SMOTE-D, UNA VERSIÓN DETERMINISTA DE SMOTE.

Por:
FREDY RODRÍGUEZ TORRES.

Tesis sometida como requisito parcial para obtener el grado de:
MAESTRO EN CIENCIAS EN EL ÁREA CIENCIAS COMPUTACIONALES.

En el:
Instituto Nacional de Astrofísica, Óptica y Electrónica.

Marzo 2017.
Tonantzintla, Puebla.

Supervisada por:
Jesús Ariel Carrasco Ochoa, INAOE
José Francisco Martínez Trinidad, INAOE

SMOTE-D, una versión determinista de SMOTE

Dedicatorias

A mis asesores de tesis, el Dr. Jesús A. Carrasco y José Fco. Martines.

A mi asesor académico, el Dr. Gustavo Rodríguez.

A todos mis profesores de maestría.

II

Agradecimientos

A mis asesores de tesis, el Dr. Jesús A. Carrasco y José Fco. Martines. A mis sinodales, el Dr. Eduardo Morales, el Dr. Luis Villaseñor y la Dra. María del P. Gómez. Y a mis profesores del INAOE por su contribución en mi formación académica.

A la Academia de Ciencias Computacioneles por su gran labor para con nosotros los estudiantes.

Al INAOE y CONACYT por su apoyo para la conclusión de mis estudios.

Resumen

En diferentes aplicaciones prácticas es común que se presente desbalance entre clases. Este problema aparece cuando la cantidad de objetos en una clase es mucho menor que en la otra. Esta diferencia en el tamaño de las clases causa que los métodos de clasificación favorezcan a la clase con mayor cantidad de objetos (mayoritaria), produciendo un mal desempeño de clasificación para la clase con menor cantidad de objetos (minoritaria).

Las soluciones propuestas en la literatura, para el problema de desbalance entre clases, pueden dividirse en dos tipos: soluciones a nivel algorítmico y soluciones a nivel de datos. Las soluciones a nivel algorítmico modifican los algoritmos de clasificación para mejorar la clasificación en la clase minoritaria. Por otra parte, las soluciones a nivel de datos utilizan métodos de re-muestreo para balancear el conjunto de datos.

Dentro de los métodos de re-muestreo podemos encontrar 3 tipos: sub-muestreo, sobre-muestreo y re-muestreo híbrido. Los métodos de sub-muestreo reducen la cantidad de objetos en la clase mayoritaria con el objetivo de balancear el conjunto de datos, esto tiene la ventaja de que los modelos de clasificación sean más simples y rápidos. Los métodos de sobre-muestreo in-

crementan la cantidad de objetos de la clase minoritaria al generar nuevos objetos sintéticos. Dicha generación tiene como objetivo balancear el conjunto de datos, resultando a la vez en una mejora de la clasificación para los objetos de la clase minoritaria, los cuales usualmente son los de mayor interés. Finalmente, los métodos de re-muestreo híbrido combinan los dos tipos de métodos anteriores, persiguiendo las ventajas de ambos.

En esta tesis se propone una versión determinista de SMOTE (Synthetic Minority Over-sampling Technique), uno de los métodos de sobre-muestreo más conocidos de la literatura, el cual genera objetos sintéticos de forma aleatoria entre dos objetos de la clase minoritaria. En esta tesis no estudiamos los métodos de sub-muestreo, ya que nuestro principal interés es mejorar la clasificación de objetos de la clase minoritaria, sin eliminar objetos de la clase mayoritaria, lo cual puede producir pérdida de información importante. Finalmente, se realizó una comparación del método propuesto contra SMOTE y algunos métodos de sobre-muestreo basados en SMOTE, utilizando conjuntos de datos con desbalance obtenidos del repositorio KEEL, mostrando las bondades del método propuesto.

Abstract

In several practical applications, it is common to find the imbalance class problem. This problem arises when the number of objects in a class is much smaller than in the other. This difference in class size makes the classification methods bias their results to the class with the largest number of objects (majority), resulting in poor performance for the class with fewer objects (minority).

The solutions proposed in the literature for the imbalance class problem can be divided into two types: algorithm-level solutions and data-level solutions. Algorithm-level solutions modify classification algorithms to improve the classification performance for the minority class. On the other hand, data-level solutions use resampling methods in order to balance the dataset.

Among resampling methods, we can find three types: subsampling, oversampling and hybrid sampling. Subsampling methods reduce the number of objects in the majority class with the aim of balancing the dataset. This reduction has as advantage that the produced classification models are simpler and faster. Oversampling methods increase the number of objects by generating synthetic objects for the minority class. This generation aims to

balance the dataset while resulting in a classification improvement for the objects in the minority class which is usually the most interesting. Finally, hybrid sampling methods combine the two above kinds of methods, trying to get of both approaches advantages.

In this thesis, we propose a deterministic version of SMOTE (Synthetic Minority Over-sampling Technique), which is a well-known oversampling method and randomly generates synthetic objects between two objects of the minority class. In this thesis we do not study subsampling methods because our main interest is to improve the classification of objects from the minority class, without deleting objects of the majority class; which could lead to loss some important information. Finally, a comparison of the proposed method against SMOTE and some oversampling methods based on SMOTE was performed. We use imbalanced datasets taken from the KEEL repository, for showing the benefits of the proposed method.

Índice general

Índice de figuras

Índice de tablas

Glosario de términos

- **AUC** (*Area Under Curve*) Área bajo la cruva
- **DT** (*Decision Tree*) Árbol de decisión
- **ENN** (*Edited Nearest Neighbor*) Edición del vecino más cercano
- **FM** (*F Measure*) Medida "F"
- **FN** (*False Negative*) Falso negativo
- **FP** (*False Positive*) Falso positivo
- **FPR** (*False Positive Rate*) Tasa de falsos positivos
- **gl** Grados de libertad
- **HVDM** (*Heterogeneous Value Difference Metric*) Métrica heterogénea de valor de diferencia
- **IR** (*Imbalance Ratio*) Grado de desbalance
- **KEEL** (*Knowledge Extraction Evolutionary Learning*) Extracción del conocimiento, aprendizaje evolutivo
- **KNN** (*K Nearest Neighbors*) Clasificador K vecinos más cercanos
- **RST** (*Rough Set Theory*) Teoría de conjuntos rugosos
- **sl** (*significance level*) Nivel de significancia
- **SMOTE** (*Synthetic Minority Oversampling Technique*) Técnica de sobre-muestreo sintético minoritario
- **SVM** (*Support Vector Machine*) Clasificador máquina de soporte vectorial
- **TN** (*True Negative*) Verdadero negativo
- **TP** (*True Positive*) Verdadero positivo

- **TPR** (*True Positive Rate*) Tasa de verdaderos positivos

Glosario de variables

- $cant_{ij}$

  Cantidad de objetos sintéticos a generar entre un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria y uno de sus k vecinos más cercanos (i, $i, ..., k$).

- d_i

  Conjunto de distancias de los k vecinos más cercano de un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria.

- d_{ij}

  Distancia de un vecino (j, $j = 1, ..., k$) más cercanos de un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria.

- dif

  Diferencia entre el número de objetos de la clase minoritaria y la clase mayoritaria.

- dif_{ij}

  Diferencia de atributos entre un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria y uno de sus k vecinos más cercanos (i, $i, ..., k$).

- dif'_{ij}

  Diferencia de atributos entre un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria y uno de sus k vecinos más cercanos (i, $i, ..., k$) dividida por la cantidad de objetos sintéticos a generar entre ellos.

- k

  Parámetro para definir la cantidad de vecinos en los métodos de sobremuestreo.

- K

  Parámetro para definir la cantidad de vecinos en el clasificador K Vecinos más Cercanos.

- m
  Subconjunto de objetos de la clase minoritaria.
- M
  Subconjunto de objetos de la clase mayoritaria.
- n
  Cantidad de objetos sintéticos a generar.
- N
  Parámetro para definir la porción a alcanzar de la diferencia entre el número de objetos entre clases, en dominio $\mathbb{R}$.
- obj_i
  Objeto (i, $i = 1, ..., |m|$) de la clase minoritaria.
- obj_{ij}
  Vecino (j, $j = 1, ..., k$) de un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria.
- p_i
  Porción de objetos sintéticos a generar para un objeto de la clase minoritaria.
- p_{ij}
  Porción de objetos sintéticos a generar para cada vecino (j, $j = 1, ..., k$) de un objeto (i, $i = 1, ..., |m|$) de la clase minoritaria.
- r
  Número de atributos de los objetos.
- std_i
  Desviación estándar (i, $i = 1, ..., |m|$) de las distancias de los k vecinos más cercanos de un objeto de la clase minoritaria.

Capítulo 1

Introducción

En este capítulo se define el problema del desbalance entre clases (en clases binarias) y algunos de los enfoques para la solución del mismo. Además, se exponen los objetivos generales y específicos de esta tesis.

En diversas áreas como la astronomía, medicina o ciencias sociales, por mencionar algunas, surge la necesidad de clasificar objetos, de una manera rápida y automatizada. Los objetos se clasifican dentro de clases definidas previamente, a esto, en el área de reconocimiento de patrones, se le conoce como clasificación supervisada.

La tarea de clasificación supervisada se vuelve más difícil de realizar satisfactoriamente cuando en los conjuntos de datos se presenta el problema del desbalance entre clases. Este problema aparece cuando en una de las dos clases se tienen menos objetos (clase minoritaria) que en la clase contraria (clase mayoritaria). El desbalance causa un sesgo a los modelos de clasificación, resultando en un buen desempeño para la clase mayoritaria y, de forma

opuesta, un mal desempeño para la clase minoritaria.

Para el problema del desbalance entre clases, en la literatura se han propuesto dos tipos de soluciones [Luengo et al., 2011, Verbiest et al., 2012, Sáez et al., 2015]: soluciones a nivel algorítmico [Ducange et al., 2010, Lin and Chen, 2013, Wang et al., 2012, Zong et al., 2013] y soluciones a nivel de datos [Sáez et al., 2015, Ramentol et al., 2012, Bunkhumpornpat et al., 2009, Han et al., 2005, Batista et al., 2004]. Además, algunos autores [López et al., 2013] reconocen un tercer tipo de soluciones al problema del desbalance entre clases conocido como clasificación sensible al costo. Quienes dividen las soluciones en dos tipos explican que este último tipo de solución puede considerarse como parte de las soluciones a nivel algorítmico, ya que la modificación algorítmica podría tener en cuenta el costo de una mala clasificación.

Las soluciones a nivel algorítmico [Barandela et al., 2003, Batuwita and Palade, 2013, Ducange et al., 2010, Lin and Chen, 2013, Wang et al., 2012, Wang et al., 2015, Zong et al., 2013] modifican las técnicas de clasificación para imponer un sesgo hacia la clase minoritaria, o bien, para mejorar el rendimiento de clasificación mediante el ajuste de los pesos para cada clase. La desventaja de este tipo de soluciones es que la solución en sí misma se limita al clasificador al cual se le han hecho adecuaciones para tratar con el desbalance, forzando a usar solamente este clasificador para resolver el problema de desbalance.

Las soluciones a nivel de datos [Batista et al., 2004, Estabrooks et al., 2004] son independientes del método de clasificación. En esta categoría, se encuentran los métodos de re-muestreo que se caracterizan por sólo tomar en cuenta a los datos, independientemente del método de clasificación que

se utilice. Los métodos de re-muestro se pueden categorizar en tres tipos: sub-muestreo, sobre-muestreo y re-muestro híbrido.

El sub-muestreo tiene como objetivo reducir la cantidad de objetos de la clase mayoritaria. Una de las ventajas de los métodos de sub-muestreo es que la reducción de la clase mayoritaria conduce a un menor costo de espacio y por ende, se obtiene un modelo de clasificación más simple. Sin embargo, la principal desventaja que se atribuye a este tipo de métodos es que son propensos a eliminar información relevante, afectando al desempeño de clasificación para la clase mayoritaria.

Por su parte, los métodos de sobre-muestreo incrementan la cantidad de objetos, al generar objetos sintéticos para la clase minoritaria. La principal ventaja de este tipo de métodos es que generalmente mejoran el desempeño de la clasificación para la clase minoritaria. Cabe mencionar que, en algunas tareas de clasificación, resulta de mayor interés el desempeño en la clase minoritaria. Por otro lado, aplicar sobre-muestreo tiene como desventaja potencial el aumento de la cantidad de objetos con ruido, ya que los objetos sintéticos pudieran generarse a partir de objetos ruidosos. Además, también se pueden generar objetos imposibles de encontrar dentro de un contexto real.

El tercer y último tipo, los métodos híbridos, combinan los dos métodos antes mencionados, reduciendo la cantidad de objetos de la clase mayoritaria a la vez que incrementan la cantidad de objetos de la clase minoritaria. Al combinar ambos métodos se busca solventar, en cierto grado, las desventajas de un método con las ventajas del otro y viceversa.

Nuestro trabajo, de la misma forma que una gran parte de trabajos

de la literatura, se centra en los métodos de sobre-muestreo, por su mejora en la clasificación de la clase minoritaria. De los métodos de sobre-muestreo propuestos en la literatura, SMOTE [Chawla et al., 2002] es uno de los más exitosos. Este método incrementa la cantidad de objetos de la clase minoritaria, generando objetos sintéticos. Una de las características de SMOTE es que genera nuevos objetos sintéticos de forma aleatoria, teniendo en cuenta un objeto de la clase minoritaria y alguno de sus vecinos más cercanos.

La generación aleatoria de objetos sintéticos de SMOTE trae consigo varios inconvenientes, dado que los objetos sintéticos son generados aleatoriamente utilizando un objeto de la clase minoritaria y cualquiera de los vecinos más cercanos de ese objeto. Esto puede causar que algún objeto de la clase minoritaria nunca sea considerado para la generación de objetos sintéticos, pudiendo desperdiciar información valiosa que dicho objeto pudiera aportar. Otro inconveniente, relacionado con el anterior, es que la mayoría o todos los objetos sintéticos a generar, para un objeto de la clase minoritaria, pueden ser generados entre él y uno solo de sus k vecinos más cercanos. Un tercer inconveniente es que los objetos sintéticos pueden ser generados aglomerados en la misma región entre el objeto de la clase minoritaria y su vecino más cercano. Esta aglomeración conduciría a que otras regiones entre el objeto y su vecino no sean utilizadas para la generación de objetos sintéticos.

Con el objetivo de evitar estos inconvenientes, en esta tesis se introduce una versión determinista de SMOTE.

1.1. Objetivos

Los objetivos planteados para este trabajo de investigación son:

1.1.1. Objetivo general

Proponer un nuevo método de sobre-muestreo, basado en la generación de objetos sintéticos en la clase minoritaria, que mejore los resultados de clasificación con respecto a los métodos de sobre-muestreo similares del estado del arte.

1.1.2. Objetivos específicos

- Proponer un criterio de selección de objetos de la clase minoritaria para la generación de objetos sintéticos.
- Proponer un criterio de combinación entre un objeto de la clase minoritaria y alguno de sus vecinos más cercanos para la generación de objetos sintéticos.
- Proponer un nuevo método de sobre-muestreo, basado en los criterios de selección y combinación propuestos en los objetivos anteriores.

La principal aportación en esta tesis es un conjunto de métodos de sobre-muestreo deterministas que permiten obtener resultados de clasificación similares a los obtenidos con métodos aleatorios reportados en la literatura sin depender del azar, evitando así la necesidad de aplicar los métodos

aleatorios varias veces para poder obtener una buena solución. El método propuesto se comparó contra SMOTE con *AUC* (*Area Under Curve*) y *F-M* (*F Measure*) como medidas de desempeño, en un total de 66 conjuntos de datos (de clases binarias) con desbalance, utilizando Arboles de Decisión, Maquinas de soporte vectorial y K-Vecinos más Cercanos como clasificadores.

1.2. Organización de la tesis

El capítulo 2 presenta los conceptos básicos necesarios para la comprensión de este trabajo de investigación. El capítulo 3 reporta los métodos del estado del arte que abordan el problema de desbalance entre clases, centrándose en métodos de sobre-muestreo, específicamente en SMOTE. El capítulo 4 describe el método de sobre-muestreo propuesto para la solución del problema de clases desbalanceadas. En el capítulo 5 se discuten los resultados experimentales obtenidos con el método propuesto. Las conclusiones y el trabajo futuro se muestran en el capítulo 6.

Capítulo 2

Marco teórico

En este capítulo se exponen los conceptos necesarios para el desarrollo y comprensión de este trabajo de investigación. Específicamente, se explican los conceptos de clases desbalanceadas (binarias), las funciones de distancia comúnmente utilizadas en este contexto para comparar objetos, las medidas que se utilizan para evaluar el desempeño de los clasificadores en problemas con clases desbalanceadas y la prueba de significancia estadística que se utiliza para validar los resultados de los experimentos de esta tesis.

2.1. Problema de desbalance entre clases

En muchas tareas de clasificación, la recolección de datos para una de las clases es más difícil de hacer y, por lo tanto, resulta escasa la muestra de objetos disponibles para ésta, en comparación con la cantidad de datos disponibles para la otra clase.

En esta situación, los clasificadores tienen un buen desempeño de clasificación con los objetos de la clase mayoritaria y, de forma contraria, un mal desempeño con los objetos de la clase minoritaria. Lo anterior se atribuye a que, dada la baja cantidad de objetos en la clase minoritaria, éstos no aportan suficiente información para el clasificador. El problema de desbalance entre clases se puede definir de la siguiente manera:

Sea T un conjunto de entrenamiento dividido en dos clases. Sea $|M|$ la cantidad de objetos en una clase y $|m|$ la cantidad de objetos en otra clase, de modo que $|T| = |M| + |m|$. El conjunto T se considera desbalanceado si:

$$|M| > |m| \tag{2.1}$$

A manera de ejemplo, imaginemos un nuevo tipo de cáncer denominado "A", del cual, por su rareza y fallas en su diagnóstico, se tienen muy pocas muestras a nivel mundial. La poca cantidad de muestras para el nuevo cáncer "A" sería demasiado pequeña al compararla con la cantidad de muestras para un cáncer similar y muy conocido denominado "B". En este ejemplo, los métodos de sobre-muestreo podrían ayudar a incrementar la cantidad de muestras (a través de muestras sintéticas) de la clase más pequeña, cáncer "A", y con ese incremento balancear el conjunto de datos. Como no hay un consenso acerca de qué tanto más grande debe ser la cantidad de objetos en una clase con respecto a otra, se han definido métricas para evaluar el nivel de desbalance. Por ejemplo el IR (*Imbalance Ratio*) [Sáez et al., 2015] que se calcula de la siguiente forma:

$$IR = \frac{|M|}{|m|} \tag{2.2}$$

Donde $|M|$ es la cantidad de objetos en la clase mayoritaria y $|m|$ es la cantidad de objetos en la clase minoritaria. El IR representa la cantidad promedio de objetos en la clase mayoritaria por cada objeto de la clase minoritaria.

A modo de ejemplo, imaginemos que la clase minoritaria tiene 100 objetos y la clase mayoritaria tiene 350 objetos, entonces IR=3.5 lo que significa que por cada objeto en la clase minoritaria se tienen en promedio 3.5 objetos en la clase mayoritaria. Con el IR podemos darnos una idea del nivel de desbalance entre las clases.

2.2. Comparación de objetos

La comparación entre objetos es una parte fundamental para muchos procesos de análisis y clasificación de objetos. Las comparaciones se utilizan cuando en algún paso de alguna tarea, como puede ser la clasificación, se necesita algún indicador de qué tan diferentes o similares son dos objetos. Esta comparación comúnmente se obtiene con alguna evaluación sobre los valores de los atributos que describen a los objetos.

A continuación, se definen las funciones de comparación entre objetos comúnmente utilizadas en trabajos del estado del arte que abordan el problema del desbalance entre clases:

Sean $x = (x_1, \ldots, x_r)$ y $y = (y_1, \ldots, y_r)$ dos objetos descritos por r atributos

- La distancia Euclidiana se define como:

$$D(x, y) = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2} \tag{2.3}$$

con: $x_i, y_i \in R; i = 1, ..., r$

- La distancia de Coseno se define como:

$$COS(x, y) = \frac{\sum_{i=1}^{r} x_i y_i}{\sqrt{\sum_{i=1}^{r} x_i^2}\sqrt{\sum_{i=1}^{r} y_i^2}} \tag{2.4}$$

con: $x_i, y_i \in R; i = 1, ..., r$

- La medida HVMD se define como:

$$HVDM(x, y) = \sqrt{\sum_{i=1}^{r}(d_i(x_i - y_i))^2} \tag{2.5}$$

Donde:

$$d_i(x_i, y_i) = \begin{cases} 1 & si\ x_i\ o\ y_i\ no\ se\ conocen \\ vdm_i(x_i, y_i) & si\ el\ atributo\ i\ es\ nominal \\ diff_i(x_i, y_i) & si\ el\ atributo\ i\ es\ numérico \end{cases} \tag{2.6}$$

con:

$$diff_i(x_i, y_i) = \frac{(x_i - y_i)}{4\sigma_i} \tag{2.7}$$

$$vdm_i(x_i, y_i) = \sum_{c=1}^{C} \left| \frac{N_{icx}}{N_{ix}} - \frac{N_{icy}}{N_{iy}} \right| \tag{2.8}$$

Donde:

- σ_i es la desviación estándar de los valores del atributo i en el conjunto de datos.
- N_{ixc} es la cantidad de objetos en el conjunto de datos que tienen el valor x en el atributo i y pertenecen a la clase c.
- N_{iyc} es la cantidad de objetos en el conjunto de datos que tienen el valor y en el atributo i y pertenecen a la clase c.
- N_{ix} es la cantidad de objetos en el conjunto de datos que tienen el valor x en el atributo i.
- N_{iy} es la cantidad de objetos en el conjunto de datos que tienen el valor y en el atributo i.

2.3. Medidas de desempeño

Para evaluar el desempeño de un clasificador supervisado se han propuesto diferentes medidas [Seiffert et al., 2010, García et al., 2012, Wang et al., 2012]. Sin embargo, no todas las medidas de desempeño son útiles cuando se presenta el problema de desbalance entre clases, ya que algunas medidas dan valores altos, aún cuando no se haya clasificado correctamente ningún objeto

de la clase minoritaria, la cual, como ya hemos mencionado, es usualmente la de mayor importancia.

Todas las medidas de desempeño pueden calcularse con base en la matriz de confusión (Tabla 2.1) obtenida a partir de un conjunto de prueba, donde un conjunto de prueba es un conjunto de datos etiquetados independiente del conjunto de entrenamiento, el cual es utilizado para evaluar el modelo de clasificación. La matriz de confusión para dos clases se ejemplifica en la Tabla 2.1; las columnas de esta matriz están asociadas a las clases asignadas por el clasificador y los renglones están asociados a la etiqueta correcta de los objetos. En la Tabla 2.1, TP (*True Positive*) y TN (*True Negative*), son el número de objetos correctamente clasificados de la clase minoritaria y mayoritaria respectivamente y, por otro lado, FP (*False Positive*) y FN (*False Negative*) son el número de objetos mal clasificados de la clase minoritaria y mayoritaria respectivamente.

Tabla 2.1: Matriz de confusión

	Predicción Positiva	Predicción Negativa
Psotivos (Clase minoritaria)	TP	FN
Negativos (Clase mayoritaria)	FP	TN

A partir de la matriz de confusión pueden calcularse diversas medidas, entre ellas:

- Tasa de verdaderos positivos
  - $TPR = \frac{TP}{TP+FN}$

  Cuanto mayor sea la proporción de TP comparada con la de FN, mayor será el valor para TPR, valores altos de TPR suponen una

buena clasificación para los objetos de la clase positiva (minoritaria) al ser correctamente clasificados en su misma clase.

- Tasa de falsos positivos
  - $FPR = \frac{FP}{FP+TN}$

  Cuanto mayor sea la proporción de FP comparada con la de TN, mayor será el valor para FPR; valores altos de FPR suponen una mala clasificación para los objetos de la clase negativa (mayoritaria) al ser equivocadamente clasificados en la clase contraria.

- AUC (*Area Under Curve*)
  - $AUC = \frac{1+TPR-FPR}{2}$

  Cuando TPR tiene un valor mayor al de FPR el resultado para AUC es mayor a 0.5, de forma contraria, cuando FPR es mayor que TPR el resultado para AUC es menor a 0.5 y, por último, cuando FPR y TPR son iguales el resultado para AUC es igual a 0.5. Se asocia el resultado de $AUC = 1$ a una clasificación perfecta, por otra parte, un $AUC = 0.5$ se asocia a una clasificación totalmente aleatoria, por último, un $AUC < 0.5$ refleja una muy mala clasificación. Esta medida es una aproximación del área bajo la curva (bidimencional) del espacio ROC (*Receiver Operating Characteristic*), donde el eje x está definido por FPR y el eje y por TPR.

- F-*Measure*
  - $F - Measure\beta = \frac{(1+\beta^2)*TP}{(1+\beta^2)*TP+\beta^2*FN+FP}$

  Cuando $\beta = 1$ (parámetro de configuración) el valor de TP es comparado con los valores de FN y FP por igual, cuanto mayor sea TP

comparado con FN y FP mayor será el valor de $F - Measure$. Por otro lado, cuando $\beta < 1$ se le da mayor importancia a FP que a FN. Finalmente, cuando $\beta > 1$ se le da mayor importancia a FN que a FP. Valores altos de $F - Measure$ suponen un buen desempeño de clasificación. Se asocia el resultado de $F - Measure = 1$, con $\beta = 1$, a una clasificación perfecta para los objetos de la clase positiva (minoritaria), por otra parte, $F - Measure = 0.5$ se asocia a una clasificación totalmente aleatoria para dicha clase, por último, un $F - Measure < 0.5$ es resultado de una mala clasificación para la clase minoritaria.

Las medidas F-*Measure* y AUC son consideradas adecuadas para evaluar resultados de clasificación en conjuntos de datos con desbalance, ya que se ven afectadas en función al tamaño de la clase en la que se haya tenido un mal desempeño. Por ejemplo, si se falla en sólo 1 de 100 objetos de la clase mayoritaria no se ven igualmente afectadas como si se hubiese clasificado mal 1 de 10 objetos de la clase minoritaria.

De forma contraria TPR y FPR no se consideran adecuadas, ya que se centran en evaluar aspectos específicos de la clasificación, por ejemplo, TPR es indiferente a la correcta clasificación de objetos de la clase negativa (clase mayoritaria), dicha indiferencia permite obtener resultados altos de TPR aún cuando no exista ningún objeto correctamente clasificado para la clase negativa.

2.4. Prueba estadística

Cuando se requiere comparar los resultados de clasificación, evaluados con alguna medida de desempeño, para dos clasificadores distintos aplicados en varios conjuntos de datos, no basta con obtener la diferencia entre las medias de los resultados, es necesario saber si esta diferencia es significativa de acuerdo a alguna prueba estadística. En esta investigación utilizaremos la prueba $t-test$ [Baldi and Long, 2001], comúnmente utilizada para comparar dos medias. Con el resultado de la prueba t-test estimaremos si la diferencia de las medias de los resultados es estadísticamente significativa o no.

La prueba estadística t-test utilizada para la comparación de medias en este trabajo se define como:

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{s_x^2}{w}+\frac{s_y^2}{z}}} \tag{2.9}$$

Donde $\overline{x}$ y $\overline{y}$ son las medias de cada conjunto de resultados, s_x y s_y son las desviaciones estándar y, por último, w y z son los tamaños de cada muestra.

Si el valor calculado de t sobrepasa los valores críticos (de la tabla de valores críticos de t) dados por los grados de libertad ($gl = w + z - 2$) y el nivel de significancia estadística deseado (por ejemplo $sl = 0.05$), el valor de t cae en la zona de rechazo. Los grados de libertad expresan la cantidad de parámetros que pueden ser asignados arbitrariamente, permitiendo introducir una corrección en los cálculos. Además, el nivel de significancia estadística nos da una idea de cuán probable es que la diferencia observada de las medias

se deba al azar.

Capítulo 3

Estado del Arte

Debido a que en esta tesis se propone un método de sobre-muestreo, en este capítulo se revisan los métodos de re-muestreo más representativos del estado del arte. Específicamente, se concentra en métodos de sobre-muestreo basados en SMOTE [Chawla et al., 2002] dado que el método propuesto en esta tesis es una versión determinista de SMOTE.

Para hacer frente al problema del desbalance de clases se han propuesto diferentes métodos de re-muestreo que tratan de equilibrar la cantidad de objetos entre las clases. Los métodos de re-muestreo pueden dividirse en 3 tipos:

- Sub-muestreo
- Sobre-muestreo
- Re-muestreo híbrido

Dentro de los métodos de sub-muestreo encontramos como los más conocidos a *Tomek links* [Tomek, 1976], *Condensed Nearest Neighbor Rule* [Hart, 1968], *One-Sided Selection* [Kubat et al., 1997], *Neighborhood Cleaning Rule* [Laurikkala, 2001], *Edited Nearest Neighbor Rule* [Wilson, 1972] y *RUS* (*Random Under Sampling*) entre otros.

Entre los métodos de sobre-muestro tenemos a SMOTE [Chawla et al., 2002], ROSE [Lunardon et al., 2014], SPIDER [Stefanowski and Wilk, 2008], GIS-C [Vivar, 2008], GIS-CJ [Vivar, 2008] y ROS (*Random Over Sampling*). SMOTE es uno de los más conocidos y exitosos [López et al., 2013], a partir del cual se han desarrollado numerosos métodos de re-muestreo [Sáez et al., 2015, Ramentol et al., 2012, Koto, 2014, Hu et al., 2009, Han et al., 2005].

Además de los métodos de sobre-muestreo y sub-muestreo, también se han propuesto métodos de re-muestreo híbrido. Entre los cuales destacan, los basados en SMOTE, como SMOTE TomekLinks y SMOTE ENN [Batista et al., 2004]. Entre otros métodos de re-muestreo híbrido están MBP+GGE [Alejo et al., 2013], RUSBoost [Seiffert et al., 2010], EasyEnsamble [Liu et al., 2009], entre otros.

Ya que el método propuesto en este trabajo de investigación es una versión determinista de SMOTE, en la siguiente sección se explica este método con detalle.

3.1. SMOTE

SMOTE [Chawla et al., 2002] es un método de sobre-muestreo que genera objetos sintéticos en puntos aleatorios entre objetos de la clase minoritaria y alguno de sus k-vecinos más cercanos elegido al azar. SMOTE trabaja como sigue:

- Calcular la cantidad de objetos sintéticos a generar (n), de acuerdo a un parámetro ($N \in \mathbb{R}$) de proporción a sobre-muestrear en la clase minoritaria (m). ($n = |m| * N$, donde $|m|$ es la cantidad de objetos en la clase minoritaria).

- Si el valor de N es:

  - $N < 1$, la cantidad de objetos a generar es menor que la cantidad de objetos en la clase minoritaria, por lo que se selecciona un subconjunto aleatorio, de tamaño n, de objetos de la clase minoritaria.

  - $N = 1$, la cantidad de objetos a generar es igual a la cantidad de objetos de la clase minoritaria, por lo que se seleccionan todos los objetos de la clase minoritaria.

  - $N > 1$, la cantidad de objetos a generar es mayor a la cantidad de objetos en la clase minoritaria, en este caso se seleccionan todos los objetos de la clase minoritaria tantas veces como el valor entero de N, además de un subconjunto aleatorio de objetos de la clase minoritaria en función de la diferencia entre N y el valor entero N.

- Para cada objeto seleccionado, de la clase minoritaria, se elige aleatoriamente uno de sus k vecinos más cercanos, donde k es un parámetro, y se prosigue con los siguientes pasos:
  - Se calcula la diferencia para cada atributo del objeto de la clase minoritaria y su vecino más cercano seleccionado.
  - Se multiplica la diferencia por un valor aleatorio entre 0 y 1.
  - Se suma el resultado de la multiplicación al objeto original de la clase minoritaria, generando de esta manera un nuevo objeto sintético.
- Se agregan los objetos sintéticos generados al conjunto de datos original.

Como se ha dicho antes, en este trabajo de investigación se propone un nuevo método de sobre-muestreo, que genera objetos sintéticos de forma similar que en SMOTE, por lo tanto, como parte del trabajo relacionado, revisaremos algunos de los métodos basados en SMOTE más exitosos y recientemente propuestos en la literatura. Dichos métodos los dividiremos en dos tipos: modificaciones de SMOTE y combinaciones de SMOTE con otros métodos.

3.2. Métodos basados en SMOTE

Los métodos de sobre-muestreo basados en SMOTE se pueden dividir en dos tipos: los métodos que modifican a SMOTE y los métodos que combinan SMOTE con algún otro tipo de pre-procesamiento de datos, como filtros de

ruido o sub-muestreo. A continuación presentamos los métodos más relevantes reportados en la literatura en cada uno de estos dos tipos.

3.2.1. Modificaciones de SMOTE

Borderline SMOTE 1 [Han et al., 2005], tiene como idea principal etiquetar algunos objetos de la clase minoritaria como bordes, de acuerdo a la clase de sus vecinos más cercanos calculados sobre el conjunto de entrenamiento completo. Si para un objeto de la clase minoritaria todos sus k vecinos más cercanos son de la clase mayoritaria se considera "ruido" y no es utilizado para el sobre-muestreo; de forma similar, si la cantidad de vecinos de la clase minoritaria es mayor que la de la clase mayoritaria, se considera "seguro" y tampoco es utilizado para generar objetos sintéticos. Los objetos de la clase minoritaria en los cuales la cantidad de vecinos más cercanos de la clase mayoritaria es mayor que la de la clase minoritaria se consideran objetos borde. Con los objetos borde se procede a generar los objetos sintéticos aplicando SMOTE.

Broderline SMOTE 2 [Han et al., 2005], este método es una variante de *Broderline* SMOTE 1, que sigue el mismo funcionamiento básico y se distingue de este último por generar objetos sintéticos entre objetos de la clase minoritaria y sus vecinos más cercanos de la clase mayoritaria. La generación de objetos sintéticos para un objeto de la clase minoritaria y un vecino cercano de la clase mayoritaria se hace a una distancia no mayor a la mitad de la distancia entre ellos.

Safe Level SMOTE [Bunkhumpornpat et al., 2009] asigna a cada objeto

de la clase minoritaria una evaluación que representa qué tan seguro es cada objeto, esta evaluación se hace con base en la cantidad de vecinos que tienen de la clase minoritaria, calculados sobre todo el conjunto de entrenamiento. El nivel de seguridad es igual a la cantidad de vecinos más cercanos de la clase minoritaria divididos sobre el total de vecinos más cercanos. La generación de objetos sintéticos es similar a SMOTE, con la diferencia de que los nuevos objetos sintéticos se generan aleatoriamente más cerca a los objetos con un mayor nivel de seguridad o en cualquier punto entre objetos con un mismo nivel de seguridad.

M-SMOTE [Hu et al., 2009] etiqueta los objetos de la clase minoritaria como ruido latente, bordes o seguros, de acuerdo a la cantidad de objetos de la clase mayoritaria en sus k vecinos más cercanos calculados sobre todo el conjunto de entrenamiento. Los objetos de la clase minoritaria etiquetados como ruido latente son aquellos objetos en los que todos sus k vecinos más cercanos pertenecen a la clase mayoritaria. Los objetos tipo borde son aquellos objetos de la clase minoritaria que tienen al menos un vecino de la clase minoritaria, pero no todos. Los objetos seguros son aquellos en los cuales ninguno de sus k vecinos más cercanos pertenecen a la clase mayoritaria. Para los objetos etiquetados como ruido latente no se generan objetos sintéticos; para los objetos etiquetados como seguros se elige cualquiera de los k vecinos más cercanos de la clase minoritaria para generar un nuevo objeto sintético; y para los objetos etiquetados como bordes sólo se utiliza al objeto más cercano de la clase minoritaria para la generación de un nuevo objeto sintético. La generación de objetos sintéticos se hace de forma similar a SMOTE.

Random SMOTE [Dong and Wang, 2011], en una primera etapa, genera N objetos sintéticos temporales en forma similar a SMOTE, en puntos alea-

torios entre 2 objetos de la clase minoritaria elegidos aleatoriamente; donde N es un parámetro de entrada. Posteriormente, para cada objeto temporal generado anteriormente se selecciona aleatoriamente un objeto de la clase minoritaria y se genera un nuevo objeto sintético entre ellos (los objetos elegidos pueden repetirse, pudiendo ser incluso el mismo objeto todas las veces).

LNE-SMOTE 1 [Maciejewski and Stefanowski, 2011] funciona de forma similar a *Safe Level* SMOTE. La diferencia con *Safe Level* SMOTE radica en que no solo se generan objetos sintéticos entre objetos de la clase minoritaria, sino también entre un objeto de la clase minoritaria y alguno de sus vecinos más cercanos en la clase mayoritaria, seleccionado aleatoriamente entre los k vecinos más cercanos considerando ambas clases.

SMOTE-OUT [Koto, 2014] genera objetos sintéticos entre objetos de la clase minoritaria y sus vecinos más cercanos de la clase mayoritaria, pero lo hace restando las diferencias de atributos entre objeto y vecino, en vez de sumarlas al objeto como se hace en SMOTE. Después de generar un objeto sintético, se encuentra su vecino más cercano de la clase minoritaria y, posteriormente, se genera otro objeto sintético entre estos dos últimos objetos; de forma similar a SMOTE. Para generar los primeros objetos sintéticos (entre objetos de la clase minoritaria y sus vecinos más cercanos de la clase mayoritaria) se utiliza un factor aleatorio entre 0 y 0.3 y para los segundos objetos sintéticos (entre los primeros objetos sintéticos y sus vecinos más cercanos de la minoritaria) se usa un factor aleatorio entre 0 y 0.5.

SMOTE-*Cosine* [Koto, 2014] genera objetos sintéticos de la misma forma que SMOTE. La diferencia de este método con respecto a SMOTE es que para el cálculo de los vecinos más cercanos se utiliza la suma de dos distan-

cias (Euclidiana y Coseno). Los autores del método señalan que al usar estas dos distancias en la generación de los objetos sintéticos se tiene en cuenta información de distancia y dirección de los objetos.

Selected SMOTE [Koto, 2014] genera objetos sintéticos de la misma forma que SMOTE. La diferencia con respecto a SMOTE es que *Selected* SMOTE sólo trabaja sobre algunos atributos elegidos por un método de selección de atributos aplicado previamente (los autores no especifican cuál). Los autores del método señalan que enfatizar en los atributos más significativos (según el método de selección de atributos) enriquece la variación de objetos de la clase minoritaria.

E-SMOTE [Deepa and Punithavalli, 2011] elige, por medio de un algoritmo genético, un subconjunto de atributos del total de atributos del conjunto de datos. El algoritmo genético inicia con individuos aleatorios y la evaluación de la aptitud de los individuos se hace con base en dos aspectos: el *accuracy* obtenido con el subconjunto de atributos seleccionados al utilizar SVM como clasificador y la porción de atributos elegidos con respecto al total de atributos. Al finalizar el algoritmo genético, se aplica SMOTE utilizando únicamente el subconjunto de atributos elegidos por el algoritmo genético.

Cabe mencionar que, tanto *Selected* SMOTE como E-SMOTE, reducen la cantidad de atributos del conjunto de entrenamiento, previo a la aplicación de SMOTE, y éste sólo actúa sobre los atributos elegidos. Por tanto, los dos métodos citados se clasifican como métodos que modifican a SMOTE, ya que el conjunto sobre-muestreado no tiene reducción en la cantidad de atributos final.

3.2.2. Combinaciones SMOTE

A continuación presentamos los principales métodos del estado del arte que combinan a SMOTE con otros métodos de distintos tipos como: sub-muestreo, filtrado de ruido o reglas de vecinos más cercanos.

SMOTE-RSB* [Ramentol et al., 2012] genera objetos sintéticos en la clase minoritaria utilizando SMOTE, después reduce el conjunto de objetos sintéticos generados por SMOTE aplicando el método RST basado en Teoría de Conjuntos Rugosos (*Rough Set Theory*). El método RST permite obtener una relación de similitud entre el conjunto de objetos sintéticos creados en la clase minoritaria y los objetos de la clase minoritaria del conjunto original, y así, con esta relación de similitud y utilizando un umbral, es posible eliminar los objetos sintéticos con menor similitud a los originales.

SMOTE Tomek *Links* [Batista et al., 2004] genera objetos sintéticos en la clase minoritaria utilizando SMOTE, después se etiquetan como Tomek todos los objetos del conjunto de entrenamiento (minoritario, minoritario sintético y mayoritario) en los cuales su vecino más cercano sea de la clase contraria y, a su vez, ese vecino también tenga como vecino más cercano a un objeto de clase contraria. Una vez etiquetados, los objetos Tomek se eliminan del conjunto de entrenamiento.

SMOTE ENN [Batista et al., 2004] genera objetos sintéticos en la clase minoritaria utilizando SMOTE, después aplica ENN (*Edited Nearest Neighbor*) para eliminar todos los objetos del conjunto de entrenamiento (minoritario, minoritario sintético y mayoritario) en los cuales 2 de sus 3 vecinos más cercanos pertenezcan a la clase contraria.

LNE-SMOTE 2 [Maciejewski and Stefanowski, 2011] genera objetos sintéticos utilizando LNE-SMOTE 1 (explicado en la sección 3.1.1), después el conjunto de entrenamiento es reducido mediante ENN y Tomek *Links*, en la misma forma que en SMOTE Tomek *Links* y SMOTE ENN.

3.2.3. Consideraciones finales

Como se mencionó en secciones previas de este capítulo, SMOTE y los métodos basados en SMOTE son aleatorios, ya sea en la forma de escoger los objetos o los vecinos de éstos, que se usarán como base para la generación de nuevos objetos sintéticos, así como en la posición donde serán generados los nuevos objetos sintéticos, con respecto al objeto base y su vecino más cercano.

Los métodos basados en SMOTE, aunque de algún modo modifican a SMOTE, conservan los pasos aleatorios de SMOTE, teniendo así el problema de producir resultados distintos cada vez que se aplican estos métodos a un mismo conjunto de datos. Con esto, existe la posibilidad de que los objetos sintéticos generados sesguen su posición hacia un objeto particular sin razón, pudiendo producir malos resultados de clasificación. Por estas razones, en esta tesis se propone un método de sobre-muestreo, el cual constituye una versión determinista de SMOTE al eliminar los pasos aleatorios del mismo.

Capítulo 4

Método propuesto

En este capítulo se explica el método propuesto, SMOTE-D, utilizando un ejemplo para ilustrar las ideas del mismo.

4.1. Método propuesto

La solución propuesta es un método de sobre-muestreo determinista, que genera objetos sintéticos para la clase minoritaria, para clases binarias y con atributos numéricos, de forma similar a SMOTE pero de manera determinista, partiendo de 3 ideas básicas: En primer lugar, calcular la cantidad de objetos sintéticos a generar para cada objeto de la clase minoritaria. Posteriormente, calcular la cantidad de objetos sintéticos a generar para cada vecino de los objetos de la clase minoritaria. Finalmente, generar los objetos sintéticos de manera uniforme entre un objeto y el vecino correspondiente. Estas ideas serán tratadas con detalle en las siguientes secciones.

4.1.1. Cantidad de objetos a generar para cada objeto de la clase minoritaria

El método propuesto toma en cuenta la desviación estándar de las distancias entre un objeto y sus k-vecinos más cercanos, para así generar una mayor cantidad de objetos sintéticos en torno a los objetos con mayor dispersión de distancias (mayor desviación estándar de las distancias), mientras que se generará una menor cantidad de objetos sintéticos en torno a aquellos objetos de la clase minoritaria con menor dispersión de distancias con sus vecinos más cercanos.

Al calcular la desviación estándar de las distancias, entre un objeto de la clase minoritaria y sus vecinos más cercanos, tendremos una idea aproximada de cuántos objetos sintéticos generar en torno a cada objeto de la clase minoritaria.

En la **figura 4.1** mostramos un ejemplo de 3 objetos de la clase minoritaria (obj_i), cada uno con 3 vecinos más cercanos y círculos punteados que representan sus desviaciones estándar de distancias (std_i) entre cada objeto y sus 3 vecinos más cercanos. Para cada vecindad se generará aproximadamente la proporción (p_i) que represente su desviación estándar (std_i) con respecto a la suma de las desviaciones estándar ($\sum_{i=1}^{|m|} std_i$) de todos los objetos de la clase minoritaria y sus k vecinos más cercanos. Lo dicho se obtiene con la ecuación 4.1:

$$p_i = \frac{std_i}{\sum_{i=1}^{|m|} std_i} \tag{4.1}$$

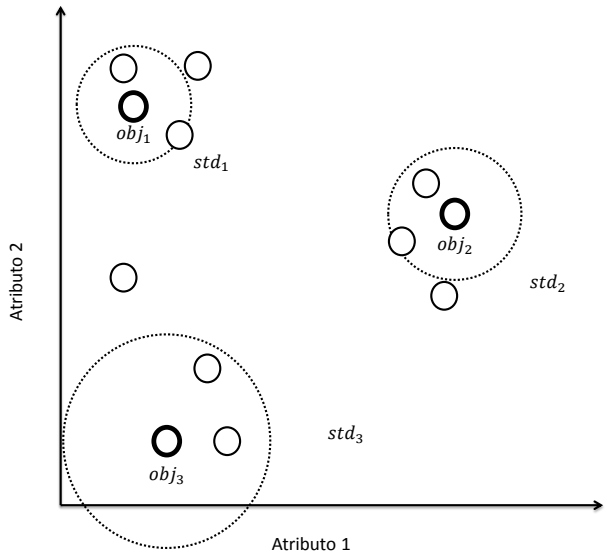

Figura 4.1: 3 objetos de la clase minoritaria con 3 vecinos más cercanos y sus desviaciones estándar indicadas con círculos punteados.

En la **figura 4.1** podemos apreciar como std_3 es más grande que std_1 y std_2, lo anterior nos permite suponer que entre los vecinos de obj_3 hay una mayor dispersión que entre los vecinos de obj_1 y obj_2. Por lo tanto, para reducir esta dispersión, se debe generar una mayor cantidad de objetos sintéticos en la vecindad de obj_3. Supongamos que los valores para std_1, std_2 y std_3 son 1.1, 1.5 y 2.6, respectivamente, y queremos generar 10 objetos sintéticos, dados estos valores de las desviaciones estándar, para la vecindad del obj_3 se generarían la mitad (5) de los 10 objetos sintéticos a generar, ya que std_3 representa la mitad del total de la suma de todas las desviaciones estándar. Para obj_1 y obj_2 se generarían 2 y 3 objetos sintéticos, respectivamente.

4.1.2. Cantidad de objetos para cada vecino

La cantidad de objetos sintéticos a generar en torno a un objeto de la clase minoritaria, y cada uno de sus k vecinos más cercanos, es calculada con base en la proporción de la distancia entre el objeto de la clase minoritaria y cada uno de sus k vecinos, al compararla con la suma de todas las distancias entre el objeto de la clase minoritaria y sus k vecinos más cercanos.

La idea de generar diferentes cantidades de objetos sintéticos, entre cada objeto de la clase minoritaria y sus k vecinos más cercanos, es aprovechar las diferentes distancias entre ellos, generando más objetos para los vecinos más lejanos.

En la **figura 4.2** tenemos un ejemplo de un objeto de la clase minoritaria (obj_i) y flechas punteadas para cada distancia con sus k vecinos más cercanos (obj_{ij}). Para cada vecino se generará aproximadamente la proporción (p_{ij}) que represente su distancia (d_{ij}), con respecto a la suma de las distancias ($\sum_{j=1}^{k} d_{ij}$) de todos los vecinos. Esto se logra con la ecuación 4.2:

$$p_{ij} = \frac{d_{ij}}{\sum_{j=1}^{k} d_{ij}} \tag{4.2}$$

En la **figura 4.2** podemos apreciar que la distancia del obj_{33} es más grande que las de obj_{31} y obj_{32}, lo anterior nos permite suponer que entre el vecino obj_{33} y obj_3 hay un mayor espacio que entre los vecinos obj_{31} y obj_{32}. Por lo tanto, hay oportunidad de generar una mayor cantidad de objetos sintéticos entre el objeto de la clase minoritaria obj_3 y su vecino cercano obj_{33}.

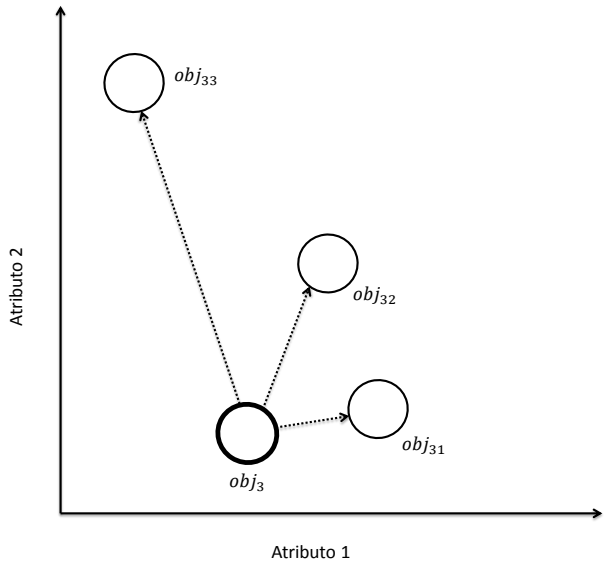

Figura 4.2: Un objeto de la clase minoritaria y las distancias con sus vecinos más cercanos, representadas como flechas punteadas.

Siguiendo con el ejemplo anterior, supongamos que tenemos para los vecinos obj_{31}, obj_{32} y obj_{33} distancias de 0.9, 2.0 y 6.9 respectivamente y queremos generar la misma cantidad de objetos sintéticos del ejemplo anterior, es decir, 5. Dado el valor de distancia para obj_{33}, se generarían para este objeto un 70 % (4) de objetos sintéticos, ya que, su distancia representa el 70 % del total de la suma de las distancias entre obj_3 y sus vecinos más cercanos. Para cada vecino más cercano, se generarán objetos sintéticos aproximadamente en la proporción que representen sus distancias con el objeto de la clase minoritaria, con respecto a la suma de todas las distancias del objeto de la clase minoritaria y todos sus vecinos más cercanos. En nuestro ejemplo 0, 1 y 4 para obj_{31}, obj_{32} y obj_{33} respectivamente.

4.1.3. Generación uniforme de objetos sintéticos

La generación de objetos sintéticos, entre un objeto de la clase minoritaria y uno de sus vecinos más cercanos, se hace de manera uniforme, es decir, uniformemente distribuidos entre ambos objetos.

La idea de generar objetos sintéticos de manera uniforme, entre un objeto de la clase minoritaria y uno de sus vecinos más cercanos, es cubrir todo el espacio posible entre los dos objetos. El cubrimiento de todo del espacio posible evita que los objetos sintéticos generados puedan estar muy cercanos entre ellos, o bien todos muy cercanos al objeto de la clase minoritaria o al vecino más cercano utilizado.

En la **figura 4.3** tenemos un ejemplo de un objeto de la clase minoritaria (obj_3) y los objetos sintéticos generados entre él y uno de sus vecinos más cercanos. Para lograr una generación uniforme de objetos sintéticos es necesario obtener la diferencia (dif'_{ij}) que tendrán los sintéticos entre ellos, lo anterior se obtiene como:

$$dif'_{ij} = \frac{obj_{ij} - obj_i}{(cant_{ij} + 1)} \tag{4.3}$$

Donde $cant_{ij} = Red(n * p_i * p_{ij})$ es la cantidad de objetos sintéticos a generar entre dos objetos, n es la cantidad de objetos sintéticos a generar para todo el conjunto y $Red()$ indica el redondeo de un número real a un entero.

En la **figura 4.3** podemos apreciar cómo la generación de 4 objetos

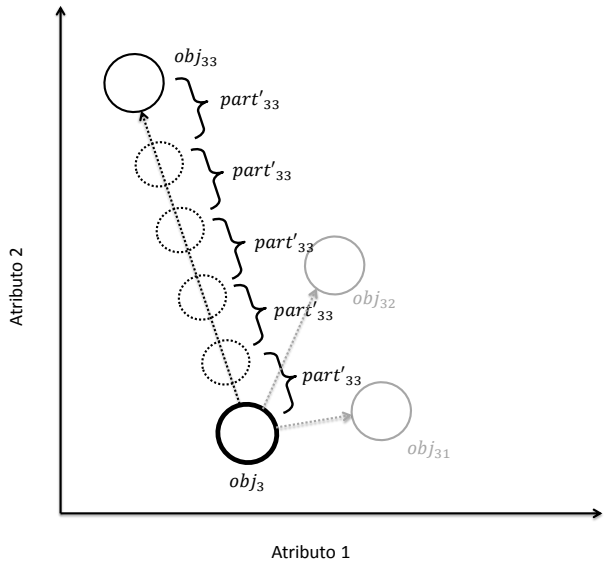

Figura 4.3: Generación de 4 objetos sintéticos (círculos punteados) entre un objeto de la clase minoritaria y uno de sus vecinos más cercanos.

sintéticos entre obj_3 y obj_{33} es uniforme entre ambos objetos, con lo que se logra un mejor cubrimiento del espacio entre ambos objetos por parte de los objetos sintéticos generados, evitando así que los objetos sintéticos generados se sesguen hacia alguno de los objetos o bien se aglomeren entre ellos.

4.2. SMOTE-D

SMOTE-D, al igual que SMOTE, en primer lugar necesita obtener los k vecinos más cercanos para los objetos de la clase minoritaria. Posteriormente, calcula la cantidad de objetos sintéticos a generar para cada objeto de la clase minoritaria, con base en la desviación estándar de las distancias de los k vecinos más cercanos de cada objeto. Después, calcula la cantidad de objetos para cada uno de los k vecinos más cercanos de cada objeto de la clase

minoritaria. Y finalmente, con la diferencia de atributos entre un objeto y uno de sus k vecinos más cercanos, teniendo en cuenta la cantidad de objetos sintéticos a generar entre ellos, se generan los nuevos objetos sintéticos de una manera uniforme entre ambos objetos.

A continuación se desarrolla un ejemplo para mostrar a detalle cómo funciona SMOTE-D.

Sea T un conjunto de datos desbalanceado dividido en dos clases. Dada la cantidad de objetos en la clase minoritaria ($|m|$) y la cantidad de objetos en la clase mayoritaria ($|M|$), calculamos la diferencia entre la cantidad de objetos en ambas clases (dif) de la siguiente forma:

$$dif = |M| - |m| \tag{4.4}$$

A manera de ejemplo, en la **figura 4.4** tenemos un conjunto de datos con $|M| = 20$ objetos en la clase mayoritaria y $|m| = 10$ objetos en la clase minoritaria, la diferencia entre la cantidad de objetos en ambas clases es $dif = 10$.

Para calcular la cantidad n de objetos a generar, SMOTE-D tiene como parámetro de entrada a N (en $\mathbb{R}$), el cual representa la porción de la diferencia dif de objetos sintéticos a generar, n está dada por:

$$n = dif * N \tag{4.5}$$

Si $N = 1$ el conjunto queda aproximadamente balanceado, en el ejemplo

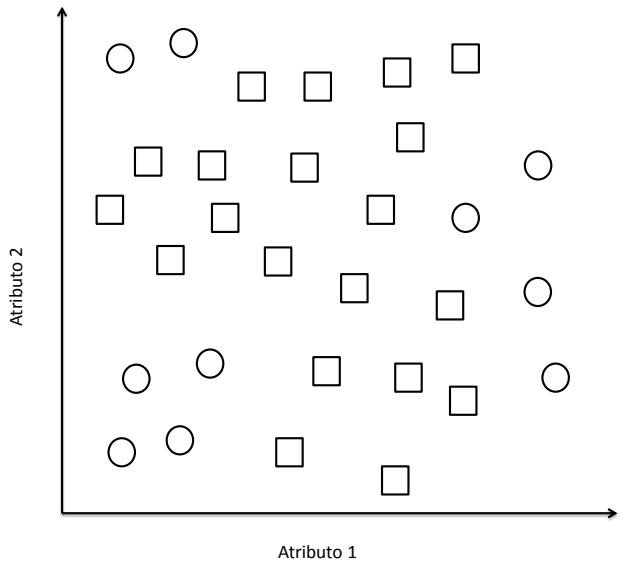

Figura 4.4: Conjunto de datos con 30 objetos, 10 en la clase minoritaria (círculos) y 20 en la clase mayoritaria (cuadrados).

$n = 10 * 1.0 = 10$.

Posteriormente, se calculan los k vecinos más cercanos, en la clase minoritaria, para todos los objetos de esta clase.

Supóngase que se tiene la misma clase minoritaria del ejemplo anterior con $|m| = 10$ objetos y considerando $k = 3$. También suponga que los objetos de la clase minoritaria tienen los vecinos mostrados en la **Tabla 4.1** con las distancias (d_{ik}) de los objetos ($obj_i, i = 1, .., 10$) de la clase minoritaria (m) y sus k vecinos más cercanos ($k = 1, 2, 3$).

A partir de las distancias de cada objeto (obj_i) de la clase minoritaria a sus k vecinos más cercanos (obj_{ij}) calculamos las desviaciones estándar ($std_i\ i = 1, \ldots, |m|$) de las distancias de los k vecinos más cercanos de cada objeto, estos valores son mostrados en la **Tabla 4.1**.

Tabla 4.1: Distancias entre cada objeto de la clase minoritaria y sus 3 vecinos más cercanos, junto con su desviación estándar.

obj_i	obj_{ij}			std_i
	obj_{i1}	obj_{i2}	obj_{i3}	
1	0.2	0.3	0.5	0.15
2	0.3	0.5	0.9	0.31
3	0.4	0.7	1.3	0.46
4	0.5	0.9	1.7	0.61
5	0.6	1.1	2.1	0.76
6	0.7	1.3	2.5	0.92
7	0.8	1.5	2.9	1.07
8	0.9	1.7	3.3	1.22
9	1.0	1.9	4.0	1.54
10	1.0	3.0	6.0	2.52

La suma de todas las std_i ($\sum_{i=1}^{|m|} std_i$) de la **Tabla 4.1** da como resultado 9.56. Esta suma nos permite obtener la fracción (p_i $i = 1, \ldots, |m|$) de objetos sintéticos a generar, para cada objeto de la clase minoritaria y sus k vecinos más cercanos, en la **Tabla 4.2** se muestra las fracciones de objetos sintéticos a generar, calculadas como:

$$p_i = \frac{std_i}{\sum_{i=1}^{|m|} std_i} \tag{4.6}$$

Las fracciones p_i, nos servirán para calcular la cantidad ($cant_{ij}$) de objetos sintéticos a generar entre los objetos obj_i y cada uno de sus k vecinos más cercanos. Para calcular las cantidades $cant_{ij}$ es necesario conocer la fracción (p_{ij} $i = 1, \ldots, |m|$ $j = 1, \ldots, k$) para cada vecino más cercano (obj_{ij}) de cada uno de los objetos de la clase minoritaria (obj_i). Las fracciones de cada vecino (p_{ij}) para cada objeto en la clase minoritaria son calculadas como sigue:

Tabla 4.2: Fracciones (p_i) de las desviaciones estándar de los 10 objetos de la clase minoritaria.

obj_i	std_i	p_i
1	0.15	0.016
2	0.31	0.032
3	0.46	0.048
4	0.61	0.064
5	0.76	0.080
6	0.92	0.096
7	1.07	0.112
8	1.22	0.128
9	1.54	0.161
10	2.52	0.263
suma	9.56	1.000

$$p_{ij} = \frac{d_{ij}}{\sum_{j=1}^{k} d_{ij}} \tag{4.7}$$

Para el ejemplo:

$$\sum_{j=1}^{k} d_{10j} = 10 \tag{4.8}$$

De esta forma las fracciones para cada vecino más cercano del obj_{10} serían igual a:

$$p_{10,1} = \frac{1}{10} = 0.1 \;\; p_{10,2} = \frac{3}{10} = 0.3 \;\; p_{10,3} = \frac{6}{10} = 0.6 \tag{4.9}$$

Al multiplicar dif por cada una de las fracciones (p_{ij}) y la fracción de la desviación estándar (p_i) del obj_{10}, podemos calcular la cantidad de objetos

sintéticos a generar ($cant_{ij}$) para cada vecino más cercano del objeto obj_{10}.

Continuando con el ejemplo anterior, si tenemos $dif = 10$, $p_{10} = 0.263$, $p_{10,1} = 0.1$, $p_{10,2} = 0.3$ y $p_{10,3} = 0.6$, se utiliza la siguiente fórmula para el cálculo de las cantidades de objetos sintéticos a generar:

$$cant_{ij} = Red(dif * p_i * p_{ij}) \tag{4.10}$$

Donde la función $Red(x)$ es el redondeo de un número real x a uno entero.

Entonces, las cantidades para obj_{10} y sus k vecinos más cercanos serían:

$$\begin{aligned} cant_{10,1} &= Red(10 * 0.263 * 0.1) = Red(0.263) = 0 \\ cant_{10,2} &= Red(10 * 0.263 * 0.3) = Red(0.789) = 1 \\ cant_{10,3} &= Red(10 * 0.263 * 0.6) = Red(1.570) = 2 \end{aligned} \tag{4.11}$$

Una vez obtenida la cantidad de objetos sintéticos a generar ($cant_{ij}$) entre un objeto de la clase minoritaria (obj_i) y cada uno de sus vecinos más cercanos (obj_{ij}), procedemos a calcular el vector de diferencias de atributos ($dif_{ij},\ i = 1, \ldots, |m|,\ j = 1, \ldots, k$) entre ambos objetos.

Supóngase que los objetos, en el conjunto de datos del ejemplo, están descritos por $r = 3$ atributos y los valores para cada atributo del objeto obj_{10}, su vecino más cercano $obj_{10,3}$ y la diferencia de atributos entre ellos ($dif_{10,3}$) son los mostrados en la **Tabla 4.3**.

Tabla 4.3: Valores de atributos para un objeto de la clase minoritaria, uno de sus k vecinos más cercanos y su diferencia.

	r_1	r_2	r_3
obj_{10}	1	3	5
$obj_{10,3}$	2	3	4.5
$dif10,3$	1	0	-0.5

Para generar objetos sintéticos entre un objeto de la clase minoritaria (obj_i) y alguno de sus vecinos más cercanos (obj_{ij}) es necesario dividir la diferencia de atributos sobre la cantidad de objetos sintéticos a generar entre alguno de sus vecinos más 1. Lo anterior nos permite generar nuevos objetos sintéticos de manera uniforme entre ambos objetos. Siguiendo con el ejemplo, en la **Tabla 4.4** se muestran los objetos sintéticos generados.

En la figura 4.5, se ilustra la generación de 3 objetos sintéticos (sin_{ijh} $i = 1, \ldots, |m|,\ j = 1, \ldots, k,\ h = 1, \ldots, r$) entre un objeto de la clase minoritaria (obj_i) y alguno de sus k vecinos más cercanos (obj_{ij}), de una manera uniforme cuando $cant_{ij} = 3$.

Tabla 4.4: Valores de atributos en la generación de 3 objetos sintéticos entre un objeto de la clase minoritaria y uno de sus vecinos más cercanos.

	r_1	r_2	r_3
obj_{10}	1.000	3.000	5.000
$sin_{10,3,1}$	1.250	3.000	4.875
$sin_{10,3,2}$	1.500	3.000	4.750
$sin_{10,3,3}$	1.750	3.000	4.625
$obj_{10,3}$	2.000	3.000	4.500

Para generar los nuevos objetos sintéticos (sin_{ijh}), es necesario sumar dif'_{ij} al objeto obj_i tantas veces como $cant_{ij}$, en cada suma se obtiene un nuevo objeto sintético. Siguiendo con el ejemplo anterior con una cantidad de

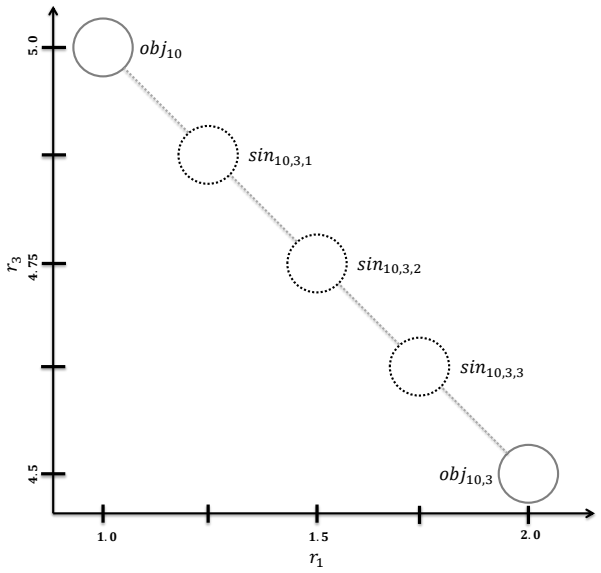

Figura 4.5: Generación uniforme de 3 objetos sintéticos (círculos punteados) entre un objeto de la clase minoritaria y uno de sus k vecinos más cercanos.

objetos sintéticos a generar $cant_{ij} = 3$ obtendríamos los 3 objetos sintéticos, mostrados en la **Tabla 4.4**.

Al finalizar, cada nuevo objeto sintético generado se agrega al subconjunto de objetos sintéticos generados (m') para la clase minoritaria (m), obteniendo así el subconjunto sobre-muestreado de objetos para la clase minoritaria ($m \cup m'$).

A continuación se presenta el algoritmo SMOTE-D:

Entrada:
- M subconjunto de objetos de la clase mayoritaria.
- m subconjunto de objetos de la clase minoritaria.
- N porcentaje de sobre-muestreo.
- k cantidad de vecinos más cercanos.

Salida:
- $D = \{M, m \cup m'\}$ conjunto de datos sobre-muestreado con los subconjuntos mayoritario (M), minoritario (m) y minoritario sintético (m').

Pasos:

1. Calcular la cantidad de objetos a generar ($n = Red((|M| - |m|) * N)$).
2. Calcular los k-vecinos más cercanos de los objetos de la clase minoritaria y sus distancias (d_{ij} $i = 1, ..., |m|$, $j = 1, ..., k$).
3. Calcular las desviaciones estándar (σ_i) de las distancias de los k-vecinos más cercanos de cada objeto.
4. Calcular las fracciones (p_i) de cada desviación estándar comparada con la suma de todas las desviaciones estándar.($p_i = \frac{d_i}{\sum_{i=1}^{|m|} d_i}$)
5. Calcular las fracciones (p_{ij}) de las distancias de los vecinos más cercanos de un objeto comparada con la suma de todas las distancias de los vecinos de ese objeto.($p_{ij} = \frac{d_{ij}}{\sum_{j=1}^{k} d_{ij}}$)
6. **Para** cada objeto de la clase minoritaria (obj_i) y cada uno de sus k vecinos más cercanos (obj_{ij}) **hacer**:
   - 6.1. Calcular la cantidad ($cant_{ij} = Red(n * p_i * p_{ij})$) de objetos sintéticos a generar entre el objeto y su vecino.
   - 6.2. Calcular la diferencia de atributos ($dif_{ij} = obj_{ij} - obj_i$) para los objetos.
   - 6.3. Calcular la diferencia dividida entre la cantidad de objetos a generar ($dif'_{ij} = dif_{ij}/(cant_{ij} + 1)$).
   - 6.4. Sumar la diferencia dividida (dif'_{ij}) al obj_i tantas veces como $cant_{ij}$. Obteniendo en cada suma un nuevo objeto sintético.
7. **Fin Para**
8. Agregar los objetos sintéticos generados al subconjunto (m').
9. **Regresar** $\{M, m \cup m'\}$.

4.3. Consideraciones finales

Con lo expuesto en las secciones previas de este capítulo, podemos decir que el método propuesto es diferente a SMOTE en 3 aspectos: la primera diferencia es que el método propuesto genera una cantidad de objetos sintéticos distinta para cada objeto de la clase minoritaria, mientras que SMOTE genera la misma cantidad de objetos para cada objeto de la clase minoritaria. Además, en SMOTE-D, la cantidad de objetos sintéticos a generar se determina a

partir de la desviación estándar de las distancias de un objeto de la clase minoritaria y sus k vecinos más cercanos.

La segunda diferencia está en que SMOTE-D genera una cantidad de objetos sintéticos, entre objetos de la clase minoritaria y cada uno de sus k vecinos, acorde con la distancia entre ellos y no de forma aleatoria como sucede en SMOTE. La distancia nos da una idea de cuánto espacio hay entre los objetos y a mayor distancia, mayor es la cantidad de objetos sintéticos a generar.

La última diferencia es que en SMOTE-D se generan objetos sintéticos uniformemente distribuidos entre un objeto de la clase minoritaria y uno de sus k vecinos más cercanos. A diferencia de SMOTE, que genera objetos sintéticos de forma aleatoria.

Con lo mencionado anteriormente, el método propuesto constituye una versión determinista de SMOTE, la cual, como mostramos en el siguiente capítulo, nos permite obtener una buena solución sin depender del azar.

Capítulo 5

Resultados experimentales

En este capítulo se muestran los resultados experimentales, en términos de calidad de clasificación, de la comparación del método propuesto (SMOTE-D) contra SMOTE. Además, se muestran los resultados experimentales de comparar métodos del estado del arte basados en SMOTE, contra sus versiones deterministas reemplazando SMOTE por SMOTE-D.

5.1. Conjuntos de datos

Para los experimentos de este trabajo de investigación, se utilizaron todos los 66 conjuntos de datos del apartado de conjuntos de datos con desbalance del repositorio KEEL [Alcalá et al., 2010], los cuales comparten las siguientes características:

- Clases binarias.

- Partición usando 5-*cross fold validation*.
- Atributos numéricos, desde un mínimo de 3 hasta un máximo de 19.

Otras de las características presentes en los conjuntos de datos utilizados son que tienen un IR (*Imbalance Ratio*) que va desde un mínimo de 1.82 hasta un máximo de 129.44 y cantidades de objetos desde un mínimo de 92 hasta un máximo de 5472.

El conjunto de datos *abalone* es el único que tiene un atributo nominal. Los valores nominales para ese atributo son F, I y M ($female, infant$ y $male$) y fueron remplazados por los atributos numéricos -1, 0 y 1 respectivamente. Esto se realizó en las versiones *abalone*19 y *abalone*9-18.

Los conjuntos de datos utilizados, así como su descripción, se muestran en la tabla A.1 del **Apéndice A**, ordenados por IR de menor a mayor.

5.2. SMOTE-D vs SMOTE

En esta sección se muestran los experimentos de la comparación, en términos de calidad de clasificación, del método propuesto (SMOTE-D) contra SMOTE.

En nuestros experimentos se utilizó la distancia Euclidiana para calcular los k vecinos más cercanos (k=5). Debido a que desde que SMOTE [Chawla et al., 2002] fue propuesto hasta los métodos de sobre-muestreo más recientes como SMOTE-IPF [Sáez et al., 2015] se ha usado k=5 (k para los métodos de sobre-muestreo), en nuestros experimentos usamos este mismo

valor. Como clasificador usamos K Vecinos más Cercanos (K-NN) con K=5 (K para encontrar la clasificación) que es el clasificador más comúnmente utilizado. Para la evaluación de los resultados se usaron las medidas Área Bajo la Curva (AUC) y F-*Measure* (F-M), además, SMOTE se aplicó un total de 10 veces por *fold*.

5.2.1. Resultados con AUC

En los resultados con AUC, mostrados en la Tabla 5.1, SMOTE-D supera a SMOTE en promedio. Además, SMOTE-D fue superior a SMOTE en 29 conjuntos de datos, mientras que SMOTE fue superior a SMOTE-D en 26 conjuntos de datos. Finalmente ambos métodos empataron en los 11 conjuntos de datos restantes.

Además, algo importante de observar es la desviación estándar en los resultados de SMOTE, por ejemplo en el conjunto de datos 23, SMOTE-D tiene un resultado de 0.888 y SMOTE un promedio de 0.874, además de que SMOTE-D supera a SMOTE por 0.014, la desviación estándar de SMOTE es igual a 0.060, más de 4 veces mayor a la diferencia. Tomando en cuenta que no hay forma de garantizar el mejor resultado de SMOTE, lo anterior resulta una desventaja, ya que sólo depende del azar el poder obtenerlo, pudiendo incluso obtener un resultado con 0.814.

5.2.2. Resultados con F-M

Por su parte, en los resultados con $F - M$ mostrados en la Tabla 5.1, SMOTE supera a SMOTE-D en promedio. Además, SMOTE-D fue superior a SMOTE en 21 conjuntos de datos, mientras que SMOTE fue superior a SMOTE-D en 34 conjuntos de datos. Ambos métodos empataron en los 11 conjuntos de datos restantes.

Por otra parte, en los resultados del conjunto de datos 23, SMOTE-D tiene un resultado de 0.791 y SMOTE un promedio de 0.788, es decir SMOTE-D supera a SMOTE por 0.003. Además, la desviación estándar de SMOTE en este conjunto de datos es igual a 0.086, más de 28 veces mayor a la diferencia. Con la desviación estándar de SMOTE se puede obtener un resultado con 0.702.

Tabla 5.1: SMOTE-D vs SMOTE usando distancia Euclidiana, K-NN y $k = 5$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.820	**0.821** ± 0.008	0.761	**0.764** ± 0.012
ecoli-0_vs_1	**0.969**	0.964 ± 0.000	**0.974**	0.971 ± 0.000
wisconsin	0.957	**0.961** ± 0.000	0.945	**0.949** ± 0.000
pima	0.652	0.652 ± 0.017	0.545	**0.552** ± 0.022
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.792	**0.796** ± 0.016	0.712	**0.715** ± 0.022
yeast1	0.639	**0.649** ± 0.009	0.488	**0.504** ± 0.012
haberman	**0.569**	0.551 ± 0.027	**0.385**	0.365 ± 0.035
vehicle2	0.923	0.923 ± 0.007	0.862	**0.863** ± 0.014
vehicle1	0.610	**0.618** ± 0.019	0.428	**0.438** ± 0.024
vehicle3	**0.639**	0.626 ± 0.019	**0.460**	0.441 ± 0.025
glass-0-1-2-3_vs_4-5-6	0.935	**0.942** ± 0.000	0.901	**0.906** ± 0.000
vehicle0	0.912	**0.922** ± 0.001	0.856	**0.864** ± 0.003
ecoli1	0.833	**0.835** ± 0.028	0.726	**0.732** ± 0.042
new-thyroid1	0.991	**0.994** ± 0.005	0.961	**0.973** ± 0.024

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
new-thyroid2	0.997	**0.998** ± 0.006	0.986	**0.994** ± 0.032
ecoli2	0.902	**0.909** ± 0.008	0.758	**0.798** ± 0.034
segment0	0.989	0.989 ± 0.000	0.975	0.975 ± 0.000
glass6	**0.933**	0.925 ± 0.009	0.852	0.852 ± 0.045
yeast3	0.847	**0.848** ± 0.012	0.676	**0.691** ± 0.022
ecoli3	**0.834**	0.825 ± 0.038	**0.626**	0.621 ± 0.050
page-blocks0	0.885	**0.895** ± 0.003	**0.794**	0.789 ± 0.005
ecoli-0-3-4_vs_5	**0.888**	0.874 ± 0.060	**0.791**	0.788 ± 0.086
yeast-2_vs_4	0.862	**0.867** ± 0.002	0.721	**0.728** ± 0.017
ecoli-0-6-7_vs_3-5	**0.877**	0.874 ± 0.000	0.685	**0.718** ± 0.000
ecoli-0-2-3-4_vs_5	0.861	**0.871** ± 0.000	0.753	**0.784** ± 0.000
glass-0-1-5_vs_2	0.624	**0.655** ± 0.008	0.241	**0.312** ± 0.014
yeast-0-3-5-9_vs_7-8	0.665	**0.676** ± 0.011	0.330	**0.348** ± 0.024
yeast-0-2-5-7-9_vs_3-6-8	0.746	**0.751** ± 0.005	0.452	**0.472** ± 0.019
yeast-0-2-5-6_vs_3-7-8-9	**0.876**	0.875 ± 0.013	0.698	**0.705** ± 0.023
ecoli-0-4-6_vs_5	**0.889**	0.877 ± 0.000	**0.796**	0.787 ± 0.000
ecoli-0-1_vs_2-3-5	**0.844**	0.841 ± 0.000	**0.693**	0.688 ± 0.000
ecoli-0-2-6-7_vs_3-5	**0.862**	0.852 ± 0.060	**0.732**	0.729 ± 0.075
glass-0-4_vs_5	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-3-4-6_vs_5	0.894	**0.896** ± 0.000	0.838	**0.848** ± 0.000
ecoli-0-3-4-7_vs_5-6	**0.882**	0.876 ± 0.000	**0.739**	0.731 ± 0.000
yeast-0-5-6-7-9_vs_4	**0.730**	0.726 ± 0.034	0.432	**0.445** ± 0.043
vowel0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-6-7_vs_5	**0.855**	0.839 ± 0.053	0.708	0.708 ± 0.075
glass-0-1-6_vs_2	0.595	0.595 ± 0.006	0.251	**0.257** ± 0.013
ecoli-0-1-4-7_vs_2-3-5-6	**0.860**	0.837 ± 0.005	**0.714**	0.688 ± 0.029
led7digit-0-2-4-5-6-7-8-9_vs_1	0.848	0.848 ± 0.000	0.520	0.520 ± 0.000
glass-0-6_vs_5	0.868	**0.869** ± 0.000	0.757	**0.766** ± 0.000
ecoli-0-1_vs_5	**0.990**	0.985 ± 0.000	**0.920**	0.914 ± 0.000
glass-0-1-4-6_vs_2	**0.680**	0.678 ± 0.006	**0.350**	0.346 ± 0.014
glass2	**0.664**	0.657 ± 0.004	**0.309**	0.293 ± 0.003
ecoli-0-1-4-7_vs_5-6	0.883	**0.893** ± 0.004	0.724	**0.752** ± 0.039
cleveland-0_vs_4	**0.554**	0.547 ± 0.012	**0.150**	0.149 ± 0.025
ecoli-0-1-4-6_vs_5	**0.892**	0.882 ± 0.004	0.798	**0.799** ± 0.040
shuttle-c0-vs-c4	**0.996**	0.995 ± 0.000	0.995	0.995 ± 0.000
yeast-1_vs_7	**0.653**	0.650 ± 0.008	0.271	**0.277** ± 0.027
glass4	**0.940**	0.939 ± 0.000	**0.826**	0.824 ± 0.000

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli4	**0.917**	0.914 ± 0.000	**0.814**	0.813 ± 0.000
page-blocks-1-3_vs_4	0.936	**0.938** ± 0.002	0.777	**0.795** ± 0.026
abalone9-18	**0.852**	0.847 ± 0.044	**0.532**	0.522 ± 0.042
glass-0-1-6_vs_5	**0.935**	0.930 ± 0.000	**0.753**	0.747 ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	**0.556**	0.541 ± 0.004	**0.122**	0.108 ± 0.005
glass5	0.890	0.890 ± 0.000	0.647	0.647 ± 0.000
yeast-2_vs_8	**0.753**	0.751 ± 0.004	0.325	**0.329** ± 0.016
yeast4	0.697	**0.715** ± 0.024	0.291	**0.321** ± 0.025
yeast-1-2-8-9_vs_7	0.564	**0.566** ± 0.003	0.105	**0.108** ± 0.006
yeast5	0.900	**0.903** ± 0.000	0.713	**0.718** ± 0.000
ecoli-0-1-3-7_vs_2-6	0.839	0.839 ± 0.000	0.513	0.513 ± 0.000
yeast6	0.789	**0.791** ± 0.002	0.344	**0.362** ± 0.018
abalone19	**0.611**	0.596 ± 0.001	**0.063**	0.057 ± 0.000
Promedio	**0.828**	0.827 ± 0.009	0.649	**0.654** ± 0.017

Fin de la Tabla 5.1.

5.2.3. Otros resultados

También se realizaron experimentos con los clasificadores Máquinas de Soporte Vectorial (SVM) y K vecinos más cercanos (KNN). En todos los casos se utilizó distancia Euclidiana y medida HVDM para calcular los k vecinos más cercanos. Finalmente, como medidas para la evaluación de los resultados, se usaron las medidas F-*Measure* (F-M) y Área Bajo la Curva (AUC).

Los experimentos restantes se encuentran en el Apéndice A "Conjuntos de Datos y Otros Resultados". En los experimentos mostrados en este capítulo y los mostrados en el Apéndice A, no se encontró diferencia estadísticamente significativa entre los resultados de SMOTE-D y SMOTE, al haber

aplicado una prueba t-test con un de nivel de significancia del 5 %. Además de aplicar *t-test*, también se aplicó una prueba de rangos con signo de *Wilcoxon* con un de nivel de significancia del 5 % con la cual tampoco fue posible encontrar diferencia estadísticamente significativa entre los resultados.

De lo anterior, podemos concluir que en términos de calidad de clasificación, ambos métodos producen el mismo resultado, pero SMOTE al ser un método aleatorio, se tiene que aplicar varias veces sobre el mismo conjunto de entrenamiento, para posteriormente seleccionar el mejor resultado, lo cual requiere un mayor tiempo.

Como podemos ver en los resultados de la Tabla 5.2 la cual contiene el resumen con los resultados promedios para SMOTE-D y SMOTE, la variación en los resultados con distintas configuraciones de la k es muy poca. Por ejemplo el resultado promedio de SMOTE-D con Árboles de Decisión y distancia Euclidiana es de 0.82 para $k = 3$ al igual que para $k = 7$, y solo varía en 0.02 para $k = 5$, un comportamiento similar se puede observar en los demás resultados. Algo notorio es que a un mayor valor para k la desviación estándar en los resultados de SMOTE crece.

Tabla 5.2: Resumen de resultados de SMOTE-D y SMOTE

Medida	k	Clasif.	AUC		FM	
			SMOTE-D	SMOTE	SMOTE-D	SMOTE
Euc	3	DT	**0.820**	0.814 ± 0.027	**0.648**	0.642 ± 0.052
		SVM	0.837	**0.844** ± 0.007	**0.633**	0.617 ± 0.010
		KNN	0.822	**0.822** ± 0.007	0.651	**0.654** ± 0.014
	5	DT	**0.822**	0.820 ± 0.034	0.644	0.644 ± 0.056
		SVM	0.832	**0.844** ± 0.007	0.616	**0.617** ± 0.016
		KNN	**0.828**	0.827 ± 0.009	0.649	**0.654** ± 0.017
	7	DT	0.820	**0.821** ± 0.036	0.640	**0.643** ± 0.061
		SVM	0.833	**0.844** ± 0.007	0.613	**0.618** ± 0.024
		KNN	0.828	**0.829** ± 0.010	0.645	**0.652** ± 0.020
HVDM	3	DT	**0.820**	0.819 ± 0.030	**0.648**	0.644 ± 0.056
		SVM	0.836	**0.845** ± 0.007	**0.635**	0.617 ± 0.015
		KNN	0.822	**0.826** ± 0.007	**0.654**	0.653 ± 0.014
	5	DT	0.817	**0.821** ± 0.031	0.633	**0.643** ± 0.057
		SVM	0.835	**0.845** ± 0.009	0.614	**0.619** ± 0.019
		KNN	0.828	**0.831** ± 0.008	0.654	**0.655** ± 0.017
	7	DT	**0.824**	0.823 ± 0.037	**0.645**	0.643 ± 0.064
		SVM	0.835	**0.845** ± 0.011	0.608	**0.618** ± 0.023
		KNN	0.828	**0.832** ± 0.009	0.648	**0.655** ± 0.018

5.3. Métodos basados en SMOTE-D vs métodos basados en SMOTE

En esta sección se muestran los experimentos de la comparación, en términos de calidad de clasificación, de métodos basados en SMOTE contra sus versiones deterministas, al reemplazar SMOTE por SMOTE-D.

En dichos experimentos también se utilizaron los 66 conjuntos de datos del repositorio KEEL [Alcalá et al., 2010], además de distancia Euclidiana y la medida HVDM para calcular los k vecinos más cercanos con $k = 5$. Como clasificadores se usaron Árboles de Decisión (DT) con poda de nodos, Máquinas de Soporte Vectorial (SVM) con un *kernel* lineal y K Vecinos más Cercanos (KNN) con $K = 5$. Todos los clasificadores se usaron con la configuración por defecto en MATLAB a excepción de KNN en el que la $K = 1$ fue reemplazado por $K = 5$, valor más usado en la literatura. Y finalmente, para la evaluación de los resultados se utilizaron las medidas F-*Measure* (F-M) y Área Bajo la Curva (AUC).

Los métodos utilizados en la comparación contra sus correspondientes versiones basadas en SMOTE-D son los siguientes:

- SMOTE *Cosine* (Cosine)
- *Borderline* SMOTE 1 (Bord. 1)
- *Borderline* SMOTE 2 (Bord. 2)
- *Safe Level* SMOTE (Safe Lev.)
- SMOTE *Out* (Out)

- M-SMOTE (M)

- *Selected* SMOTE (Selected)

5.3.1. Resultados usando distancia Euclidiana

En las Tablas 5.3, 5.4 y 5.5, donde se muestran los resultados usando distancia Euclidiana, de la comparación de los métodos basados en SMOTE contra sus versiones deterministas usando SMOTE-D, las versiones deterministas tuvieron resultados mayores a las aleatorias 18 veces (42.8 %), mientras que las versiones aleatorias tuvieron resultados mayores a las deterministas 24 veces (57.1 %), considerando los 7 métodos, los 3 clasificadores y las 2 medidas de evaluación.

.

Tabla 5.3: Métodos deterministas vs aleatorios con DT (Euclidiana)

Métodos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
Cosine	**0.822**	0.822 ± 0.032	0.642	**0.647** ± 0.060
Bord. 1	0.806	**0.811** ± 0.029	0.635	**0.642** ± 0.045
Bord. 2	0.804	**0.809** ± 0.029	0.632	**0.633** ± 0.054
Safe Lev.	**0.816**	0.804 ± 0.018	0.648	**0.648** ± 0.034
Out	**0.826**	0.818 ± 0.029	**0.649**	0.646 ± 0.047
M	**0.812**	0.812 ± 0.024	**0.645**	0.641 ± 0.044
Selected	**0.819**	0.815 ± 0.026	0.642	**0.643** ± 0.047
Promedio	**0.815**	0.813 ± 0.027	0.642	**0.643** ± 0.047

Tabla 5.4: Métodos deterministas vs aleatorios con SVM (Euclidiana)

Métodos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
Cosine	0.842	**0.850** ± 0.007	**0.639**	0.633 ± 0.016
Brod. 1	0.825	**0.839** ± 0.007	0.618	**0.621** ± 0.015
Bord. 2	0.825	**0.839** ± 0.010	0.603	**0.620** ± 0.021
Safe Lev.	**0.828**	0.811 ± 0.005	**0.628**	0.609 ± 0.012
Out	0.838	**0.851** ± 0.006	0.613	**0.634** ± 0.012
M	0.835	**0.849** ± 0.012	**0.643**	0.639 ± 0.018
Selected	0.835	**0.843** ± 0.008	**0.620**	0.614 ± 0.017
Promedio	0.833	**0.841** ± 0.008	0.623	**0.624** ± 0.016

Tabla 5.5: Métodos deterministas vs aleatorios con KNN (Euclidiana)

Métodos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
Cosine	**0.829**	0.828 ± 0.009	0.649	**0.654** ± 0.016
Bord. 1	0.819	**0.824** ± 0.007	**0.657**	0.657 ± 0.013
Bord. 2	0.821	**0.825** ± 0.009	**0.656**	0.655 ± 0.014
Safe Lev.	**0.826**	0.816 ± 0.002	0.660	**0.666** ± 0.005
Out	**0.831**	0.824 ± 0.013	**0.657**	0.656 ± 0.022
M	0.821	**0.822** ± 0.004	0.661	**0.662** ± 0.007
Selected	0.826	**0.829** ± 0.013	0.645	**0.658** ± 0.025
Promedio	**0.825**	0.824 ± 0.008	0.655	**0.658** ± 0.015

5.3.2. Resultados usando la medida HVDM

En las Tablas 5.6, 5.7 y 5.8, donde se muestran los resultados usando medida HVDM, las versiones deterministas tuvieron resultados mayores a las aleatorias 23 veces (54.7 %), mientras que las versiones aleatorias tuvieron resultados mayores a las deterministas 18 veces (42.8 %), además de 1 empate (2.3 %), considerando todos los casos.

.

Tabla 5.6: Métodos deterministas vs aleatorios con DT (HVDM)

Métodos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
Cosine	**0.822**	0.820 ± 0.033	0.642	**0.643** ± 0.063
Bord. 1	0.808	**0.810** ± 0.031	**0.637**	0.634 ± 0.053
Bord. 2	**0.815**	0.812 ± 0.035	**0.641**	0.625 ± 0.066
Safe Lev.	**0.813**	0.799 ± 0.022	**0.644**	0.637 ± 0.036
Out	**0.827**	0.823 ± 0.035	**0.641**	0.640 ± 0.062
M	**0.816**	0.805 ± 0.029	**0.648**	0.638 ± 0.054
Selected	0.814	**0.819** ± 0.026	0.628	**0.640** ± 0.047
Promedio	**0.816**	0.812 ± 0.030	**0.640**	0.637 ± 0.054

Tabla 5.7: Métodos deterministas vs aleatorios con SVM (HVDM)

Métodos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
Cosine	0.842	**0.851** ± 0.004	**0.636**	0.634 ± 0.011
Brod. 1	0.830	**0.839** ± 0.006	**0.639**	0.625 ± 0.016
Bord. 2	0.833	**0.838** ± 0.011	**0.617**	**0.617** ± 0.019
Safe Lev.	**0.817**	0.772 ± 0.004	**0.624**	0.569 ± 0.014
Out	0.839	**0.848** ± 0.007	0.618	**0.634** ± 0.011
M	0.822	**0.836** ± 0.018	**0.629**	0.627 ± 0.032
Selected	0.833	**0.845** ± 0.009	0.612	**0.616** ± 0.022
Promedio	0.831	**0.833** ± 0.008	**0.625**	0.617 ± 0.018

Tabla 5.8: Métodos deterministas vs aleatorios con KNN (HVDM)

Métodos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
Cosine	**0.829**	0.828 ± 0.009	0.649	**0.653** ± 0.017
Bord. 1	0.818	**0.826** ± 0.007	0.658	**0.659** ± 0.016
Bord. 2	0.820	**0.826** ± 0.008	**0.651**	0.650 ± 0.016
Safe Lev.	**0.822**	0.810 ± 0.003	**0.667**	0.659 ± 0.006
Out	**0.838**	0.830 ± 0.013	**0.660**	0.657 ± 0.025
M	0.817	**0.819** ± 0.003	**0.660**	0.658 ± 0.007
Selected	**0.833**	0.833 ± 0.014	0.654	**0.658** ± 0.025
Promedio	**0.825**	0.824 ± 0.008	**0.657**	0.656 ± 0.016

5.3.3. Observaciones

En los resultados de la comparación de métodos Deterministas vs Aleatorios, se pueden observar que las versiones Deterministas tienen mejores resultados promedio cuando se utiliza la distancia HVDM. De forma contraria a lo que se puede observar cuando se utiliza la distancia Euclidiana.

Los métodos basados en SMOTE, al ser aleatorios, al igual que SMOTE se tienen que aplicar varias veces para seleccionar un buen resultado, y esto requiere un mayor tiempo.

Los tiempos de ejecución de SMOTE-D y SMOTE pueden ser vistos en en la Tabla A.2, de la sección Tiempos de ejecución del apéndice A. Los tiempos de ejecución se obtuvieron al ejecutar SMOTE-D una vez y SMOTE 10 veces, aplicándolos en los 66 conjuntos de datos, además se aplicó un t-test con un 99 % de confianza sin encontrar diferencia estadísticamente significativa entre los tiempos de ejecución de cada método. Es importante resaltar que, aunque los tiempos de SMOTE-D y SMOTE son similares al ejecutarse una sola vez, en la práctica SMOTE tiene que ejecutarse varias veces para buscar un buen resultado.

Capítulo 6

Conclusiones y Trabajo Futuro

El desbalance entre clases en conjuntos de datos es un problema que afecta, en términos de calidad, a la tarea de clasificación supervisada. Por dicha afectación, la comunidad científica ha dedicado esfuerzos para resolver el problema del desbalance entre clases, dichos esfuerzos se ven reflejados en la publicación de múltiples artículos referentes al tema. En estas publicaciones encontramos muchos métodos que hacen frente al problema de desbalance entre clases con métodos de sobre-muestreo para las clases minoritarias. Dentro de los métodos de sobre-muestreo encontramos a SMOTE, como uno de los métodos mejor conocidos y referenciados.

6.1. Conclusiones

En este trabajo de tesis, se propuso un nuevo método de sobre-muestreo, su principal característica es ser una versión determinista de SMOTE, al eli-

minar los pasos aleatorios del mismo. Además, se evaluó la idea determinista de SMOTE-D, en otros métodos basados en SMOTE reportados en la literatura, reemplazando en ellos a SMOTE por SMOTE-D.

En nuestros experimentos, el método propuesto mostró ser similar a SMOTE en términos de la calidad de clasificación. Los experimentos se realizaron con 66 conjuntos de datos con distintos grados de desbalance. La experimentación con dichos conjuntos de datos nos permitió concluir que nuestro método SMOTE-D sólo necesita ejecutarse una vez para obtener una solución que permita una calidad de clasificación similar a la que se esperaría obtener al aplicar SMOTE varias veces en la búsqueda de una buena solución.

Al comparar los métodos basados en SMOTE contra sus versiones deterministas basadas en SMOTE-D podemos concluir que los métodos basados en SMOTE-D cuando se usa la medida HVDM para calcular los k vecinos más cercanos, obtienen mejores resultados que los resultados obtenidos por los métodos basados en SMOTE.

En cuanto a los experimentos variando la cantidad de vecinos más cercanos podemos concluir que tanto SMOTE-D como SMOTE presentan un comportamiento similar en los resultados de clasificación.

También podemos concluir que tanto SMOTE-D como SMOTE comparten el problema de ser sensibles al ruido, lo cual genera la posibilidad de crear objetos sintéticos para la clase minoritaria en zonas de la clase mayoritaria. Este problema no se aborda en esta tesis, aunque es parcialmente abordado en *Safe Level* SMOTE-D y M-SMOTE-D, este problema no es abordado en la presente tesis.

Finalmente, en los experimentos realizados para evaluar los tiempos de ejecución, tanto para SMOTE-D como SMOTE, podemos concluir que ambos métodos reportan un tiempo de ejecución similar. Esto refuerza la ventaja de SMOTE-D sobre SMOTE el cual usualmente se aplica varias veces.

6.2. Contribuciones

La principal contribución de esta tesis es una versión determinista (SMOTE-D) de uno de los métodos mejor conocidos en la literatura, SMOTE. SMOTE-D constituye un método de sobre-muestreo determinista que da resultados con la misma calidad que la que podría dar SMOTE al aplicarlo varias veces sin la necesidad de ello.

6.3. Trabajo futuro

Como trabajo futuro de esta tesis proponemos modificar a SMOTE-D para que pueda trabajar con atributos nominales, ya que en la versión actual solo puede trabajar con atributos numéricos.

Por otra parte, otro problema de investigación abierto consiste en proponer estrategias de re-muestreo, manteniendo el enfoque determinista propuesto, para conjuntos de datos desbalanceados con más de dos clases, y no solamente usar un enfoque de una clase contra el resto.

6.4. Publicaciones

Derivado de este trabajo de investigación, se publicó el trabajo:

Torres, F. R., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2016). **Smote-d a deterministic version of smote**. In Mexican Conference on Pattern Recognition, pages 177188. Springer.

Apéndice A

Conjuntos de Datos y Otros Resultados

En este capítulo se encuentran las tablas con los conjuntos de datos utilizados y los resultados de la comparación de SMOTE-D contra SMOTE. Los resultados se separan por la distancia Euclidiana y la medida HVDM, estos a su vez por el valor de los k vecinos más cercanos (3,5,7), y finalmente se dividen por los clasificadores Árboles de Decisión, Maquinas de Soporte Vectorial (SVM) y K Vecinos más Cercanos (K-NN).

A.1. Conjuntos de datos

Tabla A.1: Conjuntos de datos utilizados en nuestros experimentos

Índice	Nombre	# atributos	# objetos	IR
1	glass1	9	214	1.82
2	ecoli-0_vs_1	7	220	1.86
3	wisconsin	9	683	1.86
4	pima	8	768	1.87
5	iris0	4	150	2.00
6	glass0	9	214	2.06
7	yeast1	8	1484	2.46
8	haberman	3	306	2.78
9	vehicle2	18	846	2.88
10	vehicle1	18	846	2.90
11	vehicle3	18	846	2.99
12	glass-0-1-2-3_vs_4-5-6	9	214	3.20
13	vehicle0	18	846	3.25
14	ecoli1	7	336	3.36
15	new-thyroid1	5	215	5.14
16	new-thyroid2	5	215	5.14
17	ecoli2	7	336	5.46
18	segment0	19	2308	6.02
19	glass6	9	214	6.38
20	yeast3	8	1484	8.10
21	ecoli3	7	336	8.60
22	page-blocks0	10	5472	8.79
23	ecoli-0-3-4_vs_5	7	200	9.00
24	yeast-2_vs_4	8	514	9.08
25	ecoli-0-6-7_vs_3-5	7	222	9.09
26	ecoli-0-2-3-4_vs_5	7	202	9.10
27	glass-0-1-5_vs_2	9	172	9.12
28	yeast-0-3-5-9_vs_7-8	8	506	9.12

Sigue en la página siguiente.

Índice	Nombre	# atributos	# objetos	IR
29	yeast-0-2-5-7-9_vs_3-6-8	8	1004	9.14
30	yeast-0-2-5-6_vs_3-7-8-9	8	1004	9.14
31	ecoli-0-4-6_vs_5	6	203	9.15
32	ecoli-0-1_vs_2-3-5	7	244	9.17
33	ecoli-0-2-6-7_vs_3-5	7	224	9.18
34	glass-0-4_vs_5	9	92	9.22
35	ecoli-0-3-4-6_vs_5	7	205	9.25
36	ecoli-0-3-4-7_vs_5-6	7	257	9.28
37	yeast-0-5-6-7-9_vs_4	8	528	9.35
38	vowel0	13	988	9.98
39	ecoli-0-6-7_vs_5	6	220	10.00
40	glass-0-1-6_vs_2	9	192	10.29
41	ecoli-0-1-4-7_vs_2-3-5-6	7	336	10.59
42	led7digit-0-2-4-5-6-7-8-9_vs_1	7	443	10.97
43	glass-0-6_vs_5	9	108	11.00
44	ecoli-0-1_vs_5	6	240	11.00
45	glass-0-1-4-6_vs_2	9	205	11.06
46	glass2	9	214	11.59
47	ecoli-0-1-4-7_vs_5-6	6	332	12.28
48	cleveland-0_vs_4	13	177	12.62
49	ecoli-0-1-4-6_vs_5	6	280	13.00
50	shuttle-c0-vs-c4	9	1829	13.87
51	yeast-1_vs_7	7	459	14.30
52	glass4	9	214	15.47
53	ecoli4	7	336	15.80
54	page-blocks-1-3_vs_4	10	472	15.86
55	abalone9-18	8	731	16.40
56	glass-0-1-6_vs_5	9	184	19.44
57	shuttle-c2-vs-c4	9	129	20.50
58	yeast-1-4-5-8_vs_7	8	693	22.10

Sigue en la página siguiente.

Índice	Nombre	# atributos	# objetos	IR
59	glass5	9	214	22.78
60	yeast-2_vs_8	8	482	23.10
61	yeast4	8	1484	28.10
62	yeast-1-2-8-9_vs_7	8	947	30.57
63	yeast5	8	1484	32.73
64	ecoli-0-1-3-7_vs_2-6	7	281	39.14
65	yeast6	8	1484	41.40
66	abalone19	8	4174	129.44

Fin de la Tabla A.1.

A.2. Tiempos de ejecución

Tabla A.2: Tiempos de ejecución en segundos de SMOTE-D (una vez) y SMOTE (promedio y suma de diez veces).

Nombre	SMOTE-D	Promedio SMOTE	Suma SMOTE
glass1	0.000	0.001	0.006
ecoli-0_vs_1	0.006	0.000	0.003
wisconsin	0.016	0.002	0.022
pima	0.019	0.002	0.022
iris0	0.009	0.000	0.000
glass0	0.006	0.001	0.009
yeast1	0.041	0.006	0.059
haberman	0.003	0.001	0.012
vehicle2	0.025	0.004	0.044
vehicle1	0.022	0.004	0.041
vehicle3	0.019	0.004	0.041
glass-0-1-2-3_vs_4-5-6	0.006	0.001	0.009
vehicle0	0.016	0.005	0.047
ecoli1	0.003	0.002	0.016
new-thyroid1	0.009	0.001	0.009
new-thyroid2	0.000	0.002	0.016
ecoli2	0.006	0.002	0.022

Sigue en la página siguiente.

Nombre	SMOTE-D	Promedio SMOTE	Suma SMOTE
segment0	0.044	0.026	0.262
glass6	0.000	0.002	0.016
yeast3	0.016	0.012	0.125
ecoli3	0.000	0.003	0.031
page-blocks0	0.109	0.079	0.789
ecoli-0-3-4_vs_5	0.003	0.002	0.016
yeast-2_vs_4	0.003	0.004	0.044
ecoli-0-6-7_vs_3-5	0.000	0.002	0.016
ecoli-0-2-3-4_vs_5	0.003	0.002	0.016
glass-0-1-5_vs_2	0.000	0.002	0.016
yeast-0-3-5-9_vs_7-8	0.006	0.004	0.041
yeast-0-2-5-7-9_vs_3-6-8	0.006	0.008	0.078
yeast-0-2-5-6_vs_3-7-8-9	0.012	0.008	0.081
ecoli-0-4-6_vs_5	0.003	0.001	0.012
ecoli-0-1_vs_2-3-5	0.003	0.002	0.016
ecoli-0-2-6-7_vs_3-5	0.003	0.002	0.016
glass-0-4_vs_5	0.000	0.001	0.009
ecoli-0-3-4-6_vs_5	0.006	0.002	0.016
ecoli-0-3-4-7_vs_5-6	0.003	0.002	0.016
yeast-0-5-6-7-9_vs_4	0.003	0.004	0.044
vowel0	0.006	0.008	0.084
ecoli-0-6-7_vs_5	0.000	0.002	0.019
glass-0-1-6_vs_2	0.000	0.002	0.016
ecoli-0-1-4-7_vs_2-3-5-6	0.003	0.002	0.022
led7digit-0-2-4-5-6-7-8-9_vs_1	0.003	0.003	0.034
glass-0-6_vs_5	0.000	0.002	0.019
ecoli-0-1_vs_5	0.000	0.001	0.006
glass-0-1-4-6_vs_2	0.003	0.001	0.012
glass2	0.003	0.002	0.016
ecoli-0-1-4-7_vs_5-6	0.003	0.002	0.022
cleveland-0_vs_4	0.003	0.002	0.016
ecoli-0-1-4-6_vs_5	0.006	0.002	0.022
shuttle-c0-vs-c4	0.012	0.019	0.193
yeast-1_vs_7	0.003	0.004	0.041
glass4	0.000	0.002	0.016
ecoli4	0.006	0.002	0.022
page-blocks-1-3_vs_4	0.003	0.004	0.044
abalone9-18	0.003	0.007	0.069

Sigue en la página siguiente.

Nombre	SMOTE-D	Promedio SMOTE	Suma SMOTE
glass-0-1-6_vs_5	0.000	0.002	0.016
shuttle-c2-vs-c4	0.003	0.001	0.009
yeast-1-4-5-8_vs_7	0.003	0.007	0.066
glass5	0.000	0.003	0.028
yeast-2_vs_8	0.003	0.004	0.044
yeast4	0.003	0.016	0.162
yeast-1-2-8-9_vs_7	0.009	0.009	0.094
yeast5	0.006	0.017	0.165
ecoli-0-1-3-7_vs_2-6	0.000	0.002	0.025
yeast6	0.003	0.017	0.165
abalone19	0.012	0.065	0.652
Promedio	0.008	0.006	0.063

Fin de la Tabla A.2.

A.3. Utilizando la distancia Euclidiana

A.3.1. Valor de k igual a 3

Tabla A.3: SMOTE-D vs SMOTE usando distancia Euclidiana, Árboles de Decisión y $k = 3$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.725	**0.725** ± 0.036	**0.650**	0.649 ± 0.046
ecoli-0_vs_1	0.965	**0.974** ± 0.009	0.971	**0.981** ± 0.009
wisconsin	**0.952**	0.947 ± 0.005	**0.937**	0.928 ± 0.006
pima	**0.680**	0.670 ± 0.028	**0.586**	0.575 ± 0.034
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	**0.787**	0.784 ± 0.049	**0.707**	0.701 ± 0.059
yeast1	0.662	**0.662** ± 0.031	0.522	**0.523** ± 0.043
haberman	0.557	**0.564** ± 0.026	0.356	**0.365** ± 0.040
vehicle2	**0.950**	0.943 ± 0.011	**0.922**	0.911 ± 0.019
vehicle1	**0.679**	0.676 ± 0.025	**0.522**	0.515 ± 0.036
vehicle3	0.660	**0.694** ± 0.035	0.487	**0.539** ± 0.041
glass-0-1-2-3_vs_4-5-6	0.883	**0.893** ± 0.017	0.816	**0.837** ± 0.029

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
vehicle0	**0.932**	0.907 ± 0.017	**0.888**	0.852 ± 0.020
ecoli1	**0.873**	0.852 ± 0.029	**0.788**	0.760 ± 0.037
new-thyroid1	0.937	**0.951** ± 0.052	**0.909**	0.907 ± 0.068
new-thyroid2	0.920	**0.933** ± 0.031	0.844	**0.890** ± 0.022
ecoli2	0.848	**0.851** ± 0.027	**0.726**	0.715 ± 0.051
segment0	**0.994**	0.993 ± 0.000	**0.987**	0.985 ± 0.002
glass6	**0.928**	0.873 ± 0.116	**0.825**	0.761 ± 0.191
yeast3	**0.882**	0.864 ± 0.019	**0.743**	0.728 ± 0.033
ecoli3	**0.772**	0.747 ± 0.063	**0.535**	0.503 ± 0.075
page-blocks0	0.917	**0.922** ± 0.008	**0.839**	0.825 ± 0.015
ecoli-0-3-4_vs_5	0.852	**0.861** ± 0.004	0.682	**0.725** ± 0.033
yeast-2_vs_4	**0.900**	0.862 ± 0.023	**0.719**	0.707 ± 0.025
ecoli-0-6-7_vs_3-5	0.852	**0.854** ± 0.018	0.628	**0.647** ± 0.081
ecoli-0-2-3-4_vs_5	0.841	**0.849** ± 0.042	0.654	**0.716** ± 0.085
glass-0-1-5_vs_2	**0.693**	0.610 ± 0.112	**0.402**	0.265 ± 0.206
yeast-0-3-5-9_vs_7-8	0.607	**0.632** ± 0.039	0.272	**0.312** ± 0.069
yeast-0-2-5-7-9_vs_3-6-8	**0.744**	0.736 ± 0.023	**0.475**	0.474 ± 0.040
yeast-0-2-5-6_vs_3-7-8-9	**0.900**	0.896 ± 0.027	0.764	**0.770** ± 0.044
ecoli-0-4-6_vs_5	**0.861**	0.854 ± 0.043	**0.752**	0.716 ± 0.082
ecoli-0-1_vs_2-3-5	0.806	**0.810** ± 0.003	**0.641**	0.621 ± 0.015
ecoli-0-2-6-7_vs_3-5	**0.847**	0.823 ± 0.005	**0.655**	0.621 ± 0.039
glass-0-4_vs_5	**0.994**	0.994 ± 0.000	**0.933**	0.933 ± 0.000
ecoli-0-3-4-6_vs_5	0.839	**0.857** ± 0.011	**0.738**	0.733 ± 0.077
ecoli-0-3-4-7_vs_5-6	**0.862**	0.847 ± 0.008	**0.693**	0.646 ± 0.043
yeast-0-5-6-7-9_vs_4	**0.733**	0.712 ± 0.054	**0.440**	0.433 ± 0.072
vowel0	**0.938**	0.935 ± 0.002	**0.887**	0.881 ± 0.017
ecoli-0-6-7_vs_5	0.820	**0.852** ± 0.012	0.602	**0.662** ± 0.065
glass-0-1-6_vs_2	**0.640**	0.590 ± 0.010	**0.337**	0.246 ± 0.000
ecoli-0-1-4-7_vs_2-3-5-6	**0.862**	0.862 ± 0.015	**0.693**	0.675 ± 0.087
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.890**	0.887 ± 0.034	0.762	**0.769** ± 0.049
glass-0-6_vs_5	**0.906**	0.878 ± 0.009	**0.765**	0.728 ± 0.071
ecoli-0-1_vs_5	0.945	**0.945** ± 0.000	**0.866**	0.866 ± 0.000
glass-0-1-4-6_vs_2	**0.707**	0.706 ± 0.005	0.397	**0.440** ± 0.011
glass2	**0.678**	0.659 ± 0.052	**0.370**	0.347 ± 0.052
ecoli-0-1-4-7_vs_5-6	**0.867**	0.863 ± 0.035	**0.732**	0.692 ± 0.119
cleveland-0_vs_4	**0.828**	0.809 ± 0.134	0.598	**0.658** ± 0.209
ecoli-0-1-4-6_vs_5	**0.828**	0.820 ± 0.048	**0.588**	0.562 ± 0.045

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.566	**0.631** ± 0.044	0.157	**0.267** ± 0.085
glass4	0.837	**0.876** ± 0.011	0.685	**0.733** ± 0.135
ecoli4	**0.863**	0.850 ± 0.000	**0.726**	0.721 ± 0.000
page-blocks-1-3_vs_4	0.997	**0.997** ± 0.000	0.966	**0.966** ± 0.000
abalone9-18	**0.726**	0.719 ± 0.054	**0.397**	0.382 ± 0.080
glass-0-1-6_vs_5	0.838	**0.838** ± 0.000	0.660	**0.660** ± 0.000
shuttle-c2-vs-c4	0.950	0.950 ± 0.000	0.933	**0.933** ± 0.000
yeast-1-4-5-8_vs_7	**0.604**	0.508 ± 0.073	**0.169**	0.058 ± 0.067
glass5	**0.897**	0.897 ± 0.000	0.760	**0.760** ± 0.000
yeast-2_vs_8	**0.758**	0.756 ± 0.064	**0.471**	0.406 ± 0.146
yeast4	0.677	**0.679** ± 0.041	0.318	**0.319** ± 0.060
yeast-1-2-8-9_vs_7	**0.642**	0.601 ± 0.051	**0.245**	0.193 ± 0.096
yeast5	**0.912**	0.876 ± 0.035	**0.743**	0.705 ± 0.076
ecoli-0-1-3-7_vs_2-6	**0.842**	0.837 ± 0.012	**0.526**	0.501 ± 0.176
yeast6	0.754	**0.776** ± 0.002	0.360	**0.435** ± 0.014
abalone19	**0.534**	0.523 ± 0.025	**0.038**	0.026 ± 0.016
Promedio	**0.820**	0.814 ± 0.027	**0.648**	0.642 ± 0.052

Fin de la Tabla A.19.

Tabla A.4: SMOTE-D vs SMOTE usando distancia Euclidiana, SVM y $k = 3$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.603	**0.650** ± 0.000	0.578	**0.590** ± 0.000
ecoli-0_vs_1	0.969	**0.978** ± 0.000	0.974	**0.984** ± 0.000
wisconsin	**0.969**	0.968 ± 0.012	0.953	**0.955** ± 0.012
pima	0.733	**0.745** ± 0.019	0.665	**0.670** ± 0.022
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.728	**0.750** ± 0.000	0.642	**0.660** ± 0.000
yeast1	0.684	**0.706** ± 0.001	0.561	**0.581** ± 0.001
haberman	0.550	**0.623** ± 0.000	0.422	**0.434** ± 0.000
vehicle2	0.947	**0.956** ± 0.000	0.902	**0.923** ± 0.000
vehicle1	**0.812**	0.810 ± 0.010	0.665	**0.672** ± 0.017
vehicle3	0.778	**0.783** ± 0.016	0.623	**0.635** ± 0.015
glass-0-1-2-3_vs_4-5-6	0.908	**0.913** ± 0.000	0.832	**0.852** ± 0.000
vehicle0	0.958	**0.961** ± 0.004	0.920	**0.928** ± 0.012

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli1	0.904	**0.904** ± 0.019	0.773	**0.778** ± 0.018
new-thyroid1	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
new-thyroid2	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
ecoli2	**0.910**	0.907 ± 0.010	**0.731**	0.720 ± 0.040
segment0	0.992	**0.994** ± 0.004	0.987	**0.989** ± 0.004
glass6	**0.931**	0.929 ± 0.000	**0.838**	0.831 ± 0.000
yeast3	0.889	**0.896** ± 0.008	**0.678**	0.673 ± 0.013
ecoli3	0.881	**0.886** ± 0.004	**0.600**	0.586 ± 0.014
page-blocks0	0.512	**0.578** ± 0.107	0.182	**0.246** ± 0.098
ecoli-0-3-4_vs_5	0.872	**0.891** ± 0.000	0.675	**0.691** ± 0.000
yeast-2_vs_4	0.886	**0.891** ± 0.000	**0.723**	0.706 ± 0.000
ecoli-0-6-7_vs_3-5	0.837	**0.853** ± 0.000	0.584	**0.596** ± 0.000
ecoli-0-2-3-4_vs_5	0.872	**0.888** ± 0.000	0.702	**0.715** ± 0.000
glass-0-1-5_vs_2	**0.535**	0.479 ± 0.009	**0.158**	0.129 ± 0.003
yeast-0-3-5-9_vs_7-8	0.677	**0.739** ± 0.005	**0.375**	0.358 ± 0.008
yeast-0-2-5-7-9_vs_3-6-8	**0.801**	0.790 ± 0.001	**0.560**	0.520 ± 0.004
yeast-0-2-5-6_vs_3-7-8-9	**0.904**	0.901 ± 0.001	**0.739**	0.698 ± 0.007
ecoli-0-4-6_vs_5	**0.911**	0.904 ± 0.000	**0.809**	0.767 ± 0.000
ecoli-0-1_vs_2-3-5	0.860	**0.863** ± 0.006	**0.668**	0.661 ± 0.027
ecoli-0-2-6-7_vs_3-5	**0.847**	0.844 ± 0.000	**0.662**	0.645 ± 0.000
glass-0-4_vs_5	0.944	**0.969** ± 0.000	**0.893**	0.893 ± 0.000
ecoli-0-3-4-6_vs_5	**0.911**	0.902 ± 0.007	**0.814**	0.774 ± 0.025
ecoli-0-3-4-7_vs_5-6	**0.891**	0.887 ± 0.051	**0.719**	0.682 ± 0.032
yeast-0-5-6-7-9_vs_4	**0.808**	0.785 ± 0.003	**0.484**	0.441 ± 0.005
vowel0	0.961	**0.962** ± 0.001	**0.836**	0.815 ± 0.010
ecoli-0-6-7_vs_5	**0.882**	0.880 ± 0.000	**0.744**	0.732 ± 0.000
glass-0-1-6_vs_2	0.541	**0.587** ± 0.016	0.126	**0.161** ± 0.008
ecoli-0-1-4-7_vs_2-3-5-6	**0.872**	0.858 ± 0.000	**0.669**	0.602 ± 0.000
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.887**	0.883 ± 0.007	**0.693**	0.669 ± 0.063
glass-0-6_vs_5	**0.877**	0.876 ± 0.000	**0.702**	0.694 ± 0.000
ecoli-0-1_vs_5	**0.974**	0.968 ± 0.000	**0.834**	0.833 ± 0.000
glass-0-1-4-6_vs_2	0.606	**0.626** ± 0.020	**0.232**	0.199 ± 0.011
glass2	0.500	**0.622** ± 0.092	0.103	**0.177** ± 0.032
ecoli-0-1-4-7_vs_5-6	**0.890**	0.870 ± 0.008	**0.660**	0.584 ± 0.030
cleveland-0_vs_4	0.794	**0.794** ± 0.000	0.565	**0.565** ± 0.000
ecoli-0-1-4-6_vs_5	**0.896**	0.890 ± 0.005	**0.646**	0.616 ± 0.021
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
yeast-1_vs_7	0.684	**0.739** ± 0.003	**0.285**	0.282 ± 0.004
glass4	**0.925**	0.914 ± 0.007	**0.669**	0.603 ± 0.041
ecoli4	0.934	**0.942** ± 0.000	0.760	**0.769** ± 0.000
page-blocks-1-3_vs_4	**0.649**	0.639 ± 0.003	**0.322**	0.284 ± 0.021
abalone9-18	0.873	**0.889** ± 0.003	**0.665**	0.510 ± 0.016
glass-0-1-6_vs_5	**0.980**	0.978 ± 0.000	**0.726**	0.713 ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	0.623	**0.644** ± 0.002	**0.158**	0.136 ± 0.001
glass5	**0.975**	0.973 ± 0.000	**0.777**	0.751 ± 0.000
yeast-2_vs_8	**0.773**	0.766 ± 0.000	**0.668**	0.556 ± 0.000
yeast4	**0.830**	0.810 ± 0.001	**0.329**	0.287 ± 0.003
yeast-1-2-8-9_vs_7	0.659	**0.701** ± 0.004	**0.167**	0.134 ± 0.003
yeast5	**0.968**	0.965 ± 0.001	**0.492**	0.470 ± 0.009
ecoli-0-1-3-7_vs_2-6	**0.885**	0.820 ± 0.000	**0.540**	0.299 ± 0.000
yeast6	**0.885**	0.879 ± 0.000	**0.318**	0.289 ± 0.003
abalone19	0.753	**0.766** ± 0.001	**0.061**	0.050 ± 0.000
Promedio	0.837	**0.844** ± 0.007	**0.633**	0.617 ± 0.010

Fin de la Tabla A.19.

Tabla A.5: SMOTE-D vs SMOTE usando distancia Euclidiana, K-NN y $k = 3$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.807	**0.824** ± 0.000	0.744	**0.769** ± 0.000
ecoli-0_vs_1	**0.969**	0.962 ± 0.000	**0.974**	0.971 ± 0.000
wisconsin	**0.961**	0.959 ± 0.000	**0.949**	0.947 ± 0.000
pima	**0.657**	0.648 ± 0.010	**0.554**	0.546 ± 0.014
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.788	**0.789** ± 0.012	**0.706**	0.706 ± 0.017
yeast1	0.642	**0.643** ± 0.011	0.493	**0.495** ± 0.015
haberman	0.540	**0.557** ± 0.018	0.346	**0.371** ± 0.022
vehicle2	0.922	**0.923** ± 0.007	0.864	**0.866** ± 0.010
vehicle1	**0.617**	0.607 ± 0.010	**0.437**	0.421 ± 0.014
vehicle3	**0.626**	0.622 ± 0.015	**0.441**	0.435 ± 0.021
glass-0-1-2-3_vs_4-5-6	0.935	**0.944** ± 0.000	0.902	**0.910** ± 0.000
vehicle0	0.915	**0.920** ± 0.006	0.859	**0.864** ± 0.008
ecoli1	**0.843**	0.831 ± 0.028	**0.745**	0.730 ± 0.041

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
new-thyroid1	0.991	**0.995** ± 0.005	0.961	**0.976** ± 0.024
new-thyroid2	**1.000**	0.998 ± 0.000	**1.000**	0.998 ± 0.000
ecoli2	0.900	**0.907** ± 0.007	0.769	**0.798** ± 0.032
segment0	0.989	**0.989** ± 0.000	0.975	**0.976** ± 0.000
glass6	0.931	**0.934** ± 0.006	0.836	**0.863** ± 0.033
yeast3	**0.848**	0.846 ± 0.010	**0.701**	0.699 ± 0.022
ecoli3	0.795	**0.798** ± 0.037	0.587	**0.594** ± 0.039
page-blocks0	0.883	**0.891** ± 0.002	**0.796**	0.791 ± 0.005
ecoli-0-3-4_vs_5	0.866	**0.868** ± 0.000	0.777	**0.793** ± 0.000
yeast-2_vs_4	**0.875**	0.872 ± 0.000	**0.749**	0.746 ± 0.000
ecoli-0-6-7_vs_3-5	0.862	**0.865** ± 0.000	0.692	**0.710** ± 0.000
ecoli-0-2-3-4_vs_5	0.861	**0.863** ± 0.004	0.749	**0.770** ± 0.017
glass-0-1-5_vs_2	**0.662**	0.658 ± 0.013	**0.326**	0.315 ± 0.025
yeast-0-3-5-9_vs_7-8	**0.669**	0.659 ± 0.017	**0.345**	0.330 ± 0.023
yeast-0-2-5-7-9_vs_3-6-8	0.750	**0.753** ± 0.003	0.466	**0.478** ± 0.012
yeast-0-2-5-6_vs_3-7-8-9	0.867	**0.873** ± 0.001	0.714	**0.718** ± 0.010
ecoli-0-4-6_vs_5	0.864	**0.866** ± 0.000	0.742	**0.757** ± 0.000
ecoli-0-1_vs_2-3-5	**0.842**	0.840 ± 0.000	0.678	**0.692** ± 0.000
ecoli-0-2-6-7_vs_3-5	**0.842**	0.825 ± 0.064	**0.707**	0.695 ± 0.080
glass-0-4_vs_5	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-3-4-6_vs_5	0.894	**0.895** ± 0.000	0.838	**0.843** ± 0.000
ecoli-0-3-4-7_vs_5-6	**0.882**	0.869 ± 0.000	**0.739**	0.721 ± 0.000
yeast-0-5-6-7-9_vs_4	0.711	**0.721** ± 0.043	0.435	**0.452** ± 0.058
vowel0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-6-7_vs_5	0.830	**0.832** ± 0.006	0.676	**0.691** ± 0.028
glass-0-1-6_vs_2	0.604	**0.605** ± 0.000	**0.281**	0.280 ± 0.000
ecoli-0-1-4-7_vs_2-3-5-6	**0.848**	0.835 ± 0.003	**0.721**	0.705 ± 0.021
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.848**	0.848 ± 0.000	**0.520**	0.520 ± 0.000
glass-0-6_vs_5	0.868	**0.869** ± 0.000	0.757	**0.768** ± 0.000
ecoli-0-1_vs_5	0.990	**0.990** ± 0.000	0.920	**0.920** ± 0.000
glass-0-1-4-6_vs_2	**0.683**	0.677 ± 0.009	**0.358**	0.351 ± 0.024
glass2	0.644	**0.661** ± 0.012	0.281	**0.302** ± 0.015
ecoli-0-1-4-7_vs_5-6	0.863	**0.866** ± 0.004	0.698	**0.718** ± 0.039
cleveland-0_vs_4	0.535	**0.540** ± 0.014	0.130	**0.149** ± 0.036
ecoli-0-1-4-6_vs_5	0.867	**0.869** ± 0.005	0.760	**0.777** ± 0.043
shuttle-c0-vs-c4	0.995	**0.995** ± 0.000	0.991	**0.995** ± 0.000
yeast-1_vs_7	**0.659**	0.650 ± 0.003	**0.317**	0.293 ± 0.013

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass4	0.937	**0.938** ± 0.000	0.804	**0.815** ± 0.000
ecoli4	**0.920**	0.912 ± 0.000	**0.849**	0.840 ± 0.000
page-blocks-1-3_vs_4	**0.919**	0.918 ± 0.002	**0.790**	0.774 ± 0.028
abalone9-18	**0.816**	0.801 ± 0.036	**0.520**	0.500 ± 0.043
glass-0-1-6_vs_5	0.885	**0.925** ± 0.000	0.693	**0.741** ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	**0.575**	0.551 ± 0.034	**0.154**	0.123 ± 0.037
glass5	**0.890**	0.890 ± 0.000	0.647	**0.647** ± 0.000
yeast-2_vs_8	**0.769**	0.768 ± 0.003	0.417	**0.417** ± 0.014
yeast4	0.687	**0.695** ± 0.021	0.316	**0.329** ± 0.024
yeast-1-2-8-9_vs_7	0.560	**0.561** ± 0.003	0.110	**0.111** ± 0.005
yeast5	0.890	**0.897** ± 0.000	0.718	**0.726** ± 0.000
ecoli-0-1-3-7_vs_2-6	0.839	0.839 ± 0.000	0.513	**0.513** ± 0.000
yeast6	**0.781**	0.781 ± 0.000	0.386	**0.386** ± 0.006
abalone19	**0.557**	0.556 ± 0.000	**0.047**	0.046 ± 0.000
Promedio	0.822	**0.822** ± 0.007	0.651	**0.654** ± 0.014

Fin de la Tabla A.19.

A.3.2. Valor de k igual a 5

Tabla A.6: SMOTE-D vs SMOTE usando distancia Euclidiana, Árboles de decisión y $k = 5$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.735	**0.737** ± 0.057	0.659	**0.665** ± 0.073
ecoli-0_vs_1	0.972	0.972 ± 0.008	**0.982**	0.979 ± 0.008
wisconsin	0.945	0.945 ± 0.010	**0.928**	0.926 ± 0.011
pima	0.647	**0.675** ± 0.038	0.536	**0.580** ± 0.048
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.748	**0.772** ± 0.043	0.657	**0.689** ± 0.054
yeast1	0.651	**0.667** ± 0.024	0.507	**0.530** ± 0.035
haberman	**0.579**	0.578 ± 0.031	**0.396**	0.384 ± 0.042
vehicle2	0.939	**0.947** ± 0.013	0.906	**0.914** ± 0.020
vehicle1	**0.702**	0.678 ± 0.030	**0.555**	0.520 ± 0.041
vehicle3	0.674	**0.689** ± 0.054	0.509	**0.532** ± 0.073

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass-0-1-2-3_vs_4-5-6	0.863	**0.905** ± 0.012	0.801	**0.852** ± 0.034
vehicle0	0.914	0.914 ± 0.017	**0.864**	0.862 ± 0.016
ecoli1	**0.876**	0.848 ± 0.030	**0.789**	0.753 ± 0.033
new-thyroid1	0.943	**0.952** ± 0.066	0.891	**0.921** ± 0.078
new-thyroid2	**0.946**	0.936 ± 0.036	**0.900**	0.890 ± 0.047
ecoli2	0.840	**0.857** ± 0.028	0.697	**0.723** ± 0.049
segment0	**0.993**	0.991 ± 0.002	**0.986**	0.980 ± 0.003
glass6	**0.893**	0.883 ± 0.106	**0.807**	0.770 ± 0.158
yeast3	**0.883**	0.869 ± 0.013	**0.753**	0.735 ± 0.024
ecoli3	0.750	**0.771** ± 0.116	0.517	**0.538** ± 0.125
page-blocks0	0.919	**0.925** ± 0.007	**0.829**	0.824 ± 0.018
ecoli-0-3-4_vs_5	**0.938**	0.863 ± 0.064	**0.870**	0.728 ± 0.071
yeast-2_vs_4	**0.888**	0.852 ± 0.025	**0.734**	0.691 ± 0.034
ecoli-0-6-7_vs_3-5	**0.870**	0.845 ± 0.012	**0.640**	0.617 ± 0.063
ecoli-0-2-3-4_vs_5	**0.883**	0.869 ± 0.059	**0.766**	0.746 ± 0.088
glass-0-1-5_vs_2	0.545	**0.666** ± 0.080	0.182	**0.361** ± 0.136
yeast-0-3-5-9_vs_7-8	**0.669**	0.613 ± 0.051	**0.363**	0.281 ± 0.072
yeast-0-2-5-7-9_vs_3-6-8	**0.745**	0.728 ± 0.022	**0.468**	0.455 ± 0.033
yeast-0-2-5-6_vs_3-7-8-9	**0.895**	0.888 ± 0.026	0.735	**0.739** ± 0.039
ecoli-0-4-6_vs_5	0.861	**0.862** ± 0.069	**0.752**	0.716 ± 0.123
ecoli-0-1_vs_2-3-5	**0.842**	0.804 ± 0.006	**0.665**	0.605 ± 0.024
ecoli-0-2-6-7_vs_3-5	**0.840**	0.821 ± 0.014	**0.624**	0.601 ± 0.076
glass-0-4_vs_5	0.994	0.994 ± 0.000	0.933	0.933 ± 0.000
ecoli-0-3-4-6_vs_5	0.833	**0.849** ± 0.041	0.704	**0.715** ± 0.095
ecoli-0-3-4-7_vs_5-6	**0.885**	0.851 ± 0.010	**0.680**	0.646 ± 0.045
yeast-0-5-6-7-9_vs_4	**0.758**	0.714 ± 0.048	**0.486**	0.438 ± 0.067
vowel0	**0.959**	0.950 ± 0.017	**0.901**	0.887 ± 0.016
ecoli-0-6-7_vs_5	**0.867**	0.863 ± 0.008	0.667	**0.670** ± 0.051
glass-0-1-6_vs_2	0.567	**0.624** ± 0.081	0.171	**0.293** ± 0.126
ecoli-0-1-4-7_vs_2-3-5-6	0.838	**0.860** ± 0.010	0.630	**0.673** ± 0.071
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.890**	0.884 ± 0.037	**0.762**	0.760 ± 0.049
glass-0-6_vs_5	0.884	**0.889** ± 0.010	**0.741**	0.726 ± 0.078
ecoli-0-1_vs_5	0.945	0.945 ± 0.000	0.866	0.866 ± 0.000
glass-0-1-4-6_vs_2	0.611	**0.707** ± 0.000	0.250	**0.422** ± 0.000
glass2	**0.697**	0.650 ± 0.113	**0.407**	0.331 ± 0.097
ecoli-0-1-4-7_vs_5-6	**0.902**	0.871 ± 0.069	**0.738**	0.679 ± 0.116
cleveland-0_vs_4	0.819	**0.863** ± 0.136	0.547	**0.676** ± 0.290

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli-0-1-4-6_vs_5	**0.859**	0.841 ± 0.036	**0.694**	0.596 ± 0.063
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.634	**0.651** ± 0.095	0.252	**0.285** ± 0.135
glass4	**0.896**	0.881 ± 0.000	0.683	**0.754** ± 0.000
ecoli4	0.838	**0.847** ± 0.002	0.680	**0.707** ± 0.033
page-blocks-1-3_vs_4	0.997	0.997 ± 0.000	0.966	0.966 ± 0.000
abalone9-18	**0.797**	0.745 ± 0.054	**0.428**	0.405 ± 0.055
glass-0-1-6_vs_5	0.838	0.838 ± 0.000	0.660	0.660 ± 0.000
shuttle-c2-vs-c4	0.950	0.950 ± 0.000	0.933	0.933 ± 0.000
yeast-1-4-5-8_vs_7	**0.525**	0.523 ± 0.050	0.084	**0.087** ± 0.046
glass5	0.897	0.897 ± 0.000	0.760	0.760 ± 0.000
yeast-2_vs_8	**0.849**	0.768 ± 0.082	**0.482**	0.427 ± 0.178
yeast4	**0.720**	0.696 ± 0.041	**0.349**	0.333 ± 0.053
yeast-1-2-8-9_vs_7	0.562	**0.592** ± 0.051	0.124	**0.159** ± 0.066
yeast5	**0.901**	0.887 ± 0.041	**0.717**	0.716 ± 0.055
ecoli-0-1-3-7_vs_2-6	**0.844**	0.838 ± 0.011	**0.593**	0.490 ± 0.128
yeast6	0.751	**0.771** ± 0.006	0.320	**0.378** ± 0.032
abalone19	**0.570**	0.563 ± 0.033	**0.051**	0.048 ± 0.014
Promedio	**0.822**	0.820 ± 0.034	0.644	0.644 ± 0.056

Fin de la Tabla A.6.

Tabla A.7: SMOTE-D vs SMOTE usando distancia Euclidiana, SVM y $k = 5$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.588	**0.622** ± 0.000	0.569	0.569 ± 0.000
ecoli-0_vs_1	0.969	**0.979** ± 0.000	0.974	**0.986** ± 0.000
wisconsin	**0.974**	0.968 ± 0.006	**0.959**	0.956 ± 0.006
pima	0.704	**0.742** ± 0.005	0.642	**0.666** ± 0.005
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.725	**0.749** ± 0.000	0.639	**0.658** ± 0.000
yeast1	0.647	**0.707** ± 0.002	0.533	**0.582** ± 0.003
haberman	0.500	**0.624** ± 0.018	0.418	**0.437** ± 0.034
vehicle2	0.945	**0.953** ± 0.000	0.894	**0.916** ± 0.000
vehicle1	0.784	**0.809** ± 0.005	0.631	**0.670** ± 0.006
vehicle3	0.662	**0.771** ± 0.030	0.500	**0.624** ± 0.031
glass-0-1-2-3_vs_4-5-6	0.888	**0.905** ± 0.015	0.810	**0.839** ± 0.037

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
vehicle0	**0.963**	0.960 ± 0.013	0.923	**0.925** ± 0.025
ecoli1	0.900	**0.906** ± 0.000	0.764	**0.778** ± 0.000
new-thyroid1	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
new-thyroid2	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
ecoli2	0.889	**0.906** ± 0.018	0.668	**0.721** ± 0.010
segment0	**0.995**	0.994 ± 0.004	**0.991**	0.989 ± 0.004
glass6	**0.928**	0.927 ± 0.007	**0.825**	0.824 ± 0.031
yeast3	**0.900**	0.895 ± 0.001	0.673	**0.676** ± 0.004
ecoli3	0.890	**0.891** ± 0.000	**0.601**	0.590 ± 0.000
page-blocks0	0.502	**0.551** ± 0.005	0.201	**0.216** ± 0.014
ecoli-0-3-4_vs_5	0.863	**0.891** ± 0.000	0.638	**0.691** ± 0.000
yeast-2_vs_4	0.881	**0.891** ± 0.000	0.694	**0.708** ± 0.000
ecoli-0-6-7_vs_3-5	**0.855**	0.853 ± 0.000	0.596	0.596 ± 0.000
ecoli-0-2-3-4_vs_5	**0.894**	0.887 ± 0.000	0.706	0.706 ± 0.000
glass-0-1-5_vs_2	0.469	**0.506** ± 0.000	0.108	**0.164** ± 0.000
yeast-0-3-5-9_vs_7-8	0.731	**0.738** ± 0.003	**0.379**	0.359 ± 0.004
yeast-0-2-5-7-9_vs_3-6-8	**0.801**	0.791 ± 0.000	**0.548**	0.521 ± 0.000
yeast-0-2-5-6_vs_3-7-8-9	**0.905**	0.902 ± 0.003	**0.724**	0.703 ± 0.015
ecoli-0-4-6_vs_5	**0.911**	0.906 ± 0.000	**0.809**	0.777 ± 0.000
ecoli-0-1_vs_2-3-5	**0.884**	0.875 ± 0.011	**0.712**	0.693 ± 0.057
ecoli-0-2-6-7_vs_3-5	**0.845**	0.843 ± 0.007	**0.649**	0.641 ± 0.041
glass-0-4_vs_5	**0.944**	0.939 ± 0.000	**0.893**	0.875 ± 0.000
ecoli-0-3-4-6_vs_5	0.900	**0.905** ± 0.013	0.769	**0.784** ± 0.056
ecoli-0-3-4-7_vs_5-6	0.881	**0.893** ± 0.075	0.673	**0.680** ± 0.111
yeast-0-5-6-7-9_vs_4	**0.800**	0.794 ± 0.006	**0.462**	0.445 ± 0.012
vowel0	**0.959**	0.955 ± 0.002	**0.830**	0.800 ± 0.018
ecoli-0-6-7_vs_5	0.877	**0.878** ± 0.000	0.713	**0.718** ± 0.000
glass-0-1-6_vs_2	0.499	**0.560** ± 0.037	0.093	**0.160** ± 0.016
ecoli-0-1-4-7_vs_2-3-5-6	**0.867**	0.865 ± 0.004	**0.642**	0.613 ± 0.021
led7digit-0-2-4-5-6-7-8-9_vs_1	0.867	**0.883** ± 0.000	0.628	**0.672** ± 0.000
glass-0-6_vs_5	0.877	**0.878** ± 0.006	0.702	**0.708** ± 0.054
ecoli-0-1_vs_5	**0.989**	0.970 ± 0.000	**0.920**	0.833 ± 0.000
glass-0-1-4-6_vs_2	0.518	**0.617** ± 0.030	0.142	**0.194** ± 0.017
glass2	0.540	**0.609** ± 0.088	0.140	**0.177** ± 0.030
ecoli-0-1-4-7_vs_5-6	**0.885**	0.875 ± 0.000	**0.629**	0.609 ± 0.000
cleveland-0_vs_4	0.794	0.794 ± 0.000	0.565	0.565 ± 0.000
ecoli-0-1-4-6_vs_5	**0.890**	0.889 ± 0.011	**0.617**	0.608 ± 0.047

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.741	**0.749** ± 0.005	**0.316**	0.282 ± 0.007
glass4	**0.917**	0.911 ± 0.014	**0.618**	0.591 ± 0.073
ecoli4	0.934	**0.943** ± 0.000	0.760	**0.781** ± 0.000
page-blocks-1-3_vs_4	**0.730**	0.723 ± 0.011	**0.352**	0.329 ± 0.169
abalone9-18	**0.883**	0.881 ± 0.000	**0.580**	0.499 ± 0.000
glass-0-1-6_vs_5	0.977	**0.980** ± 0.008	0.686	**0.724** ± 0.096
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	0.634	**0.635** ± 0.003	**0.145**	0.131 ± 0.002
glass5	**0.978**	0.974 ± 0.000	**0.788**	0.764 ± 0.000
yeast-2_vs_8	**0.772**	0.764 ± 0.000	**0.649**	0.542 ± 0.000
yeast4	**0.813**	0.810 ± 0.001	**0.298**	0.288 ± 0.003
yeast-1-2-8-9_vs_7	0.695	**0.704** ± 0.004	**0.151**	0.136 ± 0.003
yeast5	**0.967**	0.965 ± 0.002	**0.481**	0.474 ± 0.019
ecoli-0-1-3-7_vs_2-6	**0.826**	0.819 ± 0.000	**0.340**	0.292 ± 0.000
yeast6	0.877	**0.882** ± 0.000	0.279	**0.302** ± 0.000
abalone19	0.748	**0.781** ± 0.002	0.052	0.052 ± 0.000
Promedio	0.832	**0.844** ± 0.007	0.616	**0.617** ± 0.016

Fin de la Tabla A.7.

A.3.3. Valor de k igual a 7

Tabla A.8: SMOTE-D vs SMOTE usando distancia Euclidiana, Árboles de Decisión y $k = 7$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.734**	0.717 ± 0.044	**0.654**	0.635 ± 0.057
ecoli-0_vs_1	0.966	**0.968** ± 0.009	**0.975**	0.974 ± 0.009
wisconsin	0.937	**0.943** ± 0.007	0.921	**0.924** ± 0.007
pima	0.672	**0.685** ± 0.029	0.565	**0.594** ± 0.038
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.766	**0.780** ± 0.090	0.685	**0.694** ± 0.123
yeast1	0.658	**0.668** ± 0.037	0.515	**0.531** ± 0.052
haberman	0.531	**0.582** ± 0.031	0.311	**0.390** ± 0.045
vehicle2	0.934	**0.948** ± 0.015	0.912	**0.917** ± 0.023

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
vehicle1	0.688	**0.694** ± 0.030	0.533	**0.542** ± 0.045
vehicle3	0.675	**0.698** ± 0.040	0.509	**0.543** ± 0.048
glass-0-1-2-3_vs_4-5-6	0.891	**0.899** ± 0.030	**0.846**	0.835 ± 0.056
vehicle0	0.913	**0.914** ± 0.015	**0.866**	0.863 ± 0.016
ecoli1	0.845	**0.849** ± 0.019	**0.760**	0.747 ± 0.026
new-thyroid1	0.943	**0.955** ± 0.050	0.891	**0.922** ± 0.063
new-thyroid2	**0.968**	0.955 ± 0.030	**0.955**	0.922 ± 0.023
ecoli2	0.847	**0.864** ± 0.021	0.701	**0.733** ± 0.047
segment0	**0.993**	0.991 ± 0.003	**0.984**	0.981 ± 0.004
glass6	**0.880**	0.861 ± 0.103	**0.747**	0.742 ± 0.164
yeast3	**0.889**	0.868 ± 0.015	**0.741**	0.728 ± 0.031
ecoli3	0.769	**0.775** ± 0.068	0.512	**0.535** ± 0.099
page-blocks0	0.921	**0.929** ± 0.009	**0.835**	0.823 ± 0.014
ecoli-0-3-4_vs_5	**0.913**	0.877 ± 0.056	**0.833**	0.723 ± 0.097
yeast-2_vs_4	**0.909**	0.867 ± 0.025	**0.769**	0.706 ± 0.031
ecoli-0-6-7_vs_3-5	0.837	**0.843** ± 0.012	0.580	**0.615** ± 0.057
ecoli-0-2-3-4_vs_5	**0.878**	0.876 ± 0.061	0.722	**0.744** ± 0.079
glass-0-1-5_vs_2	0.625	**0.649** ± 0.128	0.293	**0.329** ± 0.225
yeast-0-3-5-9_vs_7-8	**0.642**	0.635 ± 0.054	0.307	**0.311** ± 0.076
yeast-0-2-5-7-9_vs_3-6-8	0.705	**0.737** ± 0.021	0.411	**0.472** ± 0.032
yeast-0-2-5-6_vs_3-7-8-9	0.888	**0.893** ± 0.020	0.719	**0.756** ± 0.035
ecoli-0-4-6_vs_5	0.853	**0.862** ± 0.062	0.708	**0.716** ± 0.090
ecoli-0-1_vs_2-3-5	**0.844**	0.813 ± 0.003	**0.677**	0.603 ± 0.015
ecoli-0-2-6-7_vs_3-5	**0.840**	0.808 ± 0.009	**0.622**	0.588 ± 0.047
glass-0-4_vs_5	**0.994**	0.994 ± 0.000	**0.933**	0.933 ± 0.000
ecoli-0-3-4-6_vs_5	0.839	**0.864** ± 0.056	0.740	**0.745** ± 0.070
ecoli-0-3-4-7_vs_5-6	0.820	**0.853** ± 0.008	0.636	**0.637** ± 0.035
yeast-0-5-6-7-9_vs_4	**0.773**	0.723 ± 0.036	**0.483**	0.449 ± 0.053
vowel0	0.949	**0.963** ± 0.013	0.860	**0.893** ± 0.027
ecoli-0-6-7_vs_5	**0.867**	0.866 ± 0.035	**0.673**	0.667 ± 0.063
glass-0-1-6_vs_2	0.498	**0.630** ± 0.092	0.114	**0.298** ± 0.139
ecoli-0-1-4-7_vs_2-3-5-6	**0.865**	0.864 ± 0.009	**0.712**	0.673 ± 0.057
led7digit-0-2-4-5-6-7-8-9_vs_1	0.877	**0.885** ± 0.037	0.750	**0.772** ± 0.047
glass-0-6_vs_5	**0.881**	0.871 ± 0.009	0.724	**0.734** ± 0.072
ecoli-0-1_vs_5	0.945	**0.945** ± 0.000	**0.866**	0.866 ± 0.000
glass-0-1-4-6_vs_2	0.613	**0.687** ± 0.116	0.230	**0.385** ± 0.145
glass2	**0.723**	0.659 ± 0.142	**0.442**	0.364 ± 0.145

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli-0-1-4-7_vs_5-6	0.858	**0.863** ± 0.016	**0.673**	0.663 ± 0.115
cleveland-0_vs_4	0.791	**0.838** ± 0.117	0.541	**0.642** ± 0.245
ecoli-0-1-4-6_vs_5	**0.855**	0.845 ± 0.078	**0.655**	0.614 ± 0.110
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.667	**0.668** ± 0.039	0.309	**0.309** ± 0.055
glass4	**0.894**	0.870 ± 0.018	0.666	**0.695** ± 0.207
ecoli4	**0.867**	0.844 ± 0.003	**0.759**	0.699 ± 0.051
page-blocks-1-3_vs_4	0.997	**0.997** ± 0.000	0.966	**0.966** ± 0.000
abalone9-18	**0.816**	0.761 ± 0.060	**0.452**	0.416 ± 0.061
glass-0-1-6_vs_5	0.838	**0.838** ± 0.000	0.660	**0.660** ± 0.000
shuttle-c2-vs-c4	0.950	0.950 ± 0.000	0.933	**0.933** ± 0.000
yeast-1-4-5-8_vs_7	**0.574**	0.524 ± 0.082	**0.142**	0.087 ± 0.073
glass5	**0.897**	0.897 ± 0.000	0.760	**0.760** ± 0.000
yeast-2_vs_8	**0.770**	0.744 ± 0.038	**0.364**	0.360 ± 0.079
yeast4	0.691	**0.693** ± 0.058	0.308	**0.318** ± 0.072
yeast-1-2-8-9_vs_7	0.594	**0.603** ± 0.068	0.135	**0.166** ± 0.085
yeast5	0.879	**0.892** ± 0.045	0.681	**0.710** ± 0.044
ecoli-0-1-3-7_vs_2-6	**0.844**	0.837 ± 0.009	**0.593**	0.475 ± 0.190
yeast6	0.750	**0.767** ± 0.005	0.341	**0.373** ± 0.023
abalone19	**0.585**	0.560 ± 0.026	**0.056**	0.041 ± 0.012
Promedio	0.820	**0.821** ± 0.036	0.640	**0.643** ± 0.061

Fin de la Tabla A.19.

Tabla A.9: SMOTE-D vs SMOTE usando distancia Euclidiana, SVM y $k = 7$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.591	**0.638** ± 0.000	0.572	**0.581** ± 0.000
ecoli-0_vs_1	0.965	**0.980** ± 0.000	0.971	**0.987** ± 0.000
wisconsin	**0.973**	0.967 ± 0.006	**0.957**	0.955 ± 0.006
pima	0.634	**0.737** ± 0.019	0.591	**0.661** ± 0.022
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.718	**0.751** ± 0.000	0.633	**0.660** ± 0.000
yeast1	0.608	**0.705** ± 0.006	0.508	**0.580** ± 0.007
haberman	0.500	**0.625** ± 0.000	0.418	**0.441** ± 0.000
vehicle2	0.943	**0.956** ± 0.000	0.882	**0.922** ± 0.000
vehicle1	0.745	**0.789** ± 0.031	0.594	**0.649** ± 0.032

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
vehicle3	0.718	**0.778** ± 0.004	0.554	**0.629** ± 0.002
glass-0-1-2-3_vs_4-5-6	0.888	**0.914** ± 0.000	0.811	**0.855** ± 0.000
vehicle0	**0.966**	0.960 ± 0.003	**0.934**	0.926 ± 0.011
ecoli1	0.896	**0.906** ± 0.000	0.756	**0.780** ± 0.000
new-thyroid1	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
new-thyroid2	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
ecoli2	0.892	**0.902** ± 0.033	0.677	**0.718** ± 0.048
segment0	**0.995**	0.994 ± 0.004	0.988	**0.988** ± 0.004
glass6	0.928	**0.931** ± 0.015	0.825	**0.841** ± 0.080
yeast3	**0.902**	0.892 ± 0.001	0.672	**0.673** ± 0.004
ecoli3	**0.887**	0.885 ± 0.000	**0.590**	0.585 ± 0.000
page-blocks0	0.522	**0.549** ± 0.037	0.177	**0.219** ± 0.047
ecoli-0-3-4_vs_5	0.866	**0.916** ± 0.008	0.647	**0.733** ± 0.051
yeast-2_vs_4	**0.898**	0.890 ± 0.003	0.700	**0.704** ± 0.015
ecoli-0-6-7_vs_3-5	**0.855**	0.851 ± 0.000	**0.596**	0.586 ± 0.000
ecoli-0-2-3-4_vs_5	0.891	**0.894** ± 0.013	0.692	**0.712** ± 0.050
glass-0-1-5_vs_2	0.471	**0.490** ± 0.000	0.130	**0.157** ± 0.000
yeast-0-3-5-9_vs_7-8	**0.750**	0.743 ± 0.006	**0.377**	0.364 ± 0.009
yeast-0-2-5-7-9_vs_3-6-8	**0.802**	0.796 ± 0.016	**0.537**	0.522 ± 0.021
yeast-0-2-5-6_vs_3-7-8-9	**0.906**	0.901 ± 0.001	**0.708**	0.698 ± 0.007
ecoli-0-4-6_vs_5	**0.900**	0.894 ± 0.080	**0.750**	0.743 ± 0.144
ecoli-0-1_vs_2-3-5	**0.860**	0.859 ± 0.013	**0.668**	0.667 ± 0.065
ecoli-0-2-6-7_vs_3-5	0.842	**0.843** ± 0.007	0.638	**0.642** ± 0.041
glass-0-4_vs_5	0.944	**0.952** ± 0.000	**0.893**	0.875 ± 0.000
ecoli-0-3-4-6_vs_5	0.903	**0.906** ± 0.023	0.777	**0.790** ± 0.106
ecoli-0-3-4-7_vs_5-6	0.874	**0.899** ± 0.025	0.645	**0.689** ± 0.081
yeast-0-5-6-7-9_vs_4	0.791	**0.797** ± 0.008	0.441	**0.454** ± 0.018
vowel0	**0.953**	0.953 ± 0.016	**0.819**	0.809 ± 0.016
ecoli-0-6-7_vs_5	**0.877**	0.875 ± 0.000	**0.713**	0.704 ± 0.000
glass-0-1-6_vs_2	0.517	**0.570** ± 0.016	0.104	**0.167** ± 0.008
ecoli-0-1-4-7_vs_2-3-5-6	**0.867**	0.866 ± 0.009	**0.642**	0.620 ± 0.040
led7digit-0-2-4-5-6-7-8-9_vs_1	0.863	**0.894** ± 0.000	0.606	**0.728** ± 0.000
glass-0-6_vs_5	**0.879**	0.869 ± 0.006	**0.721**	0.697 ± 0.054
ecoli-0-1_vs_5	**0.990**	0.967 ± 0.000	**0.920**	0.806 ± 0.000
glass-0-1-4-6_vs_2	0.566	**0.609** ± 0.013	0.185	**0.189** ± 0.008
glass2	0.621	**0.630** ± 0.000	0.168	**0.185** ± 0.000
ecoli-0-1-4-7_vs_5-6	**0.884**	0.862 ± 0.004	**0.628**	0.570 ± 0.016

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
cleveland-0_vs_4	0.794	**0.794** ± 0.000	0.565	**0.565** ± 0.000
ecoli-0-1-4-6_vs_5	**0.894**	0.889 ± 0.005	**0.636**	0.611 ± 0.021
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.729	**0.741** ± 0.045	**0.297**	0.280 ± 0.024
glass4	**0.912**	0.911 ± 0.007	**0.593**	0.590 ± 0.032
ecoli4	**0.935**	0.934 ± 0.000	**0.777**	0.760 ± 0.000
page-blocks-1-3_vs_4	**0.745**	0.742 ± 0.011	**0.440**	0.348 ± 0.198
abalone9-18	**0.899**	0.890 ± 0.002	**0.554**	0.510 ± 0.009
glass-0-1-6_vs_5	**0.988**	0.981 ± 0.008	**0.826**	0.748 ± 0.096
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	**0.655**	0.634 ± 0.002	**0.148**	0.131 ± 0.001
glass5	**0.982**	0.974 ± 0.007	**0.814**	0.755 ± 0.192
yeast-2_vs_8	0.762	**0.765** ± 0.000	0.514	**0.545** ± 0.000
yeast4	0.806	**0.810** ± 0.000	0.276	**0.286** ± 0.000
yeast-1-2-8-9_vs_7	0.687	**0.698** ± 0.003	**0.142**	0.135 ± 0.002
yeast5	**0.967**	0.965 ± 0.001	**0.485**	0.474 ± 0.009
ecoli-0-1-3-7_vs_2-6	**0.826**	0.819 ± 0.005	**0.340**	0.296 ± 0.027
yeast6	0.877	**0.882** ± 0.000	0.279	**0.303** ± 0.000
abalone19	0.756	**0.765** ± 0.001	**0.052**	0.050 ± 0.000
Promedio	0.833	**0.844** ± 0.007	0.613	**0.618** ± 0.024

Fin de la Tabla A.19.

Tabla A.10: SMOTE-D vs SMOTE usando distancia Euclidiana, K-NN y $k = 7$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.823**	0.816 ± 0.015	**0.765**	0.756 ± 0.021
ecoli-0_vs_1	**0.969**	0.962 ± 0.000	**0.974**	0.970 ± 0.000
wisconsin	0.957	**0.959** ± 0.000	0.945	**0.947** ± 0.000
pima	0.650	**0.653** ± 0.028	0.542	**0.555** ± 0.036
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.781	**0.794** ± 0.014	0.696	**0.712** ± 0.019
yeast1	0.638	**0.650** ± 0.006	0.486	**0.506** ± 0.008
haberman	0.551	**0.553** ± 0.020	0.356	**0.369** ± 0.022
vehicle2	0.923	**0.924** ± 0.012	**0.865**	0.864 ± 0.018
vehicle1	**0.624**	0.621 ± 0.018	**0.444**	0.443 ± 0.024
vehicle3	0.626	**0.640** ± 0.012	0.439	**0.464** ± 0.015

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass-0-1-2-3_vs_4-5-6	0.929	**0.940** ± 0.006	0.885	**0.905** ± 0.018
vehicle0	0.917	**0.925** ± 0.002	0.865	**0.866** ± 0.005
ecoli1	0.822	**0.843** ± 0.021	0.716	**0.740** ± 0.027
new-thyroid1	0.991	**0.993** ± 0.005	0.961	**0.971** ± 0.024
new-thyroid2	0.997	**0.997** ± 0.004	0.986	**0.986** ± 0.021
ecoli2	**0.912**	0.907 ± 0.007	**0.797**	0.791 ± 0.027
segment0	0.989	**0.990** ± 0.000	0.975	**0.976** ± 0.000
glass6	**0.936**	0.920 ± 0.005	**0.865**	0.844 ± 0.027
yeast3	0.850	**0.853** ± 0.015	0.669	**0.693** ± 0.030
ecoli3	**0.834**	0.825 ± 0.046	**0.626**	0.615 ± 0.049
page-blocks0	0.885	**0.898** ± 0.005	**0.791**	0.787 ± 0.008
ecoli-0-3-4_vs_5	0.861	**0.871** ± 0.052	0.742	**0.785** ± 0.075
yeast-2_vs_4	**0.880**	0.880 ± 0.002	0.733	**0.740** ± 0.017
ecoli-0-6-7_vs_3-5	0.860	**0.863** ± 0.000	0.674	**0.693** ± 0.000
ecoli-0-2-3-4_vs_5	0.861	**0.869** ± 0.005	0.749	**0.779** ± 0.023
glass-0-1-5_vs_2	0.609	**0.652** ± 0.017	0.237	**0.303** ± 0.031
yeast-0-3-5-9_vs_7-8	0.673	**0.686** ± 0.015	0.334	**0.352** ± 0.022
yeast-0-2-5-7-9_vs_3-6-8	0.757	**0.759** ± 0.011	0.461	**0.472** ± 0.012
yeast-0-2-5-6_vs_3-7-8-9	0.870	**0.876** ± 0.002	0.666	**0.698** ± 0.013
ecoli-0-4-6_vs_5	0.864	**0.870** ± 0.000	0.746	**0.761** ± 0.000
ecoli-0-1_vs_2-3-5	**0.846**	0.839 ± 0.000	**0.700**	0.685 ± 0.000
ecoli-0-2-6-7_vs_3-5	**0.860**	0.851 ± 0.057	0.724	**0.731** ± 0.066
glass-0-4_vs_5	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-3-4-6_vs_5	0.891	**0.894** ± 0.006	0.816	**0.834** ± 0.057
ecoli-0-3-4-7_vs_5-6	**0.882**	0.882 ± 0.000	**0.739**	0.736 ± 0.000
yeast-0-5-6-7-9_vs_4	**0.745**	0.739 ± 0.030	0.445	**0.463** ± 0.043
vowel0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-6-7_vs_5	0.832	**0.837** ± 0.007	0.688	**0.708** ± 0.035
glass-0-1-6_vs_2	**0.634**	0.594 ± 0.007	**0.324**	0.249 ± 0.014
ecoli-0-1-4-7_vs_2-3-5-6	0.843	**0.849** ± 0.005	0.696	**0.700** ± 0.035
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.848**	0.848 ± 0.000	**0.520**	0.520 ± 0.000
glass-0-6_vs_5	0.865	**0.867** ± 0.000	0.741	**0.756** ± 0.000
ecoli-0-1_vs_5	0.990	**0.990** ± 0.000	0.920	**0.920** ± 0.000
glass-0-1-4-6_vs_2	0.628	**0.662** ± 0.006	0.291	**0.331** ± 0.015
glass2	0.662	**0.670** ± 0.000	0.311	**0.314** ± 0.000
ecoli-0-1-4-7_vs_5-6	0.885	**0.889** ± 0.004	0.739	**0.742** ± 0.039
cleveland-0_vs_4	0.547	**0.559** ± 0.016	0.144	**0.155** ± 0.030

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli-0-1-4-6_vs_5	**0.892**	0.881 ± 0.005	**0.800**	0.793 ± 0.043
shuttle-c0-vs-c4	**0.996**	0.995 ± 0.000	**0.995**	0.995 ± 0.000
yeast-1_vs_7	0.649	**0.658** ± 0.005	0.261	**0.281** ± 0.015
glass4	0.932	**0.935** ± 0.006	0.748	**0.781** ± 0.103
ecoli4	**0.917**	0.914 ± 0.000	0.814	**0.815** ± 0.000
page-blocks-1-3_vs_4	0.936	**0.937** ± 0.002	0.777	**0.786** ± 0.028
abalone9-18	**0.878**	0.841 ± 0.053	**0.518**	0.495 ± 0.063
glass-0-1-6_vs_5	**0.935**	0.920 ± 0.000	**0.753**	0.735 ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	**0.607**	0.592 ± 0.070	**0.152**	0.146 ± 0.053
glass5	**0.890**	0.890 ± 0.000	0.647	**0.647** ± 0.000
yeast-2_vs_8	0.746	**0.749** ± 0.003	0.300	**0.319** ± 0.013
yeast4	**0.730**	0.727 ± 0.021	0.301	**0.321** ± 0.025
yeast-1-2-8-9_vs_7	**0.602**	0.590 ± 0.043	**0.125**	0.125 ± 0.033
yeast5	0.898	**0.905** ± 0.001	0.679	**0.687** ± 0.038
ecoli-0-1-3-7_vs_2-6	0.839	0.839 ± 0.000	0.513	**0.513** ± 0.000
yeast6	0.787	**0.790** ± 0.001	0.325	**0.353** ± 0.010
abalone19	0.589	**0.591** ± 0.001	0.047	**0.050** ± 0.000
Promedio	0.828	**0.829** ± 0.010	0.645	**0.652** ± 0.020

Fin de la Tabla A.19.

A.4. Utilizando la medida HVDM

A.4.1. Valor de k igual a 3

Tabla A.11: SMOTE-D vs SMOTE usando medida HVDM, Árboles de Decisión y $k = 3$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.764**	0.747 ± 0.060	**0.698**	0.677 ± 0.070
ecoli-0_vs_1	0.972	**0.973** ± 0.009	0.978	**0.981** ± 0.009
wisconsin	0.940	**0.947** ± 0.010	0.923	**0.928** ± 0.010
pima	**0.686**	0.684 ± 0.023	**0.594**	0.593 ± 0.031

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.767	**0.788** ± 0.078	0.676	**0.710** ± 0.094
yeast1	0.647	**0.676** ± 0.012	0.503	**0.543** ± 0.017
haberman	**0.609**	0.583 ± 0.032	**0.439**	0.398 ± 0.049
vehicle2	0.950	**0.951** ± 0.011	**0.934**	0.918 ± 0.019
vehicle1	**0.703**	0.688 ± 0.050	**0.554**	0.532 ± 0.070
vehicle3	0.661	**0.702** ± 0.040	0.489	**0.551** ± 0.051
glass-0-1-2-3_vs_4-5-6	0.879	**0.894** ± 0.013	0.807	**0.837** ± 0.033
vehicle0	**0.925**	0.913 ± 0.020	**0.881**	0.860 ± 0.028
ecoli1	**0.879**	0.857 ± 0.037	**0.796**	0.763 ± 0.052
new-thyroid1	0.948	**0.955** ± 0.033	**0.914**	0.913 ± 0.046
new-thyroid2	**0.963**	0.947 ± 0.024	**0.929**	0.904 ± 0.004
ecoli2	0.842	**0.850** ± 0.028	0.714	**0.715** ± 0.052
segment0	**0.991**	0.991 ± 0.005	**0.981**	0.980 ± 0.005
glass6	0.880	**0.880** ± 0.064	0.751	**0.764** ± 0.121
yeast3	0.839	**0.879** ± 0.020	0.695	**0.742** ± 0.031
ecoli3	**0.775**	0.762 ± 0.059	**0.570**	0.517 ± 0.068
page-blocks0	0.923	**0.926** ± 0.004	**0.820**	0.800 ± 0.011
ecoli-0-3-4_vs_5	0.891	**0.895** ± 0.059	0.694	**0.745** ± 0.121
yeast-2_vs_4	**0.850**	0.841 ± 0.024	**0.689**	0.669 ± 0.035
ecoli-0-6-7_vs_3-5	**0.847**	0.843 ± 0.036	0.613	**0.620** ± 0.038
ecoli-0-2-3-4_vs_5	0.836	**0.869** ± 0.058	0.711	**0.726** ± 0.111
glass-0-1-5_vs_2	0.607	**0.611** ± 0.047	**0.287**	0.282 ± 0.090
yeast-0-3-5-9_vs_7-8	0.653	**0.659** ± 0.050	0.328	**0.341** ± 0.065
yeast-0-2-5-7-9_vs_3-6-8	0.706	**0.733** ± 0.028	0.422	**0.452** ± 0.038
yeast-0-2-5-6_vs_3-7-8-9	0.881	**0.885** ± 0.021	0.702	**0.715** ± 0.052
ecoli-0-4-6_vs_5	**0.855**	0.855 ± 0.055	**0.707**	0.687 ± 0.139
ecoli-0-1_vs_2-3-5	0.826	**0.841** ± 0.012	0.585	**0.629** ± 0.041
ecoli-0-2-6-7_vs_3-5	**0.847**	0.816 ± 0.043	**0.662**	0.609 ± 0.084
glass-0-4_vs_5	**0.994**	0.994 ± 0.000	**0.933**	0.933 ± 0.000
ecoli-0-3-4-6_vs_5	**0.856**	0.855 ± 0.019	**0.721**	0.697 ± 0.085
ecoli-0-3-4-7_vs_5-6	**0.904**	0.863 ± 0.011	**0.791**	0.697 ± 0.060
yeast-0-5-6-7-9_vs_4	**0.746**	0.706 ± 0.056	**0.483**	0.415 ± 0.084
vowel0	0.952	**0.959** ± 0.013	**0.883**	0.879 ± 0.033
ecoli-0-6-7_vs_5	**0.882**	0.861 ± 0.013	**0.759**	0.669 ± 0.083
glass-0-1-6_vs_2	0.578	**0.605** ± 0.006	**0.260**	0.259 ± 0.000
ecoli-0-1-4-7_vs_2-3-5-6	0.850	**0.859** ± 0.031	0.639	**0.687** ± 0.082

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.890**	0.877 ± 0.003	**0.770**	0.751 ± 0.029
glass-0-6_vs_5	**0.909**	0.897 ± 0.036	**0.779**	0.773 ± 0.070
ecoli-0-1_vs_5	0.945	**0.945** ± 0.000	**0.866**	0.866 ± 0.000
glass-0-1-4-6_vs_2	**0.701**	0.658 ± 0.080	**0.356**	0.344 ± 0.120
glass2	**0.697**	0.680 ± 0.099	**0.447**	0.381 ± 0.085
ecoli-0-1-4-7_vs_5-6	0.830	**0.900** ± 0.013	0.697	**0.746** ± 0.120
cleveland-0_vs_4	0.775	**0.823** ± 0.126	0.547	**0.674** ± 0.243
ecoli-0-1-4-6_vs_5	**0.848**	0.833 ± 0.008	0.610	**0.615** ± 0.049
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.634	**0.639** ± 0.090	**0.310**	0.283 ± 0.130
glass4	0.863	**0.886** ± 0.008	0.647	**0.696** ± 0.110
ecoli4	0.865	**0.873** ± 0.000	0.731	**0.754** ± 0.000
page-blocks-1-3_vs_4	0.997	**0.997** ± 0.000	0.966	**0.966** ± 0.000
abalone9-18	**0.723**	0.702 ± 0.059	**0.418**	0.378 ± 0.066
glass-0-1-6_vs_5	0.838	**0.838** ± 0.000	0.660	**0.660** ± 0.000
shuttle-c2-vs-c4	0.950	0.950 ± 0.000	0.933	**0.933** ± 0.000
yeast-1-4-5-8_vs_7	0.544	**0.545** ± 0.067	0.111	**0.118** ± 0.067
glass5	**0.897**	0.897 ± 0.000	0.760	**0.760** ± 0.000
yeast-2_vs_8	**0.759**	0.735 ± 0.065	**0.480**	0.434 ± 0.152
yeast4	0.700	**0.718** ± 0.032	0.309	**0.336** ± 0.031
yeast-1-2-8-9_vs_7	**0.636**	0.568 ± 0.042	**0.216**	0.133 ± 0.085
yeast5	**0.890**	0.887 ± 0.029	0.706	**0.710** ± 0.050
ecoli-0-1-3-7_vs_2-6	**0.844**	0.837 ± 0.006	**0.593**	0.489 ± 0.094
yeast6	**0.766**	0.754 ± 0.032	0.355	**0.393** ± 0.068
abalone19	0.493	**0.499** ± 0.001	0.000	**0.007** ± 0.000
Promedio	**0.820**	0.819 ± 0.030	**0.648**	0.644 ± 0.056

Fin de la Tabla A.19.

Tabla A.12: SMOTE-D vs SMOTE usando medida HVDM, SVM y $k = 3$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.603	**0.632** ± 0.000	**0.578**	0.577 ± 0.000
ecoli-0_vs_1	0.969	**0.979** ± 0.000	0.974	**0.986** ± 0.000
wisconsin	**0.971**	0.966 ± 0.000	**0.957**	0.953 ± 0.000
pima	0.736	**0.745** ± 0.013	0.667	**0.670** ± 0.017
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass0	0.721	**0.749** ± 0.000	0.636	**0.658** ± 0.000
yeast1	0.681	**0.709** ± 0.006	0.558	**0.585** ± 0.007
haberman	0.556	**0.627** ± 0.026	0.435	**0.444** ± 0.044
vehicle2	0.949	**0.955** ± 0.000	0.904	**0.921** ± 0.000
vehicle1	0.810	**0.811** ± 0.022	0.665	**0.673** ± 0.025
vehicle3	**0.787**	0.781 ± 0.017	0.628	**0.632** ± 0.018
glass-0-1-2-3_vs_4-5-6	0.918	**0.924** ± 0.000	0.854	**0.869** ± 0.000
vehicle0	0.957	**0.962** ± 0.005	0.920	**0.929** ± 0.017
ecoli1	0.904	**0.909** ± 0.000	0.773	**0.785** ± 0.000
new-thyroid1	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
new-thyroid2	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
ecoli2	0.894	**0.902** ± 0.005	0.722	**0.729** ± 0.020
segment0	0.994	**0.996** ± 0.004	0.987	**0.989** ± 0.004
glass6	**0.925**	0.918 ± 0.013	**0.812**	0.799 ± 0.051
yeast3	0.889	**0.891** ± 0.006	0.664	**0.664** ± 0.004
ecoli3	0.883	**0.893** ± 0.000	**0.608**	0.584 ± 0.000
page-blocks0	**0.576**	0.574 ± 0.058	0.227	**0.230** ± 0.049
ecoli-0-3-4_vs_5	0.866	**0.899** ± 0.000	0.646	**0.700** ± 0.000
yeast-2_vs_4	**0.890**	0.889 ± 0.000	**0.752**	0.713 ± 0.000
ecoli-0-6-7_vs_3-5	**0.852**	0.848 ± 0.000	**0.596**	0.569 ± 0.000
ecoli-0-2-3-4_vs_5	**0.900**	0.891 ± 0.000	**0.739**	0.692 ± 0.000
glass-0-1-5_vs_2	**0.526**	0.478 ± 0.000	**0.150**	0.125 ± 0.000
yeast-0-3-5-9_vs_7-8	0.616	**0.743** ± 0.025	0.346	**0.370** ± 0.016
yeast-0-2-5-7-9_vs_3-6-8	**0.803**	0.797 ± 0.016	**0.572**	0.522 ± 0.021
yeast-0-2-5-6_vs_3-7-8-9	**0.906**	0.904 ± 0.000	**0.748**	0.713 ± 0.000
ecoli-0-4-6_vs_5	**0.897**	0.891 ± 0.000	**0.745**	0.714 ± 0.000
ecoli-0-1_vs_2-3-5	0.882	**0.883** ± 0.006	0.699	**0.704** ± 0.038
ecoli-0-2-6-7_vs_3-5	**0.875**	0.844 ± 0.000	**0.701**	0.646 ± 0.000
glass-0-4_vs_5	**0.981**	0.964 ± 0.000	**0.893**	0.875 ± 0.000
ecoli-0-3-4-6_vs_5	**0.903**	0.897 ± 0.007	**0.777**	0.745 ± 0.025
ecoli-0-3-4-7_vs_5-6	**0.889**	0.885 ± 0.064	**0.708**	0.668 ± 0.072
yeast-0-5-6-7-9_vs_4	**0.799**	0.796 ± 0.005	**0.477**	0.450 ± 0.012
vowel0	0.944	**0.950** ± 0.001	**0.820**	0.809 ± 0.010
ecoli-0-6-7_vs_5	**0.870**	0.867 ± 0.000	**0.680**	0.662 ± 0.000
glass-0-1-6_vs_2	**0.589**	0.559 ± 0.084	**0.172**	0.135 ± 0.034
ecoli-0-1-4-7_vs_2-3-5-6	**0.880**	0.868 ± 0.004	**0.720**	0.629 ± 0.021
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.889**	0.880 ± 0.007	**0.698**	0.662 ± 0.063

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass-0-6_vs_5	**0.879**	0.868 ± 0.006	**0.711**	0.684 ± 0.054
ecoli-0-1_vs_5	**0.995**	0.972 ± 0.000	**0.960**	0.851 ± 0.000
glass-0-1-4-6_vs_2	0.578	**0.634** ± 0.020	**0.211**	0.209 ± 0.016
glass2	0.515	**0.599** ± 0.007	0.119	**0.174** ± 0.002
ecoli-0-1-4-7_vs_5-6	**0.884**	0.876 ± 0.004	**0.623**	0.592 ± 0.018
cleveland-0_vs_4	0.794	**0.794** ± 0.000	0.565	**0.565** ± 0.000
ecoli-0-1-4-6_vs_5	**0.892**	0.887 ± 0.005	**0.625**	0.601 ± 0.021
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.667	**0.761** ± 0.012	**0.298**	0.294 ± 0.016
glass4	**0.915**	0.911 ± 0.007	**0.608**	0.593 ± 0.032
ecoli4	**0.962**	0.943 ± 0.000	**0.818**	0.777 ± 0.000
page-blocks-1-3_vs_4	**0.734**	0.680 ± 0.028	**0.355**	0.308 ± 0.227
abalone9-18	0.871	**0.890** ± 0.002	**0.750**	0.527 ± 0.010
glass-0-1-6_vs_5	0.882	**0.974** ± 0.000	0.586	**0.676** ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	0.576	**0.658** ± 0.005	**0.143**	0.136 ± 0.003
glass5	**0.985**	0.975 ± 0.000	**0.848**	0.777 ± 0.000
yeast-2_vs_8	0.773	**0.773** ± 0.000	0.668	**0.668** ± 0.000
yeast4	0.814	**0.816** ± 0.001	**0.332**	0.290 ± 0.003
yeast-1-2-8-9_vs_7	0.677	**0.714** ± 0.005	**0.189**	0.144 ± 0.005
yeast5	0.960	**0.966** ± 0.001	**0.521**	0.481 ± 0.017
ecoli-0-1-3-7_vs_2-6	**0.839**	0.821 ± 0.000	**0.513**	0.306 ± 0.000
yeast6	0.861	**0.883** ± 0.000	**0.318**	0.306 ± 0.004
abalone19	0.696	**0.758** ± 0.008	**0.056**	0.049 ± 0.001
Promedio	0.836	**0.845** ± 0.007	**0.635**	0.617 ± 0.015

Fin de la Tabla A.19.

Tabla A.13: SMOTE-D vs SMOTE usando medida HVDM, K-NN y $k = 3$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.810**	0.809 ± 0.022	**0.749**	0.748 ± 0.029
ecoli-0_vs_1	**0.969**	0.967 ± 0.000	**0.974**	0.973 ± 0.000
wisconsin	0.961	**0.962** ± 0.000	0.948	**0.950** ± 0.000
pima	0.643	**0.660** ± 0.019	0.533	**0.568** ± 0.024
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.774	**0.802** ± 0.017	0.686	**0.721** ± 0.021

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
yeast1	0.640	**0.666** ± 0.017	0.489	**0.531** ± 0.022
haberman	**0.576**	0.565 ± 0.014	**0.401**	0.396 ± 0.017
vehicle2	**0.928**	0.928 ± 0.009	**0.873**	0.863 ± 0.013
vehicle1	0.631	**0.656** ± 0.013	0.454	**0.491** ± 0.016
vehicle3	0.634	**0.654** ± 0.015	0.455	**0.483** ± 0.019
glass-0-1-2-3_vs_4-5-6	0.938	**0.951** ± 0.007	0.910	**0.918** ± 0.020
vehicle0	0.916	**0.929** ± 0.007	**0.863**	0.862 ± 0.011
ecoli1	0.842	**0.858** ± 0.019	0.750	**0.759** ± 0.029
new-thyroid1	**0.994**	0.993 ± 0.005	**0.973**	0.971 ± 0.024
new-thyroid2	**1.000**	0.999 ± 0.000	**1.000**	0.997 ± 0.000
ecoli2	0.907	**0.915** ± 0.004	0.799	**0.818** ± 0.016
segment0	0.989	**0.990** ± 0.001	**0.974**	0.973 ± 0.007
glass6	0.919	**0.921** ± 0.005	0.846	**0.859** ± 0.031
yeast3	0.838	**0.853** ± 0.008	**0.691**	0.690 ± 0.013
ecoli3	0.809	**0.811** ± 0.000	**0.604**	0.601 ± 0.000
page-blocks0	0.890	**0.921** ± 0.004	**0.787**	0.786 ± 0.010
ecoli-0-3-4_vs_5	**0.863**	0.862 ± 0.000	**0.756**	0.749 ± 0.000
yeast-2_vs_4	**0.885**	0.878 ± 0.002	**0.765**	0.739 ± 0.018
ecoli-0-6-7_vs_3-5	0.840	**0.855** ± 0.000	0.678	**0.690** ± 0.000
ecoli-0-2-3-4_vs_5	0.858	**0.863** ± 0.004	0.731	**0.759** ± 0.017
glass-0-1-5_vs_2	**0.640**	0.618 ± 0.013	**0.307**	0.250 ± 0.025
yeast-0-3-5-9_vs_7-8	0.659	**0.675** ± 0.010	0.337	**0.350** ± 0.026
yeast-0-2-5-7-9_vs_3-6-8	**0.763**	0.748 ± 0.013	**0.485**	0.441 ± 0.021
yeast-0-2-5-6_vs_3-7-8-9	0.874	**0.877** ± 0.020	0.687	**0.690** ± 0.021
ecoli-0-4-6_vs_5	**0.866**	0.864 ± 0.000	**0.760**	0.744 ± 0.000
ecoli-0-1_vs_2-3-5	**0.824**	0.817 ± 0.000	**0.658**	0.646 ± 0.000
ecoli-0-2-6-7_vs_3-5	0.805	**0.811** ± 0.060	0.686	**0.689** ± 0.080
glass-0-4_vs_5	**1.000**	0.975 ± 0.000	**1.000**	0.966 ± 0.000
ecoli-0-3-4-6_vs_5	0.889	**0.892** ± 0.006	0.794	**0.818** ± 0.057
ecoli-0-3-4-7_vs_5-6	0.864	**0.865** ± 0.005	0.732	**0.736** ± 0.039
yeast-0-5-6-7-9_vs_4	**0.745**	0.739 ± 0.032	0.463	**0.463** ± 0.036
vowel0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-6-7_vs_5	**0.835**	0.834 ± 0.000	**0.702**	0.697 ± 0.000
glass-0-1-6_vs_2	**0.654**	0.640 ± 0.006	**0.353**	0.321 ± 0.013
ecoli-0-1-4-7_vs_2-3-5-6	**0.847**	0.839 ± 0.000	**0.702**	0.698 ± 0.000
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.848**	0.848 ± 0.000	**0.520**	0.520 ± 0.000
glass-0-6_vs_5	0.865	**0.867** ± 0.000	0.741	**0.756** ± 0.000

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli-0-1_vs_5	0.990	**0.990** ± 0.000	0.920	**0.920** ± 0.000
glass-0-1-4-6_vs_2	0.633	**0.647** ± 0.000	0.291	**0.304** ± 0.000
glass2	0.606	**0.629** ± 0.006	0.248	**0.265** ± 0.005
ecoli-0-1-4-7_vs_5-6	**0.871**	0.867 ± 0.000	**0.776**	0.735 ± 0.000
cleveland-0_vs_4	**0.576**	0.542 ± 0.012	**0.170**	0.150 ± 0.033
ecoli-0-1-4-6_vs_5	**0.871**	0.869 ± 0.004	**0.798**	0.774 ± 0.051
shuttle-c0-vs-c4	**0.996**	0.995 ± 0.000	**0.995**	0.995 ± 0.000
yeast-1_vs_7	0.684	**0.699** ± 0.003	0.330	**0.339** ± 0.015
glass4	**0.932**	0.932 ± 0.006	**0.748**	0.743 ± 0.070
ecoli4	0.895	**0.920** ± 0.000	0.820	**0.849** ± 0.000
page-blocks-1-3_vs_4	**0.938**	0.932 ± 0.000	**0.803**	0.777 ± 0.000
abalone9-18	0.719	**0.742** ± 0.003	0.448	**0.488** ± 0.018
glass-0-1-6_vs_5	0.835	**0.835** ± 0.000	0.633	0.633 ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	0.615	**0.616** ± 0.026	0.172	**0.173** ± 0.024
glass5	**0.890**	0.885 ± 0.000	**0.647**	0.642 ± 0.000
yeast-2_vs_8	**0.808**	0.800 ± 0.000	**0.519**	0.457 ± 0.000
yeast4	0.718	**0.753** ± 0.015	0.321	**0.343** ± 0.014
yeast-1-2-8-9_vs_7	0.583	**0.586** ± 0.044	**0.150**	0.139 ± 0.049
yeast5	**0.901**	0.901 ± 0.000	0.723	**0.723** ± 0.000
ecoli-0-1-3-7_vs_2-6	**0.840**	0.839 ± 0.000	**0.533**	0.513 ± 0.000
yeast6	**0.811**	0.806 ± 0.001	**0.427**	0.414 ± 0.010
abalone19	0.538	**0.546** ± 0.000	0.052	**0.056** ± 0.000
Promedio	0.822	**0.826** ± 0.007	**0.654**	0.653 ± 0.014

Fin de la Tabla A.19.

A.4.2. Valor de k igual a 5

Tabla A.14: SMOTE-D vs SMOTE usando medida HVDM, Árboles de Decisión y $k = 5$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.784**	0.745 ± 0.053	**0.719**	0.676 ± 0.066
ecoli-0_vs_1	0.962	**0.971** ± 0.008	0.967	**0.980** ± 0.008
wisconsin	0.938	**0.943** ± 0.009	0.920	**0.924** ± 0.009

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
pima	0.677	**0.680** ± 0.033	0.577	**0.587** ± 0.045
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	**0.787**	0.776 ± 0.058	**0.708**	0.695 ± 0.073
yeast1	0.659	**0.668** ± 0.016	0.520	**0.532** ± 0.023
haberman	0.534	**0.579** ± 0.033	0.337	**0.393** ± 0.045
vehicle2	0.946	**0.950** ± 0.013	0.915	**0.916** ± 0.018
vehicle1	0.697	0.697 ± 0.040	**0.546**	0.545 ± 0.057
vehicle3	**0.702**	0.700 ± 0.036	**0.554**	0.547 ± 0.047
glass-0-1-2-3_vs_4-5-6	0.886	**0.887** ± 0.016	0.819	**0.825** ± 0.042
vehicle0	**0.917**	0.914 ± 0.012	**0.877**	0.864 ± 0.019
ecoli1	0.839	**0.850** ± 0.020	**0.762**	0.758 ± 0.027
new-thyroid1	**0.965**	0.955 ± 0.049	**0.942**	0.913 ± 0.063
new-thyroid2	**0.965**	0.936 ± 0.031	**0.942**	0.879 ± 0.023
ecoli2	**0.867**	0.853 ± 0.027	**0.734**	0.722 ± 0.054
segment0	**0.993**	0.991 ± 0.003	**0.983**	0.980 ± 0.003
glass6	0.874	**0.875** ± 0.025	0.723	**0.762** ± 0.118
yeast3	0.858	**0.881** ± 0.017	0.707	**0.745** ± 0.026
ecoli3	0.741	**0.761** ± 0.071	0.521	**0.523** ± 0.080
page-blocks0	0.904	**0.926** ± 0.007	0.791	**0.800** ± 0.015
ecoli-0-3-4_vs_5	0.877	**0.881** ± 0.063	0.705	**0.730** ± 0.102
yeast-2_vs_4	0.829	**0.861** ± 0.037	0.661	**0.700** ± 0.049
ecoli-0-6-7_vs_3-5	**0.850**	0.841 ± 0.010	**0.623**	0.608 ± 0.067
ecoli-0-2-3-4_vs_5	0.852	**0.864** ± 0.010	0.704	**0.726** ± 0.047
glass-0-1-5_vs_2	0.548	**0.627** ± 0.072	0.160	**0.309** ± 0.155
yeast-0-3-5-9_vs_7-8	0.645	**0.673** ± 0.084	0.319	**0.350** ± 0.119
yeast-0-2-5-7-9_vs_3-6-8	**0.747**	0.746 ± 0.034	**0.497**	0.459 ± 0.063
yeast-0-2-5-6_vs_3-7-8-9	0.872	**0.885** ± 0.019	0.677	**0.704** ± 0.040
ecoli-0-4-6_vs_5	0.850	**0.857** ± 0.049	**0.690**	0.681 ± 0.076
ecoli-0-1_vs_2-3-5	0.808	**0.824** ± 0.004	0.588	**0.610** ± 0.015
ecoli-0-2-6-7_vs_3-5	**0.847**	0.812 ± 0.049	**0.659**	0.603 ± 0.078
glass-0-4_vs_5	0.994	0.994 ± 0.000	0.933	0.933 ± 0.000
ecoli-0-3-4-6_vs_5	**0.853**	0.849 ± 0.021	0.702	**0.703** ± 0.126
ecoli-0-3-4-7_vs_5-6	**0.862**	0.856 ± 0.010	**0.698**	0.661 ± 0.051
yeast-0-5-6-7-9_vs_4	0.700	**0.731** ± 0.065	0.411	**0.448** ± 0.097
vowel0	0.947	**0.957** ± 0.014	0.844	**0.863** ± 0.029
ecoli-0-6-7_vs_5	**0.867**	0.853 ± 0.008	**0.666**	0.650 ± 0.054
glass-0-1-6_vs_2	0.536	**0.625** ± 0.055	0.161	**0.295** ± 0.070

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli-0-1-4-7_vs_2-3-5-6	**0.872**	0.862 ± 0.007	0.675	**0.684** ± 0.046
led7digit-0-2-4-5-6-7-8-9_vs_1	**0.901**	0.899 ± 0.022	0.764	**0.778** ± 0.041
glass-0-6_vs_5	0.861	**0.886** ± 0.011	0.737	0.737 ± 0.086
ecoli-0-1_vs_5	0.945	0.945 ± 0.000	0.866	0.866 ± 0.000
glass-0-1-4-6_vs_2	0.691	**0.712** ± 0.116	0.352	**0.413** ± 0.171
glass2	0.633	**0.705** ± 0.124	0.334	**0.420** ± 0.101
ecoli-0-1-4-7_vs_5-6	0.840	**0.895** ± 0.066	0.666	**0.701** ± 0.147
cleveland-0_vs_4	0.709	**0.856** ± 0.150	0.383	**0.664** ± 0.311
ecoli-0-1-4-6_vs_5	**0.898**	0.841 ± 0.010	**0.665**	0.613 ± 0.060
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.621	**0.662** ± 0.054	0.255	**0.305** ± 0.062
glass4	0.815	**0.877** ± 0.015	0.533	**0.696** ± 0.171
ecoli4	0.863	**0.865** ± 0.003	0.729	**0.742** ± 0.051
page-blocks-1-3_vs_4	0.997	0.997 ± 0.000	0.966	0.966 ± 0.000
abalone9-18	**0.780**	0.704 ± 0.082	**0.423**	0.372 ± 0.095
glass-0-1-6_vs_5	0.838	0.838 ± 0.000	0.660	0.660 ± 0.000
shuttle-c2-vs-c4	0.950	0.950 ± 0.000	0.933	0.933 ± 0.000
yeast-1-4-5-8_vs_7	**0.591**	0.546 ± 0.034	**0.166**	0.116 ± 0.041
glass5	**0.947**	0.897 ± 0.000	**0.893**	0.760 ± 0.000
yeast-2_vs_8	0.721	**0.740** ± 0.004	0.322	**0.393** ± 0.000
yeast4	**0.742**	0.680 ± 0.024	**0.326**	0.280 ± 0.029
yeast-1-2-8-9_vs_7	**0.712**	0.581 ± 0.033	**0.276**	0.137 ± 0.053
yeast5	**0.878**	0.876 ± 0.033	0.668	**0.693** ± 0.049
ecoli-0-1-3-7_vs_2-6	0.840	0.840 ± 0.004	**0.540**	0.506 ± 0.080
yeast6	**0.766**	0.759 ± 0.033	0.359	**0.368** ± 0.046
abalone19	**0.519**	0.513 ± 0.039	**0.034**	0.022 ± 0.038
Promedio	0.817	**0.821** ± 0.031	0.633	**0.643** ± 0.057

Fin de la Tabla A.14.

Tabla A.15: SMOTE-D vs SMOTE usando medida HVDM, SVM y $k = 5$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.595	**0.633** ± 0.002	0.574	**0.579** ± 0.004
ecoli-0_vs_1	0.962	**0.978** ± 0.000	0.967	**0.984** ± 0.000
wisconsin	**0.970**	0.965 ± 0.000	**0.953**	0.952 ± 0.000
pima	0.666	**0.742** ± 0.021	0.612	**0.665** ± 0.027

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.704	**0.750** ± 0.020	0.622	**0.660** ± 0.020
yeast1	0.627	**0.712** ± 0.003	0.520	**0.588** ± 0.004
haberman	0.500	**0.630** ± 0.038	0.418	**0.450** ± 0.064
vehicle2	0.951	**0.957** ± 0.000	0.908	**0.923** ± 0.000
vehicle1	**0.808**	0.804 ± 0.010	0.657	**0.664** ± 0.006
vehicle3	0.693	**0.782** ± 0.006	0.532	**0.632** ± 0.011
glass-0-1-2-3_vs_4-5-6	0.918	**0.920** ± 0.000	0.844	**0.862** ± 0.000
vehicle0	**0.967**	0.960 ± 0.005	**0.932**	0.925 ± 0.017
ecoli1	0.894	**0.910** ± 0.000	0.752	**0.788** ± 0.000
new-thyroid1	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
new-thyroid2	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
ecoli2	0.901	**0.905** ± 0.005	0.705	**0.733** ± 0.020
segment0	0.994	**0.996** ± 0.004	0.987	**0.990** ± 0.004
glass6	0.922	**0.925** ± 0.000	0.801	**0.812** ± 0.000
yeast3	**0.897**	0.894 ± 0.001	0.653	**0.665** ± 0.004
ecoli3	**0.887**	0.883 ± 0.000	**0.590**	0.578 ± 0.000
page-blocks0	0.510	**0.577** ± 0.0.46	0.190	**0.234** ± 0.046
ecoli-0-3-4_vs_5	0.866	**0.883** ± 0.008	0.647	**0.677** ± 0.051
yeast-2_vs_4	**0.893**	0.889 ± 0.003	**0.721**	0.712 ± 0.015
ecoli-0-6-7_vs_3-5	0.842	**0.848** ± 0.007	0.538	**0.574** ± 0.034
ecoli-0-2-3-4_vs_5	**0.891**	0.885 ± 0.000	0.692	0.692 ± 0.000
glass-0-1-5_vs_2	**0.526**	0.509 ± 0.067	**0.172**	0.159 ± 0.038
yeast-0-3-5-9_vs_7-8	0.727	**0.745** ± 0.017	0.361	**0.375** ± 0.008
yeast-0-2-5-7-9_vs_3-6-8	0.789	**0.797** ± 0.016	0.516	**0.524** ± 0.021
yeast-0-2-5-6_vs_3-7-8-9	0.900	**0.904** ± 0.000	0.695	**0.709** ± 0.000
ecoli-0-4-6_vs_5	0.889	**0.894** ± 0.000	0.705	**0.730** ± 0.000
ecoli-0-1_vs_2-3-5	0.866	**0.883** ± 0.000	**0.706**	0.704 ± 0.000
ecoli-0-2-6-7_vs_3-5	**0.872**	0.847 ± 0.000	**0.691**	0.657 ± 0.000
glass-0-4_vs_5	0.944	**0.969** ± 0.000	0.893	0.893 ± 0.000
ecoli-0-3-4-6_vs_5	0.897	**0.899** ± 0.007	0.742	**0.755** ± 0.025
ecoli-0-3-4-7_vs_5-6	**0.881**	0.877 ± 0.025	**0.669**	0.657 ± 0.073
yeast-0-5-6-7-9_vs_4	0.780	**0.788** ± 0.028	0.428	**0.441** ± 0.027
vowel0	0.955	0.955 ± 0.016	**0.831**	0.818 ± 0.016
ecoli-0-6-7_vs_5	0.862	**0.870** ± 0.007	0.646	**0.672** ± 0.048
glass-0-1-6_vs_2	0.513	**0.580** ± 0.021	0.099	**0.136** ± 0.013
ecoli-0-1-4-7_vs_2-3-5-6	0.864	**0.878** ± 0.004	0.627	**0.632** ± 0.021

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
led7digit-0-2-4-5-6-7-8-9_vs_1	0.880	0.880 ± 0.007	0.557	**0.648** ± 0.063
glass-0-6_vs_5	0.877	0.877 ± 0.006	**0.702**	0.701 ± 0.054
ecoli-0-1_vs_5	**0.974**	0.972 ± 0.000	0.834	**0.846** ± 0.000
glass-0-1-4-6_vs_2	0.592	**0.632** ± 0.104	0.208	**0.209** ± 0.060
glass2	0.541	**0.612** ± 0.007	0.143	**0.182** ± 0.002
ecoli-0-1-4-7_vs_5-6	**0.879**	0.875 ± 0.004	**0.605**	0.586 ± 0.016
cleveland-0_vs_4	0.794	0.794 ± 0.000	0.565	0.565 ± 0.000
ecoli-0-1-4-6_vs_5	**0.890**	0.888 ± 0.011	**0.620**	0.604 ± 0.047
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.708	**0.751** ± 0.005	**0.299**	0.286 ± 0.007
glass4	**0.912**	0.910 ± 0.007	**0.593**	0.586 ± 0.032
ecoli4	**0.960**	0.952 ± 0.000	**0.806**	0.793 ± 0.000
page-blocks-1-3_vs_4	**0.761**	0.707 ± 0.024	**0.364**	0.333 ± 0.197
abalone9-18	**0.897**	0.892 ± 0.002	**0.609**	0.522 ± 0.010
glass-0-1-6_vs_5	**0.982**	0.980 ± 0.008	**0.746**	0.733 ± 0.096
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	**0.631**	0.612 ± 0.002	**0.169**	0.120 ± 0.001
glass5	**0.978**	0.974 ± 0.000	**0.788**	0.764 ± 0.000
yeast-2_vs_8	0.773	0.773 ± 0.000	0.668	0.668 ± 0.000
yeast4	**0.824**	0.819 ± 0.001	**0.307**	0.290 ± 0.003
yeast-1-2-8-9_vs_7	**0.708**	0.696 ± 0.013	**0.181**	0.133 ± 0.011
yeast5	**0.969**	0.966 ± 0.001	**0.504**	0.482 ± 0.010
ecoli-0-1-3-7_vs_2-6	**0.876**	0.820 ± 0.000	**0.393**	0.301 ± 0.000
yeast6	0.869	**0.884** ± 0.000	0.298	**0.312** ± 0.004
abalone19	0.740	**0.766** ± 0.035	**0.056**	0.050 ± 0.004
Promedio	0.835	**0.845** ± 0.009	0.614	**0.619** ± 0.019

Fin de la Tabla A.19.

Tabla A.16: SMOTE-D vs SMOTE usando medida HVDM, K-NN y $k = 5$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.817**	0.812 ± 0.021	**0.757**	0.751 ± 0.028
ecoli-0_vs_1	0.962	**0.964** ± 0.000	0.967	**0.972** ± 0.000
wisconsin	0.959	**0.962** ± 0.000	0.945	**0.949** ± 0.000
pima	0.645	**0.656** ± 0.014	0.535	**0.562** ± 0.020
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass0	0.774	**0.799** ± 0.020	0.686	**0.716** ± 0.023
yeast1	0.638	**0.667** ± 0.011	0.486	**0.532** ± 0.014
haberman	**0.579**	0.562 ± 0.035	**0.403**	0.392 ± 0.038
vehicle2	0.923	**0.930** ± 0.009	0.865	0.865 ± 0.018
vehicle1	0.623	**0.659** ± 0.022	0.446	**0.494** ± 0.027
vehicle3	0.646	**0.650** ± 0.023	0.472	**0.478** ± 0.029
glass-0-1-2-3_vs_4-5-6	0.948	0.948 ± 0.008	**0.921**	0.914 ± 0.022
vehicle0	0.921	**0.930** ± 0.007	**0.866**	0.861 ± 0.011
ecoli1	0.827	**0.858** ± 0.025	0.732	**0.756** ± 0.038
new-thyroid1	0.980	**0.991** ± 0.005	0.958	**0.961** ± 0.024
new-thyroid2	**1.000**	0.998 ± 0.000	**1.000**	0.994 ± 0.000
ecoli2	0.914	**0.919** ± 0.005	0.805	**0.822** ± 0.022
segment0	0.989	**0.990** ± 0.001	**0.974**	0.971 ± 0.006
glass6	0.919	**0.920** ± 0.008	0.846	**0.850** ± 0.045
yeast3	0.849	**0.862** ± 0.013	0.695	**0.697** ± 0.015
ecoli3	**0.838**	0.831 ± 0.002	**0.641**	0.622 ± 0.014
page-blocks0	0.898	**0.922** ± 0.004	**0.794**	0.787 ± 0.008
ecoli-0-3-4_vs_5	0.863	**0.864** ± 0.000	0.756	**0.760** ± 0.000
yeast-2_vs_4	**0.892**	0.882 ± 0.002	**0.764**	0.744 ± 0.018
ecoli-0-6-7_vs_3-5	**0.862**	0.855 ± 0.003	**0.691**	0.681 ± 0.028
ecoli-0-2-3-4_vs_5	**0.861**	0.860 ± 0.005	**0.749**	0.743 ± 0.023
glass-0-1-5_vs_2	**0.640**	0.631 ± 0.010	**0.271**	0.247 ± 0.015
yeast-0-3-5-9_vs_7-8	**0.695**	0.686 ± 0.008	**0.370**	0.359 ± 0.021
yeast-0-2-5-7-9_vs_3-6-8	**0.758**	0.749 ± 0.015	**0.458**	0.435 ± 0.027
yeast-0-2-5-6_vs_3-7-8-9	0.861	**0.881** ± 0.018	0.641	**0.679** ± 0.028
ecoli-0-4-6_vs_5	**0.869**	0.866 ± 0.000	**0.778**	0.757 ± 0.000
ecoli-0-1_vs_2-3-5	**0.840**	0.833 ± 0.000	0.653	**0.673** ± 0.000
ecoli-0-2-6-7_vs_3-5	0.795	**0.811** ± 0.065	0.636	**0.681** ± 0.092
glass-0-4_vs_5	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-3-4-6_vs_5	**0.897**	0.893 ± 0.006	**0.854**	0.827 ± 0.057
ecoli-0-3-4-7_vs_5-6	0.864	0.864 ± 0.007	**0.732**	0.731 ± 0.050
yeast-0-5-6-7-9_vs_4	0.711	**0.741** ± 0.036	0.406	**0.452** ± 0.046
vowel0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-6-7_vs_5	0.832	**0.833** ± 0.006	0.691	**0.692** ± 0.036
glass-0-1-6_vs_2	0.645	**0.649** ± 0.004	0.323	**0.324** ± 0.008
ecoli-0-1-4-7_vs_2-3-5-6	**0.862**	0.850 ± 0.003	**0.716**	0.701 ± 0.021
led7digit-0-2-4-5-6-7-8-9_vs_1	0.847	**0.848** ± 0.000	0.513	**0.520** ± 0.000

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass-0-6_vs_5	**0.870**	0.867 ± 0.000	**0.779**	0.757 ± 0.000
ecoli-0-1_vs_5	0.990	0.990 ± 0.000	0.920	0.920 ± 0.000
glass-0-1-4-6_vs_2	**0.683**	0.677 ± 0.000	**0.350**	0.334 ± 0.000
glass2	0.601	**0.638** ± 0.008	0.235	**0.273** ± 0.007
ecoli-0-1-4-7_vs_5-6	**0.870**	0.867 ± 0.000	**0.758**	0.736 ± 0.000
cleveland-0_vs_4	**0.566**	0.550 ± 0.019	**0.159**	0.156 ± 0.035
ecoli-0-1-4-6_vs_5	0.869	0.869 ± 0.006	0.776	**0.780** ± 0.063
shuttle-c0-vs-c4	0.996	0.996 ± 0.000	0.995	0.995 ± 0.000
yeast-1_vs_7	**0.708**	0.704 ± 0.005	**0.340**	0.331 ± 0.017
glass4	0.932	0.932 ± 0.003	**0.748**	0.743 ± 0.042
ecoli4	**0.920**	0.915 ± 0.000	**0.849**	0.843 ± 0.000
page-blocks-1-3_vs_4	**0.938**	0.937 ± 0.000	**0.803**	0.785 ± 0.000
abalone9-18	**0.773**	0.770 ± 0.029	0.501	**0.512** ± 0.044
glass-0-1-6_vs_5	**0.935**	0.930 ± 0.000	**0.753**	0.747 ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	0.604	**0.619** ± 0.026	0.139	**0.159** ± 0.020
glass5	0.890	0.890 ± 0.000	0.647	0.647 ± 0.000
yeast-2_vs_8	**0.799**	0.796 ± 0.000	**0.442**	0.426 ± 0.000
yeast4	0.723	**0.754** ± 0.016	0.309	**0.333** ± 0.019
yeast-1-2-8-9_vs_7	0.619	**0.620** ± 0.034	**0.165**	0.159 ± 0.028
yeast5	0.901	0.901 ± 0.000	0.717	0.717 ± 0.000
ecoli-0-1-3-7_vs_2-6	**0.840**	0.839 ± 0.000	**0.533**	0.513 ± 0.000
yeast6	**0.807**	0.797 ± 0.001	**0.392**	0.379 ± 0.008
abalone19	0.547	**0.549** ± 0.000	0.048	**0.049** ± 0.000
Promedio	0.828	**0.831** ± 0.008	0.654	**0.655** ± 0.017

Fin de la Tabla A.16.

A.4.3. Valor de k igual a 7

Tabla A.17: SMOTE-D vs SMOTE usando medida HVDM, Árboles de Decisión y $k = 7$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	**0.770**	0.739 ± 0.056	**0.704**	0.667 ± 0.068
ecoli-0_vs_1	0.958	**0.973** ± 0.009	0.964	**0.980** ± 0.009

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
wisconsin	0.941	**0.943** ± 0.012	0.922	**0.924** ± 0.012
pima	0.654	**0.682** ± 0.029	0.547	**0.590** ± 0.043
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.755	**0.792** ± 0.081	0.666	**0.715** ± 0.097
yeast1	0.645	**0.671** ± 0.026	0.496	**0.535** ± 0.036
haberman	0.548	**0.579** ± 0.032	0.345	**0.391** ± 0.047
vehicle2	0.944	**0.947** ± 0.013	**0.917**	0.911 ± 0.020
vehicle1	0.675	**0.691** ± 0.035	0.516	**0.537** ± 0.050
vehicle3	0.672	**0.697** ± 0.034	0.506	**0.543** ± 0.045
glass-0-1-2-3_vs_4-5-6	0.895	**0.898** ± 0.004	0.843	**0.847** ± 0.014
vehicle0	**0.920**	0.913 ± 0.022	**0.880**	0.864 ± 0.026
ecoli1	**0.888**	0.855 ± 0.030	**0.814**	0.761 ± 0.039
new-thyroid1	**0.960**	0.947 ± 0.054	**0.918**	0.906 ± 0.068
new-thyroid2	**0.965**	0.947 ± 0.031	**0.942**	0.905 ± 0.033
ecoli2	**0.877**	0.855 ± 0.028	**0.744**	0.722 ± 0.056
segment0	0.991	**0.992** ± 0.004	0.981	**0.982** ± 0.005
glass6	**0.928**	0.875 ± 0.079	**0.833**	0.760 ± 0.159
yeast3	0.867	**0.879** ± 0.014	0.718	**0.739** ± 0.030
ecoli3	**0.774**	0.751 ± 0.102	**0.529**	0.493 ± 0.109
page-blocks0	0.910	**0.927** ± 0.014	**0.803**	0.799 ± 0.020
ecoli-0-3-4_vs_5	0.852	**0.891** ± 0.068	0.704	**0.754** ± 0.105
yeast-2_vs_4	**0.893**	0.864 ± 0.033	**0.719**	0.697 ± 0.034
ecoli-0-6-7_vs_3-5	**0.850**	0.846 ± 0.010	**0.620**	0.612 ± 0.047
ecoli-0-2-3-4_vs_5	**0.886**	0.861 ± 0.071	**0.774**	0.726 ± 0.134
glass-0-1-5_vs_2	**0.650**	0.635 ± 0.091	**0.349**	0.305 ± 0.160
yeast-0-3-5-9_vs_7-8	**0.668**	0.648 ± 0.062	**0.342**	0.320 ± 0.100
yeast-0-2-5-7-9_vs_3-6-8	**0.743**	0.742 ± 0.031	**0.461**	0.450 ± 0.040
yeast-0-2-5-6_vs_3-7-8-9	0.882	**0.890** ± 0.019	0.669	**0.716** ± 0.044
ecoli-0-4-6_vs_5	0.878	**0.880** ± 0.037	**0.726**	0.715 ± 0.063
ecoli-0-1_vs_2-3-5	0.815	**0.816** ± 0.006	0.607	**0.611** ± 0.023
ecoli-0-2-6-7_vs_3-5	**0.840**	0.836 ± 0.067	0.620	**0.629** ± 0.117
glass-0-4_vs_5	**0.994**	0.994 ± 0.000	**0.933**	0.933 ± 0.000
ecoli-0-3-4-6_vs_5	0.853	**0.871** ± 0.018	0.698	**0.741** ± 0.097
ecoli-0-3-4-7_vs_5-6	**0.860**	0.858 ± 0.015	**0.678**	0.675 ± 0.077
yeast-0-5-6-7-9_vs_4	0.711	**0.738** ± 0.078	0.421	**0.451** ± 0.104
vowel0	0.956	**0.963** ± 0.013	0.872	**0.879** ± 0.033
ecoli-0-6-7_vs_5	**0.880**	0.868 ± 0.008	**0.731**	0.685 ± 0.064

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass-0-1-6_vs_2	0.590	**0.610** ± 0.092	0.238	**0.268** ± 0.127
ecoli-0-1-4-7_vs_2-3-5-6	**0.877**	0.866 ± 0.006	**0.699**	0.683 ± 0.040
led7digit-0-2-4-5-6-7-8-9_vs_1	0.893	**0.900** ± 0.005	0.776	**0.776** ± 0.046
glass-0-6_vs_5	**0.886**	0.873 ± 0.009	**0.758**	0.719 ± 0.070
ecoli-0-1_vs_5	0.945	**0.945** ± 0.000	**0.866**	0.866 ± 0.000
glass-0-1-4-6_vs_2	**0.719**	0.690 ± 0.123	**0.383**	0.380 ± 0.162
glass2	0.636	**0.700** ± 0.111	0.320	**0.419** ± 0.103
ecoli-0-1-4-7_vs_5-6	0.842	**0.880** ± 0.015	0.678	**0.697** ± 0.114
cleveland-0_vs_4	0.735	**0.829** ± 0.125	0.450	**0.606** ± 0.245
ecoli-0-1-4-6_vs_5	**0.876**	0.825 ± 0.080	**0.662**	0.573 ± 0.102
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.645	**0.667** ± 0.120	0.273	**0.294** ± 0.145
glass4	**0.894**	0.888 ± 0.013	0.666	**0.694** ± 0.144
ecoli4	**0.888**	0.863 ± 0.000	**0.752**	0.745 ± 0.000
page-blocks-1-3_vs_4	0.997	**0.997** ± 0.000	0.966	**0.966** ± 0.000
abalone9-18	**0.742**	0.726 ± 0.045	0.387	**0.389** ± 0.078
glass-0-1-6_vs_5	0.838	**0.838** ± 0.000	0.660	**0.660** ± 0.000
shuttle-c2-vs-c4	0.950	0.950 ± 0.000	0.933	**0.933** ± 0.000
yeast-1-4-5-8_vs_7	**0.584**	0.558 ± 0.053	**0.140**	0.124 ± 0.053
glass5	**0.947**	0.897 ± 0.000	**0.893**	0.760 ± 0.000
yeast-2_vs_8	0.730	**0.760** ± 0.065	0.392	**0.419** ± 0.160
yeast4	**0.704**	0.695 ± 0.044	**0.293**	0.290 ± 0.043
yeast-1-2-8-9_vs_7	**0.705**	0.586 ± 0.070	**0.248**	0.137 ± 0.077
yeast5	**0.877**	0.872 ± 0.036	**0.692**	0.684 ± 0.073
ecoli-0-1-3-7_vs_2-6	**0.840**	0.836 ± 0.011	**0.540**	0.472 ± 0.143
yeast6	0.763	**0.766** ± 0.023	0.331	**0.380** ± 0.055
abalone19	**0.533**	0.519 ± 0.034	**0.041**	0.025 ± 0.025
Promedio	**0.824**	0.823 ± 0.037	**0.645**	0.643 ± 0.064

Fin de la Tabla A.19.

Tabla A.18: SMOTE-D vs SMOTE usando medida HVDM, SVM y $k = 7$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.581	**0.618** ± 0.019	**0.565**	0.563 ± 0.021
ecoli-0_vs_1	0.962	**0.978** ± 0.000	0.967	**0.984** ± 0.000
wisconsin	**0.975**	0.971 ± 0.006	**0.959**	0.959 ± 0.006

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
pima	0.636	**0.742** ± 0.004	0.594	**0.665** ± 0.005
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.701	**0.746** ± 0.020	0.619	**0.656** ± 0.020
yeast1	0.549	**0.712** ± 0.001	0.473	**0.589** ± 0.002
haberman	0.500	**0.624** ± 0.006	0.418	**0.441** ± 0.007
vehicle2	0.942	**0.956** ± 0.009	0.887	**0.921** ± 0.016
vehicle1	**0.802**	0.798 ± 0.054	0.644	**0.657** ± 0.057
vehicle3	0.765	**0.772** ± 0.028	0.596	**0.620** ± 0.029
glass-0-1-2-3_vs_4-5-6	**0.924**	0.917 ± 0.000	0.848	**0.860** ± 0.000
vehicle0	0.961	**0.961** ± 0.003	0.927	**0.927** ± 0.011
ecoli1	0.892	**0.902** ± 0.038	0.748	**0.775** ± 0.038
new-thyroid1	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
new-thyroid2	0.982	0.982 ± 0.000	0.971	0.971 ± 0.000
ecoli2	0.892	**0.903** ± 0.005	0.679	**0.733** ± 0.020
segment0	0.995	**0.995** ± 0.003	0.988	**0.989** ± 0.000
glass6	0.922	**0.926** ± 0.007	0.801	**0.819** ± 0.031
yeast3	**0.910**	0.893 ± 0.009	0.651	**0.660** ± 0.011
ecoli3	0.880	**0.884** ± 0.004	0.573	**0.581** ± 0.014
page-blocks0	0.564	**0.579** ± 0.076	**0.253**	0.250 ± 0.104
ecoli-0-3-4_vs_5	0.866	**0.892** ± 0.008	0.647	**0.697** ± 0.051
yeast-2_vs_4	0.889	**0.892** ± 0.000	0.699	**0.716** ± 0.000
ecoli-0-6-7_vs_3-5	0.835	**0.845** ± 0.007	0.516	**0.561** ± 0.034
ecoli-0-2-3-4_vs_5	**0.894**	0.891 ± 0.008	**0.702**	0.696 ± 0.026
glass-0-1-5_vs_2	**0.519**	0.478 ± 0.033	**0.167**	0.145 ± 0.013
yeast-0-3-5-9_vs_7-8	**0.742**	0.737 ± 0.017	0.356	**0.366** ± 0.005
yeast-0-2-5-7-9_vs_3-6-8	0.791	**0.797** ± 0.015	0.509	**0.523** ± 0.016
yeast-0-2-5-6_vs_3-7-8-9	0.900	**0.904** ± 0.001	0.679	**0.711** ± 0.007
ecoli-0-4-6_vs_5	0.889	**0.896** ± 0.000	0.709	**0.731** ± 0.000
ecoli-0-1_vs_2-3-5	0.860	**0.871** ± 0.011	0.670	**0.672** ± 0.057
ecoli-0-2-6-7_vs_3-5	0.837	**0.843** ± 0.000	0.615	**0.641** ± 0.000
glass-0-4_vs_5	**0.944**	0.939 ± 0.000	**0.893**	0.875 ± 0.000
ecoli-0-3-4-6_vs_5	**0.897**	0.895 ± 0.000	**0.746**	0.734 ± 0.000
ecoli-0-3-4-7_vs_5-6	0.870	**0.893** ± 0.064	0.627	**0.655** ± 0.064
yeast-0-5-6-7-9_vs_4	0.787	**0.794** ± 0.003	0.431	**0.446** ± 0.006
vowel0	**0.958**	0.955 ± 0.032	**0.823**	0.820 ± 0.034
ecoli-0-6-7_vs_5	0.867	**0.876** ± 0.000	0.663	**0.709** ± 0.000
glass-0-1-6_vs_2	**0.568**	0.555 ± 0.080	**0.142**	0.135 ± 0.032

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
ecoli-0-1-4-7_vs_2-3-5-6	0.859	**0.865** ± 0.004	0.605	**0.616** ± 0.018
led7digit-0-2-4-5-6-7-8-9_vs_1	0.862	**0.880** ± 0.003	0.508	**0.652** ± 0.029
glass-0-6_vs_5	**0.879**	0.878 ± 0.006	**0.721**	0.708 ± 0.054
ecoli-0-1_vs_5	**0.984**	0.975 ± 0.000	**0.880**	0.844 ± 0.000
glass-0-1-4-6_vs_2	0.570	**0.632** ± 0.030	0.170	**0.204** ± 0.022
glass2	0.617	**0.666** ± 0.000	0.177	**0.217** ± 0.000
ecoli-0-1-4-7_vs_5-6	**0.877**	0.875 ± 0.008	**0.593**	0.587 ± 0.030
cleveland-0_vs_4	0.794	**0.794** ± 0.000	0.565	**0.565** ± 0.000
ecoli-0-1-4-6_vs_5	**0.888**	0.887 ± 0.005	**0.605**	0.599 ± 0.021
shuttle-c0-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1_vs_7	0.740	**0.744** ± 0.003	**0.313**	0.282 ± 0.004
glass4	**0.915**	0.913 ± 0.014	**0.613**	0.599 ± 0.073
ecoli4	**0.960**	0.951 ± 0.000	**0.806**	0.789 ± 0.000
page-blocks-1-3_vs_4	0.676	**0.760** ± 0.039	0.235	**0.401** ± 0.187
abalone9-18	0.889	**0.892** ± 0.002	**0.559**	0.520 ± 0.010
glass-0-1-6_vs_5	**0.985**	0.980 ± 0.008	**0.786**	0.726 ± 0.096
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	**0.673**	0.617 ± 0.010	**0.175**	0.122 ± 0.004
glass5	**0.980**	0.973 ± 0.007	**0.800**	0.742 ± 0.192
yeast-2_vs_8	**0.773**	0.773 ± 0.000	**0.668**	0.661 ± 0.000
yeast4	**0.820**	0.820 ± 0.002	**0.295**	0.293 ± 0.009
yeast-1-2-8-9_vs_7	**0.696**	0.691 ± 0.003	**0.162**	0.130 ± 0.002
yeast5	**0.969**	0.966 ± 0.001	**0.501**	0.481 ± 0.009
ecoli-0-1-3-7_vs_2-6	**0.876**	0.819 ± 0.009	**0.393**	0.294 ± 0.057
yeast6	0.864	**0.883** ± 0.001	0.278	**0.307** ± 0.007
abalone19	0.749	**0.766** ± 0.000	**0.054**	0.050 ± 0.000
Promedio	0.835	**0.845** ± 0.011	0.608	**0.618** ± 0.023

Fin de la Tabla A.19.

Tabla A.19: SMOTE-D vs SMOTE usando medida HVDM, K-NN y $k = 7$

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
glass1	0.813	**0.820** ± 0.019	0.752	**0.764** ± 0.026
ecoli-0_vs_1	0.962	**0.967** ± 0.000	0.967	**0.973** ± 0.000
wisconsin	**0.964**	0.961 ± 0.000	**0.952**	0.949 ± 0.000
pima	0.644	**0.653** ± 0.018	0.533	**0.558** ± 0.023

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
iris0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
glass0	0.785	**0.798** ± 0.024	0.698	**0.716** ± 0.032
yeast1	0.639	**0.660** ± 0.013	0.489	**0.524** ± 0.016
haberman	**0.583**	0.573 ± 0.022	**0.410**	0.402 ± 0.025
vehicle2	**0.933**	0.930 ± 0.009	**0.878**	0.866 ± 0.016
vehicle1	0.617	**0.657** ± 0.023	0.435	**0.492** ± 0.028
vehicle3	0.628	**0.650** ± 0.021	0.443	**0.478** ± 0.026
glass-0-1-2-3_vs_4-5-6	0.944	**0.949** ± 0.007	0.911	**0.915** ± 0.020
vehicle0	0.921	**0.933** ± 0.009	**0.868**	0.867 ± 0.010
ecoli1	0.813	**0.858** ± 0.025	0.708	**0.754** ± 0.035
new-thyroid1	0.994	**0.994** ± 0.004	0.973	**0.974** ± 0.018
new-thyroid2	**1.000**	0.998 ± 0.000	**1.000**	0.993 ± 0.000
ecoli2	0.903	**0.915** ± 0.004	0.785	**0.819** ± 0.019
segment0	**0.990**	0.990 ± 0.001	**0.974**	0.971 ± 0.005
glass6	**0.933**	0.921 ± 0.005	0.852	**0.859** ± 0.031
yeast3	0.847	**0.862** ± 0.018	0.688	**0.698** ± 0.023
ecoli3	**0.834**	0.817 ± 0.005	**0.626**	0.605 ± 0.028
page-blocks0	0.897	**0.921** ± 0.004	**0.797**	0.784 ± 0.011
ecoli-0-3-4_vs_5	0.861	**0.862** ± 0.000	0.742	**0.750** ± 0.000
yeast-2_vs_4	**0.890**	0.879 ± 0.001	**0.749**	0.736 ± 0.011
ecoli-0-6-7_vs_3-5	**0.860**	0.859 ± 0.003	0.674	**0.681** ± 0.028
ecoli-0-2-3-4_vs_5	**0.863**	0.861 ± 0.004	**0.771**	0.751 ± 0.017
glass-0-1-5_vs_2	0.597	**0.635** ± 0.014	0.204	**0.269** ± 0.021
yeast-0-3-5-9_vs_7-8	0.685	**0.689** ± 0.022	**0.362**	0.361 ± 0.033
yeast-0-2-5-7-9_vs_3-6-8	0.745	**0.748** ± 0.015	**0.429**	0.429 ± 0.024
yeast-0-2-5-6_vs_3-7-8-9	0.868	**0.878** ± 0.013	0.639	**0.676** ± 0.019
ecoli-0-4-6_vs_5	0.866	**0.870** ± 0.000	0.760	**0.763** ± 0.000
ecoli-0-1_vs_2-3-5	**0.842**	0.836 ± 0.000	**0.675**	0.661 ± 0.000
ecoli-0-2-6-7_vs_3-5	0.820	**0.838** ± 0.066	0.693	**0.713** ± 0.090
glass-0-4_vs_5	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-3-4-6_vs_5	0.894	**0.894** ± 0.005	**0.838**	0.836 ± 0.046
ecoli-0-3-4-7_vs_5-6	**0.884**	0.872 ± 0.006	**0.754**	0.738 ± 0.045
yeast-0-5-6-7-9_vs_4	0.718	**0.745** ± 0.036	0.407	**0.458** ± 0.045
vowel0	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
ecoli-0-6-7_vs_5	0.832	**0.834** ± 0.005	0.691	**0.697** ± 0.030
glass-0-1-6_vs_2	**0.651**	0.642 ± 0.004	**0.330**	0.317 ± 0.008
ecoli-0-1-4-7_vs_2-3-5-6	0.835	**0.847** ± 0.002	0.632	**0.681** ± 0.016

Sigue en la página siguiente.

Conjuntos de Datos	AUC		F-M	
	SMOTE-D	SMOTE	SMOTE-D	SMOTE
led7digit-0-2-4-5-6-7-8-9_vs_1	0.847	**0.848** ± 0.000	0.513	**0.520** ± 0.000
glass-0-6_vs_5	0.868	0.868 ± 0.000	0.757	**0.759** ± 0.000
ecoli-0-1_vs_5	0.990	**0.990** ± 0.000	0.920	**0.920** ± 0.000
glass-0-1-4-6_vs_2	0.625	**0.682** ± 0.014	0.273	**0.346** ± 0.033
glass2	**0.657**	0.640 ± 0.008	**0.297**	0.276 ± 0.006
ecoli-0-1-4-7_vs_5-6	**0.886**	0.884 ± 0.000	**0.757**	0.754 ± 0.000
cleveland-0_vs_4	**0.581**	0.562 ± 0.013	**0.176**	0.165 ± 0.031
ecoli-0-1-4-6_vs_5	**0.869**	0.868 ± 0.006	**0.776**	0.767 ± 0.061
shuttle-c0-vs-c4	**0.996**	0.995 ± 0.000	**0.995**	0.995 ± 0.000
yeast-1_vs_7	**0.723**	0.707 ± 0.004	**0.348**	0.326 ± 0.013
glass4	0.930	**0.931** ± 0.006	0.721	**0.738** ± 0.064
ecoli4	0.917	**0.918** ± 0.000	0.814	**0.828** ± 0.000
page-blocks-1-3_vs_4	**0.953**	0.950 ± 0.002	**0.804**	0.800 ± 0.028
abalone9-18	**0.790**	0.765 ± 0.024	0.489	**0.494** ± 0.038
glass-0-1-6_vs_5	**0.935**	0.930 ± 0.000	**0.753**	0.747 ± 0.000
shuttle-c2-vs-c4	1.000	1.000 ± 0.000	1.000	1.000 ± 0.000
yeast-1-4-5-8_vs_7	0.592	**0.621** ± 0.058	0.132	**0.157** ± 0.042
glass5	**0.890**	0.890 ± 0.000	0.647	**0.647** ± 0.000
yeast-2_vs_8	0.771	**0.784** ± 0.000	**0.421**	0.419 ± 0.000
yeast4	0.721	**0.753** ± 0.024	0.298	**0.325** ± 0.024
yeast-1-2-8-9_vs_7	**0.628**	0.624 ± 0.042	**0.158**	0.152 ± 0.036
yeast5	**0.910**	0.909 ± 0.000	0.706	**0.709** ± 0.000
ecoli-0-1-3-7_vs_2-6	**0.840**	0.839 ± 0.000	**0.533**	0.513 ± 0.000
yeast6	**0.830**	0.825 ± 0.002	0.370	**0.387** ± 0.013
abalone19	0.541	**0.541** ± 0.001	0.038	**0.038** ± 0.000
Promedio	0.828	**0.832** ± 0.009	0.648	**0.655** ± 0.018

Fin de la Tabla A.19.

Bibliografía

[Alcalá et al., 2010] Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2010). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287.

[Alejo et al., 2013] Alejo, R., Valdovinos, R. M., García, V., and Pacheco-Sanchez, J. (2013). A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters*, 34(4):380–388.

[Baldi and Long, 2001] Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519.

[Barandela et al., 2003] Barandela, R., Sánchez, J. S., Garcıa, V., and Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851.

[Batista et al., 2004] Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning

training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29.

[Batuwita and Palade, 2013] Batuwita, R. and Palade, V. (2013). Class imbalance learning methods for support vector machines. *Imbalanced learning: Foundations, algorithms, and applications*, 83.

[Bunkhumpornpat et al., 2009] Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in knowledge discovery and data mining*, pages 475–482. Springer.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357.

[Deepa and Punithavalli, 2011] Deepa, T. and Punithavalli, M. (2011). An e-smote technique for feature selection in high-dimensional imbalanced dataset. In *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, volume 2, pages 322–324. IEEE.

[Dong and Wang, 2011] Dong, Y. and Wang, X. (2011). A new over-sampling approach: random-smote for learning from imbalanced data sets. In *Knowledge Science, Engineering and Management*, pages 343–352. Springer.

[Ducange et al., 2010] Ducange, P., Lazzerini, B., and Marcelloni, F. (2010). Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets. *Soft Computing*, 14(7):713–728.

[Estabrooks et al., 2004] Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36.

[García et al., 2012] García, V., Sánchez, J. S., and Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21.

[Han et al., 2005] Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *Advances in intelligent computing*, pages 878–887. Springer.

[Hart, 1968] Hart, P. (1968). The condensed nearest neighbor rule (corresp.). *IEEE Transactions on Information Theory*, 14(3):515–516.

[Hu et al., 2009] Hu, S., Liang, Y., Ma, L., and He, Y. (2009). Msmote: improving classification performance when training data is imbalanced. In *2009 Second International Workshop on Computer Science and Engineering*, pages 13–17. IEEE.

[Koto, 2014] Koto, F. (2014). Smote-out, smote-cosine, and selected-smote: An enhancement strategy to handle imbalance in data level. In *Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on*, pages 280–284. IEEE.

[Kubat et al., 1997] Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA.

[Laurikkala, 2001] Laurikkala, J. (2001). *Improving identification of difficult small classes by balancing class distribution*. Springer.

[Lin and Chen, 2013] Lin, W.-J. and Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics*, 14(1):13–26.

[Liu et al., 2009] Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

[López et al., 2013] López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141.

[Luengo et al., 2011] Luengo, J., Fernández, A., García, S., and Herrera, F. (2011). Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling. *Soft Computing*, 15(10):1909–1936.

[Lunardon et al., 2014] Lunardon, N., Menardi, G., and Torelli, N. (2014). Rose: A package for binary imbalanced learning. *A peer-reviewed, open-access publication of the R Foundation for Statistical Computing*, page 79.

[Maciejewski and Stefanowski, 2011] Maciejewski, T. and Stefanowski, J. (2011). Local neighbourhood extension of smote for mining imbalanced data. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, pages 104–111. IEEE.

[Ramentol et al., 2012] Ramentol, E., Caballero, Y., Bello, R., and Herrera, F. (2012). Smote-rsb*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowledge and information systems*, 33(2):245–265.

[Sáez et al., 2015] Sáez, J. A., Luengo, J., Stefanowski, J., and Herrera, F. (2015). Smote–ipf: Addressing the noisy and borderline examples pro-

blem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291:184–203.

[Seiffert et al., 2010] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2010). Rusboost: A hybrid approach to alleviating class imbalance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 40(1):185–197.

[Stefanowski and Wilk, 2008] Stefanowski, J. and Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 283–292. Springer.

[Tomek, 1976] Tomek, I. (1976). Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.*, 6:769–772.

[Verbiest et al., 2012] Verbiest, N., Ramentol, E., Cornelis, C., and Herrera, F. (2012). Improving smote with fuzzy rough prototype selection to detect noise in imbalanced classification data. In *Advances in Artificial Intelligence–IBERAMIA 2012*, pages 169–178. Springer.

[Vivar, 2008] Vivar, A. I. S. (2008). Métodos de generación de instancias sintéticas mediante clustering y jittering. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, México.

[Wang et al., 2012] Wang, J., You, J., Li, Q., and Xu, Y. (2012). Extract minimum positive and maximum negative features for imbalanced binary classification. *Pattern Recognition*, 45(3):1136–1145.

[Wang et al., 2015] Wang, K.-J., Adrian, A. M., Chen, K.-H., and Wang, K.-M. (2015). A hybrid classifier combining borderline-smote with airs

algorithm for estimating brain metastasis from lung cancer: A case study in taiwan. *Computer methods and programs in biomedicine*, 119(2):63–76.

[Wilson, 1972] Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *Systems, Man and Cybernetics, IEEE Transactions on*, (3):408–421.

[Zong et al., 2013] Zong, W., Huang, G.-B., and Chen, Y. (2013). Weighted extreme learning machine for imbalance learning. *Neurocomputing*, 101:229–242.