$^-$														0/5/12	0/4/13	0/4/13	1/3/13	5/9/3
(15) SMOTE															3/14/0	2/13/2	2/13/2	12/5/0
(16) gg-SMOTE																1/15/1	1/15/1	12/5/0
(17) ncn-SMOTE																	1/15/1	11/6/0
(18) rng-SMOTE																		12/5/0
(19) b-SMOTE																		

Tabla C.17: *Prueba t−student del clasificador NBC en términos de tvp y tvn*

	tvp																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/9	0/7/10	7/1/9	0/3/14	5/4/8	1/1/15	0/8/9	0/1/16	4/5/8	0/3/14	4/5/8	0/6/11	6/2/9	3/3/11	3/3/11	2/3/12	2/3/12	3/9/5
(2) SSM$^-$		4/1/12	1/4/12	2/0/15	5/2/10	2/1/14	8/1/8	1/0/16	1/1/15	1/1/15	1/1/15	4/1/12	1/4/12	2/1/14	2/2/13	2/1/14	2/0/15	8/0/9
(3) EW$^-$			7/1/9	3/3/11	10/5/2	2/1/14	8/8/1	1/1/15	6/4/7	1/3/13	6/5/6	0/9/8	7/1/9	6/3/8	6/4/7	5/4/8	5/5/7	9/7/1
(4) EW$^-$+SSM$^-$				6/0/11	10/2/5	2/2/13	9/1/7	1/0/16	3/5/9	1/1/15	5/6/6	8/2/7	0/10/7	5/2/10	5/2/10	5/1/11	5/1/11	10/1/6
(5) RUS					14/2/1	4/3/10	13/2/2	1/2/14	8/2/7	4/3/10	9/5/3	12/2/3	10/0/7	6/9/2	6/9/2	6/10/1	8/6/3	13/4/0
(6) CNN$^-$						1/2/14	6/3/8	0/1/16	0/6/11	0/1/16	0/7/10	2/3/12	5/2/10	2/2/13	2/2/13	1/3/13	1/3/13	6/3/8
(7) TMK$_1$							14/1/2	1/1/15	11/2/4	5/5/7	12/3/2	14/1/2	12/3/2	12/3/2	11/4/2	11/3/3	11/3/3	14/2/1
(8) TMK$_2$								1/1/15	4/5/8	1/3/13	5/4/8	1/7/9	7/1/9	4/3/10	4/3/10	3/3/11	3/3/11	7/7/3
(9) CNN$^-$+TMK$_1$									14/2/1	15/2/0	15/1/1	15/1/1	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	16/1/0
(10) CNN$^-$+TMK$_2$										2/2/13	9/8/0	7/2/8	9/4/4	6/5/6	6/4/7	6/5/6	5/6/6	9/4/4
(11) TMK$_1$+CNN$^-$											15/1/1	14/2/1	15/1/1	11/4/2	10/5/2	10/4/3	10/5/2	14/3/0
(12) TMK$_2$+CNN$^-$												5/4/8	5/4/8	4/5/8	5/4/8	4/5/8	3/6/8	9/3/5
(13) EWW$^-$													7/1/9	6/4/7	6/3/8	6/2/9	6/3/8	10/6/1
(14) EWW$^-$+SSM$^-$														6/1/10	6/1/10	6/1/10	6/1/10	10/1/6
(15) SMOTE															5/12/0	5/9/3	2/12/3	10/3/4
(16) gg-SMOTE																5/9/3	1/13/3	10/4/3
(17) ncn-SMOTE																	2/12/3	11/4/2
(18) rng-SMOTE																		10/5/2
(19) b-SMOTE																		

	tvn																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	10/0/7	11/5/1	10/0/7	15/1/1	9/1/7	16/0/1	10/6/1	16/1/0	11/0/6	15/1/1	11/0/6	12/3/2	11/0/6	11/1/5	12/0/5	12/2/3	13/0/4	9/5/3
(2) SSM$^-$		10/1/6	13/0/4	12/0/5	9/2/6	12/1/4	7/1/9	16/0/1	14/2/1	14/1/2	14/2/1	11/1/5	12/2/3	11/1/5	10/2/5	11/1/5	11/1/5	7/1/9
(3) EW$^-$			9/1/7	13/3/1	5/1/11	15/0/2	2/6/9	15/1/1	9/1/7	14/1/2	8/1/8	8/8/1	11/0/6	8/2/7	8/2/7	8/4/5	9/2/6	4/5/8
(4) EW$^-$+SSM$^-$				9/0/8	8/1/8	13/0/4	7/0/10	16/0/1	11/2/4	15/0/2	11/0/6	8/1/8	9/3/5	10/0/7	10/0/7	10/0/7	10/0/7	7/0/10
(5) RUS					4/0/13	11/2/4	1/2/14	15/1/1	10/0/7	12/1/4	6/2/9	2/2/13	8/1/8	3/4/10	3/5/9	2/7/8	3/4/10	2/3/12
(6) CNN$^-$						13/0/4	10/1/6	17/0/0	13/2/2	15/1/1	11/4/2	12/1/4	10/0/7	13/0/4	13/0/4	12/1/4	13/0/4	7/3/7
(7) TMK$_1$							2/0/15	16/0/1	6/0/11	9/3/5	5/1/11	2/0/15	5/0/12	3/0/14	3/0/14	3/1/13	3/0/14	2/1/14
(8) TMK$_2$								15/1/1	11/0/6	14/1/2	10/1/6	11/3/3	11/0/6	10/1/6	11/0/6	11/2/4	12/0/5	6/5/6
(9) CNN$^-$+TMK$_1$									1/1/15	0/0/17	1/0/16	1/1/15	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16
(10) CNN$^-$+TMK$_2$										13/1/3	1/4/12	8/0/9	6/1/10	7/2/8	9/0/8	8/0/9	8/1/8	6/0/11
(11) TMK$_1$+CNN$^-$											2/1/14	2/2/13	2/0/15	1/0/16	1/2/14	1/2/14	1/3/13	1/3/13
(12) TMK$_2$+CNN$^-$												8/1/8	8/1/8	9/2/6	9/2/6	9/2/6	9/2/6	6/0/11
(13) EWW$^-$													9/1/7	7/4/6	7/3/7	7/5/5	7/3/7	4/3/10
(14) EWW$^-$+SSM$^-$														10/0/7	10/0/7	8/2/7	10/0/7	7/0/10
(15) SMOTE															6/6/5	7/5/5	7/8/2	8/0/9
(16) gg-SMOTE																4/8/5	7/6/4	6/2/9
(17) ncn-SMOTE																	5/8/4	4/3/10
(18) rng-SMOTE																		5/3/9
(19) b-SMOTE																		

Tabla C.18: *Prueba $t-$student del clasificador NBC en términos de Acc y f_1*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/5/5	8/8/1	10/0/7	12/4/1	7/3/7	16/0/1	7/9/1	16/1/0	11/0/6	15/1/1	10/1/6	8/7/2	11/0/6	9/3/5	9/3/5	11/3/3	11/3/3	10/6/1
(2) SSM$^-$		7/5/5	12/1/4	8/4/5	8/4/5	11/1/5	6/4/7	15/1/1	14/0/3	13/1/3	11/3/3	9/3/5	12/2/3	8/4/5	8/3/6	9/2/6	8/4/5	7/3/7
(3) EW$^-$			10/0/7	12/4/1	5/3/9	15/0/2	2/9/6	15/1/1	9/1/7	13/2/2	7/3/7	7/9/1	10/1/6	9/2/6	8/2/7	10/2/5	9/4/4	6/5/6
(4) EW$^-$+SSM$^-$				9/2/6	9/0/8	12/1/4	7/0/10	16/0/1	11/2/4	15/0/2	11/1/5	7/3/7	8/4/5	9/2/6	9/2/6	9/2/6	9/2/6	7/2/8
(5) RUS					4/2/11	9/5/3	1/5/11	15/1/1	9/1/7	12/1/4	6/3/8	1/6/10	8/2/7	3/5/9	2/6/9	1/8/8	2/6/9	3/5/9
(6) CNN$^-$						13/0/4	9/2/6	16/1/0	12/3/2	15/0/2	8/6/3	11/2/4	9/1/7	10/3/4	10/3/4	10/3/4	10/3/4	9/1/7
(7) TMK$_1$							2/0/15	16/0/1	6/0/11	9/4/4	6/0/11	2/1/14	5/0/12	3/1/13	3/1/13	3/1/13	3/1/13	2/2/13
(8) TMK$_2$								15/1/1	10/1/6	14/1/2	9/2/6	8/6/3	10/1/6	8/4/5	8/4/5	10/4/3	10/4/3	7/7/3
(9) CNN$^-$+TMK$_1$									2/1/14	0/0/17	1/1/15	1/1/15	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16
(10) CNN$^-$+TMK$_2$										13/0/4	1/4/12	7/1/9	6/2/9	6/3/8	7/3/7	6/3/8	8/2/7	6/2/9
(11) TMK$_1$+CNN$^-$											3/1/13	2/3/12	2/1/14	1/0/16	1/2/14	1/2/14	1/3/13	1/3/13
(12) TMK$_2$+CNN$^-$												8/2/7	6/3/8	8/3/6	8/3/6	8/3/6	8/3/6	7/1/9
(13) EWW$^-$													10/0/7	8/2/7	7/3/7	8/4/5	8/3/6	6/4/7
(14) EWW$^-$+SSM$^-$														8/2/7	9/1/7	8/3/6	8/2/7	7/1/9
(15) SMOTE															4/8/5	8/7/2	7/9/1	7/2/8
(16) gg-SMOTE																4/10/3	5/9/3	7/2/8
(17) ncn-SMOTE																	5/7/5	4/4/9
(18) rng-SMOTE																		5/4/8
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/0/10	2/8/7	4/3/10	6/4/7	5/3/9	8/5/4	2/9/6	12/3/2	5/1/11	9/4/4	5/1/11	3/7/7	4/3/10	4/5/8	3/5/9	5/4/8	5/4/8	8/5/4
(2) SSM$^-$		4/2/11	4/4/9	4/2/11	7/7/3	7/2/8	6/3/8	12/3/2	4/5/8	11/0/6	4/5/8	4/2/11	5/3/9	4/3/10	4/3/10	4/3/10	4/3/10	9/1/7
(3) EW$^-$			5/6/6	7/9/1	9/3/5	11/4/2	8/8/1	14/3/0	6/5/6	11/4/2	6/5/6	3/12/2	6/4/7	6/6/5	5/8/4	8/6/3	7/8/2	11/5/1
(4) EW$^-$+SSM$^-$				7/5/5	11/2/4	9/4/4	8/4/5	15/2/0	8/4/5	12/2/3	7/5/5	5/7/5	3/8/6	7/5/5	7/5/5	8/5/4	7/5/5	12/1/4
(5) RUS					10/3/4	9/4/4	5/7/5	14/3/0	5/8/4	11/4/2	5/8/4	1/9/7	6/4/7	2/8/7	3/7/7	2/13/2	1/10/6	10/2/5
(6) CNN$^-$						8/2/7	6/4/7	13/2/2	5/4/8	10/2/5	3/7/7	5/3/9	2/4/11	5/3/9	4/3/10	5/2/10	6/2/9	8/3/6
(7) TMK$_1$							2/6/9	14/1/2	5/3/9	6/7/4	5/3/9	2/4/11	3/5/9	4/2/11	4/3/10	4/3/10	4/4/9	8/5/4
(8) TMK$_2$								13/3/1	6/1/10	11/4/2	6/2/9	2/8/7	5/4/8	3/8/6	4/5/8	5/6/6	5/5/7	9/5/3
(9) CNN$^-$+TMK$_1$									1/2/14	1/1/15	1/3/13	0/3/14	0/2/15	0/2/15	0/3/14	1/2/14	0/2/15	4/2/11
(10) CNN$^-$+TMK$_2$										11/3/3	1/13/3	7/4/6	4/5/8	5/6/6	5/6/6	5/5/7	5/6/6	10/3/4
(11) TMK$_1$+CNN$^-$											3/3/11	3/3/11	1/5/11	1/3/13	1/4/12	1/4/12	1/4/12	6/4/7
(12) TMK$_2$+CNN$^-$												7/4/6	4/5/8	5/6/6	5/6/6	5/5/7	5/6/6	10/3/4
(13) EWW$^-$													5/6/6	7/5/5	6/4/7	9/4/4	7/6/4	12/2/3
(14) EWW$^-$+SSM$^-$														7/5/5	6/6/5	7/4/6	7/5/5	12/2/3
(15) SMOTE															4/9/4	7/8/2	5/11/1	11/2/4
(16) gg-SMOTE																5/11/1	4/11/2	12/2/3
(17) ncn-SMOTE																	1/12/4	9/3/5
(18) rng-SMOTE																		11/2/4
(19) b-SMOTE																		

Tabla C.19: *Prueba t−student del clasificador NBC en términos de $IBA_{0.1}$ y Gm*

	$IBA_{0.1}$																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	10/1/6	0/10/7	8/3/6	2/7/8	7/3/7	8/4/5	0/11/6	14/1/2	6/4/7	10/2/5	6/3/8	0/9/8	7/4/6	4/3/10	3/4/10	3/4/10	3/5/9	3/10/4
(2) SSM^-		0/3/14	3/4/10	0/3/14	5/9/3	7/1/9	3/3/11	12/3/2	3/4/10	8/1/8	3/3/11	0/3/14	4/3/10	0/5/12	0/5/12	0/5/12	0/5/12	6/2/9
(3) EW^-			9/4/4	4/9/4	11/4/2	11/3/3	8/9/0	15/2/0	9/4/4	11/3/3	8/5/4	2/12/3	9/4/4	6/7/4	6/7/4	6/7/4	6/7/4	8/8/1
(4) EW^-+SSM^-				2/6/9	9/5/3	8/3/6	5/4/8	15/2/0	4/7/6	9/3/5	4/7/6	3/5/9	2/10/5	3/6/8	3/6/8	4/5/8	3/6/8	8/2/7
(5) RUS					13/2/2	9/4/4	7/8/2	15/2/0	9/5/3	11/4/2	8/7/2	2/10/5	9/6/2	4/10/3	4/11/2	3/12/2	3/11/3	9/6/2
(6) CNN^-						7/1/9	6/2/9	14/1/2	3/4/10	8/1/8	2/6/9	2/3/12	3/6/8	2/3/12	3/2/12	3/2/12	3/2/12	5/4/8
(7) TMK_1							4/2/11	15/0/2	6/4/7	7/5/5	6/3/8	2/4/11	4/5/8	4/4/9	4/3/10	4/3/10	4/3/10	6/4/7
(8) TMK_2								15/1/1	8/4/5	11/2/4	8/2/7	0/9/8	8/4/5	4/6/7	4/6/7	3/8/6	3/7/7	6/9/2
(9) CNN^-+TMK_1									1/1/15	1/0/16	0/2/15	0/1/16	0/2/15	0/1/16	0/1/16	0/1/16	0/1/16	2/1/14
(10) CNN^-+TMK_2										9/3/5	2/11/4	4/4/9	5/7/5	3/5/9	3/5/9	5/3/9	3/5/9	7/4/6
(11) TMK_1+CNN^-											5/3/9	3/3/11	5/3/9	1/4/12	1/5/11	1/4/12	1/5/11	5/3/9
(12) TMK_2+CNN^-												4/4/9	5/7/5	2/7/8	3/6/8	4/5/8	4/5/8	8/3/6
(13) EWW^-													9/5/3	6/7/4	6/8/3	8/5/4	6/8/3	11/5/1
(14) EWW^-+SSM^-														3/6/8	3/6/8	4/5/8	3/6/8	7/3/7
(15) SMOTE															4/13/0	4/11/2	3/13/1	9/5/3
(16) gg-SMOTE																1/16/0	1/15/1	9/5/3
(17) ncn-SMOTE																	1/15/1	9/5/3
(18) rng-SMOTE																		9/5/3
(19) b-SMOTE																		

	Gm																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	10/0/7	0/10/7	7/3/7	4/5/8	7/3/7	10/2/5	1/10/6	13/2/2	6/3/8	10/2/5	7/2/8	0/10/7	8/2/7	4/3/10	3/5/9	4/3/10	4/4/9	4/9/4
(2) SSM^-		1/2/14	3/4/10	1/3/13	5/9/3	8/0/9	4/2/11	12/4/1	3/5/9	8/1/8	3/4/10	2/1/14	4/3/10	0/5/12	0/5/12	0/5/12	0/5/12	6/2/9
(3) EW^-			10/2/5	6/8/3	11/4/2	11/4/2	6/11/0	15/2/0	9/4/4	11/3/3	8/5/4	2/13/2	9/4/4	6/7/4	6/9/2	8/6/3	5/10/2	9/7/1
(4) EW^-+SSM^-				3/5/9	10/4/3	9/2/6	6/2/9	14/2/1	4/8/5	9/4/4	4/8/5	2/5/10	3/9/5	3/6/8	3/6/8	4/4/9	3/6/8	8/3/6
(5) RUS					13/2/2	9/4/4	7/6/4	15/2/0	7/8/2	11/4/2	7/8/2	3/8/6	7/7/3	3/11/3	3/12/2	4/11/2	3/12/2	9/5/3
(6) CNN^-						7/1/9	5/3/9	14/1/2	2/6/9	8/1/8	2/7/8	4/2/11	3/6/8	3/2/12	3/2/12	3/2/12	3/2/12	6/3/8
(7) TMK_1							2/4/11	15/0/2	6/3/8	7/5/5	6/3/8	2/4/11	4/4/9	4/4/9	4/3/10	4/3/10	4/3/10	5/4/8
(8) TMK_2								14/2/1	8/3/6	11/3/3	8/2/7	0/10/7	9/2/6	4/6/7	4/6/7	4/6/7	3/7/7	6/9/2
(9) CNN^-+TMK_1									1/0/16	1/0/16	0/0/17	0/2/15	0/3/14	0/2/15	0/1/16	0/1/16	0/1/16	2/1/14
(10) CNN^-+TMK_2										9/5/3	1/12/4	5/3/9	5/7/5	3/5/9	3/5/9	4/4/9	3/5/9	8/2/7
(11) TMK_1+CNN^-											4/4/9	3/3/11	2/6/9	1/4/12	1/5/11	1/4/12	1/5/11	4/3/10
(12) TMK_2+CNN^-												4/4/9	5/7/5	2/7/8	3/6/8	4/5/8	3/6/8	9/2/6
(13) EWW^-													9/4/4	7/7/3	7/6/4	9/4/4	6/8/3	8/8/1
(14) EWW^-+SSM^-														3/6/8	3/5/9	4/5/8	3/5/9	8/3/6
(15) SMOTE															4/11/2	4/11/2	3/12/2	9/5/3
(16) gg-SMOTE																2/15/0	2/13/2	10/4/3
(17) ncn-SMOTE																	1/14/2	9/4/4
(18) rng-SMOTE																		9/5/3
(19) b-SMOTE																		

Tabla C.20: *Prueba $t-$student del clasificador RBF en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/16	0/5/12	0/0/17	0/0/17	0/2/15	0/0/17	0/6/11	0/0/17	0/0/17	0/0/17	0/0/17	0/4/13	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	2/7/8
(2) SSM$^-$		10/4/3	0/3/14	3/1/13	3/6/8	5/0/12	15/2/0	1/0/16	0/1/16	1/0/16	0/1/16	9/2/6	0/2/15	2/2/13	2/2/13	2/2/13	2/2/13	13/3/1
(3) EW$^-$			0/1/16	1/2/14	5/3/9	1/0/16	10/6/1	1/0/16	0/1/16	1/0/16	0/2/15	0/8/9	0/0/17	0/2/15	1/0/16	0/1/16	0/2/15	9/7/1
(4) EW$^-$+SSM$^-$				8/1/8	9/4/4	6/1/10	17/0/0	1/0/16	3/5/9	1/0/16	7/4/6	16/0/1	0/8/9	5/5/7	6/3/8	6/4/7	5/4/8	16/1/0
(5) RUS					14/2/1	6/1/10	16/0/1	1/0/16	8/2/7	1/3/13	9/5/3	12/4/1	7/2/8	6/6/5	4/9/4	4/9/4	4/8/5	17/0/0
(6) CNN$^-$						4/0/13	14/2/1	0/0/17	0/5/12	0/0/17	0/6/11	9/1/7	4/2/11	2/2/13	1/3/13	1/2/14	1/3/13	13/0/4
(7) TMK$_1$							16/0/1	0/0/17	10/1/6	1/3/13	11/1/5	15/1/1	8/3/6	9/2/6	10/1/6	9/2/6	9/2/6	17/0/0
(8) TMK$_2$								1/0/16	0/1/16	1/0/16	0/1/16	1/4/12	0/0/17	0/1/16	1/0/16	1/0/16	1/0/16	5/9/3
(9) CNN$^-$+TMK$_1$									16/0/1	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	16/0/1	16/0/1	15/1/1	17/0/0
(10) CNN$^-$+TMK$_2$										1/0/16	10/7/0	16/0/1	6/3/8	7/5/5	6/4/7	6/4/7	6/4/7	16/0/1
(11) TMK$_1$+CNN$^-$											16/0/1	16/0/1	15/1/1	13/3/1	11/5/1	11/4/2	12/4/1	17/0/0
(12) TMK$_2$+CNN$^-$												13/3/1	5/3/9	5/2/10	4/3/10	4/4/9	5/1/11	16/0/1
(13) EWW$^-$													0/1/16	1/1/15	1/2/14	1/1/15	1/2/14	11/6/0
(14) EWW$^-$+SSM$^-$														7/2/8	7/2/8	6/4/7	6/4/7	17/0/0
(15) SMOTE															2/12/3	1/12/4	1/10/6	17/0/0
(16) gg-SMOTE																0/14/3	2/12/3	17/0/0
(17) ncn-SMOTE																	3/12/2	17/0/0
(18) rng-SMOTE																		17/0/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/1/0	12/5/0	16/1/0	17/0/0	16/1/0	17/0/0	10/7/0	17/0/0	17/0/0	17/0/0	17/0/0	13/4/0	16/1/0	17/0/0	17/0/0	17/0/0	17/0/0	10/7/0
(2) SSM$^-$		3/0/14	9/6/2	12/0/5	5/4/8	10/2/5	0/1/16	16/0/1	14/1/2	14/1/2	12/3/2	5/2/10	11/4/2	11/1/5	12/0/5	12/0/5	12/0/5	2/3/12
(3) EW$^-$			16/0/1	15/1/1	9/4/4	16/0/1	1/5/11	16/0/1	16/1/0	16/0/1	16/1/0	11/6/0	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	2/7/8
(4) EW$^-$+SSM$^-$				6/0/11	4/1/12	7/3/7	0/1/16	16/0/1	8/3/6	13/2/2	4/2/11	1/0/16	8/6/3	6/2/9	6/3/8	7/2/8	6/2/9	0/1/16
(5) RUS					2/1/14	9/1/7	1/0/16	16/0/1	8/2/7	12/4/1	5/3/9	1/1/15	11/0/6	5/8/4	6/8/3	5/9/3	5/7/5	0/1/16
(6) CNN$^-$						12/0/5	2/1/14	17/0/0	13/4/0	17/0/0	10/5/2	8/1/8	12/1/4	11/3/3	12/2/3	12/4/1	12/3/2	5/0/12
(7) TMK$_1$							1/0/16	17/0/0	8/1/8	14/2/1	7/0/10	1/0/16	9/1/7	8/1/8	8/1/8	8/1/8	7/2/8	0/1/16
(8) TMK$_2$								16/0/1	17/0/0	16/0/1	17/0/0	12/4/1	16/1/0	16/0/1	16/0/1	16/0/1	16/0/1	6/8/3
(9) CNN$^-$+TMK$_1$									1/0/16	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										15/0/2	0/6/11	1/1/15	8/3/6	6/3/8	5/5/7	5/7/5	5/6/6	1/0/16
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	3/1/13	1/3/13	1/3/13	2/2/13	2/2/13	0/0/17
(12) TMK$_2$+CNN$^-$												2/2/13	12/1/4	8/3/6	8/3/6	8/4/5	8/4/5	1/1/15
(13) EWW$^-$													16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	1/7/9
(14) EWW$^-$+SSM$^-$														6/2/9	6/1/10	6/3/8	6/2/9	0/1/16
(15) SMOTE															3/7/7	6/9/2	5/8/4	0/0/17
(16) gg-SMOTE																5/11/1	5/9/3	0/0/17
(17) ncn-SMOTE																	1/13/3	0/0/17
(18) rng-SMOTE																		0/0/17
(19) b-SMOTE																		

Tabla C.21: *Prueba t−student del clasificador RBF en términos de Acc y* f_1

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/0/1	6/9/2	16/0/1	16/1/0	12/4/1	17/0/0	2/14/1	17/0/0	16/0/1	16/1/0	15/1/1	9/7/1	16/0/1	16/1/0	16/1/0	15/2/0	15/2/0	9/8/0
(2) SSM^-		1/2/14	9/6/2	11/1/5	5/4/8	9/2/6	1/2/14	16/0/1	12/3/2	14/0/3	10/4/3	2/4/11	11/3/3	11/1/5	11/1/5	11/1/5	10/2/5	3/3/11
(3) EW^-			16/0/1	14/3/0	11/2/4	16/0/1	2/11/4	16/0/1	16/0/1	15/1/1	14/2/1	6/11/0	16/0/1	15/2/0	15/2/0	15/1/1	15/1/1	5/10/2
(4) EW^-+SSM^-				6/0/11	4/1/12	8/1/8	1/0/16	16/0/1	8/3/6	13/1/3	4/2/11	1/0/16	7/7/3	6/2/9	6/2/9	6/2/9	6/2/9	1/0/16
(5) RUS					2/1/14	8/2/7	1/0/16	16/0/1	8/3/6	12/4/1	6/3/8	1/2/14	11/0/6	5/8/4	5/9/3	4/10/3	5/7/5	0/5/12
(6) CNN^-						11/1/5	2/3/12	17/0/0	10/7/0	16/1/0	9/6/2	4/5/8	12/1/4	11/3/3	12/2/3	12/3/2	11/3/3	6/1/10
(7) TMK_1							1/0/16	17/0/0	9/1/7	14/2/1	7/1/9	1/0/16	9/0/8	8/1/8	8/1/8	8/1/8	8/1/8	0/2/15
(8) TMK_2								16/0/1	16/0/1	15/1/1	16/0/1	7/8/2	16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	8/8/1
(9) CNN^-+TMK_1									1/0/16	0/0/17	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(10) CNN^-+TMK_2										15/0/2	0/6/11	1/1/15	8/3/6	6/3/8	5/5/7	5/7/5	5/5/7	3/0/14
(11) TMK_1+CNN^-											1/0/16	1/1/15	3/2/12	1/4/12	1/4/12	2/3/12	2/3/12	0/2/15
(12) TMK_2+CNN^-												2/2/13	12/0/5	8/3/6	8/3/6	8/4/5	8/4/5	3/1/13
(13) EWW^-													16/0/1	15/1/1	15/1/1	15/1/1	15/1/1	4/9/4
(14) EWW^-+SSM^-														6/2/9	7/1/9	6/2/9	6/2/9	1/0/16
(15) SMOTE															3/8/6	6/9/2	5/8/4	1/2/14
(16) gg-SMOTE																5/11/1	5/10/2	1/3/13
(17) ncn-SMOTE																	1/13/3	1/4/12
(18) rng-SMOTE																		1/4/12
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/2/13	0/7/10	3/1/13	2/2/13	3/3/11	1/5/11	0/7/10	8/3/6	3/2/12	3/5/9	3/2/12	0/6/11	2/2/13	1/3/13	2/2/13	2/2/13	2/2/13	1/10/6
(2) SSM^-		5/3/9	2/6/9	5/2/10	7/7/3	7/2/8	8/5/4	14/2/1	6/5/6	11/1/5	5/5/7	4/2/11	2/5/10	4/3/10	4/4/9	4/4/9	4/3/10	8/7/2
(3) EW^-			6/6/5	2/8/7	8/4/5	7/5/5	7/10/0	13/1/3	8/3/6	9/3/5	6/5/6	0/14/3	6/6/5	2/8/7	3/7/7	3/9/5	2/8/7	11/5/1
(4) EW^-+SSM^-				5/5/7	8/5/4	8/2/7	10/3/4	16/1/0	7/7/3	11/2/4	6/6/5	4/6/7	3/11/3	5/6/6	5/6/6	5/5/7	5/5/7	11/3/3
(5) RUS					8/5/4	9/3/5	10/5/2	16/1/0	6/8/3	11/6/0	3/11/3	4/8/5	6/6/5	4/9/4	3/11/3	4/10/3	2/10/5	11/4/2
(6) CNN^-						6/4/7	7/6/4	13/2/2	2/12/3	11/2/4	1/11/5	4/5/8	3/6/8	3/8/6	3/8/6	3/7/7	3/8/6	7/5/5
(7) TMK_1							8/6/3	16/0/1	6/3/8	9/5/3	6/3/8	5/4/8	5/4/8	4/3/10	5/2/10	4/4/9	3/5/9	12/4/1
(8) TMK_2								10/3/4	4/5/8	8/4/5	4/4/9	0/8/9	4/3/10	1/6/10	2/5/10	2/3/12	2/4/11	7/8/2
(9) CNN^-+TMK_1									1/1/15	1/0/16	0/3/14	2/1/14	0/1/16	0/3/14	0/2/15	0/2/15	0/1/16	5/4/8
(10) CNN^-+TMK_2										11/5/1	0/13/4	5/3/9	3/6/8	3/7/7	3/8/6	3/10/4	3/8/6	10/4/3
(11) TMK_1+CNN^-											1/4/12	4/4/9	2/4/11	2/4/11	0/6/11	1/5/11	1/5/11	9/4/4
(12) TMK_2+CNN^-												5/6/6	6/5/6	4/8/5	3/9/5	4/10/3	4/10/3	10/4/3
(13) EWW^-													7/5/5	2/10/5	4/8/5	4/8/5	3/9/5	11/6/0
(14) EWW^-+SSM^-														5/6/6	4/6/7	6/4/7	5/5/7	12/3/2
(15) SMOTE															3/11/3	6/10/1	4/9/4	13/3/1
(16) gg-SMOTE																4/13/0	4/10/3	12/3/2
(17) ncn-SMOTE																	1/12/4	12/3/2
(18) rng-SMOTE																		12/3/2
(19) b-SMOTE																		

Tabla C.22: *Prueba $t-$student del clasificador RBF en términos de $IBA_{0.1}$ y Gm*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/15	0/6/11	1/2/14	0/0/17	0/3/14	2/2/13	0/7/10	9/3/5	1/0/16	5/2/10	0/1/16	0/5/12	1/1/15	0/0/17	0/0/17	0/0/17	0/0/17	2/8/7
(2) SSM$^-$		7/2/8	2/4/11	2/3/12	5/9/3	9/1/7	10/5/2	14/2/1	5/6/6	8/5/4	4/3/10	6/1/10	3/5/9	2/3/12	2/2/13	2/3/12	2/2/13	11/4/2
(3) EW$^-$			5/5/7	0/6/11	5/4/8	6/1/10	9/8/0	12/1/4	5/3/9	7/3/7	3/5/9	0/10/7	4/5/8	1/5/11	0/6/11	0/6/11	1/5/11	9/7/1
(4) EW$^-$+SSM$^-$				2/6/9	7/7/3	9/3/5	13/1/3	15/2/0	7/6/4	11/2/4	5/7/5	7/4/6	3/10/4	4/4/9	3/5/9	4/4/9	4/3/10	14/2/1
(5) RUS					13/3/1	12/3/2	15/2/0	16/1/0	12/4/1	12/4/1	10/6/1	8/8/1	9/6/2	4/9/4	4/11/2	4/11/2	2/10/5	16/1/0
(6) CNN$^-$						9/1/7	12/3/2	14/2/1	0/12/5	7/4/6	0/10/7	8/3/6	4/3/10	1/4/12	1/3/13	1/5/11	1/3/13	12/2/3
(7) TMK$_1$							11/0/6	15/0/2	5/2/10	9/3/5	6/1/10	8/3/6	4/3/10	2/3/12	2/3/12	1/4/12	1/4/12	12/2/3
(8) TMK$_2$								12/1/4	2/3/12	6/3/8	2/2/13	0/5/12	3/0/14	0/3/14	0/2/15	0/1/16	0/2/15	7/7/3
(9) CNN$^-$+TMK$_1$									0/2/15	1/1/15	0/2/15	4/1/12	0/2/15	0/1/16	0/0/17	0/0/17	0/0/17	5/1/11
(10) CNN$^-$+TMK$_2$										11/0/6	0/12/5	8/3/6	6/2/9	1/6/10	1/5/11	1/7/9	1/6/10	13/2/2
(11) TMK$_1$+CNN$^-$											3/3/11	7/2/8	4/2/11	2/2/13	0/5/12	1/4/12	1/3/13	9/3/5
(12) TMK$_2$+CNN$^-$												8/3/6	5/6/6	2/5/10	2/5/10	2/7/8	2/6/9	13/4/0
(13) EWW$^-$													6/4/7	4/3/10	2/6/9	2/6/9	2/6/9	11/6/0
(14) EWW$^-$+SSM$^-$														4/4/9	4/3/10	4/5/8	4/2/11	14/1/2
(15) SMOTE															2/14/1	4/12/1	4/9/4	16/1/0
(16) gg-SMOTE																2/14/1	2/13/2	16/1/0
(17) ncn-SMOTE																	1/13/3	17/0/0
(18) rng-SMOTE																		16/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/2/14	0/7/10	1/2/14	0/1/16	1/2/14	4/1/12	0/7/10	10/2/5	1/0/16	5/3/9	1/0/16	0/5/12	2/0/15	0/0/17	0/0/17	0/0/17	0/0/17	2/8/7
(2) SSM$^-$		7/2/8	2/7/8	2/3/12	6/8/3	9/2/6	10/5/2	15/1/1	7/4/6	9/4/4	5/3/9	6/1/10	4/5/8	3/2/12	2/2/13	3/2/12	3/2/12	10/5/2
(3) EW$^-$			5/5/7	0/7/10	5/4/8	6/2/9	8/9/0	12/2/3	5/3/9	8/2/7	5/3/9	0/10/7	6/3/8	4/2/11	1/5/11	1/5/11	2/4/11	9/8/0
(4) EW$^-$+SSM$^-$				2/4/11	7/7/3	10/2/5	12/2/3	15/1/1	7/5/5	12/1/4	5/6/6	7/4/6	3/10/4	4/4/9	4/3/10	4/4/9	4/2/11	12/3/2
(5) RUS					13/3/1	12/3/2	15/2/0	17/0/0	12/4/1	12/4/1	10/6/1	8/8/1	10/4/3	4/9/4	3/12/2	4/11/2	2/10/5	15/2/0
(6) CNN$^-$						10/0/7	11/4/2	14/1/2	1/12/4	8/4/5	0/10/7	8/3/6	4/3/10	2/4/11	1/4/12	2/4/11	2/2/13	11/4/2
(7) TMK$_1$							11/0/6	14/1/2	5/1/11	9/3/5	6/0/11	8/3/6	5/2/10	2/3/12	2/3/12	2/3/12	1/4/12	12/1/4
(8) TMK$_2$								12/2/3	2/5/10	6/3/8	2/3/12	0/6/11	3/3/11	0/4/13	0/4/13	0/2/15	0/3/14	7/8/2
(9) CNN$^-$+TMK$_1$									0/2/15	1/0/16	1/1/15	3/2/12	0/1/16	0/1/16	0/1/16	0/1/16	0/0/17	5/1/11
(10) CNN$^-$+TMK$_2$										11/1/5	0/12/5	8/2/7	6/2/9	1/6/10	1/4/12	1/5/11	1/5/11	12/3/2
(11) TMK$_1$+CNN$^-$											2/4/11	6/3/8	4/2/11	1/3/13	0/4/13	1/4/12	1/3/13	9/3/5
(12) TMK$_2$+CNN$^-$												8/3/6	6/5/6	2/5/10	2/5/10	3/6/8	2/6/9	13/3/1
(13) EWW$^-$													7/3/7	4/4/9	2/6/9	2/7/8	2/5/10	11/6/0
(14) EWW$^-$+SSM$^-$														4/4/9	4/3/10	4/4/9	4/2/11	12/3/2
(15) SMOTE															1/14/2	4/11/2	3/10/4	16/1/0
(16) gg-SMOTE																2/15/0	2/13/2	16/1/0
(17) ncn-SMOTE																	1/12/4	16/1/0
(18) rng-SMOTE																		16/1/0
(19) b-SMOTE																		

Tabla C.23: *Prueba t−student del clasificador SVM en términos de tvp y tvn*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/3/14	0/4/13	0/1/16	0/0/17	1/2/14	0/2/15	6/5/6	0/0/17	0/1/16	0/0/17	0/1/16	0/4/13	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	3/3/11
(2) SSM$^-$		12/3/2	0/5/12	5/1/11	6/4/7	7/1/9	15/2/0	1/0/16	0/2/15	1/0/16	0/2/15	10/2/5	0/1/16	5/1/11	5/1/11	5/1/11	5/1/11	13/2/2
(3) EW$^-$			0/1/16	1/1/15	2/3/12	2/1/14	10/2/5	1/0/16	0/1/16	1/0/16	0/2/15	0/6/11	0/0/17	1/1/15	1/1/15	1/1/15	1/1/15	6/5/6
(4) EW$^-$+SSM$^-$				10/0/7	10/5/2	10/2/5	16/1/0	1/1/15	2/7/8	1/1/15	8/6/3	16/1/0	0/8/9	10/0/7	10/0/7	10/0/7	10/0/7	16/1/0
(5) RUS					12/0/5	8/0/9	15/0/2	1/0/16	5/1/11	2/0/15	8/0/9	15/1/1	7/0/10	5/9/3	5/9/3	5/11/1	5/9/3	14/1/2
(6) CNN$^-$						6/1/10	14/2/1	0/1/16	0/4/13	0/1/16	0/5/12	10/3/4	2/3/12	5/1/11	5/1/11	5/1/11	5/1/11	11/3/3
(7) TMK$_1$							14/2/1	0/2/15	6/1/10	0/2/15	8/1/8	14/1/2	5/1/11	9/0/8	9/0/8	9/1/7	9/0/8	13/2/2
(8) TMK$_2$								1/0/16	0/1/16	1/0/16	0/1/16	5/2/10	0/0/17	2/0/15	2/0/15	2/0/15	2/0/15	5/6/6
(9) CNN$^-$+TMK$_1$									13/3/1	7/10/0	15/1/1	16/0/1	15/1/1	16/0/1	16/0/1	15/1/1	15/1/1	17/0/0
(10) CNN$^-$+TMK$_2$										1/3/13	13/4/0	16/1/0	6/7/4	12/2/3	12/2/3	11/3/3	12/2/3	16/1/0
(11) TMK$_1$+CNN$^-$											15/1/1	16/0/1	15/1/1	15/0/2	15/0/2	15/0/2	15/0/2	17/0/0
(12) TMK$_2$+CNN$^-$												14/3/0	3/4/10	8/3/6	8/3/6	8/3/6	8/3/6	14/2/1
(13) EWW$^-$													0/0/17	1/2/14	1/2/14	1/1/15	1/1/15	8/3/6
(14) EWW$^-$+SSM$^-$														10/2/5	10/2/5	10/2/5	10/2/5	17/0/0
(15) SMOTE															3/13/1	3/12/2	1/13/3	14/1/2
(16) gg-SMOTE																2/13/2	1/13/3	14/1/2
(17) ncn-SMOTE																	1/12/4	14/1/2
(18) rng-SMOTE																		14/1/2
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	16/1/0	13/4/0	17/0/0	17/0/0	14/2/1	14/3/0	5/5/7	17/0/0	16/1/0	17/0/0	16/1/0	13/4/0	17/0/0	17/0/0	17/0/0	17/0/0	17/0/0	9/4/4
(2) SSM$^-$		2/3/12	14/2/1	10/1/6	8/2/7	10/0/7	0/1/16	16/0/1	16/0/1	16/0/1	15/1/1	4/3/10	15/2/0	10/0/7	10/1/6	10/0/7	10/0/7	2/1/14
(3) EW$^-$			17/0/0	15/0/2	13/2/2	13/2/2	2/2/13	16/0/1	16/1/0	16/0/1	15/2/0	9/8/0	17/0/0	14/1/2	15/0/2	13/2/2	14/1/2	4/4/9
(4) EW$^-$+SSM$^-$				6/0/11	4/2/11	6/2/9	0/0/17	15/1/1	11/3/3	15/1/1	7/0/10	0/1/16	11/5/1	5/1/11	5/1/11	5/1/11	5/1/11	0/0/17
(5) RUS					6/0/11	8/1/8	1/0/16	16/0/1	13/0/4	15/1/1	11/1/5	2/0/15	11/0/6	1/5/11	0/6/11	1/5/11	1/5/11	1/2/14
(6) CNN$^-$						10/2/5	1/1/15	17/0/0	16/1/0	17/0/0	13/4/0	3/3/11	13/0/4	11/0/6	11/0/6	11/0/6	11/0/6	3/2/12
(7) TMK$_1$							1/2/14	16/1/0	9/3/5	16/1/0	8/2/7	2/2/13	10/1/6	8/0/9	8/0/9	8/0/9	8/0/9	2/2/13
(8) TMK$_2$								16/0/1	16/1/0	16/0/1	16/1/0	13/2/2	17/0/0	16/0/1	16/0/1	16/0/1	16/0/1	6/8/3
(9) CNN$^-$+TMK$_1$									1/1/15	0/8/9	1/0/16	1/0/16	1/1/15	1/0/16	1/0/16	1/0/16	1/0/16	0/0/17
(10) CNN$^-$+TMK$_2$										15/1/1	0/1/16	0/1/16	4/2/11	3/1/13	4/0/13	3/1/13	3/1/13	0/1/16
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	1/1/15	1/1/15	1/1/15	1/1/15	1/1/15	0/1/16
(12) TMK$_2$+CNN$^-$												0/2/15	10/1/6	5/1/11	5/1/11	4/2/11	5/1/11	2/1/14
(13) EWW$^-$													17/0/0	14/1/2	14/1/2	13/2/2	14/1/2	4/2/11
(14) EWW$^-$+SSM$^-$														5/1/11	5/1/11	5/1/11	5/1/11	0/0/17
(15) SMOTE															5/7/5	5/9/3	3/12/2	1/1/15
(16) gg-SMOTE																3/9/5	3/10/4	1/2/14
(17) ncn-SMOTE																	4/9/4	2/1/14
(18) rng-SMOTE																		1/1/15
(19) b-SMOTE																		

Tabla C.24: *Prueba $t-$student del clasificador SVM en términos de Acc y f_1*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	11/5/1	5/5/7	16/0/1	15/1/1	12/2/3	13/3/1	5/5/7	17/0/0	16/1/0	17/0/0	16/1/0	5/7/5	17/0/0	15/0/2	15/0/2	15/0/2	15/0/2	6/5/6
(2) SSM$^-$		4/1/12	12/3/2	8/4/5	8/2/7	10/1/6	2/4/11	16/0/1	15/1/1	16/0/1	12/4/1	2/4/11	14/1/2	7/1/9	7/2/8	7/1/9	7/1/9	4/4/9
(3) EW$^-$			15/2/0	13/2/2	11/4/2	14/1/2	6/5/6	16/0/1	16/1/0	16/0/1	15/2/0	6/11/0	17/0/0	13/2/2	12/3/2	12/2/3	12/3/2	5/7/5
(4) EW$^-$+SSM$^-$				6/0/11	6/1/10	7/2/8	1/2/14	16/0/1	12/2/3	16/0/1	7/1/9	0/2/15	11/5/1	5/1/11	5/1/11	5/1/11	5/1/11	0/3/14
(5) RUS					6/1/10	7/2/8	2/1/14	16/0/1	13/0/4	15/1/1	10/3/4	2/2/13	11/0/6	0/6/11	0/5/12	1/5/11	0/6/11	3/1/13
(6) CNN$^-$						10/2/5	2/3/12	17/0/0	15/2/0	17/0/0	13/4/0	2/4/11	11/3/3	7/3/7	8/2/7	7/3/7	7/3/7	4/4/9
(7) TMK$_1$							1/4/12	16/1/0	9/3/5	15/2/0	7/3/7	2/1/14	9/2/6	8/1/8	8/1/8	8/0/9	8/1/8	1/3/13
(8) TMK$_2$								16/0/1	16/1/0	16/0/1	16/1/0	7/6/4	17/0/0	14/0/3	14/0/3	14/0/3	14/0/3	7/8/2
(9) CNN$^-$+TMK$_1$									1/1/15	0/8/9	1/0/16	1/0/16	1/1/15	1/0/16	1/0/16	1/0/16	1/0/16	0/1/16
(10) CNN$^-$+TMK$_2$										15/1/1	0/2/15	0/1/16	4/2/11	3/1/13	4/0/13	3/1/13	4/0/13	0/1/16
(11) TMK$_1$+CNN$^-$											1/0/16	1/0/16	1/1/15	1/1/15	1/1/15	1/1/15	1/1/15	0/1/16
(12) TMK$_2$+CNN$^-$												0/2/15	10/2/5	5/0/12	4/1/12	4/1/12	4/1/12	2/1/14
(13) EWW$^-$													17/0/0	12/2/3	12/3/2	12/2/3	12/3/2	4/8/5
(14) EWW$^-$+SSM$^-$														5/0/12	5/1/11	5/1/11	4/2/11	0/1/16
(15) SMOTE															4/9/4	6/9/2	3/13/1	3/1/13
(16) gg-SMOTE																4/8/5	2/13/2	4/0/13
(17) ncn-SMOTE																	3/10/4	4/0/13
(18) rng-SMOTE																		4/0/13
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	5/4/8	0/6/11	5/3/9	4/1/12	5/6/6	5/3/9	3/7/7	9/1/7	7/2/8	9/0/8	6/3/8	0/4/13	5/2/10	3/2/12	3/2/12	3/2/12	3/2/12	2/4/11
(2) SSM$^-$		5/3/9	5/5/7	3/3/11	9/4/4	9/1/7	8/2/7	13/1/3	9/4/4	13/1/3	7/5/5	2/5/10	6/3/8	3/2/12	4/1/12	3/2/12	3/2/12	6/4/7
(3) EW$^-$			7/5/5	5/5/7	10/4/3	7/5/5	10/2/5	13/1/3	8/5/4	13/1/3	8/5/4	0/11/6	8/3/6	4/5/8	5/3/9	5/2/10	5/4/8	6/5/6
(4) EW$^-$+SSM$^-$				6/5/6	10/5/2	10/3/4	9/3/5	15/1/1	9/7/1	15/1/1	8/7/2	3/6/8	5/9/3	4/6/7	4/6/7	4/5/8	4/5/8	9/3/5
(5) RUS					14/1/2	10/2/5	10/2/5	16/1/0	11/3/3	16/1/0	10/5/2	5/6/6	8/3/6	0/9/8	0/8/9	0/9/8	0/8/9	8/4/5
(6) CNN$^-$						8/1/8	8/2/7	11/1/5	7/5/5	11/1/5	4/6/7	1/5/11	2/4/11	2/0/15	2/1/14	2/0/15	2/0/15	4/4/9
(7) TMK$_1$							9/3/5	12/1/4	4/4/9	12/1/4	5/4/8	4/4/9	4/2/11	4/2/11	5/1/11	5/1/11	4/2/11	7/5/5
(8) TMK$_2$								9/0/8	6/2/9	9/0/8	7/1/9	5/2/10	5/2/10	5/1/11	5/2/10	5/1/11	5/2/10	5/6/6
(9) CNN$^-$+TMK$_1$									1/1/15	1/9/7	1/2/14	3/0/14	1/1/15	0/1/16	0/1/16	0/1/16	0/1/16	6/2/9
(10) CNN$^-$+TMK$_2$										15/1/1	2/4/11	2/3/12	1/6/10	2/3/12	3/2/12	2/2/13	3/2/12	7/4/6
(11) TMK$_1$+CNN$^-$											1/2/14	3/0/14	1/1/15	0/1/16	0/1/16	0/1/16	0/1/16	6/2/9
(12) TMK$_2$+CNN$^-$												3/5/9	3/6/8	2/1/14	2/3/12	2/2/13	2/3/12	8/3/6
(13) EWW$^-$													10/3/4	6/6/5	6/5/6	6/4/7	6/5/6	7/5/5
(14) EWW$^-$+SSM$^-$														3/4/10	3/6/8	3/5/9	3/3/11	10/2/5
(15) SMOTE															3/13/1	3/14/0	2/14/1	11/2/4
(16) gg-SMOTE																3/12/2	1/14/2	10/3/4
(17) ncn-SMOTE																	0/13/4	10/3/4
(18) rng-SMOTE																		10/3/4
(19) b-SMOTE																		

Tabla C.25: *Prueba t−student del clasificador SVM en términos de $IBA_{0.1}$ y Gm*

	$IBA_{0.1}$																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/3/13	0/4/13	2/1/14	0/0/17	3/3/11	4/2/11	5/6/6	10/5/2	3/3/11	9/2/6	3/1/13	0/4/13	2/2/13	0/0/17	0/0/17	0/0/17	0/0/17	3/4/10
(2) SSM^-		8/5/4	5/6/6	0/4/13	8/4/5	10/2/5	10/2/5	15/1/1	9/6/2	14/1/2	4/7/6	6/5/6	5/4/8	0/3/14	1/2/14	0/3/14	1/2/14	9/5/3
(3) EW^-			5/3/9	0/4/13	5/4/8	7/3/7	11/2/4	15/1/1	6/3/8	13/1/3	5/3/9	0/8/9	5/2/10	0/3/14	0/2/15	0/1/16	0/2/15	9/3/5
(4) EW^-+SSM^-				1/6/10	10/3/4	11/3/3	10/3/4	16/1/0	9/4/4	15/1/1	7/4/6	8/4/5	5/9/3	0/4/13	0/3/14	0/5/12	0/4/13	11/2/4
(5) RUS					14/3/0	13/2/2	16/0/1	17/0/0	15/0/2	16/1/0	12/5/0	13/4/0	10/5/2	0/11/6	1/10/6	0/13/4	0/11/6	16/0/1
(6) CNN^-						9/1/7	11/2/4	15/0/2	8/4/5	12/2/3	4/5/8	6/4/7	4/3/10	0/2/15	0/2/15	0/2/15	0/2/15	9/3/5
(7) TMK_1							10/2/5	13/2/2	4/3/10	12/1/4	4/3/10	6/3/8	3/2/12	2/0/15	2/0/15	2/0/15	2/0/15	9/2/6
(8) TMK_2								11/4/2	5/2/10	10/2/5	4/1/12	4/2/11	4/2/11	1/0/16	1/0/16	1/0/16	1/0/16	6/5/6
(9) CNN^-+TMK_1									1/1/15	1/8/8	1/0/16	1/1/15	0/1/16	0/0/17	0/0/17	0/0/17	0/0/17	3/1/13
(10) CNN^-+TMK_2										15/1/1	1/5/11	5/5/7	3/5/9	2/0/15	2/0/15	1/1/15	2/0/15	8/4/5
(11) TMK_1+CNN^-											1/1/15	3/1/13	0/3/14	0/1/16	0/1/16	0/1/16	0/1/16	5/1/11
(12) TMK_2+CNN^-												8/4/5	7/2/8	1/2/14	1/3/13	0/3/14	1/3/13	13/1/3
(13) EWW^-													6/2/9	0/3/14	0/3/14	0/1/16	0/2/15	10/3/4
(14) EWW^-+SSM^-														1/4/12	1/4/12	1/4/12	1/3/13	10/3/4
(15) SMOTE															1/16/0	3/14/0	0/16/1	16/0/1
(16) gg-SMOTE																2/13/2	0/14/3	16/0/1
(17) ncn-SMOTE																	1/13/3	16/0/1
(18) rng-SMOTE																		16/0/1
(19) b-SMOTE																		

	Gm																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/3/13	0/4/13	2/2/13	0/0/17	4/2/11	4/2/11	6/5/6	10/5/2	4/3/10	9/3/5	3/2/12	0/4/13	2/2/13	0/0/17	0/0/17	0/0/17	0/0/17	3/4/10
(2) SSM^-		9/4/4	5/7/5	0/4/13	8/4/5	10/2/5	10/2/5	15/1/1	10/5/2	14/1/2	4/7/6	6/5/6	6/4/7	0/3/14	1/2/14	0/2/15	0/3/14	9/4/4
(3) EW^-			5/3/9	0/4/13	6/4/7	7/3/7	11/2/4	15/1/1	8/2/7	13/1/3	5/3/9	0/8/9	6/1/10	0/3/14	0/3/14	0/1/16	0/2/15	9/4/4
(4) EW^-+SSM^-				2/5/10	9/4/4	11/3/3	10/2/5	16/1/0	10/3/4	15/1/1	7/3/7	8/3/6	5/9/3	0/3/14	0/3/14	0/5/12	0/3/14	10/2/5
(5) RUS					14/3/0	14/1/2	15/1/1	17/0/0	15/0/2	16/1/0	12/4/1	11/6/0	11/5/1	0/11/6	0/10/7	0/12/5	0/11/6	16/0/1
(6) CNN^-						9/1/7	11/2/4	15/0/2	10/2/5	12/2/3	4/6/7	6/3/8	6/2/9	0/1/16	0/2/15	0/1/16	0/1/16	10/2/5
(7) TMK_1							10/2/5	13/2/2	5/2/10	12/2/3	4/3/10	6/3/8	3/2/12	2/0/15	2/1/14	2/1/14	2/0/15	10/1/6
(8) TMK_2								11/4/2	5/2/10	10/3/4	4/2/11	4/2/11	5/1/11	1/1/15	1/1/15	1/1/15	1/1/15	5/6/6
(9) CNN^-+TMK_1									1/1/15	1/8/8	1/0/16	1/1/15	0/1/16	0/0/17	0/0/17	0/0/17	0/0/17	2/2/13
(10) CNN^-+TMK_2										15/1/1	1/5/11	5/2/10	3/4/10	1/1/15	2/0/15	1/1/15	2/0/15	8/4/5
(11) TMK_1+CNN^-											1/0/16	3/0/14	1/1/15	0/0/17	0/1/16	0/1/16	0/0/17	5/1/11
(12) TMK_2+CNN^-												6/6/5	7/2/8	1/1/15	1/3/13	0/3/14	1/2/14	13/1/3
(13) EWW^-													7/1/9	0/5/12	0/3/14	0/3/14	0/4/13	10/3/4
(14) EWW^-+SSM^-														1/2/14	1/3/13	1/3/13	1/2/14	10/2/5
(15) SMOTE															1/16/0	3/14/0	0/16/1	16/0/1
(16) gg-SMOTE																2/13/2	0/14/3	16/0/1
(17) ncn-SMOTE																	0/14/3	16/0/1
(18) rng-SMOTE																		16/0/1
(19) b-SMOTE																		

Tabla C.26: *Prueba $t-$student del clasificador $1-NN$ en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/3/4
(2) SSM^-		2/1/5	0/0/8	1/1/6	7/1/0	0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	0/2/6	2/0/6	0/0/8	4/4/0	3/3/2	4/3/1	5/3/0	7/1/0
(3) EW^-			0/0/8	1/4/3	6/0/2	0/0/8	8/0/0	0/0/8	1/2/5	0/0/8	2/4/2	0/2/6	0/0/8	4/4/0	4/4/0	4/4/0	4/4/0	7/1/0
(4) EW^-+SSM^-				7/1/0	8/0/0	0/1/7	8/0/0	0/0/8	5/3/0	0/0/8	8/0/0	7/1/0	0/2/6	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(5) RUS					7/0/1	0/0/8	8/0/0	0/0/8	1/3/4	0/0/8	3/4/1	2/3/3	0/0/8	7/1/0	6/2/0	6/2/0	7/1/0	7/1/0
(6) CNN^-						0/0/8	5/2/1	0/0/8	0/0/8	0/0/8	0/0/8	2/0/6	0/0/8	2/4/2	1/4/3	1/5/2	2/3/3	7/0/1
(7) TMK_1							8/0/0	0/0/8	8/0/0	0/0/8	8/0/0	8/0/0	7/0/1	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(8) TMK_2								0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7	1/0/7	1/0/7	1/1/6	7/0/1
(9) CNN^-+TMK_1									8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN^-+TMK_2										0/0/8	8/0/0	2/4/2	0/0/8	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(11) TMK_1+CNN^-											8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK_2+CNN^-												2/1/5	0/0/8	7/1/0	5/3/0	6/2/0	7/1/0	7/1/0
(13) EWW^-													0/0/8	6/2/0	6/2/0	6/2/0	6/2/0	8/0/0
(14) EWW^-+SSM^-														8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(15) SMOTE															0/5/3	0/6/2	1/7/0	7/1/0
(16) gg-SMOTE																1/7/0	4/4/0	7/1/0
(17) ncn-SMOTE																	2/6/0	7/1/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	3/4/1
(2) SSM^-		2/1/5	8/0/0	6/0/2	0/4/4	8/0/0	1/0/7	8/0/0	8/0/0	8/0/0	8/0/0	3/3/2	8/0/0	0/2/6	1/2/5	0/2/6	0/1/7	0/1/7
(3) EW^-			8/0/0	6/2/0	3/3/2	8/0/0	0/0/8	8/0/0	8/0/0	8/0/0	6/1/1	7/1/0	8/0/0	2/2/4	2/3/3	1/4/3	2/2/4	0/1/7
(4) EW^-+SSM^-				0/2/6	0/0/8	7/1/0	0/0/8	8/0/0	2/3/3	8/0/0	0/0/8	0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					1/1/6	7/1/0	0/0/8	8/0/0	8/0/0	8/0/0	2/3/3	1/2/5	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(6) CNN^-						8/0/0	1/0/7	8/0/0	8/0/0	8/0/0	8/0/0	4/2/2	8/0/0	0/2/6	0/4/4	0/2/6	0/2/6	0/1/7
(7) TMK_1							0/0/8	8/0/0	1/0/7	8/0/0	0/1/7	0/0/8	1/0/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(8) TMK_2								8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	6/1/1	7/0/1	7/0/1	5/2/1	1/0/7
(9) CNN^-+TMK_1									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										8/0/0	0/0/8	0/0/8	6/2/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												2/1/5	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(13) EWW^-													8/0/0	0/3/5	0/2/6	0/3/5	0/2/6	0/1/7
(14) EWW^-+SSM^-														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															1/7/0	3/4/1	0/7/1	0/1/7
(16) gg-SMOTE																2/4/2	0/3/5	0/1/7
(17) ncn-SMOTE																	1/3/4	0/2/6
(18) rng-SMOTE																		1/0/7
(19) b-SMOTE																		

Tabla C.27: *Prueba $t-$student del clasificador $1-NN$ en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	7/0/1	8/0/0	8/0/0	8/0/0	8/0/0	4/4/0	8/0/0	8/0/0	8/0/0	8/0/0	7/0/1	8/0/0	7/1/0	6/2/0	5/3/0	5/3/0	2/5/1
(2) SSM$^-$		2/1/5	6/2/0	4/2/2	1/7/0	7/1/0	0/2/6	8/0/0	8/0/0	8/0/0	5/3/0	2/2/4	6/2/0	0/4/4	0/3/5	0/2/6	0/1/7	0/2/6
(3) EW$^-$			8/0/0	7/1/0	5/1/2	8/0/0	1/2/5	8/0/0	7/1/0	8/0/0	6/1/1	5/3/0	8/0/0	4/2/2	4/1/3	2/4/2	4/2/2	1/3/4
(4) EW$^-$+SSM$^-$				1/2/5	0/3/5	7/0/1	0/0/8	8/0/0	4/4/0	8/0/0	0/4/4	0/1/7	6/2/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/2/4	7/0/1	0/0/8	8/0/0	6/2/0	8/0/0	2/4/2	1/1/6	7/1/0	0/2/6	0/2/6	0/1/7	0/0/8	0/0/8
(6) CNN$^-$						7/1/0	1/0/7	8/0/0	8/0/0	8/0/0	6/2/0	2/1/5	6/2/0	0/3/5	0/2/6	0/2/6	0/2/6	0/2/6
(7) TMK$_1$							0/0/8	8/0/0	1/0/7	8/0/0	1/0/7	0/0/8	1/0/7	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	6/2/0	8/0/0	4/2/2	4/2/2	5/1/2	4/2/2	1/4/3
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/1/0	0/0/8	0/1/7	1/6/1	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												1/1/6	6/2/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(13) EWW$^-$													8/0/0	2/3/3	2/3/3	1/5/2	2/3/3	1/2/5
(14) EWW$^-$+SSM$^-$														0/1/7	0/1/7	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/7/1	0/7/1	0/8/0	0/3/5
(16) gg-SMOTE																1/6/1	0/7/1	0/4/4
(17) ncn-SMOTE																	0/7/1	0/3/5
(18) rng-SMOTE																		0/5/3
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/2/5	2/1/5	2/0/6	2/1/5	2/2/4	2/0/6	1/1/6	4/3/1	2/0/6	2/0/6	2/0/6	2/1/5	2/0/6	0/2/6	0/1/7	0/2/6	0/2/6	1/4/3
(2) SSM$^-$		1/3/4	2/2/4	1/4/3	2/6/0	2/2/4	1/6/1	6/1/1	2/6/0	3/3/2	1/7/0	1/2/5	2/2/4	1/4/3	0/3/5	1/3/4	1/4/3	4/2/2
(3) EW$^-$			3/3/2	4/2/2	5/2/1	4/3/1	5/1/2	5/2/1	4/3/1	5/2/1	4/4/0	1/4/3	3/1/4	3/4/1	2/4/2	2/5/1	2/5/1	4/3/1
(4) EW$^-$+SSM$^-$				2/5/1	5/1/2	5/2/1	5/1/2	7/0/1	3/5/0	6/1/1	3/3/2	3/3/2	2/5/1	5/1/2	4/2/2	4/2/2	4/2/2	5/1/2
(5) RUS					4/3/1	3/3/2	3/3/2	6/1/1	3/5/0	4/3/1	2/5/1	1/3/4	2/3/3	2/4/2	2/4/2	2/4/2	2/4/2	5/1/2
(6) CNN$^-$						2/2/4	0/6/2	5/2/1	2/1/5	2/3/3	2/3/3	1/2/5	2/1/5	0/4/4	0/2/6	0/2/6	0/2/6	2/3/3
(7) TMK$_1$							3/3/2	7/1/0	3/2/3	6/2/0	3/3/2	1/3/4	0/2/6	5/1/2	1/5/2	3/3/2	3/3/2	5/1/2
(8) TMK$_2$								5/2/1	2/4/2	2/5/1	2/4/2	2/1/5	2/1/5	0/6/2	0/6/2	0/6/2	1/4/3	4/2/2
(9) CNN$^-$+TMK$_1$									1/1/6	0/1/7	1/0/7	0/3/5	0/1/7	1/2/5	1/2/5	1/2/5	1/2/5	1/2/5
(10) CNN$^-$+TMK$_2$										3/4/1	0/6/2	1/1/6	1/3/4	1/5/2	1/4/3	1/4/3	1/5/2	5/1/2
(11) TMK$_1$+CNN$^-$											1/4/3	0/3/5	1/1/6	2/4/2	1/5/2	2/4/2	2/4/2	5/1/2
(12) TMK$_2$+CNN$^-$												1/2/5	2/2/4	1/5/2	0/6/2	2/4/2	2/4/2	5/1/2
(13) EWW$^-$													3/3/2	4/2/2	3/3/2	4/3/1	3/4/1	4/2/2
(14) EWW$^-$+SSM$^-$														6/0/2	5/1/2	5/1/2	5/1/2	5/1/2
(15) SMOTE															0/6/2	0/7/1	0/8/0	3/5/0
(16) gg-SMOTE																0/8/0	1/7/0	5/3/0
(17) ncn-SMOTE																	0/8/0	4/4/0
(18) rng-SMOTE																		4/4/0
(19) b-SMOTE																		

Tabla C.28: *Prueba $t-$student del clasificador $1-NN$ en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/0/7	0/2/6	1/1/6	1/2/5	1/1/6	4/2/2	0/0/8	7/1/0	1/2/5	6/1/1	1/0/7	0/2/6	1/1/6	0/0/8	0/0/8	0/0/8	0/1/7	1/3/4
(2) SSM^-		1/3/4	1/3/4	1/3/4	2/6/0	6/0/2	2/5/1	8/0/0	2/5/1	7/0/1	0/7/1	1/2/5	2/3/3	0/6/2	0/3/5	0/6/2	1/5/2	5/2/1
(3) EW^-			2/3/3	4/3/1	5/3/0	6/1/1	6/2/0	8/0/0	4/4/0	7/1/0	3/5/0	1/4/3	2/5/1	2/5/1	2/4/2	2/5/1	2/5/1	7/1/0
(4) EW^-+SSM^-				3/4/1	4/3/1	7/0/1	5/2/1	8/0/0	4/4/0	7/1/0	4/3/1	1/4/3	2/5/1	3/4/1	3/3/2	3/4/1	3/4/1	5/2/1
(5) RUS					4/3/1	7/0/1	3/4/1	8/0/0	2/6/0	7/0/1	2/5/1	0/4/4	1/6/1	1/5/2	1/5/2	1/5/2	2/4/2	5/2/1
(6) CNN^-						4/2/2	1/4/3	8/0/0	2/3/3	7/0/1	1/4/3	1/2/5	2/2/4	0/5/3	0/2/6	0/2/6	0/3/5	3/3/2
(7) TMK_1							2/0/6	8/0/0	1/1/6	7/1/0	1/1/6	0/1/7	1/0/7	1/2/5	0/3/5	1/2/5	1/2/5	1/3/4
(8) TMK_2								8/0/0	1/5/2	6/1/1	1/3/4	0/3/5	1/2/5	1/4/3	0/3/5	0/4/4	1/3/4	5/2/1
(9) CNN^-+TMK_1									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										7/0/1	1/4/3	0/3/5	0/3/5	0/6/2	0/6/2	0/5/3	0/6/2	3/4/1
(11) TMK_1+CNN^-											1/0/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	1/2/5
(12) TMK_2+CNN^-												1/2/5	1/4/3	1/6/1	0/6/2	0/6/2	2/5/1	6/1/1
(13) EWW^-													4/3/1	4/3/1	3/3/2	4/3/1	4/3/1	4/4/0
(14) EWW^-+SSM^-														2/4/2	3/3/2	3/3/2	3/3/2	5/2/1
(15) SMOTE															0/6/2	0/6/2	0/8/0	5/3/0
(16) gg-SMOTE																0/8/0	2/6/0	6/2/0
(17) ncn-SMOTE																	2/6/0	5/3/0
(18) rng-SMOTE																		5/3/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/0/7	0/2/6	1/1/6	1/2/5	1/2/5	5/1/2	0/1/7	8/0/0	2/2/4	6/1/1	1/0/7	0/3/5	1/3/4	0/0/8	0/0/8	0/1/7	0/1/7	1/4/3
(2) SSM^-		1/3/4	3/2/3	1/4/3	1/7/0	6/1/1	1/6/1	8/0/0	2/5/1	7/0/1	1/7/0	1/2/5	3/2/3	0/6/2	0/3/5	0/4/4	0/6/2	4/3/1
(3) EW^-			2/5/1	4/3/1	5/3/0	7/0/1	6/2/0	8/0/0	4/4/0	7/1/0	3/5/0	1/4/3	3/4/1	2/5/1	2/4/2	2/5/1	2/5/1	3/5/0
(4) EW^-+SSM^-				2/5/1	4/3/1	7/0/1	5/2/1	8/0/0	4/4/0	7/1/0	3/4/1	1/4/3	3/4/1	2/4/2	1/5/2	3/3/2	3/3/2	4/3/1
(5) RUS					4/3/1	7/0/1	2/5/1	8/0/0	3/5/0	7/0/1	2/5/1	0/4/4	1/6/1	1/5/2	1/5/2	1/5/2	2/4/2	3/4/1
(6) CNN^-						6/1/1	1/4/3	8/0/0	2/4/2	7/0/1	1/4/3	1/2/5	2/2/4	0/4/4	0/2/6	0/2/6	0/2/6	3/3/2
(7) TMK_1							1/1/6	8/0/0	1/0/7	7/1/0	1/1/6	0/1/7	1/0/7	1/1/6	0/2/6	1/1/6	0/2/6	1/2/5
(8) TMK_2								8/0/0	2/5/1	7/0/1	1/5/2	1/2/5	2/3/3	0/6/2	0/5/3	0/4/4	1/4/3	5/2/1
(9) CNN^-+TMK_1									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										7/0/1	1/4/3	0/3/5	0/4/4	0/5/3	0/3/5	0/4/4	0/5/3	2/4/2
(11) TMK_1+CNN^-											1/0/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/2/6
(12) TMK_2+CNN^-												1/2/5	2/4/2	0/6/2	0/5/3	0/5/3	0/6/2	4/3/1
(13) EWW^-													4/3/1	4/3/1	3/3/2	4/3/1	3/4/1	3/5/0
(14) EWW^-+SSM^-														2/4/2	2/4/2	2/4/2	2/4/2	3/4/1
(15) SMOTE															0/6/2	0/6/2	0/8/0	5/3/0
(16) gg-SMOTE																0/8/0	2/6/0	6/2/0
(17) ncn-SMOTE																	1/7/0	5/3/0
(18) rng-SMOTE																		5/3/0
(19) b-SMOTE																		

Tabla C.29: *Prueba $t-$student del clasificador $7-NN$ en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/2/6
(2) SSM$^-$		5/1/2	0/0/8	2/0/6	3/4/1	0/0/8	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	3/1/4	0/0/8	2/1/5	2/0/6	2/1/5	2/0/6	7/0/1
(3) EW$^-$			0/0/8	0/2/6	3/1/4	0/0/8	8/0/0	0/0/8	0/0/8	0/0/8	0/1/7	0/2/6	0/0/8	1/1/6	1/1/6	1/1/6	1/1/6	7/0/1
(4) EW$^-$+SSM$^-$				8/0/0	8/0/0	1/1/6	8/0/0	0/0/8	2/4/2	0/0/8	7/1/0	8/0/0	0/2/6	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0
(5) RUS					6/0/2	0/1/7	8/0/0	0/0/8	0/0/8	0/0/8	1/4/3	6/1/1	0/0/8	1/5/2	1/5/2	5/1/2	2/5/1	7/1/0
(6) CNN$^-$						0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	3/0/5	0/0/8	2/0/6	2/0/6	2/1/5	2/0/6	7/0/1
(7) TMK$_1$							8/0/0	0/1/7	5/2/1	0/1/7	7/1/0	8/0/0	5/2/1	7/0/1	7/0/1	7/0/1	7/0/1	7/1/0
(8) TMK$_2$								0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	6/1/1
(9) CNN$^-$+TMK$_1$									8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN$^-$+TMK$_2$										0/0/8	8/0/0	8/0/0	1/1/6	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(11) TMK$_1$+CNN$^-$											8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK$_2$+CNN$^-$												6/2/0	0/0/8	5/1/2	5/1/2	6/1/1	5/1/2	7/1/0
(13) EWW$^-$													0/0/8	2/0/6	2/0/6	1/1/6	2/0/6	7/0/1
(14) EWW$^-$+SSM$^-$														8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(15) SMOTE															0/7/1	2/6/0	0/8/0	8/0/0
(16) gg-SMOTE																4/4/0	1/6/1	8/0/0
(17) ncn-SMOTE																	0/5/3	8/0/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	5/3/0
(2) SSM$^-$		1/0/7	8/0/0	5/1/2	2/1/5	7/1/0	0/0/8	8/0/0	8/0/0	8/0/0	8/0/0	2/1/5	8/0/0	4/2/2	3/2/3	3/2/3	4/1/3	0/1/7
(3) EW$^-$			8/0/0	8/0/0	7/0/1	8/0/0	0/0/8	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	7/1/0	5/2/1	6/2/0	6/2/0	1/0/7
(4) EW$^-$+SSM$^-$				0/0/8	0/0/8	6/1/1	0/0/8	8/0/0	4/2/2	8/0/0	0/1/7	0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/1/5	7/0/1	0/0/8	8/0/0	8/0/0	8/0/0	5/3/0	0/1/7	8/0/0	0/2/6	0/2/6	0/2/6	0/1/7	0/0/8
(6) CNN$^-$						8/0/0	1/0/7	8/0/0	8/0/0	8/0/0	8/0/0	2/2/4	8/0/0	4/2/2	5/1/2	4/2/2	5/1/2	1/1/6
(7) TMK$_1$							0/0/8	8/0/0	2/0/6	8/0/0	1/0/7	0/0/8	2/1/5	1/0/7	1/0/7	1/0/7	1/0/7	0/1/7
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	1/0/7
(9) CNN$^-$+TMK$_1$									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										8/0/0	0/0/8	0/0/8	5/2/1	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												0/0/8	8/0/0	0/1/7	0/1/7	0/0/8	0/1/7	0/0/8
(13) EWW$^-$													8/0/0	5/3/0	5/2/1	5/3/0	5/2/1	1/0/7
(14) EWW$^-$+SSM$^-$														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/4/4	0/5/3	0/7/1	0/0/8
(16) gg-SMOTE																2/3/3	3/5/0	0/0/8
(17) ncn-SMOTE																	2/6/0	0/0/8
(18) rng-SMOTE																		0/0/8
(19) b-SMOTE																		

Tabla C.30: *Prueba $t-$student del clasificador $7-NN$ en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	3/5/0	8/0/0	8/0/0	8/0/0	8/0/0	3/4/1	8/0/0	8/0/0	8/0/0	8/0/0	5/3/0	8/0/0	7/0/1	7/0/1	7/0/1	7/0/1	3/4/1
(2) SSM^-		0/2/6	8/0/0	3/2/3	2/2/4	7/1/0	0/0/8	8/0/0	8/0/0	8/0/0	5/3/0	0/2/6	8/0/0	1/3/4	1/3/4	1/2/5	1/3/4	0/1/7
(3) EW^-			8/0/0	8/0/0	7/0/1	8/0/0	0/5/3	8/0/0	8/0/0	8/0/0	8/0/0	4/4/0	8/0/0	5/2/1	4/1/3	5/1/2	5/2/1	1/4/3
(4) EW^-+SSM^-				0/1/7	0/0/8	6/1/1	0/0/8	8/0/0	4/2/2	8/0/0	0/3/5	0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/2/4	7/0/1	0/0/8	8/0/0	8/0/0	8/0/0	4/4/0	0/0/8	8/0/0	0/2/6	0/1/7	0/1/7	0/1/7	0/0/8
(6) CNN^-						7/1/0	1/0/7	8/0/0	8/0/0	8/0/0	8/0/0	1/1/6	8/0/0	3/3/2	3/2/3	3/2/3	3/2/3	0/2/6
(7) TMK_1							0/0/8	8/0/0	2/1/5	8/0/0	1/0/7	0/0/8	2/1/5	1/0/7	1/0/7	1/0/7	1/0/7	0/1/7
(8) TMK_2								8/0/0	8/0/0	8/0/0	8/0/0	5/3/0	8/0/0	7/1/0	6/1/1	7/1/0	7/1/0	1/5/2
(9) CNN^-+TMK_1									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										8/0/0	0/0/8	0/0/8	2/5/1	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												0/0/8	8/0/0	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8
(13) EWW^-													8/0/0	5/1/2	4/1/3	5/1/2	4/2/2	1/2/5
(14) EWW^-+SSM^-														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/4/4	0/6/2	0/7/1	1/0/7
(16) gg-SMOTE																3/4/1	3/5/0	1/1/6
(17) ncn-SMOTE																	2/6/0	1/0/7
(18) rng-SMOTE																		1/0/7
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/0/6	0/2/6	2/0/6	1/1/6	2/0/6	2/0/6	0/1/7	5/1/2	2/0/6	3/2/3	2/0/6	1/1/6	2/0/6	0/0/8	0/0/8	0/0/8	0/0/8	0/5/3
(2) SSM^-		0/3/5	2/1/5	0/2/6	2/5/1	3/4/1	3/3/2	6/1/1	2/3/3	6/1/1	2/3/3	0/2/6	2/1/5	0/2/6	0/1/7	0/2/6	0/1/7	3/2/3
(3) EW^-			3/2/3	3/2/3	5/3/0	5/3/0	5/2/1	7/0/1	3/4/1	7/0/1	4/2/2	2/2/4	3/1/4	2/0/6	0/2/6	1/1/6	1/2/5	5/3/0
(4) EW^-+SSM^-				1/5/2	5/1/2	6/2/0	5/1/2	7/1/0	3/5/0	7/1/0	1/6/1	1/4/3	3/4/1	1/5/2	0/6/2	0/5/3	0/6/2	5/1/2
(5) RUS					7/1/0	5/2/1	5/1/2	7/0/1	3/5/0	6/1/1	3/5/0	1/3/4	2/5/1	0/5/3	0/4/4	0/5/3	0/5/3	4/2/2
(6) CNN^-						3/4/1	2/4/2	6/1/1	2/3/3	6/1/1	2/2/4	0/2/6	2/2/4	0/2/6	0/1/7	0/2/6	0/1/7	3/2/3
(7) TMK_1							2/4/2	7/1/0	1/2/5	7/1/0	1/2/5	0/1/7	1/2/5	1/2/5	0/2/6	0/2/6	0/2/6	3/3/2
(8) TMK_2								5/2/1	2/2/4	5/2/1	2/0/6	1/1/6	2/1/5	0/2/6	0/0/8	0/1/7	0/0/8	3/4/1
(9) CNN^-+TMK_1									1/0/7	0/2/6	1/0/7	0/1/7	0/1/7	0/2/6	0/1/7	0/1/7	0/1/7	0/3/5
(10) CNN^-+TMK_2										7/0/1	1/3/4	0/4/4	0/3/5	1/1/6	0/2/6	0/3/5	0/3/5	5/1/2
(11) TMK_1+CNN^-											1/0/7	0/1/7	0/1/7	0/2/6	0/1/7	0/2/6	0/2/6	0/4/4
(12) TMK_2+CNN^-												1/3/4	2/5/1	1/3/4	0/3/5	0/4/4	0/4/4	5/0/3
(13) EWW^-													3/4/1	2/3/3	1/3/4	1/4/3	1/4/3	5/2/1
(14) EWW^-+SSM^-														1/5/2	0/5/3	0/6/2	0/6/2	5/1/2
(15) SMOTE															0/4/4	0/8/0	0/8/0	6/2/0
(16) gg-SMOTE																2/6/0	3/5/0	7/1/0
(17) ncn-SMOTE																	0/8/0	6/2/0
(18) rng-SMOTE																		6/2/0
(19) b-SMOTE																		

Tabla C.31: *Prueba $t-$student del clasificador $7-NN$ en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

	$IBA_{0.1}$																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
	Gm																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/0/7	0/0/8	1/0/7	0/1/7	1/0/7	3/1/4	0/0/8	8/0/0	1/1/6	7/0/1	1/0/7	0/0/8	1/1/6	0/0/8	0/0/8	0/0/8	0/0/8	0/3/5
(2) SSM^-		3/1/4	2/2/4	0/1/7	2/6/0	6/1/1	5/2/1	8/0/0	2/5/1	8/0/0	2/2/4	1/2/5	2/3/3	0/1/7	0/1/7	0/1/7	0/1/7	4/2/2
(3) EW^-			3/2/3	1/3/4	5/1/2	6/1/1	7/1/0	8/0/0	3/3/2	8/0/0	3/2/3	0/3/5	4/1/3	2/0/6	0/2/6	1/1/6	2/0/6	7/0/1
(4) EW^-+SSM^-				0/5/3	4/3/1	5/3/0	5/2/1	8/0/0	4/3/1	8/0/0	2/3/3	3/2/3	4/4/0	0/3/5	0/3/5	0/4/4	0/5/3	4/3/1
(5) RUS					7/1/0	6/1/1	7/1/0	8/0/0	4/4/0	8/0/0	4/4/0	3/3/2	4/4/0	0/5/3	0/4/4	0/5/3	0/5/3	5/2/1
(6) CNN^-						6/1/1	5/2/1	8/0/0	3/2/3	7/1/0	2/2/4	0/2/6	3/2/3	0/2/6	0/0/8	0/2/6	0/1/7	5/1/2
(7) TMK_1							2/2/4	8/0/0	2/1/5	8/0/0	1/2/5	0/2/6	1/2/5	0/1/7	0/1/7	0/1/7	0/2/6	1/4/3
(8) TMK_2								8/0/0	1/3/4	7/1/0	1/0/7	0/1/7	1/3/4	0/1/7	0/0/8	0/1/7	0/0/8	5/2/1
(9) CNN^-+TMK_1									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										8/0/0	0/4/4	0/4/4	1/3/4	0/2/6	0/1/7	0/4/4	0/1/7	4/2/2
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												1/4/3	5/1/2	0/4/4	0/3/5	0/4/4	0/4/4	6/0/2
(13) EWW^-													4/3/1	2/1/5	0/3/5	1/3/4	2/1/5	7/1/0
(14) EWW^-+SSM^-														0/3/5	0/2/6	0/3/5	0/4/4	4/3/1
(15) SMOTE															0/7/1	1/7/0	0/8/0	7/1/0
(16) gg-SMOTE																3/5/0	3/5/0	7/1/0
(17) ncn-SMOTE																	0/7/1	7/1/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

Tabla C.32: *Prueba $t-$student del clasificador $13-NN$ en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/1/7
(2) SSM^-		6/1/1	0/0/8	2/1/5	4/2/2	0/1/7	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	4/1/3	0/0/8	2/0/6	2/0/6	2/1/5	2/1/5	7/1/0
(3) EW^-			0/0/8	0/1/7	2/2/4	0/0/8	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/2/6	0/0/8	0/1/7	0/0/8	0/2/6	0/2/6	7/0/1
(4) EW^-+SSM^-				8/0/0	8/0/0	1/2/5	8/0/0	0/0/8	2/2/4	0/0/8	7/1/0	8/0/0	0/2/6	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0
(5) RUS					6/0/2	1/0/7	8/0/0	0/0/8	0/0/8	0/0/8	1/4/3	7/0/1	0/0/8	0/5/3	0/5/3	2/6/0	1/5/2	8/0/0
(6) CNN^-						0/0/8	7/0/1	0/0/8	0/0/8	0/0/8	0/0/8	3/1/4	0/0/8	2/0/6	2/0/6	2/0/6	2/0/6	6/1/1
(7) TMK_1							8/0/0	0/1/7	6/0/2	0/1/7	6/1/1	8/0/0	5/1/2	7/0/1	7/0/1	7/0/1	7/1/0	7/1/0
(8) TMK_2								0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	6/1/1
(9) CNN^-+TMK_1									8/0/0	6/2/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN^-+TMK_2										0/0/8	8/0/0	8/0/0	1/1/6	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(11) TMK_1+CNN^-											8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK_2+CNN^-												8/0/0	0/0/8	4/3/1	4/2/2	5/3/0	5/2/1	8/0/0
(13) EWW^-													0/0/8	0/2/6	0/1/7	1/1/6	0/2/6	7/0/1
(14) EWW^-+SSM^-														8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(15) SMOTE															1/6/1	6/2/0	1/7/0	8/0/0
(16) gg-SMOTE																4/4/0	1/7/0	8/0/0
(17) ncn-SMOTE																	1/4/3	8/0/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	5/3/0
(2) SSM^-		1/0/7	8/0/0	5/1/2	3/0/5	7/1/0	0/0/8	8/0/0	8/0/0	8/0/0	8/0/0	2/1/5	8/0/0	5/1/2	5/1/2	5/1/2	5/1/2	0/1/7
(3) EW^-			8/0/0	8/0/0	7/0/1	8/0/0	0/0/8	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	1/0/7
(4) EW^-+SSM^-				0/0/8	0/0/8	5/1/2	0/0/8	8/0/0	5/2/1	8/0/0	0/2/6	0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/1/5	7/0/1	0/0/8	8/0/0	8/0/0	8/0/0	6/2/0	0/1/7	8/0/0	2/2/4	0/3/5	1/2/5	0/4/4	0/0/8
(6) CNN^-						8/0/0	1/0/7	8/0/0	8/0/0	8/0/0	8/0/0	2/2/4	8/0/0	6/0/2	6/0/2	5/1/2	6/0/2	1/1/6
(7) TMK_1							0/0/8	7/1/0	2/1/5	7/1/0	1/1/6	0/0/8	2/1/5	1/0/7	1/0/7	1/0/7	1/0/7	0/1/7
(8) TMK_2								8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	1/1/6
(9) CNN^-+TMK_1									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										8/0/0	0/0/8	0/0/8	4/2/2	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												0/0/8	8/0/0	0/2/6	0/1/7	0/1/7	0/1/7	0/0/8
(13) EWW^-													8/0/0	8/0/0	6/2/0	7/1/0	7/1/0	1/0/7
(14) EWW^-+SSM^-														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/4/4	1/3/4	1/5/2	0/0/8
(16) gg-SMOTE																2/3/3	3/5/0	0/0/8
(17) ncn-SMOTE																	3/5/0	0/0/8
(18) rng-SMOTE																		0/0/8
(19) b-SMOTE																		

Tabla C.33: *Prueba $t-$student del clasificador $13-NN$ en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	3/5/0	8/0/0	8/0/0	8/0/0	8/0/0	3/4/1	8/0/0	8/0/0	8/0/0	8/0/0	4/4/0	8/0/0	7/0/1	7/0/1	7/0/1	7/0/1	2/6/0
(2) SSM^-		0/1/7	8/0/0	4/1/3	2/3/3	7/1/0	0/0/8	8/0/0	8/0/0	8/0/0	7/1/0	0/2/6	8/0/0	3/3/2	2/3/3	3/1/4	3/2/3	0/1/7
(3) EW^-			8/0/0	8/0/0	7/0/1	8/0/0	0/5/3	8/0/0	8/0/0	8/0/0	8/0/0	5/3/0	8/0/0	7/0/1	5/2/1	6/1/1	7/0/1	1/4/3
(4) EW^-+SSM^-				0/0/8	0/0/8	5/1/2	0/0/8	8/0/0	6/1/1	8/0/0	0/2/6	0/0/8	7/1/0	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/1/5	7/0/1	0/0/8	8/0/0	8/0/0	8/0/0	5/3/0	0/1/7	8/0/0	1/3/4	1/3/4	1/4/3	1/4/3	0/0/8
(6) CNN^-						8/0/0	0/1/7	8/0/0	8/0/0	8/0/0	8/0/0	1/3/4	8/0/0	3/3/2	3/3/2	3/3/2	4/2/2	1/1/6
(7) TMK_1							0/0/8	7/1/0	2/1/5	7/1/0	1/1/6	0/0/8	2/1/5	1/0/7	1/0/7	1/0/7	1/0/7	0/1/7
(8) TMK_2								8/0/0	8/0/0	8/0/0	8/0/0	5/3/0	8/0/0	7/0/1	7/0/1	7/0/1	7/0/1	1/6/1
(9) CNN^-+TMK_1									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										8/0/0	0/0/8	0/0/8	2/3/3	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												0/0/8	8/0/0	0/2/6	0/1/7	0/1/7	0/1/7	0/0/8
(13) EWW^-													8/0/0	6/0/2	4/2/2	5/1/2	6/0/2	1/4/3
(14) EWW^-+SSM^-														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/4/4	1/5/2	1/6/1	1/0/7
(16) gg-SMOTE																1/7/0	2/6/0	1/0/7
(17) ncn-SMOTE																	1/7/0	1/0/7
(18) rng-SMOTE																		1/0/7
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/0/6	0/1/7	2/0/6	1/1/6	2/0/6	2/0/6	0/0/8	4/2/2	2/0/6	3/2/3	2/0/6	0/1/7	2/0/6	0/0/8	0/0/8	0/0/8	0/0/8	0/3/5
(2) SSM^-		1/3/4	2/2/4	0/3/5	5/3/0	4/3/1	4/2/2	6/2/0	3/3/2	6/2/0	2/2/4	0/2/6	2/2/4	0/1/7	0/2/6	0/2/6	0/2/6	4/2/2
(3) EW^-			2/2/4	3/1/4	5/3/0	5/3/0	7/1/0	7/1/0	2/5/1	6/1/1	2/2/4	1/3/4	2/3/3	0/3/5	0/3/5	1/2/5	1/2/5	6/2/0
(4) EW^-+SSM^-				0/6/2	5/1/2	6/1/1	6/0/2	8/0/0	6/2/0	7/1/0	1/6/1	2/4/2	4/4/0	1/3/4	1/4/3	1/5/2	1/4/3	6/0/2
(5) RUS					7/1/0	5/3/0	6/1/1	7/1/0	5/3/0	6/2/0	4/4/0	1/4/3	3/5/0	1/3/4	1/4/3	0/6/2	1/4/3	5/2/1
(6) CNN^-						3/3/2	5/1/2	6/1/1	2/2/4	6/1/1	2/0/6	0/2/6	2/2/4	0/1/7	0/1/7	0/0/8	0/1/7	4/1/3
(7) TMK_1							2/4/2	6/2/0	2/1/5	6/2/0	1/2/5	0/2/6	1/1/6	0/2/6	0/2/6	0/3/5	1/2/5	3/3/2
(8) TMK_2								5/2/1	2/1/5	5/2/1	2/1/5	1/0/7	2/0/6	0/1/7	0/0/8	0/0/8	0/1/7	3/4/1
(9) CNN^-+TMK_1									0/1/7	0/1/7	0/1/7	0/1/7	0/0/8	0/1/7	0/1/7	0/1/7	0/1/7	0/3/5
(10) CNN^-+TMK_2										7/1/0	0/5/3	0/6/2	0/2/6	0/4/4	0/3/5	0/4/4	1/2/5	5/1/2
(11) TMK_1+CNN^-											0/2/6	0/2/6	0/1/7	1/1/6	1/0/7	0/2/6	1/1/6	0/4/4
(12) TMK_2+CNN^-												0/6/2	2/5/1	1/3/4	1/2/5	0/5/3	1/3/4	5/1/2
(13) EWW^-													2/5/1	0/4/4	0/3/5	0/5/3	1/2/5	6/1/1
(14) EWW^-+SSM^-														1/3/4	1/2/5	1/4/3	1/4/3	6/0/2
(15) SMOTE															0/5/3	2/6/0	2/5/1	7/1/0
(16) gg-SMOTE																3/5/0	2/6/0	7/1/0
(17) ncn-SMOTE																	1/7/0	7/1/0
(18) rng-SMOTE																		7/0/1
(19) b-SMOTE																		

Tabla C.34: *Prueba $t-$student del clasificador $13-NN$ en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/7	0/0/8	1/0/7	0/0/8	1/0/7	2/2/4	0/0/8	8/0/0	1/1/6	7/1/0	1/0/7	0/0/8	1/0/7	0/0/8	0/0/8	0/0/8	0/0/8	0/2/6
(2) SSM$^-$		4/1/3	3/3/2	0/1/7	4/4/0	6/2/0	6/1/1	8/0/0	4/2/2	8/0/0	2/2/4	2/2/4	4/1/3	0/1/7	0/2/6	0/2/6	0/2/6	5/3/0
(3) EW$^-$			2/2/4	1/2/5	3/2/3	5/2/1	8/0/0	8/0/0	4/2/2	8/0/0	2/2/4	0/3/5	3/3/2	0/2/6	0/2/6	1/1/6	1/1/6	7/0/1
(4) EW$^-$+SSM$^-$				0/4/4	3/3/2	5/3/0	5/2/1	8/0/0	6/1/1	8/0/0	2/4/2	3/3/2	6/2/0	0/3/5	0/2/6	0/4/4	1/2/5	4/3/1
(5) RUS					7/1/0	6/2/0	8/0/0	8/0/0	6/2/0	8/0/0	4/4/0	5/2/1	6/2/0	0/4/4	0/5/3	0/6/2	1/4/3	7/1/0
(6) CNN$^-$						6/0/2	6/1/1	8/0/0	3/3/2	8/0/0	2/1/5	0/3/5	4/1/3	0/0/8	0/0/8	0/0/8	0/0/8	5/1/2
(7) TMK$_1$							3/1/4	7/1/0	2/1/5	7/1/0	0/3/5	1/2/5	2/1/5	0/2/6	0/1/7	0/2/6	0/2/6	2/3/3
(8) TMK$_2$								8/0/0	1/2/5	8/0/0	1/0/7	0/0/8	1/3/4	0/0/8	0/0/8	0/0/8	0/0/8	4/3/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										8/0/0	0/4/4	0/3/5	1/4/3	0/3/5	0/0/8	0/2/6	0/2/6	4/2/2
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												3/3/2	5/3/0	0/4/4	0/2/6	0/5/3	1/3/4	6/1/1
(13) EWW$^-$													3/5/0	0/4/4	0/3/5	1/1/6	1/2/5	7/0/1
(14) EWW$^-$+SSM$^-$														0/2/6	0/0/8	0/2/6	0/1/7	5/1/2
(15) SMOTE															0/6/2	3/5/0	2/5/1	7/1/0
(16) gg-SMOTE																3/5/0	1/7/0	7/1/0
(17) ncn-SMOTE																	0/7/1	7/1/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/0/7	0/0/8	1/2/5	0/0/8	1/0/7	4/1/3	0/0/8	8/0/0	2/0/6	7/1/0	1/1/6	0/0/8	2/1/5	0/0/8	0/0/8	0/0/8	0/0/8	0/2/6
(2) SSM$^-$		2/3/3	4/2/2	0/2/6	5/3/0	6/2/0	5/2/1	8/0/0	4/3/1	8/0/0	2/3/3	1/3/4	5/1/2	0/1/7	0/2/6	0/2/6	0/2/6	5/2/1
(3) EW$^-$			3/2/3	1/2/5	3/4/1	6/1/1	8/0/0	8/0/0	5/2/1	8/0/0	3/1/4	0/3/5	3/3/2	1/1/6	0/2/6	1/1/6	1/2/5	7/0/1
(4) EW$^-$+SSM$^-$				0/4/4	3/2/3	5/3/0	4/2/2	8/0/0	6/1/1	8/0/0	0/5/3	2/3/3	7/1/0	0/3/5	0/1/7	0/4/4	1/2/5	3/4/1
(5) RUS					7/1/0	6/2/0	7/1/0	8/0/0	7/1/0	8/0/0	4/4/0	4/3/1	7/1/0	1/3/4	1/4/3	0/6/2	1/4/3	7/1/0
(6) CNN$^-$						6/1/1	5/2/1	8/0/0	3/4/1	8/0/0	2/2/4	0/2/6	4/3/1	0/1/7	0/0/8	0/1/7	0/1/7	5/1/2
(7) TMK$_1$							1/3/4	7/1/0	2/1/5	7/1/0	0/3/5	1/1/6	2/1/5	0/2/6	0/1/7	0/2/6	1/1/6	1/3/4
(8) TMK$_2$								8/0/0	2/3/3	8/0/0	1/2/5	0/0/8	2/3/3	0/0/8	0/0/8	0/0/8	0/0/8	4/3/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										8/0/0	0/2/6	0/3/5	1/4/3	0/0/8	0/0/8	0/0/8	0/1/7	4/1/3
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												1/4/3	5/3/0	0/4/4	1/0/7	0/5/3	1/2/5	5/2/1
(13) EWW$^-$													5/3/0	0/4/4	0/3/5	0/4/4	1/2/5	7/0/1
(14) EWW$^-$+SSM$^-$														0/2/6	0/0/8	0/1/7	0/1/7	3/2/3
(15) SMOTE															0/6/2	3/5/0	2/5/1	7/1/0
(16) gg-SMOTE																4/4/0	1/7/0	7/1/0
(17) ncn-SMOTE																	0/7/1	7/1/0
(18) rng-SMOTE																		7/0/1
(19) b-SMOTE																		

Tabla C.35: *Prueba t−student del clasificador J48 en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

										tvp								
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/1/7	0/1/7	0/1/7	0/1/7	0/3/5
(2) SSM$^-$		4/2/2	0/0/8	2/1/5	2/6/0	0/0/8	5/2/1	0/0/8	0/0/8	0/0/8	0/4/4	2/3/3	0/0/8	3/4/1	3/2/3	3/3/2	3/5/0	7/1/0
(3) EW$^-$			0/0/8	0/3/5	2/4/2	0/0/8	7/1/0	0/0/8	0/0/8	0/0/8	1/1/6	0/2/6	0/0/8	4/1/3	4/1/3	4/1/3	4/1/3	7/0/1
(4) EW$^-$+SSM$^-$				7/1/0	8/0/0	0/1/7	8/0/0	0/0/8	3/4/1	0/0/8	7/1/0	8/0/0	0/3/5	7/1/0	7/1/0	7/1/0	8/0/0	8/0/0
(5) RUS					5/2/1	0/0/8	7/1/0	0/0/8	0/1/7	0/0/8	3/2/3	4/3/1	0/0/8	4/3/1	5/2/1	4/4/0	5/3/0	7/1/0
(6) CNN$^-$						0/0/8	6/1/1	0/0/8	0/0/8	0/0/8	0/2/6	2/3/3	0/0/8	2/3/3	2/3/3	2/5/1	2/4/2	6/2/0
(7) TMK$_1$							8/0/0	0/0/8	7/0/1	0/0/8	7/1/0	8/0/0	7/1/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0
(8) TMK$_2$								0/0/8	0/0/8	0/0/8	0/0/8	0/1/7	0/0/8	0/4/4	1/1/6	1/1/6	1/1/6	5/3/0
(9) CNN$^-$+TMK$_1$									8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN$^-$+TMK$_2$										0/0/8	8/0/0	7/1/0	0/3/5	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(11) TMK$_1$+CNN$^-$											8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK$_2$+CNN$^-$												4/2/2	0/1/7	6/2/0	6/2/0	6/2/0	7/1/0	8/0/0
(13) EWW$^-$													0/0/8	5/1/2	4/2/2	3/3/2	5/2/1	7/1/0
(14) EWW$^-$+SSM$^-$														8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(15) SMOTE															0/7/1	1/6/1	1/7/0	7/1/0
(16) gg-SMOTE																0/8/0	1/7/0	7/1/0
(17) ncn-SMOTE																	0/8/0	6/2/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

										tvn								
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	6/2/0
(2) SSM$^-$		1/2/5	8/0/0	2/5/1	0/3/5	7/0/1	1/1/6	8/0/0	8/0/0	8/0/0	2/4/2	1/3/4	8/0/0	1/1/6	1/1/6	1/1/6	1/2/5	0/2/6
(3) EW$^-$			8/0/0	6/1/1	5/2/1	8/0/0	1/0/7	8/0/0	8/0/0	8/0/0	5/3/0	5/3/0	8/0/0	3/2/3	3/2/3	3/1/4	3/2/3	1/0/7
(4) EW$^-$+SSM$^-$				0/0/8	0/0/8	6/1/1	0/0/8	8/0/0	1/4/3	8/0/0	0/1/7	0/0/8	3/5/0	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7
(5) RUS					1/3/4	8/0/0	0/1/7	8/0/0	8/0/0	8/0/0	3/4/1	1/3/4	8/0/0	1/0/7	1/0/7	0/1/7	1/0/7	0/1/7
(6) CNN$^-$						7/0/1	1/1/6	8/0/0	8/0/0	8/0/0	7/1/0	1/5/2	8/0/0	2/2/4	1/3/4	2/1/5	2/2/4	0/2/6
(7) TMK$_1$							0/0/8	8/0/0	2/0/6	8/0/0	1/1/6	0/0/8	2/0/6	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	6/1/1	6/1/1	6/2/0	5/2/1	0/5/3
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										8/0/0	0/0/8	0/0/8	3/5/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												0/3/5	8/0/0	0/1/7	0/1/7	0/0/8	0/1/7	0/0/8
(13) EWW$^-$													8/0/0	1/2/5	2/2/4	2/1/5	1/2/5	0/1/7
(14) EWW$^-$+SSM$^-$														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															1/6/1	1/6/1	0/8/0	0/2/6
(16) gg-SMOTE																1/7/0	1/6/1	0/1/7
(17) ncn-SMOTE																	0/7/1	0/1/7
(18) rng-SMOTE																		0/2/6
(19) b-SMOTE																		

Tabla C.36: *Prueba $t-$student del clasificador J48 en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	6/2/0	6/2/0	8/0/0	8/0/0	6/2/0	8/0/0	2/6/0	8/0/0	8/0/0	8/0/0	7/1/0	7/1/0	8/0/0	6/2/0	6/2/0	6/2/0	5/2/1	4/4/0
(2) SSM$^-$		2/1/5	7/1/0	2/3/3	0/4/4	7/0/1	0/3/5	8/0/0	5/3/0	8/0/0	2/4/2	2/2/4	7/1/0	1/2/5	1/2/5	1/2/5	1/1/6	0/3/5
(3) EW$^-$			8/0/0	6/2/0	5/1/2	8/0/0	1/3/4	8/0/0	7/1/0	8/0/0	5/3/0	3/5/0	8/0/0	2/3/3	2/3/3	3/2/3	2/3/3	1/4/3
(4) EW$^-$+SSM$^-$				0/1/7	0/0/8	6/1/1	0/0/8	8/0/0	1/4/3	8/0/0	0/1/7	0/0/8	2/5/1	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7
(5) RUS					1/4/3	8/0/0	0/2/6	8/0/0	7/1/0	8/0/0	3/5/0	0/4/4	8/0/0	1/0/7	1/1/6	0/1/7	1/0/7	0/1/7
(6) CNN$^-$						7/0/1	1/2/5	8/0/0	8/0/0	8/0/0	6/2/0	2/3/3	8/0/0	1/2/5	1/2/5	1/3/4	1/3/4	0/3/5
(7) TMK$_1$							0/0/8	8/0/0	2/0/6	8/0/0	1/1/6	0/0/8	2/0/6	0/1/7	0/1/7	0/0/8	0/1/7	0/0/8
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	7/1/0	7/1/0	8/0/0	3/5/0	5/1/2	2/6/0	3/4/1	1/7/0
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/1/0	0/2/6	0/1/7	3/5/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												0/3/5	7/1/0	0/2/6	0/2/6	0/0/8	0/2/6	0/0/8
(13) EWW$^-$													8/0/0	1/5/2	1/4/3	1/3/4	1/3/4	0/2/6
(14) EWW$^-$+SSM$^-$														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/8/0	1/6/1	0/8/0	0/6/2
(16) gg-SMOTE																1/7/0	1/6/1	0/7/1
(17) ncn-SMOTE																	1/6/1	0/6/2
(18) rng-SMOTE																		1/6/1
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/3/3	1/1/6	2/1/5	1/1/6	1/5/2	3/3/2	0/3/5	6/0/2	2/2/4	4/2/2	1/3/4	1/1/6	2/0/6	0/2/6	0/3/5	0/4/4	0/1/7	0/7/1
(2) SSM$^-$		1/2/5	2/3/3	1/1/6	1/5/2	2/4/2	0/6/2	7/1/0	1/4/3	4/3/1	0/4/4	1/1/6	1/4/3	0/4/4	0/5/3	0/4/4	0/4/4	1/5/2
(3) EW$^-$			5/2/1	1/6/1	3/4/1	5/2/1	0/7/1	6/2/0	3/5/0	6/2/0	3/4/1	0/7/1	4/4/0	1/6/1	2/5/1	1/6/1	1/5/2	6/1/1
(4) EW$^-$+SSM$^-$				0/4/4	1/5/2	3/4/1	0/6/2	7/1/0	0/6/2	5/3/0	1/4/3	0/3/5	1/6/1	1/3/4	1/3/4	2/1/5	2/2/4	3/3/2
(5) RUS					4/3/1	6/2/0	0/7/1	6/2/0	4/4/0	6/2/0	2/6/0	0/6/2	4/4/0	1/5/2	1/4/3	2/4/2	2/4/2	4/2/2
(6) CNN$^-$						3/2/3	1/5/2	6/1/1	2/5/1	5/2/1	2/4/2	1/2/5	3/4/1	0/4/4	0/4/4	0/5/3	0/4/4	2/5/1
(7) TMK$_1$							1/3/4	7/1/0	2/3/3	5/3/0	1/3/4	0/2/6	2/1/5	0/3/5	0/3/5	1/1/6	1/1/6	1/5/2
(8) TMK$_2$								6/2/0	3/5/0	6/2/0	1/6/1	1/5/2	2/6/0	0/6/2	0/6/2	1/5/2	0/5/3	4/4/0
(9) CNN$^-$+TMK$_1$									0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	1/0/7	0/2/6	1/0/7
(10) CNN$^-$+TMK$_2$										6/2/0	1/3/4	0/4/4	1/6/1	1/3/4	2/1/5	2/1/5	1/2/5	3/3/2
(11) TMK$_1$+CNN$^-$											0/2/6	0/2/6	0/3/5	0/2/6	0/2/6	1/1/6	1/1/6	1/4/3
(12) TMK$_2$+CNN$^-$												1/3/4	3/3/2	1/6/1	1/4/3	1/5/2	0/5/3	4/3/1
(13) EWW$^-$													5/3/0	1/5/2	2/5/1	2/5/1	3/3/2	5/2/1
(14) EWW$^-$+SSM$^-$														1/4/3	1/4/3	2/2/4	2/3/3	4/2/2
(15) SMOTE															0/8/0	2/6/0	1/7/0	5/3/0
(16) gg-SMOTE																1/7/0	0/7/1	4/4/0
(17) ncn-SMOTE																	1/6/1	3/5/0
(18) rng-SMOTE																		5/3/0
(19) b-SMOTE																		

Tabla C.37: *Prueba t−student del clasificador J48 en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/3/4	0/1/7	1/2/5	0/1/7	1/2/5	4/2/2	0/2/6	8/0/0	1/2/5	6/2/0	1/0/7	0/1/7	1/2/5	0/1/7	0/1/7	0/2/6	0/1/7	0/7/1
(2) SSM$^-$		1/3/4	2/4/2	1/1/6	1/5/2	4/2/2	1/5/2	8/0/0	1/4/3	8/0/0	0/4/4	1/1/6	2/5/1	1/3/4	1/3/4	1/3/4	0/4/4	2/5/1
(3) EW$^-$			2/6/0	2/4/2	3/4/1	6/1/1	3/5/0	8/0/0	2/5/1	7/1/0	2/5/1	0/4/4	3/5/0	2/2/4	3/1/4	2/3/3	1/4/3	6/2/0
(4) EW$^-$+SSM$^-$				0/4/4	1/5/2	5/1/2	2/5/1	8/0/0	1/5/2	7/1/0	0/3/5	0/3/5	1/6/1	1/2/5	1/2/5	2/1/5	1/2/5	4/3/1
(5) RUS					4/3/1	6/2/0	3/5/0	8/0/0	4/4/0	7/1/0	2/5/1	0/6/2	4/4/0	2/4/2	1/5/2	2/4/2	2/4/2	5/3/0
(6) CNN$^-$						4/2/2	2/4/2	8/0/0	3/4/1	7/1/0	2/3/3	1/2/5	3/3/2	2/1/5	2/0/6	2/2/4	0/4/4	3/4/1
(7) TMK$_1$							1/3/4	8/0/0	2/1/5	8/0/0	1/3/4	0/1/7	2/0/6	0/3/5	0/2/6	0/2/6	1/1/6	1/3/4
(8) TMK$_2$								8/0/0	2/5/1	7/1/0	1/4/3	0/3/5	3/4/1	0/5/3	0/3/5	1/2/5	0/5/3	5/3/0
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/1/0	0/4/4	0/3/5	1/7/0	1/3/4	1/3/4	2/2/4	2/3/3	4/3/1
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/1/7	0/0/8
(12) TMK$_2$+CNN$^-$												1/3/4	4/3/1	1/6/1	1/5/2	2/4/2	1/6/1	6/1/1
(13) EWW$^-$													3/5/0	2/4/2	2/5/1	2/5/1	3/3/2	6/2/0
(14) EWW$^-$+SSM$^-$														1/2/5	1/3/4	2/2/4	0/4/4	4/3/1
(15) SMOTE															0/8/0	1/6/1	1/7/0	6/2/0
(16) gg-SMOTE																0/8/0	1/7/0	6/2/0
(17) ncn-SMOTE																	1/7/0	5/3/0
(18) rng-SMOTE																		5/3/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/3/4	0/2/6	2/2/4	0/2/6	1/4/3	5/2/1	0/2/6	8/0/0	1/3/4	6/2/0	1/1/6	0/1/7	2/3/3	0/2/6	0/1/7	0/2/6	0/1/7	0/7/1
(2) SSM$^-$		1/2/5	3/4/1	1/2/5	1/5/2	4/2/2	1/4/3	8/0/0	3/2/3	8/0/0	0/4/4	1/1/6	3/4/1	0/4/4	0/4/4	0/4/4	0/4/4	3/4/1
(3) EW$^-$			5/3/0	2/4/2	3/4/1	7/0/1	2/6/0	8/0/0	3/5/0	8/0/0	2/5/1	0/5/3	4/4/0	1/3/4	3/1/4	2/4/2	1/5/2	6/2/0
(4) EW$^-$+SSM$^-$				0/2/6	1/5/2	5/1/2	0/5/3	8/0/0	1/5/2	7/1/0	0/4/4	0/1/7	1/6/1	1/2/5	0/2/6	0/2/6	0/3/5	3/3/2
(5) RUS					3/4/1	7/1/0	3/5/0	8/0/0	5/3/0	8/0/0	2/5/1	0/6/2	6/2/0	2/4/2	1/5/2	2/4/2	2/4/2	4/4/0
(6) CNN$^-$						4/2/2	1/4/3	8/0/0	4/3/1	8/0/0	2/4/2	1/2/5	4/3/1	2/1/5	0/3/5	0/5/3	0/4/4	3/4/1
(7) TMK$_1$							1/2/5	8/0/0	2/0/6	8/0/0	1/3/4	0/1/7	2/0/6	0/1/7	0/2/6	0/2/6	1/1/6	0/3/5
(8) TMK$_2$								8/0/0	3/5/0	8/0/0	1/6/1	0/4/4	4/4/0	0/6/2	0/4/4	0/4/4	0/5/3	5/3/0
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/1/0	0/3/5	0/3/5	1/7/0	1/2/5	1/1/6	1/1/6	0/3/5	4/1/3
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												1/4/3	4/3/1	1/6/1	1/5/2	2/4/2	0/6/2	4/3/1
(13) EWW$^-$													5/3/0	2/4/2	1/6/1	2/5/1	2/4/2	6/2/0
(14) EWW$^-$+SSM$^-$														0/3/5	0/3/5	0/3/5	0/4/4	2/3/3
(15) SMOTE															0/8/0	1/6/1	1/7/0	5/3/0
(16) gg-SMOTE																1/7/0	1/7/0	5/3/0
(17) ncn-SMOTE																	0/7/1	4/4/0
(18) rng-SMOTE																		5/3/0
(19) b-SMOTE																		

Tabla C.38: *Prueba $t-$student del clasificador MLP en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/1/7	0/0/8	0/1/7	0/1/7	0/0/8	0/3/5	0/0/8	0/0/8	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	2/2/4
(2) SSM$^-$		2/4/2	0/1/7	2/2/4	0/8/0	0/0/8	6/1/1	0/0/8	0/1/7	0/0/8	0/3/5	2/2/4	0/1/7	1/5/2	1/5/2	1/5/2	1/5/2	5/3/0
(3) EW$^-$			0/0/8	2/2/4	3/3/2	0/0/8	7/1/0	0/0/8	0/2/6	0/0/8	2/2/4	0/5/3	0/0/8	2/1/5	1/2/5	2/2/4	1/2/5	6/1/1
(4) EW$^-$+SSM$^-$				6/2/0	7/1/0	0/1/7	8/0/0	0/0/8	3/2/3	0/0/8	6/1/1	7/1/0	0/3/5	6/2/0	7/1/0	7/1/0	7/1/0	8/0/0
(5) RUS					4/2/2	0/1/7	7/1/0	0/0/8	0/3/5	0/0/8	2/3/3	3/3/2	0/0/8	3/3/2	3/3/2	3/3/2	3/3/2	7/1/0
(6) CNN$^-$						0/1/7	6/1/1	0/0/8	0/1/7	0/0/8	0/3/5	1/2/5	0/2/6	1/3/4	1/4/3	1/5/2	1/3/4	5/1/2
(7) TMK$_1$							8/0/0	0/0/8	6/1/1	0/2/6	7/1/0	8/0/0	6/1/1	7/1/0	7/1/0	7/1/0	7/1/0	8/0/0
(8) TMK$_2$								0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7	0/1/7	0/1/7	1/1/6	0/7/1
(9) CNN$^-$+TMK$_1$									8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN$^-$+TMK$_2$										0/1/7	4/4/0	6/2/0	0/3/5	6/2/0	6/2/0	6/2/0	6/2/0	8/0/0
(11) TMK$_1$+CNN$^-$											8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK$_2$+CNN$^-$												4/1/3	0/2/6	4/3/1	4/3/1	5/2/1	4/3/1	7/0/1
(13) EWW$^-$													0/0/8	3/2/3	3/1/4	3/4/1	3/1/4	8/0/0
(14) EWW$^-$+SSM$^-$														7/1/0	7/1/0	7/1/0	7/1/0	8/0/0
(15) SMOTE															1/7/0	1/7/0	0/7/1	7/1/0
(16) gg-SMOTE																0/8/0	0/8/0	7/1/0
(17) ncn-SMOTE																	0/8/0	7/1/0
(18) rng-SMOTE																		6/2/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	7/1/0	8/0/0	7/1/0	8/0/0	8/0/0	5/2/1	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	5/2/1
(2) SSM$^-$		2/0/6	7/1/0	2/4/2	1/4/3	7/0/1	0/1/7	8/0/0	6/2/0	7/1/0	4/4/0	2/3/3	6/2/0	2/3/3	2/1/5	1/4/3	2/3/3	0/2/6
(3) EW$^-$			8/0/0	5/2/1	3/3/2	8/0/0	0/1/7	8/0/0	7/1/0	8/0/0	6/2/0	5/3/0	8/0/0	5/2/1	5/1/2	4/2/2	5/2/1	1/1/6
(4) EW$^-$+SSM$^-$				0/1/7	0/1/7	6/0/2	0/0/8	8/0/0	1/3/4	7/1/0	0/2/6	0/0/8	4/4/0	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/1/5	7/0/1	0/1/7	8/0/0	7/1/0	8/0/0	3/4/1	1/2/5	8/0/0	2/3/3	1/3/4	1/3/4	1/3/4	0/1/7
(6) CNN$^-$						7/0/1	1/0/7	8/0/0	6/2/0	7/0/1	6/2/0	4/1/3	7/1/0	4/2/2	2/4/2	1/5/2	2/5/1	1/1/6
(7) TMK$_1$							0/0/8	8/0/0	2/0/6	8/0/0	1/0/7	0/0/8	2/0/6	0/1/7	0/1/7	0/1/7	0/1/7	0/0/8
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	7/1/0	7/1/0	7/1/0	7/1/0	1/7/0
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										6/2/0	0/3/5	0/1/7	5/3/0	0/2/6	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											1/0/7	0/0/8	0/2/6	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												1/3/4	6/2/0	1/2/5	1/2/5	1/2/5	1/2/5	0/1/7
(13) EWW$^-$													8/0/0	3/2/3	2/3/3	3/2/3	3/2/3	0/2/6
(14) EWW$^-$+SSM$^-$														0/1/7	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/7/1	0/6/2	0/8/0	0/1/7
(16) gg-SMOTE																0/8/0	1/7/0	0/1/7
(17) ncn-SMOTE																	1/7/0	0/1/7
(18) rng-SMOTE																		0/1/7
(19) b-SMOTE																		

Tabla C.39: *Prueba t−student del clasificador MLP en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	6/2/0	8/0/0	7/1/0	7/1/0	8/0/0	2/6/0	8/0/0	8/0/0	8/0/0	8/0/0	6/2/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	6/2/0
(2) SSM$^-$		1/1/6	7/1/0	2/3/3	1/5/2	6/1/1	0/1/7	8/0/0	4/4/0	7/0/1	1/7/0	0/4/4	6/2/0	0/5/3	0/3/5	0/4/4	0/5/3	0/2/6
(3) EW$^-$			8/0/0	7/1/0	4/3/1	8/0/0	0/3/5	8/0/0	7/1/0	8/0/0	6/2/0	5/3/0	8/0/0	4/3/1	5/2/1	4/3/1	5/3/0	1/3/4
(4) EW$^-$+SSM$^-$				0/2/6	1/0/7	6/0/2	0/0/8	8/0/0	1/5/2	7/0/1	0/2/6	0/0/8	3/5/0	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					2/3/3	7/0/1	0/1/7	8/0/0	7/1/0	8/0/0	4/3/1	1/2/5	7/1/0	0/4/4	0/4/4	1/3/4	0/5/3	0/2/6
(6) CNN$^-$						7/0/1	0/2/6	8/0/0	5/3/0	7/0/1	3/5/0	2/3/3	6/2/0	1/5/2	1/5/2	1/4/3	2/4/2	1/1/6
(7) TMK$_1$							0/0/8	8/0/0	2/0/6	8/0/0	1/1/6	0/0/8	2/0/6	1/0/7	1/0/7	0/1/7	0/1/7	0/1/7
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	7/1/0	6/2/0	7/1/0	8/0/0	1/7/0
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										6/2/0	0/4/4	0/1/7	3/5/0	0/2/6	0/0/8	0/0/8	0/1/7	0/0/8
(11) TMK$_1$+CNN$^-$											1/0/7	0/0/8	1/1/6	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												0/3/5	6/2/0	1/2/5	1/0/7	1/1/6	1/3/4	0/1/7
(13) EWW$^-$													8/0/0	3/5/0	2/4/2	3/2/3	3/4/1	0/2/6
(14) EWW$^-$+SSM$^-$														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															0/7/1	0/8/0	0/8/0	0/2/6
(16) gg-SMOTE																0/8/0	0/8/0	0/4/4
(17) ncn-SMOTE																	0/8/0	0/4/4
(18) rng-SMOTE																		0/2/6
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/3/2	0/3/5	3/2/3	0/4/4	2/4/2	3/2/3	0/6/2	6/0/2	2/3/3	5/1/2	2/4/2	0/3/5	3/2/3	1/3/4	0/3/5	1/3/4	1/2/5	1/5/2
(2) SSM$^-$		0/3/5	1/3/4	0/4/4	0/8/0	5/0/3	1/4/3	6/0/2	0/7/1	5/0/3	0/8/0	0/1/7	0/5/3	0/2/6	0/2/6	0/2/6	0/3/5	1/4/3
(3) EW$^-$			4/4/0	3/3/2	5/3/0	5/1/2	5/3/0	6/2/0	6/2/0	6/2/0	4/4/0	0/6/2	4/4/0	2/4/2	1/5/2	2/5/1	3/3/2	5/3/0
(4) EW$^-$+SSM$^-$				1/3/4	3/3/2	5/1/2	2/3/3	7/1/0	1/6/1	5/2/1	2/3/3	0/3/5	0/8/0	1/3/4	1/3/4	1/3/4	1/3/4	2/4/2
(5) RUS					3/5/0	6/0/2	3/5/0	7/0/1	3/5/0	6/1/1	4/4/0	2/3/3	3/4/1	0/7/1	0/6/2	0/7/1	0/7/1	4/4/0
(6) CNN$^-$						4/1/3	2/2/4	6/1/1	0/6/2	5/1/2	0/7/1	0/2/6	0/6/2	0/4/4	0/3/5	0/4/4	0/4/4	1/5/2
(7) TMK$_1$							3/1/4	8/0/0	3/0/5	7/1/0	3/0/5	0/3/5	2/1/5	2/1/5	2/0/6	2/0/6	1/1/6	3/2/3
(8) TMK$_2$								6/1/1	3/4/1	6/0/2	2/4/2	0/4/4	3/3/2	0/5/3	0/4/4	1/4/3	1/4/3	0/7/1
(9) CNN$^-$+TMK$_1$									0/2/6	0/0/8	0/2/6	0/0/8	0/0/8	1/0/7	0/1/7	0/2/6	1/0/7	0/2/6
(10) CNN$^-$+TMK$_2$										5/2/1	0/8/0	0/2/6	1/5/2	1/3/4	0/3/5	0/4/4	1/2/5	2/4/2
(11) TMK$_1$+CNN$^-$											3/0/5	0/2/6	1/2/5	1/1/6	1/1/6	1/1/6	1/1/6	1/1/6
(12) TMK$_2$+CNN$^-$												0/4/4	1/5/2	0/6/2	0/3/5	0/5/3	0/5/3	2/4/2
(13) EWW$^-$													3/5/0	2/5/1	2/5/1	3/4/1	2/5/1	5/3/0
(14) EWW$^-$+SSM$^-$														1/3/4	1/4/3	1/3/4	1/3/4	4/1/3
(15) SMOTE															0/8/0	0/8/0	0/7/1	4/4/0
(16) gg-SMOTE																0/8/0	0/8/0	4/4/0
(17) ncn-SMOTE																	0/8/0	4/4/0
(18) rng-SMOTE																		4/4/0
(19) b-SMOTE																		

Tabla C.40: *Prueba $t-$student del clasificador MLP en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/3/3	0/2/6	2/1/5	0/2/6	2/3/3	4/2/2	0/4/4	8/0/0	2/1/5	6/1/1	2/2/4	0/1/7	2/1/5	0/2/6	0/2/6	1/1/6	0/2/6	2/4/2
(2) SSM$^-$		0/3/5	1/5/2	0/3/5	1/7/0	5/1/2	2/4/2	8/0/0	0/7/1	5/2/1	0/7/1	0/1/7	0/7/1	0/3/5	0/2/6	0/3/5	0/3/5	1/5/2
(3) EW$^-$			3/5/0	2/3/3	5/3/0	6/0/2	6/2/0	8/0/0	4/4/0	7/1/0	4/4/0	0/6/2	4/4/0	1/5/2	1/5/2	2/4/2	2/4/2	7/1/0
(4) EW$^-$+SSM$^-$				1/4/3	3/5/0	5/1/2	2/4/2	8/0/0	1/7/0	6/1/1	1/6/1	0/2/6	0/8/0	1/4/3	0/4/4	0/4/4	1/3/4	2/4/2
(5) RUS					5/3/0	5/2/1	4/4/0	8/0/0	4/4/0	7/1/0	4/4/0	2/4/2	3/4/1	0/7/1	0/6/2	0/8/0	0/7/1	5/3/0
(6) CNN$^-$						5/1/2	2/3/3	8/0/0	0/7/1	5/2/1	0/7/1	0/1/7	0/6/2	0/4/4	0/4/4	0/3/5	0/3/5	1/6/1
(7) TMK$_1$							2/1/5	8/0/0	2/1/5	8/0/0	2/1/5	1/1/6	2/1/5	1/2/5	1/2/5	1/2/5	1/2/5	2/1/5
(8) TMK$_2$								8/0/0	2/3/3	7/0/1	2/4/2	0/2/6	2/4/2	1/3/4	0/3/5	0/4/4	1/2/5	0/7/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										6/1/1	0/8/0	0/3/5	0/8/0	0/4/4	0/2/6	0/3/5	2/1/5	3/2/3
(11) TMK$_1$+CNN$^-$											1/2/5	0/0/8	1/1/6	0/1/7	0/1/7	0/1/7	0/1/7	0/2/6
(12) TMK$_2$+CNN$^-$												0/4/4	0/8/0	0/6/2	0/3/5	0/5/3	0/4/4	3/3/2
(13) EWW$^-$													7/1/0	2/5/1	2/5/1	3/4/1	2/5/1	8/0/0
(14) EWW$^-$+SSM$^-$														1/3/4	0/2/6	0/3/5	1/1/6	1/5/2
(15) SMOTE															0/8/0	0/8/0	0/7/1	4/4/0
(16) gg-SMOTE																0/8/0	0/8/0	4/4/0
(17) ncn-SMOTE																	0/8/0	4/4/0
(18) rng-SMOTE																		5/3/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/3/3	0/2/6	2/3/3	0/2/6	2/3/3	5/1/2	0/5/3	8/0/0	2/3/3	6/1/1	2/2/4	0/2/6	2/4/2	1/1/6	0/2/6	1/1/6	1/2/5	2/4/2
(2) SSM$^-$		0/3/5	1/5/2	0/3/5	1/7/0	5/1/2	2/4/2	8/0/0	0/7/1	5/2/1	0/7/1	0/2/6	0/7/1	0/4/4	0/2/6	0/2/6	0/3/5	1/4/3
(3) EW$^-$			6/2/0	2/4/2	5/3/0	6/0/2	4/4/0	8/0/0	7/1/0	7/1/0	4/4/0	0/6/2	7/1/0	1/6/1	1/5/2	2/4/2	2/4/2	7/1/0
(4) EW$^-$+SSM$^-$				1/4/3	2/5/1	6/0/2	1/5/2	8/0/0	1/6/1	7/0/1	1/5/2	0/2/6	0/8/0	0/4/4	0/3/5	0/3/5	1/1/6	1/5/2
(5) RUS					5/3/0	5/2/1	4/4/0	8/0/0	5/3/0	7/1/0	4/4/0	2/3/3	6/1/1	0/7/1	0/6/2	0/7/1	0/7/1	4/4/0
(6) CNN$^-$						5/1/2	2/3/3	8/0/0	1/6/1	5/2/1	0/7/1	0/3/5	0/6/2	0/5/3	0/4/4	0/2/6	0/3/5	1/6/1
(7) TMK$_1$							2/1/5	8/0/0	2/1/5	8/0/0	2/1/5	1/1/6	2/0/6	1/2/5	1/2/5	1/2/5	1/1/6	2/1/5
(8) TMK$_2$								8/0/0	3/5/0	7/0/1	2/4/2	0/3/5	2/5/1	1/4/3	0/4/4	0/4/4	1/4/3	0/7/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										6/1/1	0/8/0	0/2/6	1/6/1	0/4/4	0/2/6	0/3/5	0/3/5	1/4/3
(11) TMK$_1$+CNN$^-$											1/2/5	0/0/8	1/1/6	0/1/7	0/0/8	0/1/7	0/1/7	0/1/7
(12) TMK$_2$+CNN$^-$												0/4/4	3/5/0	0/6/2	0/3/5	0/4/4	0/5/3	1/5/2
(13) EWW$^-$													8/0/0	2/5/1	2/5/1	3/4/1	2/5/1	6/2/0
(14) EWW$^-$+SSM$^-$														0/3/5	0/2/6	0/2/6	1/1/6	1/5/2
(15) SMOTE															0/8/0	0/8/0	0/7/1	4/4/0
(16) gg-SMOTE																0/8/0	0/8/0	4/4/0
(17) ncn-SMOTE																	0/8/0	4/4/0
(18) rng-SMOTE																		4/4/0
(19) b-SMOTE																		

Tabla C.41: *Prueba t−student del clasificador NBC en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

	tvp																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/0/6	0/0/8	2/0/6	0/1/7	2/0/6	0/0/8	0/0/8	0/0/8	1/1/6	0/0/8	1/1/6	0/0/8	1/1/6	1/0/7	1/0/7	1/0/7	1/0/7	3/3/2
(2) SSM^-		1/1/6	1/0/7	1/0/7	3/1/4	0/0/8	5/1/2	0/0/8	0/0/8	0/0/8	0/0/8	1/1/6	1/0/7	1/0/7	1/1/6	1/1/6	1/0/7	5/0/3
(3) EW^-			2/0/6	3/1/4	5/2/1	0/0/8	7/1/0	0/0/8	3/0/5	0/0/8	3/1/4	0/2/6	2/0/6	3/1/4	3/1/4	3/1/4	3/1/4	8/0/0
(4) EW^-+SSM^-				5/0/3	7/0/1	0/1/7	6/0/2	0/0/8	1/2/5	0/0/8	3/3/2	5/1/2	0/2/6	4/1/3	4/1/3	4/1/3	4/1/3	7/0/1
(5) RUS					7/0/1	0/0/8	7/0/1	0/0/8	2/0/6	0/0/8	3/3/2	4/1/3	2/0/6	1/5/2	1/5/2	3/4/1	2/3/3	6/2/0
(6) CNN^-						0/0/8	5/0/3	0/0/8	0/0/8	0/0/8	0/0/8	1/1/6	1/0/7	1/0/7	1/0/7	1/0/7	1/0/7	5/0/3
(7) TMK_1							8/0/0	1/0/7	6/0/2	2/1/5	7/1/0	8/0/0	6/2/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(8) TMK_2								0/0/8	1/1/6	0/0/8	2/0/6	0/1/7	2/0/6	1/0/7	1/0/7	1/0/7	1/0/7	6/1/1
(9) CNN^-+TMK_1									7/1/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN^-+TMK_2										1/0/7	7/1/0	5/0/3	5/1/2	4/3/1	4/2/2	5/2/1	4/3/1	7/0/1
(11) TMK_1+CNN^-											8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK_2+CNN^-												3/2/3	1/1/6	2/3/3	3/2/3	3/2/3	2/3/3	7/0/1
(13) EWW^-													2/0/6	3/1/4	3/1/4	4/0/4	3/1/4	8/0/0
(14) EWW^-+SSM^-														5/0/3	5/0/3	5/1/2	5/1/2	7/0/1
(15) SMOTE															1/7/0	4/4/0	1/7/0	6/1/1
(16) gg-SMOTE																4/4/0	1/7/0	6/1/1
(17) ncn-SMOTE																	0/5/3	6/1/1
(18) rng-SMOTE																		6/1/1
(19) b-SMOTE																		

	tvn																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/0/1	7/0/1	7/0/1	7/0/1	6/1/1	8/0/0	7/0/1	8/0/0	8/0/0	8/0/0	8/0/0	7/0/1	7/0/1	7/0/1	7/0/1	7/0/1	7/0/1	4/1/3
(2) SSM^-		4/1/3	7/0/1	6/0/2	3/1/4	6/1/1	1/1/6	8/0/0	7/1/0	7/1/0	7/1/0	5/1/2	7/0/1	6/0/2	5/1/2	6/0/2	6/0/2	1/1/6
(3) EW^-			6/1/1	5/2/1	3/1/4	8/0/0	1/0/7	8/0/0	6/1/1	8/0/0	5/1/2	6/2/0	7/0/1	4/2/2	5/1/2	4/2/2	4/2/2	1/1/6
(4) EW^-+SSM^-				3/0/5	2/0/6	7/0/1	1/0/7	8/0/0	4/2/2	8/0/0	4/0/4	2/1/5	6/1/1	4/0/4	4/0/4	4/0/4	4/0/4	1/0/7
(5) RUS					2/0/6	8/0/0	0/1/7	8/0/0	7/0/1	8/0/0	3/2/3	2/1/5	5/1/2	2/4/2	3/4/1	2/4/2	3/3/2	1/1/6
(6) CNN^-						6/0/2	3/1/4	8/0/0	7/0/1	7/1/0	7/1/0	5/1/2	7/0/1	6/0/2	6/0/2	6/0/2	6/0/2	1/2/5
(7) TMK_1							0/0/8	8/0/0	3/0/5	6/1/1	2/1/5	0/0/8	2/0/6	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(8) TMK_2								8/0/0	8/0/0	8/0/0	7/1/0	7/0/1	7/0/1	7/0/1	7/0/1	7/0/1	7/0/1	2/1/5
(9) CNN^-+TMK_1									0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										6/1/1	0/1/7	2/0/6	3/1/4	2/1/5	3/0/5	2/0/6	2/1/5	0/0/8
(11) TMK_1+CNN^-											0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												2/1/5	5/1/2	3/2/3	3/2/3	3/2/3	3/2/3	0/0/8
(13) EWW^-													6/1/1	4/2/2	5/1/2	4/2/2	4/2/2	1/0/7
(14) EWW^-+SSM^-														4/0/4	4/0/4	2/2/4	4/0/4	1/0/7
(15) SMOTE															2/5/1	0/4/4	0/7/1	2/0/6
(16) gg-SMOTE																0/4/4	1/4/3	2/0/6
(17) ncn-SMOTE																	3/5/0	2/0/6
(18) rng-SMOTE																		2/0/6
(19) b-SMOTE																		

Tabla C.42: *Prueba $t-$student del clasificador NBC en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/4/0	5/2/1	7/0/1	5/2/1	4/3/1	8/0/0	4/3/1	8/0/0	8/0/0	8/0/0	7/1/0	5/2/1	7/0/1	5/2/1	5/2/1	5/2/1	5/2/1	5/2/1
(2) SSM$^-$		2/4/2	7/0/1	3/3/2	2/4/2	6/0/2	1/3/4	7/1/0	7/0/1	7/0/1	4/3/1	4/2/2	7/0/1	3/3/2	3/3/2	4/2/2	3/3/2	2/2/4
(3) EW$^-$			7/0/1	6/1/1	3/3/2	8/0/0	1/2/5	8/0/0	6/1/1	8/0/0	4/3/1	5/3/0	7/0/1	5/2/1	5/1/2	5/1/2	5/2/1	3/0/5
(4) EW$^-$+SSM$^-$				3/2/3	2/0/6	6/1/1	1/0/7	8/0/0	4/2/2	8/0/0	4/1/3	1/3/4	5/2/1	3/2/3	3/2/3	3/2/3	3/2/3	1/2/5
(5) RUS					2/2/4	6/2/0	0/3/5	8/0/0	6/1/1	8/0/0	3/3/2	1/3/4	5/2/1	2/5/1	2/5/1	1/5/2	2/5/1	2/3/3
(6) CNN$^-$						6/0/2	2/2/4	7/1/0	7/0/1	7/0/1	5/2/1	4/2/2	6/1/1	4/2/2	4/2/2	4/2/2	4/2/2	3/0/5
(7) TMK$_1$							0/0/8	8/0/0	3/0/5	6/1/1	3/0/5	0/1/7	2/0/6	0/1/7	0/1/7	0/0/8	0/1/7	0/1/7
(8) TMK$_2$								8/0/0	7/1/0	8/0/0	6/2/0	6/1/1	7/0/1	5/3/0	5/3/0	5/3/0	5/3/0	3/3/2
(9) CNN$^-$+TMK$_1$									1/1/6	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										6/0/2	0/1/7	1/1/6	3/2/3	2/1/5	2/2/4	1/2/5	2/2/4	0/2/6
(11) TMK$_1$+CNN$^-$											1/1/6	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												2/2/4	3/3/2	3/2/3	3/2/3	3/2/3	3/2/3	1/1/6
(13) EWW$^-$													7/0/1	5/1/2	5/1/2	4/2/2	5/1/2	2/1/5
(14) EWW$^-$+SSM$^-$														2/2/4	3/1/4	2/3/3	2/2/4	1/1/6
(15) SMOTE															1/6/1	1/6/1	0/8/0	2/1/5
(16) gg-SMOTE																0/6/2	0/6/2	2/1/5
(17) ncn-SMOTE																	3/4/1	2/1/5
(18) rng-SMOTE																		2/1/5
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/0/5	1/1/6	2/1/5	1/1/6	3/1/4	3/3/2	1/2/5	6/1/1	3/0/5	5/1/2	3/0/5	1/2/5	2/1/5	1/2/5	1/3/4	1/2/5	1/2/5	4/2/2
(2) SSM$^-$		0/2/6	1/2/5	0/2/6	3/5/0	2/2/4	2/3/3	5/2/1	1/2/5	5/0/3	1/2/5	0/2/6	2/1/5	0/3/5	0/3/5	0/3/5	0/3/5	4/1/3
(3) EW$^-$			2/4/2	3/4/1	6/2/0	6/1/1	6/2/0	7/1/0	3/4/1	6/1/1	3/4/1	1/6/1	3/2/3	3/4/1	3/5/0	4/4/0	3/5/0	7/1/0
(4) EW$^-$+SSM$^-$				2/4/2	5/1/2	4/2/2	4/2/2	7/1/0	3/2/3	5/2/1	3/2/3	1/5/2	1/5/2	3/3/2	3/3/2	4/3/1	3/3/2	7/0/1
(5) RUS					7/1/0	5/1/2	4/3/1	7/1/0	2/6/0	6/1/1	2/6/0	1/4/3	3/3/2	1/7/0	2/6/0	2/6/0	1/7/0	7/0/1
(6) CNN$^-$						3/1/4	1/3/4	5/2/1	1/2/5	5/0/3	1/2/5	0/2/6	1/1/6	1/1/6	0/2/6	1/1/6	1/1/6	4/0/4
(7) TMK$_1$							1/3/4	8/0/0	3/1/4	4/4/0	3/1/4	1/1/6	1/4/3	2/1/5	2/2/4	2/2/4	2/2/4	4/3/1
(8) TMK$_2$								6/1/1	3/0/5	6/1/1	3/1/4	0/3/5	2/2/4	1/3/4	2/3/3	1/4/3	1/3/4	4/2/2
(9) CNN$^-$+TMK$_1$									1/1/6	0/0/8	1/2/5	0/1/7	0/1/7	0/1/7	0/2/6	1/1/6	0/1/7	2/1/5
(10) CNN$^-$+TMK$_2$										5/1/2	0/7/1	2/3/3	2/3/3	1/4/3	1/4/3	1/4/3	1/4/3	5/1/2
(11) TMK$_1$+CNN$^-$											2/1/5	1/1/6	0/3/5	1/1/6	1/1/6	1/1/6	1/1/6	3/2/3
(12) TMK$_2$+CNN$^-$												2/3/3	2/3/3	1/4/3	1/4/3	1/4/3	1/4/3	5/1/2
(13) EWW$^-$													2/4/2	4/3/1	4/2/2	5/3/0	4/2/2	7/0/1
(14) EWW$^-$+SSM$^-$														3/3/2	3/3/2	4/1/3	3/3/2	6/1/1
(15) SMOTE															2/6/0	2/6/0	0/8/0	7/0/1
(16) gg-SMOTE																1/7/0	0/7/1	6/1/1
(17) ncn-SMOTE																	0/7/1	5/2/1
(18) rng-SMOTE																		7/0/1
(19) b-SMOTE																		

Tabla C.43: *Prueba t−student del clasificador NBC en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/0/5	0/2/6	2/1/5	1/1/6	3/1/4	4/2/2	0/3/5	7/0/1	2/1/5	6/0/2	2/1/5	0/2/6	2/1/5	1/1/6	1/0/7	1/1/6	1/1/6	3/3/2
(2) SSM$^-$		0/2/6	1/2/5	0/2/6	3/5/0	4/0/4	3/2/3	6/1/1	1/1/6	5/0/3	1/1/6	0/2/6	2/1/5	0/3/5	0/3/5	0/3/5	0/3/5	4/1/3
(3) EW$^-$			3/1/4	3/3/2	6/2/0	6/1/1	7/1/0	7/1/0	4/2/2	6/1/1	3/3/2	1/4/3	3/1/4	3/4/1	3/3/2	3/4/1	3/3/2	7/1/0
(4) EW$^-$+SSM$^-$				2/4/2	6/1/1	5/1/2	5/1/2	8/0/0	3/3/2	6/1/1	3/3/2	2/3/3	2/4/2	3/3/2	3/3/2	4/2/2	3/3/2	6/1/1
(5) RUS					7/1/0	5/1/2	5/2/1	7/1/0	4/3/1	6/1/1	3/5/0	1/4/3	3/3/2	1/6/1	1/7/0	1/7/0	1/6/1	6/1/1
(6) CNN$^-$						4/0/4	4/0/4	6/1/1	1/1/6	5/0/3	1/1/6	0/2/6	1/2/5	0/2/6	0/2/6	0/2/6	0/2/6	4/0/4
(7) TMK$_1$							2/0/6	8/0/0	3/1/4	5/2/1	3/1/4	1/1/6	1/2/5	2/1/5	2/1/5	2/1/5	2/1/5	3/1/4
(8) TMK$_2$								7/0/1	3/1/4	6/0/2	3/0/5	0/2/6	2/1/5	1/3/4	2/1/5	1/4/3	1/3/4	5/2/1
(9) CNN$^-$+TMK$_1$									1/1/6	0/0/8	0/2/6	0/0/8	0/0/8	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7
(10) CNN$^-$+TMK$_2$										5/1/2	1/5/2	2/2/4	2/3/3	1/3/4	1/3/4	3/1/4	1/3/4	5/1/2
(11) TMK$_1$+CNN$^-$											2/1/5	1/1/6	0/2/6	1/1/6	1/1/6	1/1/6	1/1/6	2/1/5
(12) TMK$_2$+CNN$^-$												2/2/4	2/3/3	1/4/3	1/4/3	2/3/3	1/4/3	6/0/2
(13) EWW$^-$													3/2/3	3/3/2	3/4/1	4/3/1	3/3/2	8/0/0
(14) EWW$^-$+SSM$^-$														3/2/3	3/2/3	4/1/3	3/2/3	6/1/1
(15) SMOTE															2/6/0	2/6/0	1/7/0	6/1/1
(16) gg-SMOTE																1/7/0	0/7/1	6/1/1
(17) ncn-SMOTE																	0/7/1	5/2/1
(18) rng-SMOTE																		6/1/1
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/0/5	0/2/6	2/1/5	1/1/6	3/1/4	6/0/2	1/2/5	7/0/1	2/1/5	6/0/2	3/0/5	0/3/5	3/0/5	1/1/6	1/1/6	1/1/6	1/1/6	3/3/2
(2) SSM$^-$		0/2/6	1/2/5	0/2/6	3/5/0	4/0/4	3/2/3	6/2/0	1/2/5	5/0/3	1/1/6	1/1/6	2/1/5	0/3/5	0/3/5	0/3/5	0/3/5	4/1/3
(3) EW$^-$			3/1/4	3/4/1	6/2/0	6/1/1	5/3/0	8/0/0	4/2/2	6/1/1	3/3/2	1/5/2	3/2/3	3/4/1	3/5/0	4/4/0	3/5/0	7/1/0
(4) EW$^-$+SSM$^-$				2/3/3	6/1/1	5/1/2	5/0/3	8/0/0	3/3/2	6/1/1	3/3/2	1/4/3	2/4/2	3/3/2	3/3/2	4/1/3	3/3/2	6/1/1
(5) RUS					7/1/0	5/1/2	5/2/1	7/1/0	3/5/0	6/1/1	2/6/0	2/3/3	3/3/2	1/6/1	1/7/0	2/6/0	1/7/0	6/1/1
(6) CNN$^-$						4/0/4	3/1/4	6/1/1	1/2/5	5/0/3	1/2/5	1/1/6	1/2/5	1/1/6	1/1/6	1/1/6	1/1/6	4/0/4
(7) TMK$_1$							1/1/6	8/0/0	3/1/4	5/2/1	3/1/4	1/1/6	1/2/5	2/1/5	2/1/5	2/1/5	2/1/5	2/1/5
(8) TMK$_2$								7/0/1	3/1/4	6/1/1	3/0/5	0/3/5	3/0/5	1/3/4	2/2/4	1/4/3	1/3/4	5/2/1
(9) CNN$^-$+TMK$_1$									1/0/7	0/0/8	0/0/8	0/0/8	0/0/8	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7
(10) CNN$^-$+TMK$_2$										5/1/2	0/6/2	2/2/4	2/3/3	1/3/4	1/3/4	2/2/4	1/3/4	5/0/3
(11) TMK$_1$+CNN$^-$											2/1/5	1/1/6	0/2/6	1/1/6	1/1/6	1/1/6	1/1/6	1/1/6
(12) TMK$_2$+CNN$^-$												2/2/4	2/3/3	1/4/3	1/4/3	2/3/3	1/4/3	6/0/2
(13) EWW$^-$													3/2/3	4/3/1	4/2/2	5/2/1	4/2/2	6/2/0
(14) EWW$^-$+SSM$^-$														3/2/3	3/2/3	4/1/3	3/2/3	6/1/1
(15) SMOTE															2/6/0	2/6/0	1/7/0	6/1/1
(16) gg-SMOTE																1/7/0	0/7/1	7/0/1
(17) ncn-SMOTE																	0/7/1	5/2/1
(18) rng-SMOTE																		6/1/1
(19) b-SMOTE																		

Tabla C.44: *Prueba $t-$student del clasificador RBF en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/2/5
(2) SSM$^-$		2/4/2	0/0/8	2/1/5	2/3/3	1/0/7	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	2/1/5	0/0/8	1/2/5	1/2/5	1/2/5	1/2/5	6/2/0
(3) EW$^-$			0/0/8	1/1/6	4/2/2	0/0/8	8/0/0	0/0/8	0/0/8	0/0/8	0/1/7	0/2/6	0/0/8	0/2/6	1/0/7	0/1/7	0/2/6	7/1/0
(4) EW$^-$+SSM$^-$				7/1/0	6/2/0	2/1/5	8/0/0	0/0/8	2/2/4	0/0/8	6/0/2	8/0/0	0/3/5	4/4/0	5/3/0	5/3/0	4/4/0	8/0/0
(5) RUS					5/2/1	0/0/8	8/0/0	0/0/8	0/2/6	0/0/8	1/5/2	5/2/1	0/1/7	1/4/3	0/5/3	1/5/2	0/5/3	8/0/0
(6) CNN$^-$						1/0/7	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	2/1/5	0/1/7	1/2/5	1/2/5	1/1/6	1/2/5	6/0/2
(7) TMK$_1$							8/0/0	0/0/8	6/0/2	1/1/6	6/1/1	8/0/0	5/1/2	7/1/0	8/0/0	7/1/0	7/1/0	8/0/0
(8) TMK$_2$								0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	3/4/1
(9) CNN$^-$+TMK$_1$									8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN$^-$+TMK$_2$										0/0/8	8/0/0	8/0/0	2/1/5	5/3/0	5/3/0	5/3/0	5/3/0	8/0/0
(11) TMK$_1$+CNN$^-$											8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK$_2$+CNN$^-$												5/3/0	1/1/6	3/1/4	3/2/3	3/2/3	4/0/4	8/0/0
(13) EWW$^-$													0/0/8	1/1/6	1/2/5	1/1/6	1/2/5	8/0/0
(14) EWW$^-$+SSM$^-$														6/2/0	6/2/0	5/3/0	5/3/0	8/0/0
(15) SMOTE															0/8/0	1/7/0	1/6/1	8/0/0
(16) gg-SMOTE																0/7/1	0/8/0	8/0/0
(17) ncn-SMOTE																	0/7/1	8/0/0
(18) rng-SMOTE																		8/0/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/1/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0
(2) SSM$^-$		3/0/5	6/1/1	6/0/2	2/2/4	7/0/1	0/1/7	8/0/0	8/0/0	8/0/0	7/1/0	4/2/2	6/1/1	6/0/2	6/0/2	6/0/2	6/0/2	1/3/4
(3) EW$^-$			7/0/1	7/0/1	2/3/3	8/0/0	0/0/8	8/0/0	7/1/0	8/0/0	7/1/0	7/1/0	7/0/1	6/1/1	6/1/1	6/1/1	6/1/1	1/1/6
(4) EW$^-$+SSM$^-$				1/0/7	2/0/6	5/1/2	0/1/7	8/0/0	5/1/2	8/0/0	2/1/5	1/0/7	4/4/0	2/1/5	2/1/5	2/1/5	2/1/5	0/1/7
(5) RUS					1/1/6	7/0/1	0/0/8	8/0/0	6/1/1	7/1/0	4/2/2	1/0/7	7/0/1	3/5/0	3/5/0	2/6/0	3/4/1	0/0/8
(6) CNN$^-$						7/0/1	1/0/7	8/0/0	7/1/0	8/0/0	6/2/0	6/1/1	6/0/2	6/1/1	6/1/1	6/1/1	6/1/1	3/0/5
(7) TMK$_1$							0/0/8	8/0/0	2/1/5	7/0/1	2/0/6	0/0/8	3/0/5	1/1/6	1/1/6	1/1/6	0/2/6	0/0/8
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	4/3/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/0/1	0/2/6	0/1/7	3/2/3	2/2/4	1/3/4	1/3/4	1/3/4	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												1/2/5	6/0/2	4/1/3	4/1/3	4/0/4	4/1/3	0/1/7
(13) EWW$^-$													7/0/1	6/1/1	6/1/1	6/1/1	6/1/1	1/1/6
(14) EWW$^-$+SSM$^-$														2/1/5	1/1/6	1/2/5	1/2/5	0/1/7
(15) SMOTE															0/5/3	0/6/2	0/5/3	0/0/8
(16) gg-SMOTE																0/7/1	0/7/1	0/0/8
(17) ncn-SMOTE																	1/7/0	0/0/8
(18) rng-SMOTE																		0/0/8
(19) b-SMOTE																		

Tabla C.45: *Prueba t−student del clasificador RBF en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/0/1	4/3/1	7/0/1	7/1/0	5/2/1	8/0/0	0/8/0	8/0/0	7/0/1	7/1/0	7/0/1	6/1/1	7/0/1	7/1/0	7/1/0	6/2/0	6/2/0	6/2/0
(2) SSM$^-$		1/2/5	6/1/1	5/1/2	2/2/4	6/0/2	1/2/5	8/0/0	7/1/0	8/0/0	6/2/0	2/3/3	6/0/2	6/0/2	5/1/2	5/1/2	5/1/2	2/3/3
(3) EW$^-$			7/0/1	6/2/0	3/2/3	8/0/0	1/4/3	8/0/0	7/0/1	7/1/0	5/2/1	4/4/0	7/0/1	6/2/0	6/2/0	6/1/1	6/1/1	3/4/1
(4) EW$^-$+SSM$^-$				1/0/7	2/0/6	5/0/3	1/0/7	8/0/0	5/1/2	8/0/0	2/1/5	1/0/7	3/5/0	2/1/5	2/1/5	2/1/5	2/1/5	1/0/7
(5) RUS					1/1/6	6/1/1	0/0/8	8/0/0	6/1/1	7/1/0	4/2/2	1/1/6	7/0/1	3/5/0	3/5/0	2/6/0	3/4/1	0/4/4
(6) CNN$^-$						6/1/1	1/2/5	8/0/0	6/2/0	7/1/0	6/2/0	3/4/1	6/0/2	6/1/1	6/1/1	6/0/2	6/0/2	4/1/3
(7) TMK$_1$							0/0/8	8/0/0	3/0/5	7/0/1	2/1/5	0/0/8	3/0/5	2/0/6	1/1/6	1/1/6	1/1/6	0/1/7
(8) TMK$_2$								8/0/0	7/0/1	7/1/0	7/0/1	5/2/1	7/0/1	7/1/0	7/1/0	7/1/0	7/1/0	6/2/0
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/0/1	0/2/6	1/0/7	3/2/3	2/2/4	1/3/4	1/3/4	1/3/4	2/0/6
(11) TMK$_1$+CNN$^-$											0/0/8	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7
(12) TMK$_2$+CNN$^-$												1/2/5	6/0/2	4/1/3	4/1/3	4/0/4	4/1/3	2/1/5
(13) EWW$^-$													7/0/1	6/1/1	6/1/1	6/1/1	6/1/1	3/3/2
(14) EWW$^-$+SSM$^-$														2/1/5	2/1/5	1/2/5	2/1/5	1/0/7
(15) SMOTE															0/7/1	0/6/2	0/5/3	1/1/6
(16) gg-SMOTE																0/7/1	0/7/1	1/2/5
(17) ncn-SMOTE																	1/7/0	1/3/4
(18) rng-SMOTE																		1/3/4
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/1/7	0/0/8	0/0/8	1/0/7	0/2/6	0/1/7	4/2/2	1/0/7	1/3/4	1/0/7	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/3/4
(2) SSM$^-$		1/2/5	0/2/6	1/1/6	3/5/0	3/2/3	4/4/0	6/2/0	2/3/3	6/0/2	2/2/4	1/0/7	0/2/6	1/2/5	1/2/5	1/2/5	1/2/5	4/4/0
(3) EW$^-$			2/5/1	0/5/3	4/3/1	5/2/1	5/3/0	7/1/0	4/2/2	5/2/1	2/4/2	0/6/2	2/5/1	0/5/3	0/5/3	0/7/1	0/5/3	8/0/0
(4) EW$^-$+SSM$^-$				1/5/2	5/3/0	4/2/2	6/2/0	8/0/0	3/5/0	6/2/0	3/4/1	1/4/3	1/6/1	2/4/2	2/4/2	2/3/3	2/3/3	7/1/0
(5) RUS					5/2/1	5/1/2	5/3/0	7/1/0	4/3/1	5/3/0	1/6/1	0/6/2	2/5/1	1/7/0	1/7/0	1/6/1	1/6/1	6/2/0
(6) CNN$^-$						3/2/3	3/4/1	6/1/1	0/6/2	5/1/2	0/6/2	1/3/4	0/3/5	1/4/3	1/4/3	1/2/5	1/3/4	4/2/2
(7) TMK$_1$							4/3/1	8/0/0	3/1/4	6/1/1	3/1/4	1/1/6	2/2/4	2/0/6	2/0/6	1/1/6	1/1/6	6/2/0
(8) TMK$_2$								5/2/1	1/3/4	4/3/1	1/2/5	0/2/6	0/2/6	0/3/5	0/3/5	0/1/7	0/2/6	5/2/1
(9) CNN$^-$+TMK$_1$									1/0/7	0/0/8	0/2/6	0/0/8	0/0/8	0/2/6	0/1/7	0/1/7	0/1/7	1/3/4
(10) CNN$^-$+TMK$_2$										5/2/1	0/7/1	1/3/4	0/4/4	1/4/3	1/4/3	1/4/3	1/4/3	5/2/1
(11) TMK$_1$+CNN$^-$											1/2/5	0/3/5	0/2/6	0/3/5	0/3/5	0/3/5	0/3/5	4/3/1
(12) TMK$_2$+CNN$^-$												1/4/3	3/2/3	2/4/2	1/5/2	2/4/2	2/5/1	5/2/1
(13) EWW$^-$													3/4/1	0/7/1	0/7/1	1/6/1	1/6/1	8/0/0
(14) EWW$^-$+SSM$^-$														2/4/2	1/4/3	2/3/3	2/3/3	7/1/0
(15) SMOTE															0/8/0	0/7/1	0/6/2	7/1/0
(16) gg-SMOTE																0/8/0	0/8/0	7/1/0
(17) ncn-SMOTE																	0/8/0	7/1/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

Tabla C.46: *Prueba $t-$student del clasificador RBF en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/0/8	0/1/7	0/1/7	0/0/8	0/1/7	2/2/4	0/1/7	7/1/0	1/0/7	5/1/2	0/1/7	0/1/7	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	1/3/4
(2) SSM$^-$		1/2/5	1/1/6	1/1/6	3/5/0	5/0/3	4/4/0	8/0/0	3/4/1	6/2/0	2/1/5	1/0/7	2/2/4	1/2/5	1/1/6	1/1/6	1/1/6	5/3/0
(3) EW$^-$			3/4/1	0/5/3	4/3/1	5/1/2	7/1/0	8/0/0	4/2/2	6/2/0	2/4/2	0/4/4	2/5/1	0/5/3	0/5/3	0/5/3	1/4/3	7/1/0
(4) EW$^-$+SSM$^-$				1/4/3	4/3/1	5/1/2	7/0/1	8/0/0	4/3/1	8/0/0	3/3/2	1/4/3	2/6/0	2/3/3	2/3/3	2/3/3	2/3/3	7/1/0
(5) RUS					6/1/1	6/0/2	6/2/0	8/0/0	4/3/1	7/1/0	3/4/1	1/6/1	4/3/1	1/7/0	1/7/0	1/7/0	1/6/1	7/1/0
(6) CNN$^-$						5/0/3	5/2/1	8/0/0	0/7/1	5/2/1	0/4/4	1/3/4	2/2/4	1/3/4	1/2/5	1/2/5	1/1/6	5/2/1
(7) TMK$_1$							3/0/5	8/0/0	3/0/5	7/0/1	3/0/5	1/2/5	2/1/5	2/0/6	2/0/6	1/1/6	1/1/6	3/2/3
(8) TMK$_2$								8/0/0	1/2/5	5/2/1	1/1/6	0/1/7	1/0/7	0/2/6	0/2/6	0/1/7	0/2/6	5/2/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/0/1	0/5/3	1/3/4	3/1/4	1/4/3	1/4/3	1/4/3	1/4/3	5/1/2
(11) TMK$_1$+CNN$^-$											0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	1/2/5
(12) TMK$_2$+CNN$^-$												1/3/4	2/4/2	2/4/2	2/4/2	2/4/2	2/4/2	5/3/0
(13) EWW$^-$													4/3/1	3/3/2	2/4/2	2/4/2	2/4/2	8/0/0
(14) EWW$^-$+SSM$^-$														2/2/4	2/2/4	2/3/3	2/2/4	6/1/1
(15) SMOTE															0/8/0	0/7/1	0/6/2	7/1/0
(16) gg-SMOTE																0/8/0	0/8/0	7/1/0
(17) ncn-SMOTE																	1/7/0	8/0/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/7	0/1/7	0/1/7	0/1/7	1/0/7	4/1/3	0/1/7	7/1/0	1/0/7	5/2/1	1/0/7	0/1/7	1/0/7	0/0/8	0/0/8	0/0/8	0/0/8	2/2/4
(2) SSM$^-$		1/2/5	1/4/3	1/1/6	3/5/0	5/0/3	4/4/0	8/0/0	4/3/1	6/2/0	2/2/4	1/0/7	2/3/3	1/2/5	1/1/6	1/1/6	1/2/5	4/4/0
(3) EW$^-$			3/4/1	0/6/2	4/3/1	5/2/1	6/2/0	8/0/0	4/2/2	7/1/0	4/2/2	0/4/4	4/3/1	3/2/3	1/4/3	1/4/3	2/3/3	7/1/0
(4) EW$^-$+SSM$^-$				1/2/5	4/3/1	5/1/2	6/1/1	8/0/0	4/2/2	8/0/0	3/2/3	1/4/3	2/6/0	2/3/3	2/2/4	2/3/3	2/2/4	5/2/1
(5) RUS					6/1/1	6/0/2	6/2/0	8/0/0	4/3/1	7/1/0	3/4/1	1/6/1	6/1/1	1/7/0	1/7/0	1/7/0	1/6/1	6/2/0
(6) CNN$^-$						5/0/3	4/3/1	8/0/0	1/6/1	5/2/1	0/4/4	1/3/4	2/2/4	2/3/3	1/3/4	1/2/5	2/0/6	4/2/2
(7) TMK$_1$							3/0/5	8/0/0	3/0/5	7/0/1	3/0/5	1/2/5	3/0/5	2/0/6	2/0/6	1/1/6	1/1/6	3/1/4
(8) TMK$_2$								8/0/0	1/4/3	5/2/1	1/2/5	0/2/6	1/3/4	0/4/4	0/4/4	0/2/6	0/3/5	5/2/1
(9) CNN$^-$+TMK$_1$									0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/0/1	0/5/3	1/2/5	3/1/4	1/4/3	1/3/4	1/3/4	1/3/4	4/2/2
(11) TMK$_1$+CNN$^-$											0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	1/2/5
(12) TMK$_2$+CNN$^-$												1/3/4	3/3/2	2/4/2	2/4/2	2/3/3	2/4/2	5/2/1
(13) EWW$^-$													5/2/1	3/4/1	2/4/2	2/5/1	2/3/3	8/0/0
(14) EWW$^-$+SSM$^-$														2/1/5	2/1/5	2/2/4	2/1/5	5/2/1
(15) SMOTE															0/8/0	0/6/2	0/6/2	7/1/0
(16) gg-SMOTE																0/8/0	0/8/0	7/1/0
(17) ncn-SMOTE																	1/7/0	7/1/0
(18) rng-SMOTE																		7/1/0
(19) b-SMOTE																		

Tabla C.47: *Prueba t−student del clasificador SVM en términos de tvp y tvn: bases de datos moderadamente desbalanceadas*

									tvp										
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/7	0/0/8	0/0/8	0/0/8	1/0/7	0/0/8	3/2/3	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7
(2) SSM$^-$		5/1/2	0/0/8	3/1/4	4/1/3	1/0/7	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	4/0/4	0/0/8	3/1/4	3/1/4	3/1/4	3/1/4	6/0/2
(3) EW$^-$			0/0/8	1/0/7	1/1/6	0/0/8	6/0/2	0/0/8	0/0/8	0/0/8	0/1/7	0/1/7	0/0/8	1/0/7	1/0/7	1/0/7	1/0/7	2/2/4
(4) EW$^-$+SSM$^-$				8/0/0	7/1/0	3/1/4	8/0/0	0/1/7	1/4/3	0/1/7	6/2/0	8/0/0	0/4/4	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(5) RUS					5/0/3	1/0/7	7/0/1	0/0/8	0/0/8	0/0/8	2/0/6	7/0/1	0/0/8	0/6/2	1/5/2	1/7/0	1/5/2	5/1/2
(6) CNN$^-$						1/0/7	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	4/2/2	0/1/7	3/0/5	3/0/5	3/0/5	3/0/5	5/1/2
(7) TMK$_1$							8/0/0	0/1/7	5/0/3	0/1/7	6/0/2	8/0/0	4/1/3	7/0/1	7/0/1	7/1/0	7/0/1	7/0/1
(8) TMK$_2$								0/0/8	0/0/8	0/0/8	0/0/8	2/0/6	0/0/8	1/0/7	1/0/7	1/0/7	1/0/7	2/2/4
(9) CNN$^-$+TMK$_1$									6/2/0	1/7/0	8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(10) CNN$^-$+TMK$_2$										0/2/6	8/0/0	8/0/0	2/4/2	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(11) TMK$_1$+CNN$^-$											8/0/0	8/0/0	7/1/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(12) TMK$_2$+CNN$^-$												7/1/0	1/0/7	5/1/2	5/1/2	5/1/2	5/1/2	6/1/1
(13) EWW$^-$													0/0/8	1/1/6	1/1/6	1/0/7	1/0/7	3/1/4
(14) EWW$^-$+SSM$^-$														8/0/0	8/0/0	8/0/0	8/0/0	8/0/0
(15) SMOTE															2/6/0	2/6/0	1/6/1	5/1/2
(16) gg-SMOTE																1/6/1	0/7/1	5/1/2
(17) ncn-SMOTE																	0/6/2	5/1/2
(18) rng-SMOTE																		5/1/2
(19) b-SMOTE																		

									tvn									
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/0	8/0/0	8/0/0	8/0/0	7/0/1	8/0/0	4/2/2	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	7/1/0
(2) SSM$^-$		2/1/5	8/0/0	4/1/3	2/1/5	7/0/1	0/0/8	8/0/0	8/0/0	8/0/0	7/1/0	3/1/4	8/0/0	4/0/4	4/1/3	4/0/4	4/0/4	2/0/6
(3) EW$^-$			8/0/0	7/0/1	6/1/1	8/0/0	1/0/7	8/0/0	8/0/0	8/0/0	7/1/0	6/2/0	8/0/0	6/1/1	7/0/1	5/2/1	6/1/1	4/1/3
(4) EW$^-$+SSM$^-$				0/0/8	0/0/8	5/1/2	0/0/8	7/1/0	4/3/1	7/1/0	1/0/7	0/0/8	5/2/1	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(5) RUS					3/0/5	7/1/0	0/0/8	8/0/0	8/0/0	8/0/0	7/0/1	1/0/7	8/0/0	1/3/4	0/4/4	0/3/5	1/2/5	1/2/5
(6) CNN$^-$						7/0/1	0/0/8	8/0/0	8/0/0	8/0/0	8/0/0	2/1/5	8/0/0	5/0/3	5/0/3	5/0/3	5/0/3	2/1/5
(7) TMK$_1$							0/0/8	7/1/0	2/2/4	7/1/0	2/0/6	0/0/8	2/1/5	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	7/0/1	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	4/3/1
(9) CNN$^-$+TMK$_1$									0/1/7	0/7/1	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/1/0	0/0/8	0/0/8	2/2/4	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												0/1/7	7/0/1	1/0/7	1/0/7	0/1/7	1/0/7	2/0/6
(13) EWW$^-$													8/0/0	6/1/1	6/1/1	5/2/1	6/1/1	4/0/4
(14) EWW$^-$+SSM$^-$														0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(15) SMOTE															1/4/3	0/5/3	0/7/1	1/1/6
(16) gg-SMOTE																0/5/3	0/6/2	1/2/5
(17) ncn-SMOTE																	3/5/0	2/1/5
(18) rng-SMOTE																		1/1/6
(19) b-SMOTE																		

Tabla C.48: *Prueba $t-$student del clasificador SVM en términos de Acc y f_1: bases de datos moderadamente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	5/3/0	3/0/5	8/0/0	6/1/1	5/0/3	7/1/0	3/1/4	8/0/0	8/0/0	8/0/0	8/0/0	3/2/3	8/0/0	6/0/2	6/0/2	6/0/2	6/0/2	5/1/2
(2) SSM$^-$		2/0/6	8/0/0	3/3/2	2/1/5	6/1/1	1/2/5	8/0/0	8/0/0	8/0/0	6/2/0	2/1/5	8/0/0	2/1/5	2/2/4	2/1/5	2/1/5	3/2/3
(3) EW$^-$			8/0/0	5/2/1	5/2/1	8/0/0	4/1/3	8/0/0	8/0/0	8/0/0	7/1/0	4/4/0	8/0/0	5/2/1	4/3/1	4/2/2	4/3/1	5/2/1
(4) EW$^-$+SSM$^-$				0/0/8	1/0/7	5/1/2	0/1/7	8/0/0	5/2/1	8/0/0	1/1/6	0/0/8	6/1/1	0/0/8	0/0/8	0/0/8	0/0/8	0/1/7
(5) RUS					3/1/4	6/2/0	1/1/6	8/0/0	8/0/0	8/0/0	6/2/0	1/2/5	8/0/0	0/4/4	0/3/5	0/3/5	0/3/5	3/1/4
(6) CNN$^-$						7/0/1	1/1/6	8/0/0	8/0/0	8/0/0	8/0/0	1/2/5	7/1/0	2/2/4	3/1/4	2/2/4	2/2/4	3/2/3
(7) TMK$_1$							0/1/7	7/1/0	2/2/4	7/1/0	2/0/6	0/0/8	2/1/5	0/1/7	0/1/7	0/0/8	0/1/7	1/0/7
(8) TMK$_2$								8/0/0	8/0/0	8/0/0	8/0/0	4/2/2	8/0/0	6/0/2	6/0/2	6/0/2	6/0/2	6/2/0
(9) CNN$^-$+TMK$_1$									0/1/7	0/7/1	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN$^-$+TMK$_2$										7/1/0	0/0/8	0/0/8	2/2/4	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(11) TMK$_1$+CNN$^-$											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK$_2$+CNN$^-$												0/1/7	7/0/1	1/0/7	0/1/7	0/1/7	0/1/7	2/0/6
(13) EWW$^-$													8/0/0	4/2/2	4/3/1	4/2/2	4/3/1	4/3/1
(14) EWW$^-$+SSM$^-$														0/0/8	0/0/8	0/0/8	0/0/8	0/1/7
(15) SMOTE															0/6/2	0/6/2	0/8/0	3/1/4
(16) gg-SMOTE																0/5/3	0/8/0	4/0/4
(17) ncn-SMOTE																	2/6/0	4/0/4
(18) rng-SMOTE																		4/0/4
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/2/5	0/0/8	2/0/6	0/0/8	1/3/4	3/0/5	2/3/3	4/1/3	3/0/5	4/0/4	2/1/5	0/0/8	2/0/6	0/0/8	0/0/8	0/0/8	0/0/8	1/1/6
(2) SSM$^-$		3/1/4	3/2/3	0/3/5	3/3/2	5/0/3	6/0/2	6/1/1	4/2/2	6/1/1	2/4/2	1/2/5	3/2/3	0/2/6	1/1/6	0/2/6	0/2/6	3/2/3
(3) EW$^-$			2/3/3	0/4/4	4/2/2	3/2/3	6/0/2	6/1/1	3/2/3	6/1/1	3/2/3	0/4/4	3/2/3	0/4/4	0/3/5	0/2/6	0/4/4	3/2/3
(4) EW$^-$+SSM$^-$				1/5/2	4/2/2	6/1/1	6/0/2	7/1/0	4/3/1	7/1/0	3/3/2	2/3/3	2/5/1	0/5/3	0/5/3	0/4/4	0/4/4	5/1/2
(5) RUS					7/1/0	5/2/1	6/0/2	7/1/0	5/3/0	7/1/0	4/4/0	2/6/0	4/3/1	0/6/2	0/6/2	0/5/3	0/5/3	4/3/1
(6) CNN$^-$						5/0/3	5/1/2	6/0/2	3/3/2	6/0/2	3/1/4	0/3/5	2/2/4	0/0/8	0/1/7	0/0/8	0/0/8	2/2/4
(7) TMK$_1$							5/1/2	6/1/1	1/2/5	6/1/1	1/2/5	2/2/4	1/1/6	1/1/6	1/1/6	1/1/6	1/1/6	3/3/2
(8) TMK$_2$								3/0/5	2/0/6	3/0/5	2/0/6	2/0/6	2/0/6	2/0/6	2/0/6	2/0/6	2/0/6	2/2/4
(9) CNN$^-$+TMK$_1$									0/1/7	0/7/1	0/1/7	1/0/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	3/2/3
(10) CNN$^-$+TMK$_2$										7/1/0	0/2/6	1/2/5	1/4/3	0/2/6	0/2/6	0/1/7	0/2/6	4/2/2
(11) TMK$_1$+CNN$^-$											0/1/7	1/0/7	0/1/7	0/1/7	0/1/7	0/1/7	0/1/7	3/2/3
(12) TMK$_2$+CNN$^-$												2/3/3	3/3/2	0/1/7	0/3/5	0/2/6	0/3/5	4/2/2
(13) EWW$^-$													4/2/2	0/6/2	0/5/3	0/4/4	0/5/3	3/3/2
(14) EWW$^-$+SSM$^-$														0/3/5	0/4/4	0/3/5	0/2/6	5/1/2
(15) SMOTE															0/8/0	0/8/0	0/8/0	6/1/1
(16) gg-SMOTE																0/7/1	0/8/0	5/2/1
(17) ncn-SMOTE																	0/8/0	5/2/1
(18) rng-SMOTE																		5/2/1
(19) b-SMOTE																		

Tabla C.49: *Prueba t−student del clasificador SVM en términos de $IBA_{0.1}$ y Gm: bases de datos moderadamente desbalanceadas*

	$IBA_{0.1}$																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/6	0/0/8	2/0/6	0/0/8	2/1/5	3/0/5	2/3/3	6/2/0	2/2/4	6/2/0	2/0/6	0/0/8	2/2/4	0/0/8	0/0/8	0/0/8	0/0/8	1/1/6
(2) SSM^-		2/3/3	4/1/3	0/3/5	4/2/2	5/0/3	6/0/2	8/0/0	6/2/0	8/0/0	2/3/3	1/3/4	4/2/2	0/2/6	1/1/6	0/2/6	1/1/6	4/2/2
(3) EW^-			4/2/2	0/4/4	3/2/3	4/2/2	6/0/2	8/0/0	4/2/2	8/0/0	3/2/3	0/3/5	4/2/2	0/3/5	0/2/6	0/1/7	0/2/6	5/0/3
(4) EW^-+SSM^-				1/2/5	4/1/3	5/1/2	5/1/2	8/0/0	5/2/1	8/0/0	2/2/4	1/3/4	5/2/1	0/2/6	0/1/7	0/2/6	0/2/6	5/0/3
(5) RUS					6/2/0	5/2/1	7/0/1	8/0/0	8/0/0	8/0/0	4/4/0	4/4/0	5/2/1	0/6/2	0/6/2	0/6/2	0/6/2	7/0/1
(6) CNN^-						5/0/3	5/1/2	8/0/0	5/2/1	8/0/0	3/1/4	1/3/4	3/3/2	0/1/7	0/1/7	0/1/7	0/1/7	4/1/3
(7) TMK_1							5/0/3	7/1/0	2/2/4	7/1/0	1/2/5	2/1/5	2/1/5	1/0/7	1/0/7	1/0/7	1/0/7	4/0/4
(8) TMK_2								6/2/0	3/1/4	6/2/0	2/0/6	2/0/6	2/1/5	1/0/7	1/0/7	1/0/7	1/0/7	3/1/4
(9) CNN^-+TMK_1									0/1/7	0/7/1	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										7/1/0	0/1/7	0/3/5	1/3/4	0/0/8	0/0/8	0/0/8	0/0/8	1/3/4
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												3/2/3	5/1/2	0/2/6	0/3/5	0/2/6	0/3/5	6/0/2
(13) EWW^-													5/2/1	0/3/5	0/3/5	0/1/7	0/2/6	5/1/2
(14) EWW^-+SSM^-														0/1/7	0/1/7	0/1/7	0/1/7	3/2/3
(15) SMOTE															0/8/0	0/8/0	0/8/0	7/0/1
(16) gg-SMOTE																0/7/1	0/8/0	7/0/1
(17) ncn-SMOTE																	0/8/0	7/0/1
(18) rng-SMOTE																		7/0/1
(19) b-SMOTE																		

	Gm																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/6	0/0/8	2/1/5	0/0/8	3/0/5	3/0/5	3/2/3	6/2/0	3/2/3	6/2/0	2/1/5	0/0/8	2/2/4	0/0/8	0/0/8	0/0/8	0/0/8	1/1/6
(2) SSM^-		3/2/3	4/2/2	0/3/5	4/2/2	5/0/3	6/0/2	8/0/0	7/1/0	8/0/0	2/2/4	1/3/4	5/1/2	0/2/6	1/1/6	0/1/7	0/2/6	4/2/2
(3) EW^-			4/2/2	0/4/4	3/3/2	4/2/2	6/0/2	8/0/0	6/1/1	8/0/0	3/2/3	0/3/5	5/1/2	0/3/5	0/3/5	0/1/7	0/2/6	5/1/2
(4) EW^-+SSM^-				1/2/5	3/2/3	5/1/2	5/1/2	8/0/0	6/1/1	8/0/0	2/1/5	1/2/5	5/2/1	0/0/8	0/0/8	0/1/7	0/0/8	4/1/3
(5) RUS					6/2/0	6/1/1	6/1/1	8/0/0	8/0/0	8/0/0	4/4/0	3/5/0	6/2/0	0/6/2	0/6/2	0/5/3	0/6/2	7/0/1
(6) CNN^-						5/0/3	5/1/2	8/0/0	7/0/1	8/0/0	3/2/3	1/2/5	5/1/2	0/0/8	0/1/7	0/0/8	0/0/8	5/0/3
(7) TMK_1							5/0/3	7/1/0	3/1/4	7/1/0	1/2/5	2/1/5	2/1/5	1/0/7	1/1/6	1/1/6	1/0/7	4/0/4
(8) TMK_2								6/2/0	3/1/4	6/2/0	2/1/5	2/0/6	2/1/5	1/1/6	1/1/6	1/1/6	1/1/6	2/2/4
(9) CNN^-+TMK_1									0/1/7	0/7/1	0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(10) CNN^-+TMK_2										7/1/0	0/1/7	0/0/8	1/3/4	0/0/8	0/0/8	0/0/8	0/0/8	1/3/4
(11) TMK_1+CNN^-											0/0/8	0/0/8	0/1/7	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8
(12) TMK_2+CNN^-												1/4/3	5/1/2	0/1/7	0/3/5	0/2/6	0/2/6	6/0/2
(13) EWW^-													6/1/1	0/5/3	0/3/5	0/3/5	0/4/4	5/1/2
(14) EWW^-+SSM^-														0/0/8	0/0/8	0/0/8	0/0/8	3/2/3
(15) SMOTE															0/8/0	0/8/0	0/8/0	7/0/1
(16) gg-SMOTE																0/7/1	0/8/0	7/0/1
(17) ncn-SMOTE																	0/8/0	7/0/1
(18) rng-SMOTE																		7/0/1
(19) b-SMOTE																		

Tabla C.50: *Prueba $t-$student del clasificador $1-NN$ en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/3/6	0/2/7	0/1/8	0/0/9	0/1/8	0/0/9	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/1/8	0/1/8	0/1/8	0/0/9	0/1/8	0/1/8	1/7/1
(2) SSM^-		2/5/2	0/2/7	0/0/9	1/5/3	0/0/9	5/3/1	0/0/9	0/2/7	0/0/9	0/3/6	1/4/4	0/1/8	1/2/6	0/1/8	0/2/7	1/3/5	5/3/1
(3) EW^-			0/1/8	0/0/9	2/4/3	1/0/8	6/3/0	1/0/8	1/1/7	1/0/8	2/0/7	0/5/4	0/1/8	1/1/7	1/0/8	1/0/8	1/3/5	4/5/0
(4) EW^-+SSM^-				0/3/6	5/2/2	1/3/5	8/1/0	1/0/8	2/5/2	1/0/8	3/4/2	4/4/1	0/5/4	2/3/4	1/4/4	1/4/4	4/2/3	7/2/0
(5) RUS					8/1/0	2/5/2	9/0/0	1/1/7	7/2/0	1/4/4	8/1/0	8/1/0	4/5/0	6/3/0	6/3/0	6/3/0	7/2/0	9/0/0
(6) CNN^-						0/2/7	4/4/1	0/0/9	0/2/7	0/0/9	0/4/5	3/3/3	1/2/6	1/3/5	0/2/7	0/4/5	2/4/3	7/1/1
(7) TMK_1							9/0/0	0/2/7	5/3/1	0/3/6	7/2/0	6/2/1	4/4/1	7/2/0	6/3/0	5/4/0	7/2/0	9/0/0
(8) TMK_2								0/0/9	0/0/9	0/0/9	0/0/9	0/1/8	0/1/8	1/0/8	1/0/8	1/0/8	1/1/7	3/5/1
(9) CNN^-+TMK_1									9/0/0	7/2/0	9/0/0	8/0/1	8/0/1	9/0/0	8/1/0	8/1/0	9/0/0	9/0/0
(10) CNN^-+TMK_2										1/0/8	5/4/0	5/2/2	1/5/3	2/5/2	1/5/3	1/5/3	3/5/1	8/1/0
(11) TMK_1+CNN^-											9/0/0	7/1/1	7/1/1	9/0/0	8/1/0	8/1/0	9/0/0	9/0/0
(12) TMK_2+CNN^-												3/3/3	1/4/4	2/4/3	1/4/4	1/4/4	3/3/3	7/1/1
(13) EWW^-													0/2/7	2/2/5	2/2/5	2/1/6	2/3/4	7/2/0
(14) EWW^-+SSM^-														3/3/3	2/4/3	2/4/3	5/2/2	8/1/0
(15) SMOTE															0/6/3	0/7/2	4/5/0	9/0/0
(16) gg-SMOTE																2/7/0	6/3/0	9/0/0
(17) ncn-SMOTE																	6/3/0	9/0/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	7/2/0	9/0/0	9/0/0	9/0/0	9/0/0	6/3/0	9/0/0	9/0/0	9/0/0	9/0/0	8/1/0	9/0/0	8/1/0	9/0/0	9/0/0	7/2/0	3/5/1
(2) SSM^-		1/1/7	5/4/0	8/1/0	5/2/2	6/2/1	1/1/7	9/0/0	8/1/0	9/0/0	6/2/1	3/0/6	7/2/0	2/2/5	3/2/4	2/3/4	1/1/7	0/1/8
(3) EW^-			9/0/0	9/0/0	8/0/1	8/0/1	1/3/5	9/0/0	8/0/1	8/0/1	8/0/1	5/4/0	9/0/0	6/1/2	7/1/1	7/1/1	6/1/2	1/3/5
(4) EW^-+SSM^-				8/1/0	3/2/4	5/1/3	0/1/8	9/0/0	6/2/1	8/0/1	4/2/3	1/0/8	6/3/0	2/0/7	2/0/7	2/0/7	1/1/7	0/1/8
(5) RUS					0/0/9	2/0/7	0/0/9	8/0/1	0/2/7	6/1/2	0/0/9	0/0/9	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN^-						6/2/1	1/1/7	9/0/0	9/0/0	9/0/0	6/3/0	2/0/7	6/0/3	2/2/5	3/2/4	2/3/4	1/1/7	0/1/8
(7) TMK_1							0/0/9	9/0/0	4/1/4	9/0/0	1/2/6	1/0/8	4/0/5	0/1/8	0/1/8	0/1/8	0/1/8	0/0/9
(8) TMK_2								9/0/0	9/0/0	9/0/0	9/0/0	8/1/0	9/0/0	8/0/1	8/0/1	8/0/1	6/2/1	1/4/4
(9) CNN^-+TMK_1									0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(10) CNN^-+TMK_2										8/0/1	0/0/9	1/1/7	5/1/3	0/2/7	1/1/7	0/2/7	0/0/9	0/0/9
(11) TMK_1+CNN^-											0/0/9	1/0/8	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(12) TMK_2+CNN^-												2/0/7	6/0/3	1/1/7	1/1/7	1/2/6	1/1/7	0/0/9
(13) EWW^-													9/0/0	4/3/2	5/2/2	6/1/2	4/2/3	0/1/8
(14) EWW^-+SSM^-														0/1/8	1/0/8	0/0/9	0/0/9	0/0/9
(15) SMOTE															5/4/0	4/5/0	1/5/3	0/0/9
(16) gg-SMOTE																0/6/3	0/2/7	0/0/9
(17) ncn-SMOTE																	0/2/7	0/0/9
(18) rng-SMOTE																		0/2/7
(19) b-SMOTE																		

Tabla C.51: *Prueba $t-$student del clasificador $1-NN$ en términos de Acc y f_1: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/1/0	5/2/2	8/0/1	9/0/0	9/0/0	9/0/0	5/2/2	9/0/0	9/0/0	9/0/0	9/0/0	4/4/1	8/1/0	5/3/1	5/4/0	5/4/0	5/3/1	0/8/1
(2) SSM$^-$		1/0/8	4/3/2	8/1/0	5/3/1	6/1/2	1/1/7	9/0/0	7/1/1	9/0/0	6/2/1	2/1/6	5/4/0	1/1/7	2/1/6	1/2/6	1/1/7	0/1/8
(3) EW$^-$			8/1/0	9/0/0	8/0/1	8/0/1	2/3/4	9/0/0	8/0/1	8/0/1	8/0/1	4/5/0	9/0/0	5/2/2	5/2/2	5/2/2	4/3/2	0/4/5
(4) EW$^-$+SSM$^-$				8/1/0	4/3/2	5/1/3	1/0/8	9/0/0	5/3/1	8/0/1	4/3/2	1/0/8	6/3/0	2/0/7	2/0/7	2/0/7	2/0/7	1/0/8
(5) RUS					0/2/7	2/0/7	0/0/9	8/0/1	0/2/7	6/1/2	0/1/8	0/0/9	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN$^-$						6/1/2	1/1/7	9/0/0	8/1/0	9/0/0	4/5/0	1/1/7	4/2/3	1/1/7	2/1/6	1/2/6	1/1/7	0/0/9
(7) TMK$_1$							0/0/9	9/0/0	3/3/3	8/1/0	2/2/5	1/0/8	4/0/5	0/1/8	0/1/8	0/1/8	0/1/8	0/0/9
(8) TMK$_2$								9/0/0	9/0/0	9/0/0	7/2/0	4/4/1	8/1/0	5/2/2	5/3/1	5/3/1	5/2/2	2/3/4
(9) CNN$^-$+TMK$_1$									0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(10) CNN$^-$+TMK$_2$										8/0/1	0/1/8	1/1/7	3/3/3	0/0/9	1/0/8	0/0/9	0/0/9	0/0/9
(11) TMK$_1$+CNN$^-$											0/0/9	1/0/8	1/1/7	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(12) TMK$_2$+CNN$^-$												1/1/7	4/2/3	1/1/7	1/1/7	1/1/7	0/1/8	0/0/9
(13) EWW$^-$													8/1/0	3/3/3	4/3/2	3/4/2	3/4/2	0/5/4
(14) EWW$^-$+SSM$^-$														0/1/8	1/1/7	0/1/8	0/1/8	0/0/9
(15) SMOTE															3/6/0	2/7/0	1/7/1	1/2/6
(16) gg-SMOTE																0/6/3	0/6/3	0/4/5
(17) ncn-SMOTE																	0/7/2	0/4/5
(18) rng-SMOTE																		1/2/6
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	5/3/1	3/3/3	5/2/2	6/3/0	6/3/0	5/2/2	1/5/3	7/2/0	5/4/0	6/2/1	5/3/1	2/2/5	5/2/2	2/4/3	2/6/1	2/5/2	3/4/2	0/8/1
(2) SSM$^-$		0/1/8	2/2/5	5/3/1	5/3/1	4/0/5	1/1/7	7/2/0	4/3/2	6/2/1	4/3/2	0/0/9	3/3/3	1/1/7	1/1/7	1/1/7	0/3/6	0/3/6
(3) EW$^-$			6/2/1	8/1/0	8/1/0	5/4/0	3/4/2	8/1/0	8/1/0	7/2/0	8/1/0	3/5/1	6/3/0	3/4/2	3/4/2	3/4/2	3/5/1	1/5/3
(4) EW$^-$+SSM$^-$				7/2/0	5/4/0	5/2/2	2/1/6	9/0/0	6/3/0	7/2/0	6/2/1	0/2/7	4/5/0	2/1/6	3/1/5	2/1/6	2/2/5	2/2/5
(5) RUS					0/6/3	2/2/5	1/2/6	9/0/0	1/4/4	5/3/1	0/5/4	0/0/9	1/1/7	0/2/7	0/1/8	0/1/8	0/2/7	1/2/6
(6) CNN$^-$						4/1/4	0/2/7	8/1/0	4/5/0	6/1/2	1/7/1	0/0/9	2/2/5	0/2/7	1/1/7	0/2/7	0/2/7	0/2/7
(7) TMK$_1$							2/2/5	9/0/0	4/2/3	7/2/0	3/2/4	0/2/7	2/2/5	1/2/6	1/3/5	1/2/6	1/3/5	1/3/5
(8) TMK$_2$								8/1/0	7/2/0	6/1/2	6/2/1	2/3/4	6/1/2	4/2/3	3/4/2	4/3/2	4/3/2	3/3/3
(9) CNN$^-$+TMK$_1$									0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/1/8	0/1/8	0/1/8	0/1/8	0/1/8
(10) CNN$^-$+TMK$_2$										6/1/2	0/3/6	0/0/9	2/2/5	0/1/8	0/3/6	0/2/7	0/2/7	0/3/6
(11) TMK$_1$+CNN$^-$											1/2/6	0/1/8	1/1/7	1/0/8	1/1/7	1/1/7	1/2/6	1/2/6
(12) TMK$_2$+CNN$^-$												0/0/9	2/2/5	1/1/7	2/1/6	1/1/7	1/1/7	1/2/6
(13) EWW$^-$													7/2/0	3/4/2	4/4/1	3/5/1	3/4/2	4/3/2
(14) EWW$^-$+SSM$^-$														2/0/7	2/1/6	2/0/7	2/0/7	2/2/5
(15) SMOTE															2/7/0	1/8/0	1/8/0	1/6/2
(16) gg-SMOTE																0/7/2	0/7/2	0/7/2
(17) ncn-SMOTE																	0/9/0	0/7/2
(18) rng-SMOTE																		1/5/3
(19) b-SMOTE																		

Tabla C.52: *Prueba $t-$student del clasificador $1-NN$ en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/4/4	3/1/5	1/4/4	2/2/5	1/6/2	3/1/5	3/3/3	6/2/1	1/3/5	5/1/3	1/4/4	2/0/7	1/3/5	0/1/8	1/0/8	0/1/8	0/1/8	0/7/2
(2) SSM$^-$		2/3/4	1/2/6	2/1/6	1/6/2	3/2/4	5/3/1	7/2/0	2/2/5	5/0/4	1/4/4	1/3/5	1/2/6	0/1/8	1/0/8	0/1/8	0/3/6	3/4/2
(3) EW$^-$			2/3/4	2/2/5	2/5/2	3/1/5	4/4/1	7/2/0	2/2/5	5/0/4	2/4/3	1/4/4	2/3/4	0/1/8	0/0/9	0/1/8	0/2/7	3/4/2
(4) EW$^-$+SSM$^-$				3/4/2	4/3/2	3/2/4	4/3/2	8/1/0	3/4/2	4/2/3	2/5/2	1/5/3	1/5/3	0/2/7	0/3/6	0/3/6	1/3/5	4/4/1
(5) RUS					6/1/2	2/3/4	5/2/2	8/0/1	2/5/2	4/3/2	4/3/2	1/6/2	2/3/4	0/3/6	1/2/6	1/2/6	1/3/5	5/2/2
(6) CNN$^-$						4/2/3	3/5/1	6/3/0	3/1/5	5/0/4	2/4/3	2/3/4	2/3/4	0/1/8	1/0/8	0/1/8	0/3/6	2/4/3
(7) TMK$_1$							5/1/3	9/0/0	3/3/3	5/3/1	3/2/4	3/3/3	2/4/3	2/2/5	1/3/5	1/3/5	2/3/4	5/1/3
(8) TMK$_2$								6/3/0	2/2/5	6/0/3	1/5/3	2/1/6	2/2/5	0/0/9	0/1/8	0/0/9	0/1/8	2/4/3
(9) CNN$^-$+TMK$_1$									0/0/9	0/0/9	0/1/8	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	1/1/7
(10) CNN$^-$+TMK$_2$										4/1/4	2/4/3	2/4/3	2/4/3	0/2/7	0/2/7	0/2/7	0/2/7	3/5/1
(11) TMK$_1$+CNN$^-$											3/1/5	2/2/5	2/2/5	1/2/6	1/2/6	1/2/6	2/1/6	3/1/5
(12) TMK$_2$+CNN$^-$												2/4/3	2/4/3	0/0/9	0/2/7	0/1/8	0/2/7	3/4/2
(13) EWW$^-$													2/5/2	1/2/6	1/1/7	1/1/7	1/3/5	6/1/2
(14) EWW$^-$+SSM$^-$														1/2/6	1/3/5	1/3/5	1/4/4	4/4/1
(15) SMOTE															1/8/0	0/8/1	5/4/0	8/0/1
(16) gg-SMOTE																0/7/2	5/3/1	8/0/1
(17) ncn-SMOTE																	5/4/0	8/0/1
(18) rng-SMOTE																		7/1/1
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/4/3	3/0/6	1/3/5	2/2/5	1/5/3	3/2/4	3/3/3	6/2/1	2/2/5	6/0/3	1/3/5	2/0/7	1/3/5	0/1/8	1/0/8	0/1/8	0/1/8	0/7/2
(2) SSM$^-$		2/2/5	1/2/6	3/0/6	1/6/2	4/1/4	3/4/2	7/2/0	2/2/5	5/0/4	1/4/4	1/3/5	1/2/6	0/1/8	1/0/8	0/1/8	0/3/6	2/2/5
(3) EW$^-$			2/3/4	4/0/5	2/6/1	3/1/5	4/3/2	7/2/0	3/2/4	5/0/4	2/5/2	2/3/4	2/3/4	0/1/8	0/0/9	0/0/9	0/2/7	3/4/2
(4) EW$^-$+SSM$^-$				3/3/3	4/4/1	3/2/4	5/2/2	7/2/0	3/4/2	4/2/3	2/5/2	0/6/3	1/5/3	0/2/7	0/2/7	0/3/6	1/3/5	4/4/1
(5) RUS					6/0/3	3/2/4	5/1/3	8/0/1	2/5/2	5/3/1	4/2/3	1/4/4	3/1/5	0/3/6	1/2/6	1/2/6	2/1/6	5/2/2
(6) CNN$^-$						4/2/3	2/6/1	7/2/0	3/1/5	5/1/3	2/4/3	1/4/4	1/5/3	0/1/8	1/1/7	0/1/8	0/2/7	1/5/3
(7) TMK$_1$							4/2/3	8/1/0	3/3/3	6/2/1	3/2/4	3/3/3	2/4/3	1/3/5	1/3/5	1/3/5	2/3/4	4/2/3
(8) TMK$_2$								6/3/0	2/2/5	6/0/3	1/4/4	2/1/6	2/2/5	0/0/9	0/1/8	0/0/9	0/1/8	2/4/3
(9) CNN$^-$+TMK$_1$									0/1/8	0/0/9	0/2/7	0/2/7	0/1/8	0/0/9	0/1/8	0/1/8	0/1/8	0/2/7
(10) CNN$^-$+TMK$_2$										4/2/3	2/4/3	1/4/4	2/4/3	0/2/7	0/2/7	0/2/7	0/2/7	3/4/2
(11) TMK$_1$+CNN$^-$											3/1/5	3/1/5	2/2/5	2/1/6	2/1/6	2/1/6	2/1/6	3/1/5
(12) TMK$_2$+CNN$^-$												1/5/3	2/4/3	0/0/9	0/2/7	0/1/8	0/1/8	3/4/2
(13) EWW$^-$													2/6/1	1/1/7	1/1/7	1/1/7	1/3/5	6/1/2
(14) EWW$^-$+SSM$^-$														1/3/5	1/2/6	1/2/6	1/4/4	5/3/1
(15) SMOTE															1/8/0	0/9/0	5/4/0	8/0/1
(16) gg-SMOTE																0/7/2	5/3/1	8/0/1
(17) ncn-SMOTE																	4/5/0	8/0/1
(18) rng-SMOTE																		7/1/1
(19) b-SMOTE																		

Tabla C.53: *Prueba $t-$student del clasificador $7-NN$ en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

	tvp																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/8	0/3/6	0/1/8	0/0/9	0/0/9	0/0/9	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/2/7	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/4/5
(2) SSM$^-$		7/1/1	0/4/5	1/0/8	1/3/5	2/2/5	7/2/0	1/0/8	0/1/8	1/0/8	0/1/8	4/3/2	0/3/6	1/1/7	1/1/7	1/1/7	1/1/7	6/2/1
(3) EW$^-$			0/1/8	0/1/8	1/1/7	1/0/8	4/4/1	1/0/8	0/0/9	1/0/8	0/0/9	0/3/6	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	4/4/1
(4) EW$^-$+SSM$^-$				1/1/7	6/0/3	3/2/4	8/1/0	1/0/8	2/2/5	1/0/8	3/3/3	5/3/1	0/5/4	1/2/6	1/2/6	1/2/6	1/2/6	6/3/0
(5) RUS					7/1/1	7/0/2	8/0/1	1/1/7	6/1/2	1/4/4	6/1/2	8/1/0	7/1/1	5/4/0	4/5/0	6/3/0	5/4/0	9/0/0
(6) CNN$^-$						2/3/4	7/1/1	0/0/9	0/2/7	0/0/9	0/3/6	6/1/2	3/0/6	2/1/6	1/1/7	1/2/6	1/2/6	6/2/1
(7) TMK$_1$							8/0/1	0/1/8	3/1/5	0/1/8	3/2/4	8/0/1	3/2/4	3/0/6	3/0/6	3/0/6	3/0/6	8/1/0
(8) TMK$_2$								1/0/8	0/0/9	1/0/8	0/0/9	1/2/6	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	3/4/2
(9) CNN$^-$+TMK$_1$									7/1/1	4/5/0	8/0/1	8/0/1	8/0/1	7/1/1	7/1/1	7/1/1	7/1/1	9/0/0
(10) CNN$^-$+TMK$_2$										1/1/7	6/3/0	8/0/1	3/2/4	3/2/4	3/2/4	3/2/4	3/3/3	8/1/0
(11) TMK$_1$+CNN$^-$											8/0/1	8/0/1	8/0/1	7/1/1	7/1/1	7/1/1	7/1/1	9/0/0
(12) TMK$_2$+CNN$^-$												7/1/1	3/0/6	3/2/4	2/2/5	2/2/5	2/3/4	7/2/0
(13) EWW$^-$													0/1/8	1/1/7	1/1/7	1/1/7	1/1/7	6/3/0
(14) EWW$^-$+SSM$^-$														3/2/4	1/4/4	2/2/5	2/3/4	8/1/0
(15) SMOTE															1/8/0	1/8/0	1/8/0	9/0/0
(16) gg-SMOTE																2/6/1	1/8/0	9/0/0
(17) ncn-SMOTE																	1/7/1	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

	tvn																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	7/2/0	9/0/0	9/0/0	9/0/0	9/0/0	5/4/0	9/0/0	9/0/0	9/0/0	9/0/0	8/1/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	4/5/0
(2) SSM$^-$		0/2/7	6/3/0	7/1/1	5/3/1	5/1/3	0/0/9	8/0/1	8/1/0	8/0/1	7/2/0	2/1/6	8/1/0	4/0/5	4/0/5	4/0/5	4/0/5	1/1/7
(3) EW$^-$			9/0/0	8/0/1	7/1/1	8/0/1	1/3/5	8/0/1	9/0/0	8/0/1	9/0/0	7/2/0	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/1/6
(4) EW$^-$+SSM$^-$				7/1/1	3/2/4	3/2/4	0/0/9	8/0/1	7/1/1	8/0/1	4/2/3	1/0/8	7/2/0	4/0/5	4/0/5	4/0/5	4/0/5	0/1/8
(5) RUS					3/1/5	2/0/7	1/0/8	8/0/1	5/0/4	7/1/1	5/0/4	1/0/8	1/1/7	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN$^-$						5/1/3	1/0/8	9/0/0	9/0/0	9/0/0	7/2/0	2/1/6	5/1/3	5/0/4	5/0/4	5/0/4	5/0/4	0/3/6
(7) TMK$_1$							1/0/8	9/0/0	5/2/2	9/0/0	4/1/4	1/0/8	7/0/2	3/1/5	4/0/5	4/1/4	4/1/4	0/1/8
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	7/1/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	3/3/3
(9) CNN$^-$+TMK$_1$									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(10) CNN$^-$+TMK$_2$										8/0/1	0/0/9	0/1/8	4/1/4	3/1/5	4/0/5	4/0/5	3/1/5	0/0/9
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(12) TMK$_2$+CNN$^-$												0/2/7	5/1/3	4/0/5	4/0/5	4/0/5	4/0/5	0/1/8
(13) EWW$^-$													9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/2/7
(14) EWW$^-$+SSM$^-$														2/1/6	2/1/6	2/1/6	2/1/6	0/0/9
(15) SMOTE															3/3/3	3/4/2	3/4/2	0/0/9
(16) gg-SMOTE																2/4/3	2/4/3	0/0/9
(17) ncn-SMOTE																	1/7/1	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

Tabla C.54: *Prueba $t-$student del clasificador $7-NN$ en términos de Acc y f_1: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/2/0	5/3/1	8/1/0	9/0/0	7/2/0	9/0/0	1/8/0	9/0/0	9/0/0	9/0/0	8/1/0	6/3/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	1/8/0
(2) SSM^-		1/0/8	5/3/1	7/0/2	5/3/1	4/2/3	0/1/8	8/0/1	8/1/0	8/0/1	7/2/0	3/0/6	8/1/0	4/0/5	4/0/5	4/0/5	4/0/5	1/2/6
(3) EW^-			8/1/0	8/0/1	7/1/1	8/0/1	2/3/4	8/0/1	9/0/0	8/0/1	8/1/0	4/5/0	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/2/5
(4) EW^-+SSM^-				7/1/1	4/1/4	3/2/4	0/0/9	8/0/1	6/2/1	8/0/1	5/1/3	0/2/7	7/2/0	4/0/5	4/0/5	4/0/5	4/0/5	0/1/8
(5) RUS					3/1/5	2/0/7	1/0/8	8/0/1	5/0/4	7/1/1	5/0/4	1/0/8	1/1/7	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN^-						3/3/3	1/1/7	9/0/0	9/0/0	9/0/0	5/4/0	1/2/6	5/1/3	5/0/4	5/0/4	5/0/4	5/0/4	0/4/5
(7) TMK_1							1/0/8	9/0/0	5/2/2	9/0/0	4/2/3	1/0/8	6/1/2	4/0/5	4/0/5	4/0/5	4/0/5	0/1/8
(8) TMK_2								8/0/1	9/0/0	8/0/1	8/1/0	5/3/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	1/7/1
(9) CNN^-+TMK_1									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(10) CNN^-+TMK_2										8/0/1	0/0/9	0/1/8	3/2/4	3/1/5	3/1/5	3/1/5	2/2/5	0/0/9
(11) TMK_1+CNN^-											1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(12) TMK_2+CNN^-												0/2/7	5/1/3	4/0/5	4/0/5	4/0/5	4/0/5	0/2/7
(13) EWW^-													9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/4/5
(14) EWW^-+SSM^-														2/1/6	2/1/6	2/1/6	2/1/6	0/0/9
(15) SMOTE															3/3/3	3/5/1	2/6/1	0/0/9
(16) gg-SMOTE																2/4/3	3/3/3	0/0/9
(17) ncn-SMOTE																	1/6/2	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/2/3	2/4/3	5/1/3	6/1/2	5/2/2	5/2/2	0/6/3	7/0/2	5/2/2	7/0/2	5/2/2	3/3/3	5/2/2	6/1/2	5/2/2	6/1/2	6/1/2	0/8/1
(2) SSM^-		2/1/6	3/3/3	5/1/3	6/3/0	4/2/3	2/0/7	8/0/1	5/3/1	8/0/1	4/5/0	2/2/5	5/2/2	2/1/6	2/2/5	2/1/6	2/1/6	2/3/4
(3) EW^-			6/1/2	7/0/2	7/0/2	6/2/1	2/3/4	8/0/1	7/1/1	8/0/1	7/0/2	2/6/1	7/1/1	5/2/2	5/3/1	7/0/2	7/1/1	3/2/4
(4) EW^-+SSM^-				6/2/1	5/3/1	3/3/3	1/2/6	8/1/0	6/1/2	8/1/0	4/4/1	0/3/6	6/3/0	3/0/6	3/1/5	3/0/6	3/0/6	1/3/5
(5) RUS					6/1/2	3/1/5	2/0/7	9/0/0	5/1/3	8/1/0	5/1/3	1/1/7	1/2/6	0/1/8	1/1/7	1/0/8	1/0/8	1/1/7
(6) CNN^-						3/1/5	1/1/7	7/1/1	6/1/2	7/0/2	3/3/3	1/1/7	2/2/5	2/1/6	2/1/6	2/1/6	2/1/6	0/3/6
(7) TMK_1							1/2/6	8/0/1	3/3/3	7/1/1	4/1/4	1/1/7	5/1/3	2/1/6	2/1/6	2/1/6	2/1/6	0/4/5
(8) TMK_2								8/0/1	7/1/1	8/0/1	6/2/1	4/4/1	7/1/1	6/2/1	5/3/1	7/1/1	6/2/1	1/7/1
(9) CNN^-+TMK_1									0/0/9	0/1/8	0/1/8	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	1/1/7
(10) CNN^-+TMK_2										9/0/0	2/0/7	1/1/7	3/2/4	2/2/5	2/3/4	2/2/5	2/2/5	2/2/5
(11) TMK_1+CNN^-											0/1/8	0/1/8	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	1/1/7
(12) TMK_2+CNN^-												1/1/7	2/3/4	2/1/6	2/2/5	2/1/6	2/1/6	1/3/5
(13) EWW^-													7/2/0	6/1/2	5/3/1	6/1/2	6/2/1	1/5/3
(14) EWW^-+SSM^-														2/1/6	2/2/5	2/1/6	2/1/6	1/3/5
(15) SMOTE															3/4/2	3/5/1	2/7/0	2/1/6
(16) gg-SMOTE																1/4/4	3/4/2	2/2/5
(17) ncn-SMOTE																	1/7/1	2/1/6
(18) rng-SMOTE																		2/1/6
(19) b-SMOTE																		

Tabla C.55: *Prueba $t-$student del clasificador $7-NN$ en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/1/5	1/2/6	3/1/5	2/0/7	3/1/5	2/0/7	0/4/5	7/0/2	3/2/4	5/0/4	3/2/4	1/1/7	3/1/5	0/1/8	0/0/9	0/0/9	0/1/8	0/4/5
(2) SSM$^-$		2/2/5	1/4/4	2/1/6	4/4/1	3/2/4	3/2/4	9/0/0	4/2/3	6/2/1	4/2/3	1/2/6	1/3/5	0/1/8	0/1/8	0/1/8	0/1/8	3/2/4
(3) EW$^-$			3/2/4	2/1/6	6/0/3	3/2/4	4/4/1	8/0/1	5/1/3	6/1/2	5/1/3	1/2/6	3/2/4	1/1/7	1/0/8	1/0/8	1/1/7	4/3/2
(4) EW$^-$+SSM$^-$				2/2/5	7/1/1	3/2/4	4/2/3	8/0/1	4/3/2	6/1/2	5/3/1	2/3/4	2/4/3	0/2/7	0/2/7	0/2/7	0/2/7	3/3/3
(5) RUS					8/1/0	4/2/3	6/1/2	9/0/0	7/2/0	7/2/0	8/1/0	5/2/2	3/4/2	1/2/6	1/2/6	1/2/6	1/2/6	7/0/2
(6) CNN$^-$						3/1/5	3/2/4	8/0/1	4/2/3	5/2/2	1/5/3	1/2/6	1/1/7	0/1/8	0/1/8	0/1/8	0/1/8	2/3/4
(7) TMK$_1$							6/0/3	8/0/1	5/1/3	6/2/1	4/2/3	3/2/4	4/1/4	1/0/8	1/1/7	1/1/7	1/1/7	6/1/2
(8) TMK$_2$								8/0/1	4/1/4	6/1/2	4/1/4	1/2/6	3/2/4	1/1/7	1/0/8	1/0/8	1/1/7	3/4/2
(9) CNN$^-$+TMK$_1$									1/0/8	1/0/8	0/1/8	0/0/9	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	1/0/8
(10) CNN$^-$+TMK$_2$										6/2/1	2/2/5	3/0/6	1/3/5	0/1/8	0/2/7	0/2/7	0/2/7	3/2/4
(11) TMK$_1$+CNN$^-$											1/2/6	2/0/7	1/2/6	0/1/8	0/1/8	0/1/8	0/1/8	3/1/5
(12) TMK$_2$+CNN$^-$												2/1/6	1/1/7	0/1/8	0/1/8	0/1/8	0/1/8	2/4/3
(13) EWW$^-$													3/3/3	1/2/6	2/0/7	2/0/7	1/2/6	6/2/1
(14) EWW$^-$+SSM$^-$														0/2/7	0/2/7	1/1/7	0/2/7	5/1/3
(15) SMOTE															2/7/0	4/5/0	2/7/0	8/1/0
(16) gg-SMOTE																2/5/2	0/8/1	9/0/0
(17) ncn-SMOTE																	1/7/1	9/0/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/1/5	1/2/6	3/2/4	2/2/5	4/0/5	2/1/6	0/4/5	7/0/2	5/0/4	5/1/3	4/1/4	1/1/7	3/2/4	0/1/8	0/0/9	0/0/9	0/1/8	0/4/5
(2) SSM$^-$		2/2/5	1/4/4	2/1/6	3/5/1	3/2/4	3/2/4	7/2/0	4/2/3	5/3/1	4/2/3	1/2/6	2/2/5	0/1/8	0/1/8	1/0/8	0/1/8	2/3/4
(3) EW$^-$			3/3/3	2/2/5	6/0/3	4/1/4	4/4/1	8/0/1	5/1/3	6/1/2	5/1/3	1/2/6	3/2/4	0/2/7	1/0/8	1/0/8	1/1/7	4/3/2
(4) EW$^-$+SSM$^-$				2/3/4	7/1/1	3/2/4	4/2/3	8/0/1	4/3/2	6/1/2	5/3/1	2/2/5	2/4/3	0/2/7	0/2/7	0/2/7	0/2/7	2/4/3
(5) RUS					8/1/0	5/1/3	5/2/2	9/0/0	6/3/0	7/2/0	8/1/0	4/3/2	2/5/2	1/2/6	1/3/5	1/2/6	1/3/5	5/2/2
(6) CNN$^-$						3/1/5	3/1/5	7/1/1	5/1/3	5/2/2	2/4/3	1/2/6	1/1/7	0/1/8	0/1/8	0/1/8	0/1/8	2/2/5
(7) TMK$_1$							5/1/3	8/0/1	5/1/3	6/1/2	5/1/3	3/2/4	4/1/4	1/1/7	1/1/7	1/1/7	1/1/7	5/2/2
(8) TMK$_2$								8/0/1	5/0/4	6/1/2	5/0/4	1/2/6	3/2/4	0/2/7	1/0/8	0/1/8	1/1/7	3/4/2
(9) CNN$^-$+TMK$_1$									1/0/8	1/0/8	1/0/8	0/0/9	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	1/0/8
(10) CNN$^-$+TMK$_2$										6/2/1	2/2/5	3/0/6	1/3/5	0/1/8	0/1/8	0/2/7	0/2/7	3/1/5
(11) TMK$_1$+CNN$^-$											2/1/6	2/0/7	1/2/6	0/2/7	0/2/7	0/2/7	0/2/7	3/0/6
(12) TMK$_2$+CNN$^-$												2/1/6	1/1/7	0/1/8	0/1/8	0/1/8	0/1/8	2/3/4
(13) EWW$^-$													4/2/3	1/2/6	2/0/7	2/0/7	2/1/6	6/2/1
(14) EWW$^-$+SSM$^-$														0/2/7	0/2/7	1/1/7	0/3/6	4/2/3
(15) SMOTE															2/7/0	4/5/0	2/7/0	8/1/0
(16) gg-SMOTE																1/6/2	0/8/1	9/0/0
(17) ncn-SMOTE																	1/7/1	9/0/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Tabla C.56: *Prueba $t-$student del clasificador $13-NN$ en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/8	0/3/6	0/0/9	0/0/9	0/1/8	0/1/8	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/3/6	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	0/5/4
(2) SSM$^-$		7/2/0	0/2/7	1/1/7	2/1/6	2/3/4	7/2/0	1/0/8	0/1/8	1/0/8	0/1/8	6/1/2	0/2/7	1/1/7	1/1/7	1/1/7	1/1/7	7/0/2
(3) EW$^-$			0/0/9	0/1/8	1/2/6	1/1/7	5/3/1	1/0/8	0/0/9	1/0/8	0/0/9	0/4/5	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	5/2/2
(4) EW$^-$+SSM$^-$				1/2/6	6/0/3	3/2/4	9/0/0	1/0/8	1/3/5	1/0/8	2/4/3	8/1/0	0/5/4	1/3/5	1/3/5	1/3/5	1/3/5	7/2/0
(5) RUS					6/1/2	7/0/2	8/0/1	1/1/7	3/3/3	1/2/6	5/1/3	8/1/0	6/2/1	5/4/0	4/5/0	6/3/0	5/3/1	9/0/0
(6) CNN$^-$						3/4/2	7/1/1	0/2/7	0/3/6	0/2/7	0/3/6	6/1/2	3/0/6	2/1/6	2/0/7	2/1/6	2/1/6	6/2/1
(7) TMK$_1$							7/1/1	0/1/8	1/2/6	0/1/8	2/2/5	7/1/1	3/2/4	2/1/6	2/1/6	3/0/6	2/1/6	7/1/1
(8) TMK$_2$								1/0/8	0/0/9	1/0/8	0/0/9	1/3/5	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	2/5/2
(9) CNN$^-$+TMK$_1$									6/2/1	3/6/0	6/2/1	8/0/1	8/0/1	7/1/1	7/1/1	7/1/1	7/1/1	8/1/0
(10) CNN$^-$+TMK$_2$										1/2/6	6/3/0	8/1/0	3/4/2	3/3/3	3/3/3	3/3/3	3/3/3	9/0/0
(11) TMK$_1$+CNN$^-$											6/2/1	8/0/1	8/0/1	7/1/1	7/1/1	7/1/1	7/1/1	8/1/0
(12) TMK$_2$+CNN$^-$												8/1/0	3/1/5	3/2/4	3/0/6	3/2/4	3/1/5	7/2/0
(13) EWW$^-$													0/0/9	1/1/7	1/1/7	1/1/7	1/1/7	6/2/1
(14) EWW$^-$+SSM$^-$														2/2/5	2/2/5	2/2/5	2/2/5	9/0/0
(15) SMOTE															3/4/2	2/7/0	1/8/0	9/0/0
(16) gg-SMOTE																2/6/1	2/5/2	9/0/0
(17) ncn-SMOTE																	1/7/1	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	5/4/0	9/0/0	9/0/0	9/0/0	9/0/0	5/4/0	9/0/0	9/0/0	9/0/0	9/0/0	7/2/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	4/5/0
(2) SSM$^-$		0/1/8	8/1/0	7/0/2	6/2/1	5/0/4	0/0/9	8/0/1	9/0/0	8/0/1	9/0/0	2/0/7	9/0/0	4/0/5	4/0/5	4/0/5	4/0/5	1/1/7
(3) EW$^-$			9/0/0	8/0/1	8/0/1	8/0/1	1/4/4	8/0/1	9/0/0	8/0/1	9/0/0	4/5/0	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/3/4
(4) EW$^-$+SSM$^-$				7/0/2	3/1/5	3/1/5	0/0/9	8/0/1	8/0/1	8/0/1	4/3/2	0/1/8	8/1/0	4/0/5	4/0/5	4/0/5	4/0/5	0/0/9
(5) RUS					4/0/5	2/0/7	1/0/8	8/0/1	5/0/4	7/1/1	5/0/4	1/0/8	3/0/6	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN$^-$						4/2/3	1/0/8	9/0/0	9/0/0	9/0/0	7/2/0	2/0/7	5/1/3	5/0/4	5/0/4	5/0/4	5/0/4	1/2/6
(7) TMK$_1$							1/0/8	9/0/0	6/1/2	9/0/0	4/3/2	1/0/8	7/0/2	4/1/4	4/1/4	4/1/4	4/1/4	0/1/8
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	6/2/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/4/3
(9) CNN$^-$+TMK$_1$									1/0/8	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8
(10) CNN$^-$+TMK$_2$										8/0/1	0/0/9	0/0/9	4/0/5	2/2/5	3/1/5	3/1/5	3/1/5	0/0/9
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8
(12) TMK$_2$+CNN$^-$												0/1/8	5/1/3	4/0/5	4/0/5	4/0/5	4/0/5	0/0/9
(13) EWW$^-$													9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	1/1/7
(14) EWW$^-$+SSM$^-$														2/0/7	2/0/7	2/0/7	2/0/7	0/0/9
(15) SMOTE															2/4/3	3/5/1	3/5/1	0/0/9
(16) gg-SMOTE																2/4/3	3/3/3	0/0/9
(17) ncn-SMOTE																	3/5/1	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

Tabla C.57: *Prueba t−student del clasificador* 13−*NN en términos de Acc y* f_1*: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/1/0	4/5/0	9/0/0	9/0/0	7/2/0	8/1/0	2/7/0	9/0/0	9/0/0	9/0/0	9/0/0	5/4/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	2/7/0
(2) SSM$^-$		0/2/7	7/2/0	7/0/2	7/1/1	5/0/4	0/2/7	8/0/1	9/0/0	8/0/1	9/0/0	2/1/6	9/0/0	4/0/5	4/0/5	4/0/5	4/0/5	0/2/7
(3) EW$^-$			8/1/0	8/0/1	7/1/1	8/0/1	1/5/3	8/0/1	9/0/0	8/0/1	8/1/0	3/5/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	1/5/3
(4) EW$^-$+SSM$^-$				6/1/2	3/3/3	3/2/4	0/0/9	8/0/1	8/0/1	8/0/1	4/3/2	0/1/8	8/1/0	4/0/5	4/0/5	4/0/5	4/0/5	0/0/9
(5) RUS					4/0/5	2/0/7	1/0/8	8/0/1	5/0/4	7/1/1	5/0/4	1/0/8	3/1/5	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN$^-$						4/2/3	1/0/8	9/0/0	9/0/0	9/0/0	7/2/0	1/2/6	5/1/3	5/0/4	5/0/4	5/0/4	5/0/4	0/3/6
(7) TMK$_1$							1/0/8	9/0/0	6/1/2	8/1/0	4/3/2	1/0/8	6/1/2	4/0/5	4/0/5	4/0/5	4/0/5	0/1/8
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	4/4/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	1/6/2
(9) CNN$^-$+TMK$_1$									1/0/8	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8
(10) CNN$^-$+TMK$_2$										8/0/1	0/0/9	0/0/9	4/0/5	2/2/5	3/1/5	3/1/5	3/1/5	0/0/9
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8
(12) TMK$_2$+CNN$^-$												0/1/8	5/1/3	4/0/5	4/0/5	4/0/5	4/0/5	0/0/9
(13) EWW$^-$													9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/4/5
(14) EWW$^-$+SSM$^-$														2/0/7	2/0/7	2/0/7	2/0/7	0/0/9
(15) SMOTE															2/4/3	4/4/1	3/5/1	0/0/9
(16) gg-SMOTE																3/3/3	3/4/2	0/0/9
(17) ncn-SMOTE																	3/4/2	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	5/2/2	1/6/2	5/2/2	7/0/2	5/3/1	5/2/2	0/7/2	7/0/2	5/2/2	7/0/2	5/2/2	3/4/2	6/1/2	6/1/2	6/1/2	7/0/2	6/1/2	0/7/2
(2) SSM$^-$		1/3/5	4/4/1	5/1/3	6/3/0	4/1/4	1/2/6	8/0/1	6/1/2	8/0/1	6/1/2	1/3/5	7/1/1	2/1/6	2/1/6	2/1/6	2/2/5	2/1/6
(3) EW$^-$			6/2/1	7/0/2	6/3/0	7/1/1	0/6/3	8/0/1	7/1/1	8/0/1	7/1/1	1/7/1	7/1/1	7/0/2	7/1/1	7/1/1	7/1/1	1/6/2
(4) EW$^-$+SSM$^-$				5/2/2	5/3/1	5/1/3	1/2/6	8/1/0	6/3/0	8/1/0	4/5/0	1/2/6	7/2/0	3/0/6	3/1/5	3/0/6	3/1/5	1/3/5
(5) RUS					6/0/3	4/0/5	2/0/7	8/1/0	5/1/3	8/1/0	6/0/3	2/0/7	3/2/4	1/0/8	1/0/8	1/0/8	1/0/8	2/0/7
(6) CNN$^-$						3/2/4	0/3/6	7/1/1	7/0/2	7/0/2	3/4/2	0/4/5	2/2/5	3/0/6	3/0/6	3/0/6	3/0/6	0/3/6
(7) TMK$_1$							1/2/6	7/0/2	5/0/4	7/0/2	3/3/3	1/0/8	4/1/4	2/0/7	2/0/7	2/0/7	2/0/7	0/3/6
(8) TMK$_2$								8/0/1	8/0/1	8/0/1	7/1/1	4/5/0	7/1/1	7/1/1	7/1/1	7/1/1	8/0/1	1/7/1
(9) CNN$^-$+TMK$_1$									0/1/8	0/2/7	1/0/8	1/0/8	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	1/1/7
(10) CNN$^-$+TMK$_2$										8/1/0	2/0/7	1/0/8	2/2/5	1/2/6	2/2/5	2/1/6	2/2/5	2/0/7
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	1/1/7
(12) TMK$_2$+CNN$^-$												1/3/5	2/3/4	2/1/6	2/2/5	3/0/6	2/2/5	1/2/6
(13) EWW$^-$													7/1/1	7/0/2	7/1/1	7/1/1	7/1/1	1/4/4
(14) EWW$^-$+SSM$^-$														2/0/7	2/1/6	2/0/7	2/1/6	2/1/6
(15) SMOTE															3/4/2	2/6/1	4/5/0	2/0/7
(16) gg-SMOTE																2/5/2	3/5/1	2/0/7
(17) ncn-SMOTE																	2/6/1	2/0/7
(18) rng-SMOTE																		2/0/7
(19) b-SMOTE																		

Tabla C.58: *Prueba $t-$student del clasificador $13-NN$ en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/1/4	0/3/6	5/0/4	1/3/5	4/1/4	2/1/6	0/5/4	7/1/1	4/1/4	6/0/3	4/0/5	1/2/6	5/1/3	0/0/9	0/0/9	0/0/9	0/0/9	0/6/3
(2) SSM$^-$		2/2/5	2/3/4	2/1/6	5/2/2	3/2/4	2/2/5	8/1/0	5/1/3	7/1/1	4/2/3	1/2/6	3/3/3	0/1/8	0/1/8	0/1/8	0/1/8	3/0/6
(3) EW$^-$			5/0/4	2/2/5	5/2/2	3/1/5	4/5/0	8/1/0	5/0/4	7/0/2	5/1/3	0/4/5	6/0/3	1/0/8	1/0/8	1/0/8	1/0/8	4/3/2
(4) EW$^-$+SSM$^-$				2/3/4	6/2/1	3/2/4	4/0/5	8/0/1	5/1/3	7/1/1	4/4/1	3/1/5	3/3/3	1/1/7	1/1/7	1/1/7	1/1/7	2/2/5
(5) RUS					7/2/0	4/1/4	5/3/1	9/0/0	5/3/1	7/2/0	7/2/0	4/2/3	4/2/3	1/2/6	1/2/6	1/2/6	1/2/6	4/4/1
(6) CNN$^-$						2/3/4	3/1/5	7/1/1	4/1/4	6/1/2	2/2/5	2/1/6	2/1/6	0/1/8	0/1/8	0/2/7	0/2/7	2/2/5
(7) TMK$_1$							5/1/3	7/1/1	4/1/4	6/1/2	4/1/4	3/3/3	5/0/4	1/0/8	1/0/8	1/1/7	1/1/7	5/1/3
(8) TMK$_2$								8/1/0	5/0/4	7/0/2	5/0/4	1/3/5	5/1/3	1/0/8	1/0/8	1/0/8	1/0/8	2/5/2
(9) CNN$^-$+TMK$_1$									1/0/8	0/2/7	0/1/8	0/1/8	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	0/1/8
(10) CNN$^-$+TMK$_2$										6/2/1	2/3/4	3/1/5	2/3/4	0/2/7	0/2/7	0/2/7	0/2/7	3/1/5
(11) TMK$_1$+CNN$^-$											1/2/6	2/0/7	1/0/8	0/1/8	0/1/8	0/1/8	0/1/8	2/1/6
(12) TMK$_2$+CNN$^-$												3/0/6	2/2/5	0/2/7	0/2/7	0/2/7	0/2/7	2/3/4
(13) EWW$^-$													6/0/3	1/1/7	1/1/7	2/0/7	2/0/7	6/1/2
(14) EWW$^-$+SSM$^-$														1/0/8	0/1/8	1/0/8	1/0/8	3/0/6
(15) SMOTE															2/6/1	2/7/0	2/7/0	8/1/0
(16) gg-SMOTE																2/5/2	2/6/1	8/1/0
(17) ncn-SMOTE																	1/8/0	8/1/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/2/3	0/3/6	5/0/4	3/1/5	4/1/4	2/0/7	0/5/4	7/1/1	4/1/4	6/0/3	4/1/4	1/2/6	5/1/3	0/0/9	0/0/9	0/0/9	0/0/9	0/6/3
(2) SSM$^-$		2/2/5	2/3/4	2/1/6	5/2/2	4/0/5	2/2/5	7/2/0	5/1/3	6/2/1	4/2/3	1/2/6	4/2/3	0/1/8	0/1/8	1/0/8	0/1/8	2/1/6
(3) EW$^-$			5/0/4	3/2/4	5/2/2	3/1/5	3/6/0	8/1/0	5/1/3	7/0/2	5/1/3	0/4/5	6/0/3	0/1/8	1/0/8	1/1/7	1/0/8	4/3/2
(4) EW$^-$+SSM$^-$				2/2/5	6/2/1	4/1/4	4/0/5	8/0/1	5/1/3	7/0/2	4/4/1	3/1/5	3/3/3	1/1/7	1/1/7	1/1/7	1/1/7	2/2/5
(5) RUS					7/2/0	4/1/4	5/1/3	9/0/0	5/3/1	7/2/0	6/2/1	3/3/3	4/3/2	1/2/6	1/2/6	1/2/6	1/2/6	4/2/3
(6) CNN$^-$						2/2/5	2/2/5	7/2/0	5/1/3	6/1/2	2/3/4	2/1/6	2/1/6	0/1/8	0/1/8	0/2/7	0/2/7	2/2/5
(7) TMK$_1$							6/0/3	8/0/1	5/0/4	6/1/2	3/3/3	4/1/4	5/0/4	1/0/8	1/1/7	1/1/7	1/1/7	5/2/2
(8) TMK$_2$								8/1/0	5/0/4	7/0/2	5/0/4	1/3/5	5/1/3	0/1/8	1/0/8	1/0/8	1/0/8	2/5/2
(9) CNN$^-$+TMK$_1$									1/0/8	0/2/7	0/1/8	0/1/8	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	0/1/8
(10) CNN$^-$+TMK$_2$										6/2/1	2/2/5	3/1/5	2/3/4	0/2/7	0/2/7	0/2/7	0/2/7	3/1/5
(11) TMK$_1$+CNN$^-$											1/2/6	2/0/7	1/0/8	0/1/8	0/1/8	0/1/8	0/1/8	2/1/6
(12) TMK$_2$+CNN$^-$												3/1/5	2/2/5	0/2/7	0/2/7	0/2/7	0/2/7	2/3/4
(13) EWW$^-$													6/0/3	0/2/7	1/1/7	2/0/7	2/0/7	5/2/2
(14) EWW$^-$+SSM$^-$														1/0/8	1/0/8	1/0/8	1/0/8	3/0/6
(15) SMOTE															2/5/2	2/7/0	2/7/0	8/1/0
(16) gg-SMOTE																2/5/2	2/6/1	8/1/0
(17) ncn-SMOTE																	1/8/0	8/1/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Tabla C.59: *Prueba t−student del clasificador J48 en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

	tvp																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/7	0/4/5	0/1/8	0/0/9	1/2/6	0/0/9	0/7/2	0/0/9	0/2/7	0/0/9	1/1/7	0/4/5	0/1/8	0/1/8	0/0/9	0/0/9	0/1/8	1/7/1
(2) SSM^-		5/2/2	0/2/7	1/2/6	1/5/3	2/2/5	6/1/2	1/0/8	0/4/5	1/0/8	0/4/5	5/1/3	0/3/6	3/3/3	2/4/3	3/2/4	3/4/2	7/0/2
(3) EW^-			1/1/7	1/0/8	3/2/4	1/0/8	4/4/1	1/0/8	1/2/6	1/0/8	2/1/6	0/6/3	0/1/8	1/3/5	1/1/7	1/1/7	1/5/3	5/4/0
(4) EW^-+SSM^-				1/4/4	4/3/2	3/2/4	7/1/1	1/0/8	3/3/3	1/0/8	3/4/2	7/1/1	0/6/3	5/2/2	4/3/2	4/3/2	5/2/2	7/2/0
(5) RUS					7/0/2	5/2/2	8/0/1	1/1/7	6/0/3	1/3/5	6/0/3	8/0/1	4/4/1	9/0/0	8/1/0	7/2/0	9/0/0	9/0/0
(6) CNN^-						1/2/6	5/1/3	0/0/9	0/6/3	0/0/9	0/7/2	4/2/3	2/2/5	3/2/4	3/1/5	3/1/5	3/3/3	6/2/1
(7) TMK_1							8/0/1	0/0/9	4/3/2	0/2/7	5/2/2	8/0/1	2/5/2	4/4/1	5/3/1	4/3/2	5/3/1	9/0/0
(8) TMK_2								1/0/8	0/3/6	1/0/8	1/2/6	2/3/4	0/1/8	1/2/6	1/1/7	1/1/7	1/2/6	2/7/0
(9) CNN^-+TMK_1									8/0/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	9/0/0	8/0/1	8/0/1	9/0/0
(10) CNN^-+TMK_2										1/0/8	3/6/0	6/2/1	2/3/4	4/4/1	4/4/1	4/3/2	5/3/1	8/1/0
(11) TMK_1+CNN^-											8/0/1	8/0/1	6/2/1	8/0/1	9/0/0	8/0/1	8/0/1	9/0/0
(12) TMK_2+CNN^-												5/3/1	2/2/5	5/2/2	3/4/2	4/2/3	5/2/2	7/2/0
(13) EWW^-													0/0/9	2/3/4	1/3/5	1/2/6	2/3/4	5/4/0
(14) EWW^-+SSM^-														7/2/0	6/3/0	5/4/0	7/2/0	9/0/0
(15) SMOTE															1/4/4	0/7/2	1/8/0	8/1/0
(16) gg-SMOTE																0/8/1	3/5/1	9/0/0
(17) ncn-SMOTE																	3/6/0	9/0/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

	tvn																	
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/7	0/4/5	0/1/8	0/0/9	1/2/6	0/0/9	0/7/2	0/0/9	0/2/7	0/0/9	1/1/7	0/4/5	0/1/8	0/1/8	0/0/9	0/0/9	0/1/8	1/7/1
(2) SSM^-		5/2/2	0/2/7	1/2/6	1/5/3	2/2/5	6/1/2	1/0/8	0/4/5	1/0/8	0/4/5	5/1/3	0/3/6	3/3/3	2/4/3	3/2/4	3/4/2	7/0/2
(3) EW^-			1/1/7	1/0/8	3/2/4	1/0/8	4/4/1	1/0/8	1/2/6	1/0/8	2/1/6	0/6/3	0/1/8	1/3/5	1/1/7	1/1/7	1/5/3	5/4/0
(4) EW^-+SSM^-				1/4/4	4/3/2	3/2/4	7/1/1	1/0/8	3/3/3	1/0/8	3/4/2	7/1/1	0/6/3	5/2/2	4/3/2	4/3/2	5/2/2	7/2/0
(5) RUS					7/0/2	5/2/2	8/0/1	1/1/7	6/0/3	1/3/5	6/0/3	8/0/1	4/4/1	9/0/0	8/1/0	7/2/0	9/0/0	9/0/0
(6) CNN^-						1/2/6	5/1/3	0/0/9	0/6/3	0/0/9	0/7/2	4/2/3	2/2/5	3/2/4	3/1/5	3/1/5	3/3/3	6/2/1
(7) TMK_1							8/0/1	0/0/9	4/3/2	0/2/7	5/2/2	8/0/1	2/5/2	4/4/1	5/3/1	4/3/2	5/3/1	9/0/0
(8) TMK_2								1/0/8	0/3/6	1/0/8	1/2/6	2/3/4	0/1/8	1/2/6	1/1/7	1/1/7	1/2/6	2/7/0
(9) CNN^-+TMK_1									8/0/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	9/0/0	8/0/1	8/0/1	9/0/0
(10) CNN^-+TMK_2										1/0/8	3/6/0	6/2/1	2/3/4	4/4/1	4/4/1	4/3/2	5/3/1	8/1/0
(11) TMK_1+CNN^-											8/0/1	8/0/1	6/2/1	8/0/1	9/0/0	8/0/1	8/0/1	9/0/0
(12) TMK_2+CNN^-												5/3/1	2/2/5	5/2/2	3/4/2	4/2/3	5/2/2	7/2/0
(13) EWW^-													0/0/9	2/3/4	1/3/5	1/2/6	2/3/4	5/4/0
(14) EWW^-+SSM^-														7/2/0	6/3/0	5/4/0	7/2/0	9/0/0
(15) SMOTE															1/4/4	0/7/2	1/8/0	8/1/0
(16) gg-SMOTE																0/8/1	3/5/1	9/0/0
(17) ncn-SMOTE																	3/6/0	9/0/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Tabla C.60: *Prueba $t-$student del clasificador J48 en términos de Acc y f_1: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	2/7/0	9/0/0	9/0/0	9/0/0	9/0/0	1/8/0	9/0/0	9/0/0	9/0/0	9/0/0	5/4/0	9/0/0	7/1/1	7/1/1	7/1/1	5/3/1	3/6/0
(2) SSM$^-$		0/1/8	4/5/0	4/1/4	2/5/2	2/2/5	0/0/9	8/0/1	7/1/1	8/0/1	3/5/1	0/2/7	6/3/0	0/2/7	0/2/7	0/2/7	0/1/8	0/2/7
(3) EW$^-$			9/0/0	8/0/1	7/1/1	7/1/1	1/7/1	8/0/1	9/0/0	8/0/1	8/1/0	4/5/0	9/0/0	4/3/2	4/3/2	5/2/2	3/4/2	1/7/1
(4) EW$^-$+SSM$^-$				4/1/4	2/2/5	2/1/6	0/0/9	8/0/1	2/5/2	7/1/1	2/3/4	0/1/8	4/5/0	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(5) RUS					2/1/6	2/0/7	1/0/8	8/0/1	4/1/4	5/2/2	4/0/5	1/0/8	5/0/4	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(6) CNN$^-$						3/2/4	1/0/8	9/0/0	7/2/0	9/0/0	2/7/0	1/2/6	6/1/2	1/2/6	0/3/6	1/2/6	1/2/6	0/2/7
(7) TMK$_1$							1/0/8	9/0/0	5/2/2	9/0/0	5/2/2	1/2/6	6/1/2	1/1/7	0/2/7	1/2/6	1/1/7	0/1/8
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	4/4/1	9/0/0	6/1/2	6/1/2	6/1/2	3/4/2	3/5/1
(9) CNN$^-$+TMK$_1$									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9	1/0/8	1/0/8	0/0/9
(10) CNN$^-$+TMK$_2$										7/1/1	0/4/5	0/1/8	3/5/1	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	2/0/7	1/0/8	0/0/9	1/0/8	1/0/8	0/0/9
(12) TMK$_2$+CNN$^-$												0/2/7	4/4/1	0/2/7	0/2/7	0/2/7	0/0/9	0/2/7
(13) EWW$^-$													8/1/0	3/2/4	3/3/3	4/1/4	1/4/4	0/5/4
(14) EWW$^-$+SSM$^-$														0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(15) SMOTE															3/5/1	2/6/1	0/8/1	1/4/4
(16) gg-SMOTE																3/5/1	1/4/4	1/4/4
(17) ncn-SMOTE																	0/4/5	1/3/5
(18) rng-SMOTE																		1/5/3
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	7/1/1	0/6/3	7/1/1	7/1/1	6/3/0	5/3/1	0/8/1	8/0/1	8/0/1	6/2/1	7/1/1	1/6/2	7/1/1	1/5/3	2/4/3	2/4/3	1/5/3	1/8/0
(2) SSM$^-$		0/1/8	1/5/3	0/4/5	3/4/2	3/0/6	0/1/8	8/0/1	4/3/2	7/1/1	2/4/3	0/1/8	3/3/3	0/0/9	1/0/8	0/0/9	0/0/9	1/1/7
(3) EW$^-$			7/2/0	7/1/1	9/0/0	6/3/0	1/8/0	9/0/0	7/2/0	8/1/0	8/1/0	2/7/0	7/2/0	2/5/2	4/4/1	4/3/2	2/5/2	5/4/0
(4) EW$^-$+SSM$^-$				2/2/5	3/5/1	3/0/6	0/2/7	8/0/1	4/3/2	7/0/2	2/6/1	0/2/7	2/5/2	0/1/8	1/1/7	0/0/9	0/0/9	1/1/7
(5) RUS					4/3/2	3/0/6	1/1/7	9/0/0	5/3/1	6/2/1	5/2/2	1/1/7	5/1/3	1/0/8	1/0/8	0/1/8	1/0/8	1/2/6
(6) CNN$^-$						2/0/7	0/0/9	7/0/2	4/3/2	6/1/2	0/8/1	0/0/9	2/3/4	0/0/9	0/0/9	0/0/9	0/0/9	0/4/5
(7) TMK$_1$							0/3/6	8/1/0	6/0/3	7/2/0	6/0/3	1/2/6	5/1/3	0/3/6	0/4/5	0/3/6	0/3/6	3/1/5
(8) TMK$_2$								9/0/0	8/1/0	7/2/0	8/1/0	1/7/1	7/2/0	1/6/2	3/5/1	2/5/2	1/6/2	3/6/0
(9) CNN$^-$+TMK$_1$									0/0/9	1/0/8	0/1/8	0/1/8	0/1/8	0/0/9	1/0/8	0/0/9	0/0/9	1/1/7
(10) CNN$^-$+TMK$_2$										6/1/2	1/6/2	0/2/7	1/4/4	0/0/9	1/0/8	0/0/9	0/0/9	1/1/7
(11) TMK$_1$+CNN$^-$											1/2/6	1/0/8	1/1/7	0/1/8	1/1/7	0/1/8	0/1/8	2/1/6
(12) TMK$_2$+CNN$^-$												0/2/7	1/6/2	0/1/8	1/0/8	0/0/9	0/0/9	1/2/6
(13) EWW$^-$													6/3/0	2/4/3	4/3/2	3/2/4	2/4/3	3/6/0
(14) EWW$^-$+SSM$^-$														0/2/7	1/1/7	0/1/8	0/1/8	1/3/5
(15) SMOTE															3/5/1	1/7/1	0/9/0	5/3/1
(16) gg-SMOTE																2/5/2	0/6/3	4/3/2
(17) ncn-SMOTE																	0/8/1	4/3/2
(18) rng-SMOTE																		5/3/1
(19) b-SMOTE																		

Tabla C.61: *Prueba t−student del clasificador J48 en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	3/2/4	0/4/5	4/1/4	0/3/6	2/4/3	1/1/7	0/7/2	6/1/2	2/4/3	3/0/6	2/4/3	0/4/5	3/1/5	0/1/8	0/0/9	0/2/7	0/1/8	1/7/1
(2) SSM^-		0/3/6	1/4/4	0/0/9	1/6/2	2/1/6	2/2/5	7/1/1	1/6/2	4/1/4	0/6/3	0/3/6	3/2/4	0/0/9	1/0/8	0/0/9	0/1/8	4/1/4
(3) EW^-			6/0/3	0/3/6	5/4/0	2/1/6	4/5/0	7/2/0	4/5/0	4/1/4	4/5/0	1/7/1	4/2/3	1/5/3	1/2/6	0/4/5	1/5/3	6/3/0
(4) EW^-+SSM^-				0/1/8	4/4/1	1/3/5	3/1/5	7/0/2	2/4/3	3/4/2	2/5/2	1/3/5	0/6/3	1/1/7	1/2/6	0/2/7	1/1/7	4/2/3
(5) RUS					8/1/0	2/6/1	6/3/0	9/0/0	7/2/0	6/2/1	8/1/0	5/4/0	7/2/0	3/5/1	2/5/2	3/4/2	3/5/1	7/2/0
(6) CNN^-						1/1/7	2/3/4	6/1/2	0/7/2	3/1/5	0/7/2	0/4/5	1/3/5	0/1/8	0/0/9	0/1/8	1/0/8	3/4/2
(7) TMK_1							5/2/2	8/0/1	6/1/2	4/5/0	6/1/2	5/2/2	4/4/1	2/3/4	1/5/3	3/2/4	4/1/4	7/1/1
(8) TMK_2								8/1/0	4/3/2	4/1/4	4/3/2	1/3/5	4/1/4	0/3/6	1/1/7	0/3/6	1/2/6	2/7/0
(9) CNN^-+TMK_1									0/1/8	1/0/8	0/2/7	0/2/7	1/0/8	0/1/8	1/1/7	0/0/9	0/1/8	2/1/6
(10) CNN^-+TMK_2										4/1/4	1/6/2	1/3/5	2/4/3	0/2/7	1/1/7	0/2/7	1/2/6	5/2/2
(11) TMK_1+CNN^-											4/1/4	2/3/4	1/4/4	2/2/5	1/4/4	1/3/5	2/2/5	5/1/3
(12) TMK_2+CNN^-												0/5/4	2/3/4	0/2/7	1/0/8	0/1/8	1/2/6	4/3/2
(13) EWW^-													4/2/3	2/3/4	1/3/5	0/4/5	2/3/4	5/4/0
(14) EWW^-+SSM^-														2/1/6	3/1/5	2/1/6	2/1/6	6/0/3
(15) SMOTE															1/4/4	0/7/2	1/8/0	8/1/0
(16) gg-SMOTE																0/8/1	3/5/1	9/0/0
(17) ncn-SMOTE																	3/6/0	8/1/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/1/4	0/4/5	4/1/4	0/3/6	3/3/3	1/1/7	0/8/1	6/1/2	2/4/3	3/1/5	3/3/3	0/4/5	3/2/4	0/1/8	0/0/9	0/2/7	0/1/8	1/7/1
(2) SSM^-		0/2/7	2/3/4	0/0/9	2/5/2	2/1/6	2/1/6	7/1/1	2/5/2	4/1/4	0/6/3	0/2/7	3/2/4	0/0/9	1/0/8	0/0/9	0/0/9	4/0/5
(3) EW^-			6/1/2	1/3/5	6/3/0	2/2/5	4/5/0	7/2/0	5/4/0	4/2/3	5/4/0	0/8/1	4/3/2	0/5/4	1/2/6	0/4/5	1/5/3	6/3/0
(4) EW^-+SSM^-				0/1/8	4/4/1	1/3/5	3/1/5	7/0/2	2/4/3	5/1/3	2/5/2	0/4/5	0/7/2	1/1/7	0/3/6	0/2/7	1/1/7	4/2/3
(5) RUS					8/1/0	3/3/3	6/3/0	9/0/0	8/1/0	6/2/1	8/1/0	4/4/1	7/2/0	2/5/2	2/3/4	2/5/2	3/4/2	6/3/0
(6) CNN^-						1/1/7	2/2/5	6/1/2	0/7/2	3/2/4	0/7/2	0/3/6	1/3/5	0/1/8	0/0/9	0/0/9	0/1/8	3/3/3
(7) TMK_1							5/2/2	8/0/1	7/0/2	5/3/1	6/1/2	5/2/2	3/5/1	2/3/4	1/5/3	3/2/4	3/2/4	7/1/1
(8) TMK_2								7/2/0	3/5/1	4/1/4	4/3/2	1/3/5	4/2/3	0/3/6	1/1/7	0/2/7	0/3/6	2/7/0
(9) CNN^-+TMK_1									0/3/6	1/0/8	1/1/7	0/2/7	1/1/7	0/1/8	1/0/8	0/1/8	0/2/7	3/0/6
(10) CNN^-+TMK_2										4/2/3	1/6/2	1/3/5	2/4/3	0/2/7	1/1/7	0/2/7	0/2/7	5/1/3
(11) TMK_1+CNN^-											3/1/5	2/2/5	2/3/4	1/3/5	2/2/5	1/2/6	1/3/5	5/1/3
(12) TMK_2+CNN^-												0/4/5	2/3/4	0/2/7	1/0/8	0/0/9	0/2/7	4/2/3
(13) EWW^-													4/4/1	1/4/4	1/3/5	0/4/5	2/3/4	5/4/0
(14) EWW^-+SSM^-														2/2/5	2/2/5	2/2/5	2/2/5	5/1/3
(15) SMOTE															1/4/4	0/7/2	1/8/0	8/1/0
(16) gg-SMOTE																0/8/1	3/5/1	9/0/0
(17) ncn-SMOTE																	3/6/0	8/1/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Tabla C.62: *Prueba $t-$student del clasificador MLP en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/3/6	0/6/3	0/1/8	0/0/9	0/5/4	0/0/9	1/7/1	0/0/9	0/1/8	0/0/9	0/2/7	0/3/6	0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	0/7/2
(2) SSM$^-$		5/4/0	0/5/4	1/1/7	1/6/2	2/2/5	6/3/0	1/0/8	0/5/4	1/0/8	0/6/3	5/2/2	0/3/6	1/3/5	1/3/5	1/3/5	1/4/4	6/2/1
(3) EW$^-$			0/3/6	1/0/8	1/5/3	1/1/7	2/6/1	1/0/8	0/2/7	1/0/8	0/3/6	0/5/4	0/1/8	1/0/8	1/0/8	1/0/8	1/1/7	3/4/2
(4) EW$^-$+SSM$^-$				1/2/6	3/5/1	2/2/5	8/1/0	1/0/8	2/6/1	1/0/8	2/6/1	5/3/1	0/5/4	1/5/3	2/4/3	1/5/3	2/6/1	6/3/0
(5) RUS					7/2/0	2/5/2	8/0/1	1/0/8	6/2/1	1/3/5	4/4/1	7/1/1	4/4/1	5/4/0	6/3/0	6/3/0	7/2/0	9/0/0
(6) CNN$^-$						1/3/5	5/3/1	0/0/9	1/4/4	0/1/8	1/7/1	4/2/3	1/3/5	1/4/4	1/2/6	1/3/5	2/2/5	5/3/1
(7) TMK$_1$							8/0/1	0/0/9	5/2/2	0/2/7	4/3/2	7/1/1	3/3/3	3/4/2	3/4/2	4/3/2	5/2/2	9/0/0
(8) TMK$_2$								1/0/8	0/1/8	1/0/8	0/0/9	1/3/5	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	1/6/2
(9) CNN$^-$+TMK$_1$									8/0/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	8/0/1	8/0/1	8/0/1	9/0/0
(10) CNN$^-$+TMK$_2$										1/0/8	1/8/0	6/2/1	1/3/5	2/3/4	2/5/2	3/4/2	2/6/1	8/0/1
(11) TMK$_1$+CNN$^-$											8/0/1	7/1/1	5/3/1	8/0/1	8/0/1	7/1/1	8/0/1	9/0/0
(12) TMK$_2$+CNN$^-$												6/2/1	1/4/4	2/4/3	2/3/4	2/4/3	3/5/1	6/2/1
(13) EWW$^-$													0/2/7	1/1/7	1/1/7	2/1/6	2/1/6	5/4/0
(14) EWW$^-$+SSM$^-$														3/5/1	3/5/1	3/5/1	5/3/1	8/1/0
(15) SMOTE															1/7/1	0/8/1	2/7/0	8/1/0
(16) gg-SMOTE																0/8/1	3/5/1	8/1/0
(17) ncn-SMOTE																	2/7/0	7/2/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	3/6/0	9/0/0	9/0/0	8/1/0	9/0/0	1/8/0	9/0/0	9/0/0	9/0/0	9/0/0	7/2/0	9/0/0	8/1/0	9/0/0	9/0/0	8/1/0	3/6/0
(2) SSM$^-$		0/1/8	5/4/0	3/2/4	0/4/5	3/1/5	0/0/9	8/0/1	2/4/3	7/1/1	1/4/4	1/0/8	5/3/1	1/0/8	1/1/7	1/1/7	1/0/8	0/1/8
(3) EW$^-$			9/0/0	8/0/1	7/1/1	8/0/1	1/5/3	8/0/1	9/0/0	8/0/1	9/0/0	5/4/0	9/0/0	7/1/1	8/0/1	8/0/1	7/1/1	2/4/3
(4) EW$^-$+SSM$^-$				3/1/5	0/3/6	2/2/5	0/0/9	8/0/1	0/4/5	6/1/2	0/3/6	0/1/8	2/6/1	0/0/9	0/0/9	0/1/8	0/0/9	0/0/9
(5) RUS					1/3/5	3/0/6	1/0/8	8/0/1	2/3/4	7/0/2	2/3/4	1/0/8	5/1/3	0/1/8	0/0/9	0/1/8	0/1/8	0/0/9
(6) CNN$^-$						5/2/2	1/0/8	9/0/0	3/6/0	9/0/0	1/8/0	2/0/7	7/2/0	1/2/6	2/3/4	2/3/4	1/2/6	0/2/7
(7) TMK$_1$							1/0/8	9/0/0	3/2/4	9/0/0	3/2/4	1/0/8	6/1/2	2/1/6	3/0/6	2/2/5	2/1/6	0/0/9
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	5/3/1	9/0/0	7/1/1	8/0/1	8/0/1	7/1/1	3/5/1
(9) CNN$^-$+TMK$_1$									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8	1/0/8	1/0/8	0/0/9
(10) CNN$^-$+TMK$_2$										8/0/1	0/5/4	0/1/8	7/2/0	0/2/7	0/3/6	0/3/6	0/1/8	0/0/9
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	1/2/6	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(12) TMK$_2$+CNN$^-$												1/0/8	7/2/0	0/2/7	1/2/6	1/3/5	0/2/7	0/1/8
(13) EWW$^-$													9/0/0	4/3/2	7/0/2	7/0/2	4/3/2	0/4/5
(14) EWW$^-$+SSM$^-$														0/0/9	0/0/9	0/0/9	0/0/9	0/0/9
(15) SMOTE															6/2/1	6/3/0	1/8/0	0/2/7
(16) gg-SMOTE																3/4/2	1/2/6	0/1/8
(17) ncn-SMOTE																	0/2/7	0/0/9
(18) rng-SMOTE																		0/3/6
(19) b-SMOTE																		

Tabla C.63: *Prueba t−student del clasificador MLP en términos de Acc y f_1: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/0/1	2/6/1	8/0/1	8/0/1	8/0/1	8/1/0	2/7/0	9/0/0	8/0/1	8/0/1	8/0/1	5/3/1	8/0/1	5/2/2	7/1/1	6/2/1	5/2/2	2/7/0
(2) SSM$^-$		1/1/7	5/3/1	3/2/4	0/4/5	4/1/4	1/0/8	8/0/1	2/4/3	7/1/1	1/4/4	2/1/6	5/3/1	0/1/8	0/2/7	1/1/7	0/1/8	1/2/6
(3) EW$^-$			8/0/1	7/0/2	7/0/2	8/0/1	2/6/1	8/0/1	8/0/1	7/0/2	8/0/1	4/5/0	8/0/1	3/3/3	5/2/2	5/2/2	3/3/3	1/6/2
(4) EW$^-$+SSM$^-$				3/1/5	0/3/6	3/2/4	1/0/8	8/0/1	0/4/5	6/1/2	0/3/6	1/1/7	1/7/1	0/0/9	0/0/9	0/2/7	0/0/9	1/0/8
(5) RUS					1/3/5	4/0/5	2/0/7	8/0/1	2/3/4	7/0/2	2/3/4	2/0/7	5/1/3	0/1/8	0/0/9	0/1/8	0/1/8	1/0/8
(6) CNN$^-$						6/2/1	2/0/7	9/0/0	3/6/0	9/0/0	1/8/0	3/1/5	7/2/0	1/2/6	1/3/5	1/3/5	1/2/6	1/3/5
(7) TMK$_1$							1/1/7	9/0/0	2/3/4	8/0/1	2/2/5	1/0/8	6/0/3	1/1/7	2/0/7	1/2/6	1/1/7	0/1/8
(8) TMK$_2$								8/0/1	8/0/1	7/0/2	8/0/1	4/3/2	8/0/1	3/3/3	6/1/2	5/2/2	3/3/3	1/6/2
(9) CNN$^-$+TMK$_1$									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8	1/0/8	1/0/8	0/0/9
(10) CNN$^-$+TMK$_2$										8/0/1	0/6/3	1/1/7	7/2/0	0/1/8	0/2/7	0/3/6	0/1/8	1/0/8
(11) TMK$_1$+CNN$^-$											1/0/8	2/0/7	1/2/6	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8
(12) TMK$_2$+CNN$^-$												2/1/6	7/2/0	0/2/7	0/3/6	0/3/6	0/2/7	1/2/6
(13) EWW$^-$													8/0/1	1/4/4	3/3/3	4/2/3	2/2/5	1/4/4
(14) EWW$^-$+SSM$^-$														0/0/9	0/0/9	0/0/9	0/0/9	1/0/8
(15) SMOTE															4/4/1	6/3/0	1/8/0	2/3/4
(16) gg-SMOTE																3/5/1	1/3/5	1/3/5
(17) ncn-SMOTE																	0/3/6	1/2/6
(18) rng-SMOTE																		2/3/4
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	6/0/3	0/7/2	6/0/3	6/0/3	6/1/2	5/2/2	1/7/1	7/1/1	6/0/3	6/0/3	5/1/3	1/5/3	6/0/3	1/4/4	2/4/3	2/4/3	2/3/4	0/8/1
(2) SSM$^-$		2/1/6	1/6/2	2/2/5	1/4/4	4/2/3	2/1/6	8/1/0	0/5/4	6/3/0	0/5/4	1/3/5	2/4/3	0/1/8	0/3/6	0/2/7	0/2/7	2/1/6
(3) EW$^-$			6/1/2	6/0/3	6/1/2	7/1/1	1/8/0	7/2/0	6/1/2	6/1/2	5/2/2	2/6/1	6/1/2	1/4/4	2/5/2	4/2/3	1/4/4	3/5/1
(4) EW$^-$+SSM$^-$				0/4/5	1/4/4	4/2/3	3/0/6	7/2/0	1/4/4	6/2/1	1/4/4	1/3/5	0/8/1	0/1/8	0/2/7	0/2/7	0/1/8	2/1/6
(5) RUS					2/3/4	5/0/4	3/0/6	9/0/0	2/5/2	8/0/1	2/3/4	2/1/6	5/4/0	0/1/8	1/0/8	0/2/7	0/0/9	2/1/6
(6) CNN$^-$						4/2/3	2/1/6	8/0/1	1/7/1	7/1/1	0/8/1	1/2/6	5/3/1	0/3/6	0/3/6	0/3/6	0/3/6	1/2/6
(7) TMK$_1$							1/3/5	8/0/1	1/3/5	6/2/1	1/3/5	0/3/6	5/0/4	0/2/7	1/1/7	0/3/6	0/3/6	1/3/5
(8) TMK$_2$								7/1/1	6/1/2	6/1/2	5/2/2	1/6/2	6/0/3	1/4/4	1/6/2	3/3/3	1/4/4	1/6/2
(9) CNN$^-$+TMK$_1$									0/1/8	0/1/8	0/1/8	0/1/8	0/1/8	0/0/9	0/1/8	0/0/9	0/0/9	1/0/8
(10) CNN$^-$+TMK$_2$										7/2/0	0/8/1	1/3/5	4/4/1	0/2/7	0/3/6	0/3/6	0/2/7	2/1/6
(11) TMK$_1$+CNN$^-$											0/1/8	1/2/6	0/2/7	0/0/9	0/0/9	0/1/8	0/1/8	2/1/6
(12) TMK$_2$+CNN$^-$												1/3/5	6/2/1	0/3/6	0/4/5	1/2/6	0/3/6	3/1/5
(13) EWW$^-$													6/2/1	1/3/5	1/7/1	1/5/3	1/4/4	2/5/2
(14) EWW$^-$+SSM$^-$														0/0/9	0/1/8	0/1/8	0/0/9	2/1/6
(15) SMOTE															3/6/0	6/3/0	1/7/1	4/5/0
(16) gg-SMOTE																3/4/2	0/5/4	3/4/2
(17) ncn-SMOTE																	0/4/5	2/4/3
(18) rng-SMOTE																		4/4/1
(19) b-SMOTE																		

Tabla C.64: *Prueba $t-$student del clasificador MLP en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/2/5	0/6/3	1/3/5	0/2/7	2/3/4	1/1/7	1/7/1	6/1/2	1/3/5	2/2/5	1/3/5	0/3/6	1/2/6	0/0/9	0/0/9	0/0/9	0/1/8	0/7/2
(2) SSM$^-$		3/3/3	1/6/2	0/1/8	1/4/4	3/1/5	3/4/2	8/0/1	1/3/5	4/4/1	1/4/4	2/2/5	1/5/3	0/1/8	0/2/7	0/1/8	0/1/8	5/2/2
(3) EW$^-$			3/2/4	0/2/7	2/4/3	2/2/5	3/6/0	7/2/0	2/4/3	3/2/4	2/4/3	0/6/3	2/2/5	0/0/9	1/0/8	0/0/9	0/1/8	3/4/2
(4) EW$^-$+SSM$^-$				0/2/7	3/3/3	3/2/4	5/2/2	7/1/1	0/5/4	4/3/2	1/4/4	2/3/4	0/6/3	0/1/8	0/1/8	0/2/7	0/2/7	4/3/2
(5) RUS					7/2/0	5/3/1	8/1/0	9/0/0	7/2/0	6/3/0	6/3/0	5/3/1	6/3/0	1/5/3	1/5/3	0/6/3	1/4/4	6/3/0
(6) CNN$^-$						2/2/5	4/3/2	7/1/1	0/8/1	4/3/2	0/7/2	3/2/4	2/3/4	0/1/8	0/1/8	0/1/8	0/2/7	4/3/2
(7) TMK$_1$							7/0/2	8/0/1	3/3/3	6/1/2	3/3/3	4/3/2	4/2/3	0/4/5	2/1/6	2/1/6	1/2/6	7/1/1
(8) TMK$_2$								7/1/1	2/2/5	3/1/5	2/3/4	0/3/6	2/1/6	0/0/9	0/1/8	0/0/9	0/0/9	1/6/2
(9) CNN$^-$+TMK$_1$									0/2/7	1/0/8	0/2/7	0/1/8	0/2/7	0/0/9	0/1/8	0/0/9	0/1/8	1/1/7
(10) CNN$^-$+TMK$_2$										4/5/0	0/8/1	3/2/4	3/4/2	0/1/8	0/4/5	1/1/7	0/3/6	5/2/2
(11) TMK$_1$+CNN$^-$											1/3/5	3/2/4	2/2/5	1/1/7	0/3/6	1/2/6	1/2/6	5/2/2
(12) TMK$_2$+CNN$^-$												3/2/4	4/3/2	0/1/8	0/2/7	1/1/7	0/3/6	5/2/2
(13) EWW$^-$													2/4/3	0/3/6	1/2/6	0/3/6	0/3/6	4/5/0
(14) EWW$^-$+SSM$^-$														0/2/7	0/2/7	0/2/7	0/2/7	4/4/1
(15) SMOTE															2/6/1	2/5/2	2/6/1	8/1/0
(16) gg-SMOTE																1/7/1	1/7/1	8/1/0
(17) ncn-SMOTE																	1/8/0	7/2/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/2/5	0/6/3	2/2/5	1/1/7	2/3/4	1/1/7	1/7/1	6/1/2	1/3/5	3/1/5	1/3/5	0/3/6	1/2/6	0/0/9	0/0/9	0/0/9	0/1/8	0/7/2
(2) SSM$^-$		3/2/4	1/6/2	0/1/8	1/4/4	3/1/5	3/4/2	7/2/0	1/3/5	4/3/2	1/3/5	2/2/5	1/5/3	0/1/8	0/2/7	0/1/8	0/1/8	5/1/3
(3) EW$^-$			3/3/3	0/3/6	4/2/3	2/3/4	3/6/0	7/1/1	2/4/3	4/2/3	2/4/3	0/6/3	2/2/5	0/0/9	1/1/7	0/1/8	0/1/8	3/4/2
(4) EW$^-$+SSM$^-$				0/2/7	3/3/3	2/4/3	5/2/2	6/1/2	0/5/4	4/2/3	0/4/5	2/3/4	0/6/3	0/1/8	0/1/8	0/1/8	0/2/7	4/2/3
(5) RUS					7/2/0	5/3/1	7/2/0	9/0/0	6/3/0	7/2/0	6/3/0	5/3/1	6/3/0	1/4/4	1/5/3	1/5/3	1/4/4	6/2/1
(6) CNN$^-$						2/2/5	4/3/2	7/1/1	0/8/1	4/3/2	0/7/2	3/2/4	2/3/4	0/1/8	0/1/8	0/1/8	0/2/7	4/2/3
(7) TMK$_1$							7/0/2	8/0/1	3/4/2	6/1/2	2/4/3	3/3/3	3/3/3	0/4/5	2/1/6	2/1/6	1/2/6	7/1/1
(8) TMK$_2$								6/2/1	2/2/5	3/2/4	2/3/4	0/3/6	2/2/5	0/0/9	0/1/8	0/0/9	0/0/9	1/5/3
(9) CNN$^-$+TMK$_1$									1/1/7	0/1/8	0/2/7	0/2/7	1/1/7	0/0/9	0/1/8	0/1/8	0/0/9	1/2/6
(10) CNN$^-$+TMK$_2$										4/5/0	0/8/1	3/2/4	3/4/2	0/1/8	0/2/7	1/1/7	0/2/7	5/2/2
(11) TMK$_1$+CNN$^-$											0/5/4	3/2/4	3/1/5	1/1/7	0/3/6	1/2/6	1/2/6	5/1/3
(12) TMK$_2$+CNN$^-$												3/2/4	4/3/2	0/1/8	0/2/7	1/1/7	0/3/6	5/2/2
(13) EWW$^-$													3/4/2	0/3/6	1/2/6	0/3/6	0/2/7	4/5/0
(14) EWW$^-$+SSM$^-$														0/2/7	0/2/7	0/2/7	0/2/7	4/4/1
(15) SMOTE															3/6/0	2/5/2	2/6/1	8/1/0
(16) gg-SMOTE																1/7/1	1/7/1	8/1/0
(17) ncn-SMOTE																	1/7/1	7/2/0
(18) rng-SMOTE																		8/1/0
(19) b-SMOTE																		

Tabla C.65: *Prueba t−student del clasificador RBF en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/1/8	0/5/4	0/0/9	0/0/9	0/2/7	0/0/9	0/6/3	0/0/9	0/0/9	0/0/9	0/0/9	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	1/5/3
(2) SSM$^-$		8/0/1	0/3/6	1/0/8	1/3/5	4/0/5	8/1/0	1/0/8	0/1/8	1/0/8	0/1/8	7/1/1	0/2/7	1/0/8	1/0/8	1/0/8	1/0/8	7/1/1
(3) EW$^-$			0/1/8	0/1/8	1/1/7	1/0/8	2/6/1	1/0/8	0/1/8	1/0/8	0/1/8	0/6/3	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	2/6/1
(4) EW$^-$+SSM$^-$				1/0/8	3/2/4	4/0/5	9/0/0	1/0/8	1/3/5	1/0/8	1/4/4	8/0/1	0/5/4	1/1/7	1/0/8	1/1/7	1/0/8	8/1/0
(5) RUS					9/0/0	6/1/2	8/0/1	1/0/8	8/0/1	1/3/5	8/0/1	7/2/0	7/1/1	5/2/2	4/4/1	3/4/2	4/3/2	9/0/0
(6) CNN$^-$						3/0/6	7/1/1	0/0/9	0/5/4	0/0/9	0/6/3	7/0/2	4/1/4	1/0/8	0/1/8	0/1/8	0/1/8	7/0/2
(7) TMK$_1$							8/0/1	0/0/9	4/1/4	0/2/7	5/0/4	7/1/1	3/2/4	2/1/6	2/1/6	2/1/6	2/1/6	9/0/0
(8) TMK$_2$								1/0/8	0/1/8	1/0/8	0/1/8	1/4/4	0/0/9	0/1/8	1/0/8	1/0/8	1/0/8	2/5/2
(9) CNN$^-$+TMK$_1$									8/0/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	8/0/1	8/0/1	7/1/1	9/0/0
(10) CNN$^-$+TMK$_2$										1/0/8	2/7/0	8/0/1	4/2/3	2/2/5	1/1/7	1/1/7	1/1/7	8/0/1
(11) TMK$_1$+CNN$^-$											8/0/1	8/0/1	8/0/1	5/3/1	3/5/1	3/4/2	4/4/1	9/0/0
(12) TMK$_2$+CNN$^-$												8/0/1	4/2/3	2/1/6	1/1/7	1/2/6	1/1/7	8/0/1
(13) EWW$^-$													0/1/8	0/0/9	0/0/9	0/0/9	0/0/9	3/6/0
(14) EWW$^-$+SSM$^-$														1/0/8	1/0/8	1/1/7	1/1/7	9/0/0
(15) SMOTE															2/4/3	0/5/4	0/4/5	9/0/0
(16) gg-SMOTE																0/7/2	2/4/3	9/0/0
(17) ncn-SMOTE																	3/5/1	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	4/5/0	9/0/0	9/0/0	8/1/0	9/0/0	3/6/0	9/0/0	9/0/0	9/0/0	9/0/0	5/4/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	3/6/0
(2) SSM$^-$		0/0/9	3/5/1	6/0/3	3/2/4	3/2/4	0/0/9	8/0/1	6/1/2	6/1/2	5/2/2	1/0/8	5/3/1	5/1/3	6/0/3	6/0/3	6/0/3	1/0/8
(3) EW$^-$			9/0/0	8/1/0	7/1/1	8/0/1	1/5/3	8/0/1	9/0/0	8/0/1	9/0/0	4/5/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	1/6/2
(4) EW$^-$+SSM$^-$				5/0/4	2/1/6	2/2/5	0/0/9	8/0/1	3/2/4	5/2/2	2/1/6	0/0/9	4/2/3	4/1/4	4/2/3	5/1/3	4/1/4	0/0/9
(5) RUS					1/0/8	2/1/6	1/0/8	8/0/1	2/1/6	5/3/1	1/1/7	0/1/8	4/0/5	2/3/4	3/3/3	3/3/3	2/3/4	0/1/8
(6) CNN$^-$						5/0/4	1/1/7	9/0/0	6/3/0	9/0/0	4/3/2	2/0/7	6/1/2	5/2/2	6/1/2	6/3/0	6/2/1	2/0/7
(7) TMK$_1$							1/0/8	9/0/0	6/0/3	7/2/0	5/0/4	1/0/8	6/1/2	7/0/2	7/0/2	7/0/2	7/0/2	0/1/8
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	4/4/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/5/2
(9) CNN$^-$+TMK$_1$									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(10) CNN$^-$+TMK$_2$										8/0/1	0/4/5	1/0/8	5/1/3	4/1/4	4/2/3	4/4/1	4/3/2	1/0/8
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	3/0/6	1/3/5	1/3/5	2/2/5	2/2/5	0/0/9
(12) TMK$_2$+CNN$^-$												1/0/8	6/1/2	4/2/3	4/2/3	4/4/1	4/3/2	1/0/8
(13) EWW$^-$													9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	0/6/3
(14) EWW$^-$+SSM$^-$														4/1/4	5/0/4	5/1/3	5/0/4	0/0/9
(15) SMOTE															3/2/4	6/3/0	5/3/1	0/0/9
(16) gg-SMOTE																5/4/0	5/2/2	0/0/9
(17) ncn-SMOTE																	0/6/3	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

Tabla C.66: *Prueba $t-$student del clasificador RBF en términos de Acc y f_1: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	9/0/0	2/6/1	9/0/0	9/0/0	7/2/0	9/0/0	2/6/1	9/0/0	9/0/0	9/0/0	8/1/0	3/6/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	3/6/0
(2) SSM$^-$		0/0/9	3/5/1	6/0/3	3/2/4	3/2/4	0/0/9	8/0/1	5/2/2	6/0/3	4/2/3	0/1/8	5/3/1	5/1/3	6/0/3	6/0/3	5/1/3	1/0/8
(3) EW$^-$			9/0/0	8/1/0	8/0/1	8/0/1	1/7/1	8/0/1	9/0/0	8/0/1	9/0/0	2/7/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	2/6/1
(4) EW$^-$+SSM$^-$				5/0/4	2/1/6	3/1/5	0/0/9	8/0/1	3/2/4	5/1/3	2/1/6	0/0/9	4/2/3	4/1/4	4/1/4	4/1/4	4/1/4	0/0/9
(5) RUS					1/0/8	2/1/6	1/0/8	8/0/1	2/2/5	5/3/1	2/1/6	0/1/8	4/0/5	2/3/4	2/4/3	2/4/3	2/3/4	0/1/8
(6) CNN$^-$						5/0/4	1/1/7	9/0/0	4/5/0	9/0/0	3/4/2	1/1/7	6/1/2	5/2/2	6/1/2	6/3/0	5/3/1	2/0/7
(7) TMK$_1$							1/0/8	9/0/0	6/1/2	7/2/0	5/0/4	1/0/8	6/0/3	6/1/2	7/0/2	7/0/2	7/0/2	0/1/8
(8) TMK$_2$								8/0/1	9/0/0	8/0/1	9/0/0	2/6/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/6/1
(9) CNN$^-$+TMK$_1$									1/0/8	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(10) CNN$^-$+TMK$_2$										8/0/1	0/4/5	0/1/8	5/1/3	4/1/4	4/2/3	4/4/1	4/2/3	1/0/8
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	3/1/5	1/3/5	1/3/5	2/2/5	2/2/5	0/1/8
(12) TMK$_2$+CNN$^-$												1/0/8	6/0/3	4/2/3	4/2/3	4/4/1	4/3/2	1/0/8
(13) EWW$^-$													9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	1/6/2
(14) EWW$^-$+SSM$^-$														4/1/4	5/0/4	5/0/4	4/1/4	0/0/9
(15) SMOTE															3/1/5	6/3/0	5/3/1	0/1/8
(16) gg-SMOTE																5/4/0	5/3/1	0/1/8
(17) ncn-SMOTE																	0/6/3	0/1/8
(18) rng-SMOTE																		0/1/8
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	2/2/5	0/6/3	3/1/5	2/2/5	2/3/4	1/3/5	0/6/3	4/1/4	2/2/5	2/2/5	2/2/5	0/5/4	2/2/5	1/3/5	2/2/5	2/2/5	2/2/5	0/7/2
(2) SSM$^-$		4/1/4	2/4/3	4/1/4	4/2/3	4/0/5	4/1/4	8/0/1	4/2/3	5/1/3	3/3/3	3/2/4	2/3/4	3/1/5	3/2/4	3/2/4	3/1/5	4/3/2
(3) EW$^-$			4/1/4	2/3/4	4/1/4	2/3/4	2/7/0	6/0/3	4/1/4	4/1/4	4/1/4	0/8/1	4/1/4	2/3/4	3/2/4	3/2/4	2/3/4	3/5/1
(4) EW$^-$+SSM$^-$				4/0/5	3/2/4	4/0/5	4/1/4	8/1/0	4/2/3	5/0/4	3/2/4	3/2/4	2/5/2	3/2/4	3/2/4	3/2/4	3/2/4	4/2/3
(5) RUS					3/3/3	4/2/3	5/2/2	9/0/0	2/5/2	6/3/0	2/5/2	4/2/3	4/1/4	3/2/4	2/4/3	3/4/2	1/4/4	5/2/2
(6) CNN$^-$						3/2/4	4/2/3	7/1/1	2/6/1	6/1/2	1/5/3	3/2/4	3/3/3	2/4/3	2/4/3	2/5/2	2/5/2	3/3/3
(7) TMK$_1$							4/3/2	8/0/1	3/2/4	3/4/2	3/2/4	4/3/2	3/2/4	2/3/4	3/2/4	3/3/3	2/4/3	6/2/1
(8) TMK$_2$								5/1/3	3/2/4	4/1/4	3/2/4	0/6/3	4/1/4	1/3/5	2/2/5	2/2/5	2/2/5	2/6/1
(9) CNN$^-$+TMK$_1$									0/1/8	1/0/8	0/1/8	2/1/6	0/1/8	0/1/8	0/1/8	0/1/8	0/0/9	4/1/4
(10) CNN$^-$+TMK$_2$										6/3/0	0/6/3	4/0/5	3/2/4	2/3/4	2/4/3	2/6/1	2/4/3	5/2/2
(11) TMK$_1$+CNN$^-$											0/2/7	4/1/4	2/2/5	2/1/6	0/3/6	1/2/6	1/2/6	5/1/3
(12) TMK$_2$+CNN$^-$												4/2/3	3/3/3	2/4/3	2/4/3	2/6/1	2/5/2	5/2/2
(13) EWW$^-$													4/1/4	2/3/4	4/1/4	3/2/4	2/3/4	3/6/0
(14) EWW$^-$+SSM$^-$														3/2/4	3/2/4	4/1/4	3/2/4	5/2/2
(15) SMOTE															3/3/3	6/3/0	4/3/2	6/2/1
(16) gg-SMOTE																4/5/0	4/2/3	5/2/2
(17) ncn-SMOTE																	1/4/4	5/2/2
(18) rng-SMOTE																		5/2/2
(19) b-SMOTE																		

Tabla C.67: *Prueba t−student del clasificador RBF en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/7	0/5/4	1/1/7	0/0/9	0/2/7	0/0/9	0/6/3	2/2/5	0/0/9	0/1/8	0/0/9	0/4/5	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	1/5/3
(2) SSM$^-$		6/0/3	1/3/5	1/2/6	2/4/3	4/1/4	6/1/2	6/2/1	2/2/5	2/3/4	2/2/5	5/1/3	1/3/5	1/1/7	1/1/7	1/2/6	1/1/7	6/1/2
(3) EW$^-$			2/1/6	0/1/8	1/1/7	1/0/8	2/7/0	4/1/4	1/1/7	1/1/7	1/1/7	0/6/3	2/0/7	1/0/8	0/1/8	0/1/8	0/1/8	2/6/1
(4) EW$^-$+SSM$^-$				1/2/6	3/4/2	4/2/3	6/1/2	7/2/0	3/3/3	3/2/4	2/4/3	6/0/3	1/4/4	2/1/6	1/2/6	2/1/6	2/0/7	7/1/1
(5) RUS					7/2/0	6/3/0	9/0/0	8/1/0	8/1/0	5/3/1	7/2/0	7/2/0	5/3/1	3/2/4	3/4/2	3/4/2	1/4/4	9/0/0
(6) CNN$^-$						4/1/4	7/1/1	6/2/1	0/5/4	2/2/5	0/6/3	7/0/2	2/1/6	0/1/8	0/1/8	0/3/6	0/2/7	7/0/2
(7) TMK_1							8/0/1	7/0/2	2/2/5	2/3/4	3/1/5	7/1/1	2/2/5	0/3/6	0/3/6	0/3/6	0/3/6	9/0/0
(8) TMK_2								4/1/4	1/1/7	1/1/7	1/1/7	0/4/5	2/0/7	0/1/8	0/0/9	0/0/9	0/0/9	2/5/2
(9) CNN$^-$+TMK_1									0/2/7	1/1/7	0/2/7	4/1/4	0/2/7	0/1/8	0/0/9	0/0/9	0/0/9	5/1/3
(10) CNN$^-$+TMK_2										4/0/5	0/7/2	7/0/2	3/1/5	0/2/7	0/1/8	0/3/6	0/2/7	8/1/0
(11) TMK_1+CNN$^-$											3/2/4	7/1/1	4/1/4	2/1/6	0/4/5	1/3/5	1/2/6	8/1/0
(12) TMK_2+CNN$^-$												7/0/2	3/2/4	0/1/8	0/1/8	0/3/6	0/2/7	8/1/0
(13) EWW$^-$													2/1/6	1/0/8	0/2/7	0/2/7	0/2/7	3/6/0
(14) EWW$^-$+SSM$^-$														2/2/5	2/1/6	2/2/5	2/0/7	8/0/1
(15) SMOTE															2/6/1	4/5/0	4/3/2	9/0/0
(16) gg-SMOTE																2/6/1	2/5/2	9/0/0
(17) ncn-SMOTE																	0/6/3	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	1/1/7	0/6/3	1/1/7	0/0/9	0/2/7	0/0/9	0/6/3	3/1/5	0/0/9	0/1/8	0/0/9	0/4/5	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	0/6/3
(2) SSM$^-$		6/0/3	1/3/5	1/2/6	3/3/3	4/2/3	6/1/2	7/1/1	3/1/5	3/2/4	3/1/5	5/1/3	2/2/5	2/0/7	1/1/7	2/1/6	2/0/7	6/1/2
(3) EW$^-$			2/1/6	0/1/8	1/1/7	1/0/8	2/7/0	4/2/3	1/1/7	1/1/7	1/1/7	0/6/3	2/0/7	1/0/8	0/1/8	0/1/8	0/1/8	2/7/0
(4) EW$^-$+SSM$^-$				1/2/6	3/4/2	5/1/3	6/1/2	7/1/1	3/3/3	4/1/4	2/4/3	6/0/3	1/4/4	2/1/6	2/1/6	2/1/6	2/0/7	7/1/1
(5) RUS					7/2/0	6/3/0	9/0/0	9/0/0	8/1/0	5/3/1	7/2/0	7/2/0	4/3/2	3/2/4	2/5/2	3/4/2	1/4/4	9/0/0
(6) CNN$^-$						5/0/4	7/1/1	6/1/2	0/6/3	3/2/4	0/6/3	7/0/2	2/1/6	0/1/8	0/1/8	1/2/6	0/2/7	7/2/0
(7) TMK_1							8/0/1	6/1/2	2/1/6	2/3/4	3/0/6	7/1/1	2/2/5	0/3/6	0/3/6	1/2/6	0/3/6	9/0/0
(8) TMK_2								4/2/3	1/1/7	1/1/7	1/1/7	0/4/5	2/0/7	0/0/9	0/0/9	0/0/9	0/0/9	2/6/1
(9) CNN$^-$+TMK_1									0/2/7	1/0/8	1/1/7	3/2/4	0/1/8	0/1/8	0/1/8	0/1/8	0/0/9	5/1/3
(10) CNN$^-$+TMK_2										4/1/4	0/7/2	7/0/2	3/1/5	0/2/7	0/1/8	0/2/7	0/2/7	8/1/0
(11) TMK_1+CNN$^-$											2/3/4	6/2/1	4/1/4	1/2/6	0/3/6	1/3/5	1/2/6	8/1/0
(12) TMK_2+CNN$^-$												7/0/2	3/2/4	0/1/8	0/1/8	1/3/5	0/2/7	8/1/0
(13) EWW$^-$													2/1/6	1/0/8	0/2/7	0/2/7	0/2/7	3/6/0
(14) EWW$^-$+SSM$^-$														2/3/4	2/2/5	2/2/5	2/1/6	7/1/1
(15) SMOTE															1/6/2	4/5/0	3/4/2	9/0/0
(16) gg-SMOTE																2/7/0	2/5/2	9/0/0
(17) ncn-SMOTE																	0/5/4	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

Tabla C.68: *Prueba $t-$student del clasificador SVM en términos de tvp y tvn: bases de datos fuertemente desbalanceadas*

tvp

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/2/7	0/4/5	0/1/8	0/0/9	0/2/7	0/2/7	3/3/3	0/0/9	0/1/8	0/0/9	0/1/8	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	2/3/4
(2) SSM^-		7/2/0	0/5/4	2/0/7	2/3/4	6/1/2	7/2/0	1/0/8	0/2/7	1/0/8	0/2/7	6/2/1	0/1/8	2/0/7	2/0/7	2/0/7	2/0/7	7/2/0
(3) EW^-			0/1/8	0/1/8	1/2/6	2/1/6	4/2/3	1/0/8	0/1/8	1/0/8	0/1/8	0/5/4	0/0/9	0/1/8	0/1/8	0/1/8	0/1/8	4/3/2
(4) EW^-+SSM^-				2/0/7	3/4/2	7/1/1	8/1/0	1/0/8	1/3/5	1/0/8	2/4/3	8/1/0	0/4/5	2/0/7	2/0/7	2/0/7	2/0/7	8/1/0
(5) RUS					7/0/2	7/0/2	8/0/1	1/0/8	5/1/3	2/0/7	6/0/3	8/1/0	7/0/2	5/3/1	4/4/1	4/4/1	4/4/1	9/0/0
(6) CNN^-						5/1/3	7/1/1	0/1/8	0/4/5	0/1/8	0/5/4	6/1/2	2/2/5	2/1/6	2/1/6	2/1/6	2/1/6	6/2/1
(7) TMK_1							6/2/1	0/1/8	1/1/7	0/1/8	2/1/6	6/1/2	1/0/8	2/0/7	2/0/7	2/0/7	2/0/7	6/2/1
(8) TMK_2								1/0/8	0/1/8	1/0/8	0/1/8	3/2/4	0/0/9	1/0/8	1/0/8	1/0/8	1/0/8	3/4/2
(9) CNN^-+TMK_1									7/1/1	6/3/0	7/1/1	8/0/1	8/0/1	8/0/1	8/0/1	7/1/1	7/1/1	9/0/0
(10) CNN^-+TMK_2										1/1/7	5/4/0	8/1/0	4/3/2	4/2/3	4/2/3	3/3/3	4/2/3	8/1/0
(11) TMK_1+CNN^-											7/1/1	8/0/1	8/0/1	7/0/2	7/0/2	7/0/2	7/0/2	9/0/0
(12) TMK_2+CNN^-												7/2/0	2/4/3	3/2/4	3/2/4	3/2/4	3/2/4	8/1/0
(13) EWW^-													0/0/9	0/1/8	0/1/8	0/1/8	0/1/8	5/2/2
(14) EWW^-+SSM^-														2/2/5	2/2/5	2/2/5	2/2/5	9/0/0
(15) SMOTE															1/7/1	1/6/2	0/7/2	9/0/0
(16) gg-SMOTE																1/7/1	1/6/2	9/0/0
(17) ncn-SMOTE																	1/6/2	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

tvn

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	8/1/0	5/4/0	9/0/0	9/0/0	7/2/0	6/3/0	1/3/5	9/0/0	8/1/0	9/0/0	8/1/0	5/4/0	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	2/3/4
(2) SSM^-		0/2/7	6/2/1	6/0/3	6/1/2	3/0/6	0/1/8	8/0/1	8/0/1	8/0/1	8/0/1	1/2/6	7/2/0	6/0/3	6/0/3	6/0/3	6/0/3	0/1/8
(3) EW^-			9/0/0	8/0/1	7/1/1	5/2/2	1/2/6	8/0/1	8/1/0	8/0/1	8/1/0	3/6/0	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/3/6
(4) EW^-+SSM^-				6/0/3	4/2/3	1/1/7	0/0/9	8/0/1	7/0/2	8/0/1	6/0/3	0/1/8	6/3/0	5/1/3	5/1/3	5/1/3	5/1/3	0/0/9
(5) RUS					3/0/6	1/0/8	1/0/8	8/0/1	5/0/4	7/1/1	4/1/4	1/0/8	3/0/6	0/2/7	0/2/7	1/2/6	0/3/6	0/0/9
(6) CNN^-						3/2/4	1/1/7	9/0/0	8/1/0	9/0/0	5/4/0	1/2/6	5/0/4	6/0/3	6/0/3	6/0/3	6/0/3	1/1/7
(7) TMK_1							1/2/6	9/0/0	7/1/1	9/0/0	6/2/1	2/2/5	8/0/1	8/0/1	8/0/1	8/0/1	8/0/1	1/2/6
(8) TMK_2								8/0/1	8/1/0	8/0/1	8/1/0	6/2/1	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	2/5/2
(9) CNN^-+TMK_1									1/0/8	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/0/9
(10) CNN^-+TMK_2										8/0/1	0/1/8	0/1/8	2/0/7	3/1/5	4/0/5	3/1/5	3/1/5	0/1/8
(11) TMK_1+CNN^-											1/0/8	1/0/8	1/0/8	1/1/7	1/1/7	1/1/7	1/1/7	0/1/8
(12) TMK_2+CNN^-												0/1/8	3/1/5	4/1/4	4/1/4	4/1/4	4/1/4	0/1/8
(13) EWW^-													9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/2/7
(14) EWW^-+SSM^-														5/1/3	5/1/3	5/1/3	5/1/3	0/0/9
(15) SMOTE															4/3/2	5/4/0	3/5/1	0/0/9
(16) gg-SMOTE																3/4/2	3/4/2	0/0/9
(17) ncn-SMOTE																	1/4/4	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

Tabla C.69: *Prueba t−student del clasificador SVM en términos de Acc y f_1: bases de datos fuertemente desbalanceadas*

Acc

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	6/2/1	2/5/2	8/0/1	9/0/0	7/2/0	6/2/1	2/4/3	9/0/0	8/1/0	9/0/0	8/1/0	2/5/2	9/0/0	9/0/0	9/0/0	9/0/0	9/0/0	1/4/4
(2) SSM$^-$		2/1/6	4/3/2	5/1/3	6/1/2	4/0/5	1/2/6	8/0/1	7/1/1	8/0/1	6/2/1	0/3/6	6/1/2	5/0/4	5/0/4	5/0/4	5/0/4	1/2/6
(3) EW$^-$			7/2/0	8/0/1	6/2/1	6/1/2	2/4/3	8/0/1	8/1/0	8/0/1	8/1/0	2/7/0	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/5/4
(4) EW$^-$+SSM$^-$				6/0/3	5/1/3	2/1/6	1/1/7	8/0/1	7/0/2	8/0/1	6/0/3	0/2/7	5/4/0	5/1/3	5/1/3	5/1/3	5/1/3	0/2/7
(5) RUS					3/0/6	1/0/8	1/0/8	8/0/1	5/0/4	7/1/1	4/1/4	1/0/8	3/0/6	0/2/7	0/2/7	1/2/6	0/3/6	0/0/9
(6) CNN$^-$						3/2/4	1/2/6	9/0/0	7/2/0	9/0/0	5/4/0	1/2/6	4/2/3	5/1/3	5/1/3	5/1/3	5/1/3	1/2/6
(7) TMK$_1$							1/3/5	9/0/0	7/1/1	8/1/0	5/3/1	2/1/6	7/1/1	8/0/1	8/0/1	8/0/1	8/0/1	0/3/6
(8) TMK$_2$								8/0/1	8/1/0	8/0/1	8/1/0	3/4/2	9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	1/6/2
(9) CNN$^-$+TMK$_1$									1/0/8	0/1/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	1/0/8	0/1/8
(10) CNN$^-$+TMK$_2$										8/0/1	0/2/7	0/1/8	2/0/7	3/1/5	4/0/5	3/1/5	4/0/5	0/1/8
(11) TMK$_1$+CNN$^-$											1/0/8	1/0/8	1/0/8	1/1/7	1/1/7	1/1/7	1/1/7	0/1/8
(12) TMK$_2$+CNN$^-$												0/1/8	3/2/4	4/0/5	4/0/5	4/0/5	4/0/5	0/1/8
(13) EWW$^-$													9/0/0	8/0/1	8/0/1	8/0/1	8/0/1	0/5/4
(14) EWW$^-$+SSM$^-$														5/0/4	5/1/3	5/1/3	4/2/3	0/0/9
(15) SMOTE															4/3/2	6/3/0	3/5/1	0/0/9
(16) gg-SMOTE																4/3/2	2/5/2	0/0/9
(17) ncn-SMOTE																	1/4/4	0/0/9
(18) rng-SMOTE																		0/0/9
(19) b-SMOTE																		

f_1

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	4/2/3	0/6/3	3/3/3	4/1/4	4/3/2	2/3/4	1/4/4	5/0/4	4/2/3	5/0/4	4/2/3	0/4/5	3/2/4	3/2/4	3/2/4	3/2/4	3/2/4	1/3/5
(2) SSM$^-$		2/2/5	2/3/4	3/0/6	6/1/2	4/1/4	2/2/5	7/0/2	5/2/2	7/0/2	5/1/3	1/3/5	3/1/5	3/0/6	3/0/6	3/0/6	3/0/6	3/2/4
(3) EW$^-$			5/2/2	5/1/3	6/2/1	4/3/2	4/2/3	7/0/2	5/3/1	7/0/2	5/3/1	0/7/2	5/1/3	4/1/4	5/0/4	5/0/4	5/0/4	3/3/3
(4) EW$^-$+SSM$^-$				5/0/4	6/3/0	4/2/3	3/3/3	8/0/1	5/4/0	8/0/1	5/4/0	1/3/5	3/4/2	4/1/4	4/1/4	4/1/4	4/1/4	4/2/3
(5) RUS					7/0/2	5/0/4	4/2/3	9/0/0	6/0/3	9/0/0	6/1/2	3/0/6	4/0/5	0/3/6	0/2/7	0/4/5	0/3/6	4/1/4
(6) CNN$^-$						3/1/5	3/1/5	5/1/3	4/2/3	5/1/3	1/5/3	1/2/6	0/2/7	2/0/7	2/0/7	2/0/7	2/0/7	2/2/5
(7) TMK$_1$							4/2/3	6/0/3	3/2/4	6/0/3	4/2/3	2/2/5	3/1/5	3/1/5	4/0/5	4/0/5	3/1/5	4/2/3
(8) TMK$_2$								6/0/3	4/2/3	6/0/3	5/1/3	3/2/4	3/2/4	3/1/5	3/2/4	3/1/5	3/2/4	3/4/2
(9) CNN$^-$+TMK$_1$									1/0/8	1/2/6	1/1/7	2/0/7	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	3/0/6
(10) CNN$^-$+TMK$_2$										8/0/1	2/2/5	1/1/7	0/2/7	2/1/6	3/0/6	2/1/6	3/0/6	3/2/4
(11) TMK$_1$+CNN$^-$											1/1/7	2/0/7	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	3/0/6
(12) TMK$_2$+CNN$^-$												1/2/6	0/3/6	2/0/7	2/0/7	2/0/7	2/0/7	4/1/4
(13) EWW$^-$													6/1/2	6/0/3	6/0/3	6/0/3	6/0/3	4/2/3
(14) EWW$^-$+SSM$^-$														3/1/5	3/2/4	3/2/4	3/1/5	5/1/3
(15) SMOTE															3/5/1	3/6/0	2/6/1	5/1/3
(16) gg-SMOTE																3/5/1	1/6/2	5/1/3
(17) ncn-SMOTE																	0/5/4	5/1/3
(18) rng-SMOTE																		5/1/3
(19) b-SMOTE																		

Tabla C.70: *Prueba $t-$student del clasificador SVM en términos de $IBA_{0.1}$ y Gm: bases de datos fuertemente desbalanceadas*

$IBA_{0.1}$

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/2/7	0/4/5	0/1/8	0/0/9	1/2/6	1/2/6	3/3/3	4/3/2	1/1/7	3/0/6	1/1/7	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	2/3/4
(2) SSM$^-$		6/2/1	1/5/3	0/1/8	4/2/3	5/2/2	4/2/3	7/1/1	3/4/2	6/1/2	2/4/3	5/2/2	1/2/6	0/1/8	0/1/8	0/1/8	0/1/8	5/3/1
(3) EW$^-$			1/1/7	0/0/9	2/2/5	3/1/5	5/2/2	7/1/1	2/1/6	5/1/3	2/1/6	0/5/4	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	4/3/2
(4) EW$^-$+SSM$^-$				0/4/5	6/2/1	6/2/1	5/2/2	8/1/0	4/2/3	7/1/1	5/2/2	7/1/1	0/7/2	0/2/7	0/2/7	0/3/6	0/2/7	6/2/1
(5) RUS					8/1/0	8/0/1	9/0/0	9/0/0	7/0/2	8/1/0	8/1/0	9/0/0	5/3/1	0/5/4	1/4/4	0/7/2	0/5/4	9/0/0
(6) CNN$^-$						4/1/4	6/1/2	7/0/2	3/2/4	4/2/3	1/4/4	5/1/3	1/0/8	0/1/8	0/1/8	0/1/8	0/1/8	5/2/2
(7) TMK$_1$							5/2/2	6/1/2	2/1/6	5/0/4	3/1/5	4/2/3	1/1/7	1/0/8	1/0/8	1/0/8	1/0/8	5/2/2
(8) TMK$_2$								5/2/2	2/1/6	4/0/5	2/1/6	2/2/5	2/1/6	0/0/9	0/0/9	0/0/9	0/0/9	3/4/2
(9) CNN$^-$+TMK$_1$									1/0/8	1/1/7	1/0/8	1/1/7	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	3/1/5
(10) CNN$^-$+TMK$_2$										8/0/1	1/4/4	5/2/2	2/2/5	2/0/7	2/0/7	1/1/7	2/0/7	7/1/1
(11) TMK$_1$+CNN$^-$											1/1/7	3/1/5	0/2/7	0/1/8	0/1/8	0/1/8	0/1/8	5/1/3
(12) TMK$_2$+CNN$^-$												5/2/2	2/1/6	1/0/8	1/0/8	0/1/8	1/0/8	7/1/1
(13) EWW$^-$													1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	5/2/2
(14) EWW$^-$+SSM$^-$														1/3/5	1/3/5	1/3/5	1/2/6	7/1/1
(15) SMOTE															1/8/0	3/6/0	0/8/1	9/0/0
(16) gg-SMOTE																2/6/1	0/6/3	9/0/0
(17) ncn-SMOTE																	1/5/3	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

Gm

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Original	0/2/7	0/4/5	0/1/8	0/0/9	1/2/6	1/2/6	3/3/3	4/3/2	1/1/7	3/1/5	1/1/7	0/4/5	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	2/3/4
(2) SSM$^-$		6/2/1	1/5/3	0/1/8	4/2/3	5/2/2	4/2/3	7/1/1	3/4/2	6/1/2	2/5/2	5/2/2	1/3/5	0/1/8	0/1/8	0/1/8	0/1/8	5/2/2
(3) EW$^-$			1/1/7	0/0/9	3/1/5	3/1/5	5/2/2	7/1/1	2/1/6	5/1/3	2/1/6	0/5/4	1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	4/3/2
(4) EW$^-$+SSM$^-$				1/3/5	6/2/1	6/2/1	5/1/3	8/1/0	4/2/3	7/1/1	5/2/2	7/1/1	0/7/2	0/3/6	0/3/6	0/4/5	0/3/6	6/1/2
(5) RUS					8/1/0	8/0/1	9/0/0	9/0/0	7/0/2	8/1/0	8/0/1	8/1/0	5/3/1	0/5/4	0/4/5	0/7/2	0/5/4	9/0/0
(6) CNN$^-$						4/1/4	6/1/2	7/0/2	3/2/4	4/2/3	1/4/4	5/1/3	1/1/7	0/1/8	0/1/8	0/1/8	0/1/8	5/2/2
(7) TMK$_1$							5/2/2	6/1/2	2/1/6	5/1/3	3/1/5	4/2/3	1/1/7	1/0/8	1/0/8	1/0/8	1/0/8	6/1/2
(8) TMK$_2$								5/2/2	2/1/6	4/1/4	2/1/6	2/2/5	3/0/6	0/0/9	0/0/9	0/0/9	0/0/9	3/4/2
(9) CNN$^-$+TMK$_1$									1/0/8	1/1/7	1/0/8	1/1/7	0/0/9	0/0/9	0/0/9	0/0/9	0/0/9	2/2/5
(10) CNN$^-$+TMK$_2$										8/0/1	1/4/4	5/2/2	2/1/6	1/1/7	2/0/7	1/1/7	2/0/7	7/1/1
(11) TMK$_1$+CNN$^-$											1/0/8	3/0/6	1/0/8	0/0/9	0/1/8	0/1/8	0/0/9	5/1/3
(12) TMK$_2$+CNN$^-$												5/2/2	2/1/6	1/0/8	1/0/8	0/1/8	1/0/8	7/1/1
(13) EWW$^-$													1/0/8	0/0/9	0/0/9	0/0/9	0/0/9	5/2/2
(14) EWW$^-$+SSM$^-$														1/2/6	1/3/5	1/3/5	1/2/6	7/0/2
(15) SMOTE															1/8/0	3/6/0	0/8/1	9/0/0
(16) gg-SMOTE																2/6/1	0/6/3	9/0/0
(17) ncn-SMOTE																	0/6/3	9/0/0
(18) rng-SMOTE																		9/0/0
(19) b-SMOTE																		

Referencias bibliográficas

Aha, D.W.; Kibler, D. y Albert, M.K.: «Instance-Based Learning Algorithms». *Machine Learning*, 1991, **6(1)**, pp. 37–66. 65

Al-Haddad, L.; Morris, C.W. y Boddy, L.: «Training radial basis function neural networks: effects of training set size and imbalanced training sets». *Journal of Microbiological Methods*, 2000, **43(1)**, pp. 33–44. 30, 95

Alpaydin, E.: *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004. 3

Asuncion, A. y Newman, D.J.: «UCI Machine Learning Repository», 2007. http://www.ics.uci.edu/mlearn/MLRepository.html 51, 65

Barandela, R.; Cortés, N. y Palacios., A.: «The nearest neighbor rule and the reduction of the training sample size». En: *Proceedings of the 9th Spanish Symposium on Pattern Recognition and Image Analysis I*, pp. 103–108, 2001. 72

Barandela, R.; Ferri, F.J. y Sánchez, J.S.: «Decision Boundary Preserving Prototype Selection for Nearest Neighbor Classification». *International Journal of Pattern Recognition and Artificial Intelligence*, 2005a, **19(6)**, pp. 787–806. 36, 70, 72

Barandela, R.; Hernández, J.K.; Sánchez, J.S. y Ferri, F.J.: «Imbalanced Training Set Reduction and Feature Selection Through Genetic Optimization». En: B. López; J. Meléndez; P. Radeva y J. Vitrià (Eds.), *CCIA*, volumen 131 de *Frontiers in Artificial Intelligence and Applications*, pp. 215–222, 2005b. 37

Barandela, R.; Rangel, E.; Sánchez, J.S. y Ferri, F.J.: «Restricted Decontamination for the Imbalanced Training Sample Problem». En: A. Sanfeliu y J. Ruiz-Shulcloper (Eds.), *CIARP*, volumen 2905 de *LNCS*, pp. 424–431, 2003a. 39

Barandela, R.; Sánchez, J.S. y Valdovinos, R.M.: «New applications of ensembles of classifiers». *Pattern Analysis and Applications*, 2003b, **6(3)**, pp. 245–256. 44

Barandela, R.; Valdovinos, R. M.; Sánchez, J.S. y Ferri, F.J.: «The Imbalanced Training Sample Problem: Under or over Sampling?» En: A.L.N. Fred; T. Caelli; R.P.W. Duin; A.C. Campilho y D. de Ridder (Eds.), *SSPR/SPR*, volumen 3138 de *LNCS*, pp. 806–814, 2004. 42, 45, 83, 84

Batista, G.E.A.P.A.; de Carvalho, A.C.P.L.F. y Monard, M.C.: «Applying One-Sided Selection to Unbalanced Datasets». En: O. Cairó; L.E. Sucar y F.J. Cantu (Eds.), *MICAI*, volumen 1793 de *LNCS*, pp. 315–325, 2000. 37

Batista, G.E.A.P.A.; Prati, R.C. y Monard, M.C.: «A study of the behavior of several methods for balancing machine learning training data». *SIGKDD Explorations*, 2004, **6(1)**, pp. 20–29. 45, 74, 83, 84

Batista, G.E.A.P.A.; Prati, R.C. y Monard, M.C.: «Balancing Strategies and Class Overlapping». En: A.F. Famili; J.N. Kok; J.M. Peña; A. Siebes y A.J. Feelders (Eds.), *IDA*, volumen 3646 de *LNCS*, pp. 24–35, 2005. 95

Bradley, A.P.: «The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms». *Pattern Recognition*, 1997, **30(7)**, pp. 1145–1159. 49, 54

Cárdenas, A.A. y Baras, J.S.: «B-ROC Curves for the Assessment of Classifiers over Imbalanced Data Sets». En: *AAAI*, , 2006. 48, 54

Cardie, C. y Nowe, N.: «Improving Minority Class Prediction Using Case-Specific Feature Weights». En: D.H. Fisher (Ed.), *ICML*, pp. 57–65, 1997. 41

Chan, P.K. y Stolfo, S.J.: «Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection». En: *KDD*, pp. 164–168, 1998. 29, 30, 40, 44

Chaudhuri, B.B.: «A new definition of neighborhood of a point in multi-dimensional space». *Pattern Recognition Letters*, 1996, **17(1)**, pp. 11 – 17. 80

Chawla, N.V.: «C4.5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, an decision tree estructure». En: *Proceedings of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II*, , 2003. 42, 83, 84

Chawla, N.V.: «Data Mining for Imbalanced Datasets: An Overview». En: O. Maimon y L. Rokach (Eds.), *The Data Mining and Knowledge Discovery Handbook*, pp. 853–867, 2005. 31

Chawla, N.V.; Bowyer, K.W.; Hall, L.O. y Kegelmeyer, W.P.: «SMOTE: Synthetic Minority Over-Sampling Technique». *Journal of Artificial Intelligence Research*, 2002, **16**, pp. 321–357. 33, 34, 84

Chawla, N.V.; Japkowicz, N. y Kolcz, A.: «ICML: Workshop on Learning from Imbalanced Data Sets II». En: *ICML*, , 2003a. 31

Chawla, N.V.; Japkowicz, N. y Kotcz, A.: «Editorial: special issue on learning from imbalanced data sets». *SIGKDD Explorations*, 2004, **6(1)**, pp. 1–6. 31

Chawla, N.V.; Lazarevic, A.; Hall, L.O. y Bowyer, K.W.: «SMOTEBoost: Improving Prediction of the Minority Class in Boosting». En: N. Lavrac; D. Gamberger; H. Blockeel y L. Todorovski (Eds.), *PKDD*, volumen 2838 de *LNCS*, pp. 107–119, 2003b. 44

Chawla, N.V. y Sylvester, J.: «Exploiting Diversity in Ensembles: Improving the Performance on Unbalanced Datasets». En: M. Haindl; J. Kittler y F. Roli (Eds.), *MCS*, volumen 4472 de *LNCS*, pp. 397–406, 2007. 45

Ciraco, M.; Rogalewski, M. y Weiss, G.: «Improving classifier utility by altering the misclassification cost ratio». En: G.M. Weiss; M. Saar-Tsechansky y B. Zadrozny (Eds.), *UBDM*, pp. 46–52, 2005. 41

Cohen, G.; Hilario, M.; Sax, H.; Hugonnet, S. y Geissbuhler, A.: «Learning from Imbalanced Data in Surveillance of Nosocomial Infection». *Artificial Intelligence in Medicine*, 2006, **37(1)**, pp. 7–18. 3, 30, 35, 38, 41

Cortes, C. y Vapnik, V.: «Support–vector networks». *Machine Learning*, 1995, **20(3)**, pp. 273–297. 18

Dasarathy, B.: *Nearest neighbour norms:* NN *pattern recognition techniques*. IEEE Computer Society Press, 1991. 96

Daskalaki, S.; Kopanas, I. y Avouris, N.: «Evaluation of Classifiers for an Uneven Class Distribution Problem». *Applied Artificial Intelligence*, 2006, **20(5)**, pp. 381–417. 46, 51, 54

del Castillo, M.D. y Serrano, J.S.: «A multistrategy approach for digital text categorization from imbalanced documents». *SIGKDD Explorations*, 2004, **6(1)**, pp. 70–79. 30

Demšar, J.: «Statistical Comparisons of Classifiers over Multiple Data Sets». *Journal Machine Learning Research*, 2006, **7**, pp. 1–30. 46, 53, 75

Devijver, P. A. y Kittler, J.: *Pattern recognition: A statistical approach*. Prentice Hall, 1982. 96

Dietterich, T.G.: «Machine Learning». *Annual Review of Computer Science*, 1990, **4**, pp. 64–69. 29

Dietterich, T.G.; Margineantu, D.; Provost, F. y Turney, P.: «Proceedings of the ICML 2000 Workshop on Cost-Sensitive Learning», 2000. 31

Domingos, P.: «MetaCost: A General Method for Making Classifiers Cost-Sensitive». En: *KDD*, pp. 155–164, 1999. 41

Drummond, C.; Elazmeh, W. y Japkowicz, N.: «Evaluation Methods For Machine Learning». *Informe técnico WS-06-06*, Association for the Advancement of Artificial Intelligente, 2006. 54

Drummond, C. y Holte, R.: «C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling», 2003. 40, 42, 83

Drummond, C. y Holte, R.C.: «Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria». En: P. Langley (Ed.), *ICML*, pp. 239–246, 2000. 40

Drummond, C. y Holte, R.C.: «Cost curves: An improved method for visualizing classifier performance». *Machine Learning*, 2006, **65(1)**, pp. 95–130. 49, 55

Duda, R.O.; Hart, P.E. y Stork, D.G.: *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2001. 3, 65

Eitrich, T. y Lang, B.: «Parallel Tuning of Support Vector Machine Learning Parameters for Large and Unbalanced Data Sets». En: M.R. Berthold; R.C. Glen; K. Diederichs; O. Kohlbacher y I. Fischer (Eds.), *CompLife*, volumen 3695 de *LNCS*, pp. 253–264, 2005. 41

Elkan, C.: «The Foundations of Cost-Sensitive Learning». En: B. Nebel (Ed.), *IJCAI*, pp. 973–978, 2001. 40, 41

Ertekin, S.; Huang, J.; Bottou, L. y Giles, L.C.: «Learning on the border: active learning in imbalanced data classification». En: M.J. Silva; A.H.F. Laender; R.A. Baeza-Yates; D.L. McGuinness; B. Olstad; Ø.H. Olsen y A.O. Falcão (Eds.), *CIKM*, pp. 127–136, 2007. 39

Estabrooks, A. y Japkowicz, N.: «A Mixture-of-Experts Framework for Learning from Imbalanced Data Sets». En: F. Hoffmann; D.J. Hand; N.M. Adams; D.H. Fisher y G. Guimarães (Eds.), *IDA*, volumen 2189 de *LNCS*, pp. 34–43, 2001. 43

Estabrooks, A.; Jo, T. y Japkowicz, N.: «A multiple resampling method for learning from imbalanced data sets». *Computational Intelligence*, 2004, **20(1)**, pp. 18–36. 30, 43

Ezawa, K.J.; Singh, M. y Norton, S.W.: «Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management». En: *ICML*, pp. 139–147, 1996. 41

Fawcett, T.: «An Introduction to ROC Analysis». *Pattern Recognition Letters*, 2006, **27**, pp. 861–874. 47, 54

Fawcett, T. y Provost, F.: «Adaptive Fraud Detection». *Data Minining and Knowledge Discovery*, 1997, **1(3)**, pp. 291–316. 3, 29, 54

Fernández, A.; del Jesus, M.J. y Herrera, F.: «On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets». *Expert Systems with Applications*, 2009, **36(6)**, pp. 9805 – 9812. 95

Ferri, C.; Hernández-Orallo, J. y Modroiu, R.: «An experimental comparison of performance measures for classification». *Pattern Recognition Letters*, 2009, **30(1)**, pp. 27–38. 53, 60

Ferri, C.; Hernández-Orallo, J. y Salido, M.A.: «Volume under the ROC Surface for Multi-class Problems». En: N. Lavrac; D. Gamberger; L. Todorovski y H. Blockeel (Eds.), *ECML*, volumen 2837 de *LNCS*, pp. 108–120, 2003. 50

Fix, E. y Hodges, J.: «Discriminatory Analysis: nonparametric discrimination». *Informe técnico project 21 - 49 - 004, 4*, USAF School of Aviation Medicine, 1951. 24

Folleco, A.; Khoshgoftaar, T.M. y Napolitano, A.: «Comparison of Four Performance Metrics for Evaluating Sampling Techniques for Low Quality Class-Imbalanced Data». En: M.A. Wani; X.-W. Chen; D. Casasent; L.A. Kurgan; T.H. y K. Hafeez (Eds.), *ICMLA*, pp. 153–158, 2008. 54

Gabriel, K.R. y Sokal, R.R.: «A New Statistical Approach to Geographic Variation Analysis». *Systematic Zoology*, 1969, **18(3)**, pp. 259–278. 121

García, S. y Herrera, F.: «Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy». *Evolutionary Computation*, 2009, **17(3)**, pp. 275–306. 38, 74

Gu, Q.; Cai, Z.; Zhu, L. y Huang, B.: «Data Mining on Imbalanced Data Sets». En: X.-X. Zhou y W.Z. Li and (Eds.), *ICACTE*, pp. 1020–1024, 2008. 31

Guo, H. y Viktor, H.L.: «Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach». *SIGKDD Explorations*, 2004, **6(1)**, pp. 30–39. 44

Han, H.; Wang, W. y Mao, B.: «Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning». En: D.-S. Huang; X.-P. Zhang y G.-B. Huang (Eds.), *ICIC*, volumen 3644 de *LNCS*, pp. 878–887, 2005. 34

Hand, D.J. y Vinciotti, V.: «Choosing k for two-class nearest neighbour classifiers with unbalanced classes». *Pattern Recognition Letters*, 2003, **24(9-10)**, pp. 1555–1562. 30, 96

Hansen, L.K. y Salamon, P.: «Neural Network Ensembles». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, **12(10)**, pp. 993–1001. 43

Hart, P.E.: «The condensed nearest neighbour rule». *IEEE Transactions on Information Theory*, 1968, **14**, pp. 515–516. 36, 71

Haykin, S.: *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, 1998. 18, 22

He, G.; Han, H. y Wang, W.: «An Over-sampling Expert System for Learning from Imbalanced Data Sets». En: M. Zhao y Z. Shi (Eds.), *ICNN&B*, volumen 1, pp. 537–541, 2005. 44

He, H. y Garcia, E.A.: «Learning from Imbalanced Data». *IEEE Transactions on Knowledge and Data Engineering*, 2009, **21(9)**, pp. 1263–1284. 74

Hernández-Orallo, J.; Ferri, C.; Lachiche, N. y Flach, P.A.: «The 1st workshop on ROC analysis in artificial intelligence (ROCAI-2004)». *SIGKDD Explorations*, 2004, **6(2)**, pp. 159–161. 54

Holte, R.; Acker, L. y Porter, B.: «Concept Learning and the Problem of Small Disjuncts». *Informe técnico*, University of Texas at Austin, Austin, TX, USA, 1989. 32

Huang, J. y Ling, C.X.: «Constructing New and Better Evaluation Measures for Machine Learning». En: M. M. Veloso (Ed.), *IJCAI*, pp. 859–864, 2007. 61

Huang, Y.-M.; Hung, C.-M. y Jiau, H.C.: «Evaluation of Neural Networks and Data Mining Methods on a Credit Assesssment Task for Class Imbalance Problem». *Nonlinear Analysis: Real World Applications*, 2006, **7(4)**, pp. 720–757. 30

Hulse, J.V. y Khoshgoftaar, T.: «Knowledge discovery from imbalanced and noisy data». *Data & Knowledge Engineering*, 2009, **In Press, Corrected Proof**, pp.–. doi: DOI:10.1016/j.datak.2009.08.005. 95

Hulse, J.V.; Khoshgoftaar, T.M. y Napolitano, A.: «Experimental perspectives on learning from imbalanced data». En: Z. Ghahramani (Ed.), *ICML*, volumen 227 de *ACM International Conference Proceeding Series*, pp. 935–942, 2007. 42, 54, 74, 83, 84

Imam, T.; Ting, K.M. y Kamruzzaman, J.: «z-SVM: An SVM for Improved Classification of Imbalanced Data». En: A. Sattar y B.H. Kang (Eds.), *ACAI*, volumen 4304 de *LNCS*, pp. 264–273, 2006. 41

Japkowicz, N.: «Learning from Imbalanced Data Sets». *Informe técnico WS–00-05*, AAAI, 2000. 31

Japkowicz, N.: «Concept-Learning in the Presence of Between-Class and Within-Class Imbalances». En: E. Stroulia y S. Matwin (Eds.), *Canadian Conference on AI*, volumen 2056 de *LNCS*, pp. 67–77, 2001. 31, 35

Japkowicz, N.; Myers, C. y Gluck, M.A.: «A Novelty Detection Approach to Classification». En: *IJCAI*, pp. 518–523, 1995. 42

Japkowicz, N. y Stephen, S.: «The class imbalance problem: A systematic study». *Intelligent Data Analysis*, 2002, **6(5)**, pp. 429–449. 30, 33, 40, 42, 51, 83, 84, 95, 96

Jaromczyk, J.W. y Toussaint, G.T.: «Relative Neighborhood Graphs and Their Relatives». *Proceedings of the IEEE*, 1992, **80(9)**, pp. 1502–1517. 80

Jin, H. y C.X.Ling: «Using auc and accuracy in evaluating learning algorithms». *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17(3)**, pp. 299–310. 54

Jo, T. y Japkowicz, N.: «Class imbalances versus small disjuncts». *SIGKDD Explorations*, 2004, **6(1)**, pp. 40–49. 32, 35, 52, 95

Joshi, M.V.; Kumar, V. y Agarwal, R.C.: «Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements». En: N. Cercone; T.Y. Lin y X. Wu (Eds.), *ICDM*, pp. 257–264, 2001. 44

Juszczak, P. y Duin, R.P.W.: «Uncertainty sampling for one-class classifiers». En: *ICML: Workshop on Learning with Imbalanced Data Sets II*, pp. 81–88, 2003. 39, 42

Kotsiantis, S. y Pintelas, P.: «Mixture of expert agents for handling imbalanced data sets». *Annals of Mathematics and Computing & TeleInformatics*, 2003, **1(1)**, pp. 46–55. 44

Kubat, M.; Holte, R.C. y Matwin, S.: «Learning When Negative Examples Abound». En: M.V. Someren y G. Widmer (Eds.), *ECML*, volumen 1224 de *LNCS*, pp. 146–153, 1997. 95, 96

Kubat, M.; Holte, R.C. y Matwin, S.: «Machine Learning for the Detection of Oil Spills in Satellite Radar Images». *Machine Learning*, 1998, **30(2-3)**, pp. 195–215. 3, 30, 41, 54

Kubat, M. y Matwin, S.: «Addressing the Curse of Imbalanced Training Sets: One-Sided Selection». En: D.H. Fisher (Ed.), *ICML*, pp. 179–186, 1997. 6, 36, 50, 54, 70, 121

Kukar, M. y Kononenko, I.: «Cost-Sensitive Learning with Neural Networks». En: *ECAI*, pp. 445–449, 1998. 40

Kuncheva, L. y Jain, L.C.: «Nearest neighbor classifier: Simultaneous editing and feature selection». *Pattern Recognition Letters*, 1999, **20(11-13)**, pp. 1149–1156. 37

Landgrebe, T.; Paclík, P. y Duin, R.P.W.: «Precision-recall operating characteristic (P-ROC) curves in imprecise environments». En: *ICPR (4)*, pp. 123–127, 2006. 48, 50, 54

Landgrebe, T.; Paclík, P.; Tax, D.M.J.; Verzakov, S. y Duin, R.P.W.: «Cost-Based Classifier Evaluation for Imbalanced Problems». En: A.L.N. Fred; T. Caelli; R.P.W. Duin; A.C. Campilho y D. de Ridder (Eds.), *SSPR/SPR*, volumen 3138 de *LNCS*, pp. 762–770, 2004. 49

Laurikkala, J.: «Instance-based data reduction for improved identification of difficult small classes». *Intelligent Data Analysis*, 2002, **6(4)**, pp. 311–322. 37

Lee, H. y Cho, S.: «The Novelty Detection Approach for Different Degrees of Class Imbalance». En: I. King; J. Wang; L. Chan y D.L. Wang (Eds.), *ICONIP (2)*, volumen 4233 de *LNCS*, pp. 21–30, 2006. 42

Lewis, D.D. y Catlett, J.: «Heterogenous Uncertainty Sampling for Supervised Learning». En: W.W. Cohen y H. Hirsch (Eds.), *ICML*, pp. 148–156, 1994. 30

Liao, T.W.: «Classification of weld flaws with imbalanced class data». *Expert Systems with Applications*, 2008, **35(3)**, pp. 1041–1052. 74

Ling, C.X. y Li, C.: «Data Mining for Direct Marketing: Problems and Solutions». En: *KDD*, pp. 73–79, 1998. 33, 43

Ling, C.X. y Sheng, V.S.: «Cost-Sensitive Learning and the Class Imbalance Problem». En: Claude Sammut (Ed.), *Encyclopedia of Machine Learning (to appear)*, Springer, 2010. 40

Ling, C.X.; Yang, Q.; Wang, J. y Zhang, S.: «Decision trees with minimal costs». En: C.E. Brodley (Ed.), *ICML*, volumen 69 de *ACM International Conference Proceeding Series*, 2004. 40

Liu, A.; Ghosh, J. y Martin, C.: «Framework for analyzing skew in evaluation metrics». *Informe técnico WS-07-05*, AAAI Workshop on Evaluation Methods for Machine Learning II, 2007. 54

Liu, X.-Y.; Wu, J. y Zhou, Z.-H.: «Exploratory Under-Sampling for Class-Imbalance Learning». En: *ICDM*, pp. 965–969, 2006a. 45

Liu, X.-Y. y Zhou, Z.-H.: «The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study». En: *ICDM*, pp. 970–974, 2006. 41, 45

Liu, Y.; An, A. y Huang, X.: «Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles». En: W.K. Ng; M. Kitsuregawa; J. Li y K. Chang (Eds.), *PAKDD*, volumen 3918 de *LNCS*, pp. 107–118, 2006b. 45

Liu, Y.; Loh, H.-T. y Sun, A.: «Imbalanced text classification: A term weighting approach». *Expert Systems with Applications*, 2009, **36(1)**, pp. 690 – 701. 30

Maloof, M.A.: «Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown». En: *ICML: Workshop on Learning from Imbalanced Data Sets II*, , 2003. 40

Mazurowski, M.A.; Habas, P.A.; Zurada, J.M.; Lo, J.Y.; Baker, J.A. y Tourassi, G.D.: «Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance». *Neural Networks*, 2008, **21(2-3)**, pp. 427 – 436. 95

McCarthy, K.; Zabar, B. y Weiss, G.M.: «Does cost-sensitive learning beat sampling for classifying rare classes?» En: *UDBM*, pp. 69–77, 2005. 40, 42, 88

McCulloch, W.S. y Pitts, W.A.: «A logical calculus of the ideas immanent in nervous activity». *Bulletin of Mathematical Biophysics*, 1943, **5**, pp. 115–133. 20

Meng, H.-H.; Li, G.-Z.; Wang, R.-S.; Zhao, X.-M. y Chen, L.: «The imbalanced problem in mass–spectrometry data analysis». En: D.-Z. Du y X.-S. Zhang (Eds.), *Optimization and Systems Biology*, volumen 9 de *LNOR*, pp. 136–143, 2008. 30

Minsky, M. y Papert, S.: *Perceptrons*. MIT Press, 1969. 20

Molinara, M.; Ricamato, M.T. y Tortorella, F.: «Facing Imbalanced Classes through Aggregation of Classifiers». En: R. Cucchiara (Ed.), *ICIAP*, pp. 43–48, 2007. 45

Murphey, Y.L.; Guo, H. y Feldkamp, L.A.: «Neural Learning from Unbalanced Data». *Applied Intelligent*, 2004, **21(2)**, pp. 117–128. 30, 95

Nickerson, A.S.; Japkowicz, N. y Milios, E.: «Using unsupervised learning to guide resampling in imbalanced data sets». En: *International Workshop on Artificial Intelligence and Statistics*, , 2001. 35

Orriols-Puig, A. y Bernadó-Mansilla, E.: «The class imbalance problem in learning classifier systems: a preliminary study». En: F. Rothlauf (Ed.), *GECCO Workshops*, pp. 74–78, 2005. 30

Orriols-Puig, A. y Bernadó-Mansilla, E.: «Evolutionary rule-based systems for imbalanced data sets». *Soft Computing*, 2009, **13(3)**, pp. 213–225. 84, 88

Pazzani, M.J.; Merz, C.J.; Murphy, P.M.; K.A.; Hume, T. y Brunk, C.: «Reducing Misclassification Costs». En: *ICML*, pp. 217–225, 1994. 40, 41

Platt, J.: «Machines using Sequential Minimal Optimization». En: B. Schoelkopf; C. Burges y A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1998. 65

Prati, R.C.; Batista, G.E.A.P.A. y Monard, M.C.: «Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior». En: R. Monroy; G. Arroyo-Figueroa; L.E. Sucar y J.H.S. Azuela (Eds.), *MICAI*, volumen 2972 de *LNCS*, pp. 312–321, 2004a. 42, 52, 95

Prati, R.C.; Batista, G.E.A.P.A. y Monard, M.C.: «Learning with Class Skews and Small Disjuncts». En: A.L.C. Bazzan y S. Labidi (Eds.), *SBIA*, volumen 3171 de *LNCS*, pp. 296–306, 2004b. 32, 52

Provost, F.: «Machine learning from imbalanced data sets 101 (Extended Abstract)». En: *AAAI: Workshop on Learning with Imbalanced Data Sets*, , 2003. 39

Provost, F. y Fawcett, T.: «Robust Classification for Imprecise Environments». *Machine Learning*, 2001, **42(3)**, pp. 203–231. 54

Provost, F.J. y Fawcett, T.: «Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions». En: *KDD*, pp. 43–48, 1997. 46, 47, 54

Quang, A.T.; Zhang, Q.L. y Li, X.: «Evolving support vector machine parameters». En: *ICMLC*, pp. 548–551, 2002. 41

Quinlan, J.R.: *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. 65

Radivojac, P.; Korad, U.; Sivalingam, K.M. y Obradovic, Z.: «Learning from class-imbalanced data in wireless sensor networks». En: *VTC*, volumen 5, pp. 3030–3034, 2003. 30

Ranawana, R. y Palade, V.: «Optimized Precision - A New Measure for Classifier Performance Evaluation». En: *CEC*, pp. 2254–2261, 2006. 47, 51, 53, 55

Raskutti, B. y Kowalczyk, A.: «Extreme re-balancing for SVMs: a case study». *SIGKDD Explorations*, 2004, **6(1)**, pp. 60–69. 42

Ritter, G.L.; Woodruff, H.B.; Lowry, S.R. y Isenhour, T.J.: «An algorithm for Selective Nearest Neighbor Decision Rule». *IEEE Transactions on Information Theory*, 1975, **24**, pp. 665–669. 70, 71, 72

Rosenblatt, F.: «The Percepron: A probabilistic model for information storage and organization in the brain». *Psychological Review*, 1958, **65**, pp. 386–408. 20

Rumelhart, D.E.; Hinton, G.E. y Williams, R.J.: «Learning representations by backpropagation error». *Nature*, 1986, **323**, pp. 533–536. 21

Seiffert, C.; Khoshgoftaar, T.M.; Hulse, J.V. y Napolitano, A.: «A Comparative Study of Data Sampling and Cost Sensitive Learning». En: *ICDM Workshops*, pp. 46–52, 2008. 88

Sánchez, J.S.; Pla, F. y Ferri, F.J.: «On the use of neighbourhood-based non-parametric classifiers». *Pattern Recognition Letters*, 1997a, **18(11-13)**, pp. 1179 – 1186. 79, 80

Sánchez, J.S.; Pla, F. y Ferri, F.J.: «Prototype selection for the nearest neighbour rule through proximity graphs». *Pattern Recognition Letters*, 1997b, **18(6)**, pp. 507 – 513. 80

Sokolova, M.: «Assessing invariance properties of evaluation measures». En: *NIPS: Workshop on Testing Deployable Learning and Decision Systems*, , 2006. 53, 62

Sokolova, M.; Japkowicz, N. y Szpakowicz, S.: «Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation». En: A. Sattar y B.H. Kang (Eds.), *Australian Conference on Artificial Intelligence*, volumen 4304 de *LNCS*, pp. 1015–1021, 2006. 49, 62, 63

Sokolova, M. y Lapalme, G.: «Performance Measures in Classification of Human Communications». En: Z. Kobti y D. Wu (Eds.), *Canadian Conference on AI*, volumen 4509 de *LNCS*, pp. 159–170, 2007. 62

Sokolova, M. y Lapalme, G.: «A systematic analysis of performance measures for classification tasks». *Information Processing & Management*, 2009, **45(4)**, pp. 427–437. 62

Sousa, A.V.; Mendonça, A.M. y Campilho, A.C.: «Minimizing the Imbalance Problem in Chromatographic Profile Classification with One-Class Classifiers». En: A.C. Campilho y M.S. Kamel (Eds.), *ICIAR*, volumen 5112 de *LNCS*, pp. 413–422, 2008. 30

Spackman, K.A.: «Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning». En: A.M. Segre (Ed.), *ML*, pp. 160–163, 1989. 47

Sun, Y.; Kamel, M.S.; Wong, A.K.C. y Wang, Y.: «Cost-sensitive boosting for classification of imbalanced data». *Pattern Recognition*, 2007, **40(12)**, pp. 3358 – 3378. 45

Swets, J. A.: «Measuring the accuracy of diagnostic systems». *Science*, 1988, **240(4857)**, pp. 1285–1293. 47

Taft, L.M.; Evans, R.S.; Shyu, C.R.; Egger, M.J.; Chawla, N.V.; Mitchell, J.A.; Thornton, S.N.; Bray, B. y Varner, M.: «Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery». *Journal of Biomedical Informatics*, 2009, **42(2)**, pp. 356 – 364. 30

Tan, P.-N.; Kumar, V. y Srivastava, J.: «Selecting the right interestingness measure for association patterns». En: *KDD*, pp. 32–41, 2002. 60, 62

Tan, S.: «Neighbor-weighted k-nearest neighbor for unbalanced text corpus». *Expert Systems with Applications*, 2005, **28(4)**, pp. 667–671. 30, 41, 54

Tax, D.: *One-class classification: Concept-learning in the absence of counter-examples*. Tesis doctoral, Delft University of Technology, 2001. 42

Tomek, I.: «An Experiment with the Edited Nearest-Neighbor Rule». *IEEE Transactions on Systems, Man and Cybernetics*, 1976a, **6(6)**, pp. 448–452. 73

Tomek, I.: «Two Modifications of CNN». *IEEE Transactions on Systems, Man and Cybernetics*, 1976b, **6(11)**, pp. 769–772. 6, 36, 37, 74, 121

Toussaint, G.T.: «A counterexample to Tomek's consistency theorem for a condensed nearest neighbor decision rule». *Pattern Recognition Letters*, 1994, **15(8)**, pp. 797–801. 37, 121

Turney, P.: «Types of cost in inductive concept learning». En: *ICML*, pp. 15–21, 2000. 40

Vapnik, V.: *Statistical learning theory*. John Wiley & Sons, 1998. 18

Veropoulos, K.; Campbell, C. y Cristianini, N.: «Controlling the sensitivity of support vector machines». En: *IJCAI*, pp. 55–60, 1999. 40

Viola, P.A. y Jones, M.J.: «Rapid Object Detection using a Boosted Cascade of Simple Features». En: *CVPR (1)*, pp. 511–518, 2001. 45

Visa, S. y Ralescu, A.: «Learning Imbalanced and Overlapping Classes using Fuzzy Sets». En: *ICML: Workshop on Learning from Imbalanced Datasets*, pp. 97–104, 2003. 52, 95, 96

Visa, S. y Ralescu, A.: «Issues in mining imbalanced data sets - a review paper». En: *Midwest Artificial Intelligence and Cognitive Science Conference*, pp. 393–400, 2005. 31

Wang, J.; Xu, M.; Wang, H. y Zhang, J.: «Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding». En: *ICSP*, volumen 3, pp. 16–20, 2006. 34

Weiss, G.M.: «Learning with rare cases and small disjuncts». En: *ICML*, pp. 558–565, 1995. 52

Weiss, G.M.: *The effect of small disjuncts and class distribution on decision tree learning*. Tesis doctoral, Rutgers University, 2003. 29, 95

Weiss, G.M.: «Mining with rarity: a unifying framework». *SIGKDD Explorations*, 2004, **6(1)**, pp. 7–19. 30, 32, 40

Weiss, G.M. y Hrish, H.: «Learning to predict rare events in event sequences». En: *KDD*, pp. 359–363, 1998. 30

Weiss, G.M.; McCarthy, K. y Zabar, B.: «Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?» En: R. Stahlbock; S.F. Crone y S. Lessmann (Eds.), *DMIN*, pp. 35–41, 2007. 41, 95

Weiss, G.M. y Provost, F.J.: «Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction». *Journal of Artificial Intelligence Research*, 2003, **19**, pp. 315–354. 95

Weng, C.G. y Poon, J.: «A New Evaluation Measure for Imbalanced Datasets». En: J.F. Roddick; J. Li; P. Christen y P.J. Kennedy (Eds.), *AusDM*, volumen 87 de *CRPIT*, pp. 27–32, 2008. 50, 55

Wilson, D.L: «Asymptotic properties of nearest neighbour rules using edited data». *IEEE Transactions on Systems, Man and Cybernetics*, 1972, **2**, pp. 408–421. 37, 70

Witten, I.H. y Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. 3, 65, 75, 98

Wong, A.K.S.; Lee, J.W.T. y Yeung, D.S.: «Improving Text Classifier Performance based on AUC». En: *ICPR (3)*, pp. 268–271, 2006. 54

Wu, G. y Chang, E.Y.: «Class-Boundary Alignment for Imbalanced Dataset Learning». En: *ICML: Workshop on Learning from Imbalanced Datasets*, pp. 49–56, 2003. 30

Xie, J. y Qiu, Z.: «The effect of imbalanced data sets on LDA: A theoretical and empirical analysis». *Pattern Recognition*, 2007, **40(2)**, pp. 557–562. 30

Yang, C.-Y.; Wang, J.; Yang, J.-S. y Yu, G.-D.: «Imbalanced SVM Learning with Margin Compensation». En: F. Sun; J.Z.; Y. Tan; J.C. y W. Yu (Eds.), *ISNN (1)*, volumen 5263 de *LNCS*, pp. 636–644, 2008. 40

Yen, S.-J.; Lee, Y.-S.; Lin, C.-H. y Ying, J.-C.: «Investigating the Effect of Sampling Methods for Imbalanced Data Distributions». En: *SMC*, volumen 5, pp. 4163–4168, 2006. 38

Zadrozny, B.; Langford, J. y Abe, N.: «Cost-Sensitive Learning by Cost-Proportionate Example Weighting». En: *ICDM*, pp. 435–442, 2003. 40

Zhang, J.; Bloedorn, E.; Rosen, L. y Venese, D.: «Learning Rules from Highly Unbalanced Data Sets». En: *ICDM*, pp. 571–574, 2004. 41

Zhang, J. y Mani, I.: «kNN approach to unbalanced data distributions: a case study involving information extraction». En: *ICML: Workshop on Learning from Imbalanced Datasets II*, , 2003. 30, 38, 54, 95, 96

Zhang, J.; Yim, Y.-E. y Yang, J.: «Intelligent Selection of Instances for Prediction Functions in Lazy Learning Algorithms». *Artificial Intelligence Review*, 1997, **11(1-5)**, pp. 175–191. 80

Zhang, X.; Wang, X.; Guo, H.; Guo, Z.; Wu, X. y Su, Z.: «Floatcascade learning for fast imbalanced web mining». En: J. Huai; R. Chen; H.-W. Hon; Yunhao Liu; W.-Y. Ma; A. Tomkins y X. Zhang (Eds.), *WWW*, pp. 71–80, 2008. 30

Zhuang, L. y Dai, H.: «Parameter Optimization of Kernel-based One-class Classifier on Imbalance Learning». *Journal of Computers*, 2006, **1(7)**, pp. 32–40. 41