# Método de Selección de Atributos por Clase

por

**Bárbara Berenice Pineda Bautista**

Tesis sometida como requisito parcial para obtener el grado de

**MAESTRO EN CIENCIAS EN EL ÁREA DE CIENCIAS COMPUTACIONALES**

en el

**Instituto Nacional de Astrofísica, Óptica y Electrónica**

2009

Tonantzintla, Puebla

Supervisada por:

**Dr. Jesús Ariel Carrasco Ochoa, INAOE**
**Dr. José Francisco Martínez Trinidad, INAOE**

©INAOE 2009
El autor otorga al INAOE el permiso de reproducir y distribuir copias en su totalidad o en partes de esta tesis

# Resumen

En muchos dominios se requiere resolver problemas de clasificación donde los objetos de estudio están descritos con una gran cantidad de atributos, muchos de los cuales pueden ser atributos redundantes y/o irrelevantes, con el objetivo de mejorar la calidad de clasificación es conveniente eliminar este tipo de atributos. La selección de atributos ha sido ampliamente utilizada para la eliminación de atributos redundantes y/o irrelevantes y puede realizarse de 2 maneras:

1. Selección de atributos para todas las clases.
2. Selección de atributos por clase.

La selección de atributos por clase surge con la idea de que cada clase de un problema de clasificación puede exhibir propiedades diferentes y por lo tanto, ser descrita mejor por un subconjunto diferente de atributos. En esta tesis, se propone un método de selección de atributos por clase. El método propuesto para la selección de atributos por clase permite, aplicando la técnica de binarización *uno-contra-todos*, utilizar los selectores de atributos que realizan una selección de atributos para todas las clases. Además, debido a que los clasificadores supervisados no permiten un subconjunto de atributos diferente para cada clase, se propone utilizar la selección de atributos por clase mediante un ensamble de clasificadores y una estrategia de decisión para clasificar nuevos objetos. Los resultados experimentales muestran que en la mayoría de los casos se mejora la calidad de clasificación cuando se realiza selección de atributos por clase comparado con la selección de atributos para todas las clases o con la opción de no hacer selección de atributos.

# Abstract

In many domains it is required to solve classification problems where the objects of study are described with a large number of features, many of which can be redundant and/or irrelevant. In order to improve the quality in classification it is necessary to eliminate this kind of features. Feature selection has been widely used for the elimination of redundant and/or irrelevant features. There are two types for feature selection:

1. Feature selection for all classes.
2. Feature selection by class.

Feature selection by class emerges with the idea that each class of a classification problem may have different properties and it should be described by a different feature subset. In this thesis a method of feature selection by class. The proposed method allows, by applying the *one-against-all* class binarization technique, the use of conventional feature selectors. Because supervised classifiers do not allow using a different feature subset for each class, it is also proposed to use a classifier ensemble and a new strategy decision for taking advantage of feature selection by class. The experimental results showed that in most cases the classification accuracy is improved when feature selection by class is used, compared against feature selection for all classes or without feature selection.

# Dedicatoria

*A mi madre Bárbara Bautista Cruzado, por su gran fortaleza y sabiduría, por ser la mujer que más admiro en el mundo.*

*Má cuando crezca quiero ser como tú.*

*A mi padre Marcos Pineda Márquez, por todas sus enseñanzas y quien un día sabiamente me dijo: “No importa lo que quieras ser, lo que importa es que seas la mejor en lo que hagas”.*

*Pá créeme lo sigo intentando.*

# Agradecimientos

*Al Instituto Nacional de Astrofísica, Óptica y Electrónica, por el excelente trato y apoyo brindado durante mi estancia en esta institución.*

*Al Consejo Nacional de Ciencia y Tecnología (CONACYT) de México, por el apoyo económico proporcionado para mis estudios de maestría.*

*A mis asesores, Dr. Jesús Ariel Carrasco Ochoa y Dr. José Francisco Martínez Trinidad por su paciencia, apoyo y conocimientos transmitidos.*

*A mis sinodales Dr. Eduardo Francisco Morales Manzanares, Dr. Jesús Antonio García Bernal y Dr. Manuel Montes y Gómez, por sus valiosos comentarios y aportaciones para mejorar este trabajo.*

*A mis padres, por todo su amor y por creer en mí, por enseñarme todo lo que sé de la vida, porque no me alcanzan las palabras para decirles cuanto los amo y cuanto agradezco a Dios que sean mis padres.*

*A mis hermanos a quienes admiro, gracias por toda una vida de amor, consejos y apoyo, por ayudarme a seguir cuando siento que ya no puedo.*

*A mi abue por ser mi ángel de la guarda, porque sé que donde quiera que este comparte mis logros.*

*A mi amiga Ale por ser una hermana más para mí, por todo su cariño, consejos y apoyo, por esa gran amistad que ha soportado las ausencias y ha estado incondicionalmente para mí.*

*A Roger por su amor, paciencia y apoyo, por ser mi soporte y enseñarme lo simple que puede ser la vida, gracias por coincidir conmigo y sobre todo gracias por hacerme tan feliz.*

*A todos mis amigos, porque todos son igualmente importantes para mí, porque de cada uno he aprendido sin fin de cosas y sobre todo con quienes he pasado momentos inolvidables, gracias por su apoyo y sobre todo gracias por permitirme ser su amiga.*

# Contenido

**1. Introducción** **1**

**2. Conceptos preliminares** **5**

2.1. Introducción . . . . . 6

2.2. Selección de atributos . . . . . 7

2.2.1. Clasificación de los selectores de atributos . . . . . 7

2.3. Selección de atributos por clase . . . . . 8

2.4. Binarización de clases . . . . . 9

2.4.1. Técnicas de binarización de clases . . . . . 9

2.5. Clases desbalanceadas . . . . . 10

2.5.1. SMOTE (Synthetic Minority Over-sampling Technique) . . 12

2.6. Ensamble de clasificadores . . . . . 13

2.6.1. Técnicas para ensamble de clasificadores . . . . . 15

2.7. Resumen . . . . . 17

**3. Estado del arte** **19**

3.1. Métodos de selección de atributos para todas las clases . . . . . 19

3.2. Métodos de selección de atributos por clase . . . . . 20
3.3. Discusión . . . . . 23
3.4. Resumen . . . . . 25

**4. Método propuesto 27**
4.1. Solución propuesta . . . . . 27
4.1.1. Binarización de clases: *uno-contra-todos* . . . . . 28
4.1.2. Selección de atributos por clase . . . . . 29
4.1.3. Clasificación con la selección de atributos por clase . . . 31
4.2. Resumen . . . . . 34

**5. Experimentación y resultados 35**
5.1. Bases de datos . . . . . 36
5.2. Estimación del error . . . . . 38
5.3. Selectores de atributos y clasificadores utilizados . . . . . 38
5.4. Experimentos realizados . . . . . 40
5.5. Comparación entre la selección de atributos para todas las clases y no hacer selección de atributos . . . . . 41
5.6. Resultados del método propuesto . . . . . 48
5.7. Resumen . . . . . 61

**6. Conclusiones y trabajo futuro 63**
6.1. Conclusiones . . . . . 63
6.2. Trabajo futuro . . . . . 65

**Referencias 67**

**A. Apéndice 73**

# Lista de Figuras

4.1. Binarización de un problema de 3 clases con la técnica *uno-contra-todos*. . . . . 29

4.2. Selección de atributos por clase. . . . . 31

4.3. Proceso de clasificación utilizando la selección de atributos por clase. . . . . 33

5.1. Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador Bayes. . . . . 42

5.2. Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador J48. . . . . 43

5.3. Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=1. . . . . 43

5.4. Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=3. . . . . 43

5.5. Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador MLP. . . . . 44

5.6. Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador Bayes. . . . . 45

5.7. Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador J48. . . . . 45

5.8. Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=1. . . . . 45

5.9. Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=3. . . . . 46

5.10. Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador MLP. . . . . 46

5.11. Comparación del promedio del porcentaje de error de 18 bases de datos entre WEKA con y sin selección de atributos para el selector de atributos e2b4. . . . . 47

5.12. Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador Bayes. . . . . 48

5.13. Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador J48. . . . . 49

5.14. Comparación del promedio de porcentaje de error entre de 14 bases de datos el método propuesto y WEKA para el Clasificador IBk con k=1. . . . . 49

5.15. Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador IBk con k=3. . . . . 49

5.16. Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador MLP. . . . . 50

5.17. Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el Clasificador Bayes. . . . . 52

5.18. Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el Clasificador J48. . . . . 53

5.19. Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=1. . . . . 53

5.20. Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=3. . . . . 53

5.21. Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el clasificador MLP. . . . . 54

5.22. Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el Clasificador Bayes. . . . . 55

5.23. Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el Clasificador J48. . . . . 56

5.24. Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=1. . . . . 56

5.25. Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=3. . . . . 56

5.26. Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el clasificador MLP. . . . . 57

5.27. Comparación del promedio de porcentaje de error de 18 bases de datos entre el método propuesto y WEKA para el selector de atributos e2b4. . . . . 58

# Lista de Tablas

4.1. Ejemplo del criterio de decisión propuesto para la clasificación. . 33

5.1. Resumen de las características de las bases de datos utilizadas. 36

5.2. Número y proporción de objetos por clase en las bases de datos utilizadas. . . . . 37

5.3. Selectores de atributos para todas las clases utilizados en el método propuesto. . . . . 39

5.4. Comparación entre los métodos probados por pares para 14 bases de datos. . . . . 51

5.5. Comparación entre los métodos probados por pares. . . . . 59

5.6. Promedio por clasificador de los selectores de atributos utilizados en los métodos probados. . . . . 60

5.7. Significancia estadística de los mejores resultados del método propuesto con repetición, para cada base de datos utilizada. . . 61

A.2. Promedio del porcentaje de error de 14 bases de datos con los métodos probados por clasificador. . . . . 74

A.4. Promedio del porcentaje de error de 19 bases de datos con los métodos probados por clasificador, los * señalan los selectores de atributos en los que se probaron menos de 19 bases de datos. 75

A.6. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda BestFirst (e1b1). . . . . 77

A.8. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda ExhaustiveSearch (e1b2). . . . . 78

A.10. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda GeneticSearch (e1b3). . . . . 79

A.12. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda GreedyStepwise (e1b4). . . . . 80

A.14. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda RandomSearch (e1b5). . . . . 82

A.16. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda BestFirst (e2b1). . . . . 83

A.18. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda ExhaustiveSearch (e2b2). . . . . 84

A.20. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda GeneticSearch (e2b3). . . . . 85

A.22. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda GreedyStepwise (e2b4). . . . . 87

A.24. Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda RandomSearch (e2b5). . . . . 88

# Capítulo 1

# Introducción

En diversas áreas del conocimiento humano existen problemas de clasificación supervisada (en donde las clases son conocidas), por ejemplo dentro de la medicina una tarea típica consiste en separar a los pacientes saludables de aquellos que padecen cierta enfermedad. De manera general en un problema de clasificación supervisada se cuenta con un conjunto de objetos (patrones) que sirven como información base para decidir la clase a la que pertenece un nuevo objeto. Es decir, la información base disponible es una muestra previamente clasificada en donde cada objeto está relacionado con una clase en particular y será utilizada para entrenar a un clasificador que permita clasificar (etiquetar) nuevos objetos.

Determinar los atributos que se utilizarán para describir a los objetos de un problema de clasificación supervisada es importante, ya que de esto depende la buena clasificación de nuevos objetos. En la práctica no se conocen *a priori* los atributos relevantes para un problema de clasificación supervisada, por lo tanto, una gran cantidad de atributos suelen ser incorporados para tratar de describir mejor al problema. En la actualidad es muy fácil obtener una gran cantidad de atributos, esto se debe principalmente a los avances en la adquisición y almacenamiento de información. Se podría pensar que utilizando un

mayor número de atributos se puede mejorar la clasificación de objetos, desafortunadamente no siempre es así, algunos pueden ser atributos redundantes y/o irrelevantes que afectan la calidad de clasificación.

En el intento por descartar atributos redundantes y/o irrelevantes con la finalidad de aumentar la calidad de clasificación en los problemas de clasificación supervisada, se han aplicado técnicas como la selección de atributos, la cuál se ha vuelto un campo de investigación muy importante para áreas como el reconocimiento de patrones, el aprendizaje automático y la minería de datos.

La selección de atributos consiste en encontrar un subconjunto de atributos del conjunto de atributos originales, de tal manera que aumente la calidad de clasificación, facilite la visualización y entendimiento de los datos, minimice los requerimientos de almacenamiento y reduzca los tiempos de ejecución. Algunos métodos de selección de atributos ponen mayor énfasis en algún aspecto que en otro, esta tesis se enfoca en proponer un método de selección de atributos que permita aumentar la calidad de clasificación.

Comúnmente los métodos de selección de atributos obtienen un solo subconjunto de atributos para representar a todas las clases de un problema de clasificación supervisada. Sin embargo, existen problemas en donde es difícil encontrar un buen subconjunto de atributos para todas las clases, además diferentes atributos pueden tener un poder de discriminación diferente para cada clase; basándose en esta idea actualmente se han propuesto trabajos como [OLS99], [FW05] y [Nan06] dentro de la selección de atributos en los que un subconjunto de atributos posiblemente diferente es elegido para cada clase. Por lo tanto la selección de atributos se puede dividir en:

1. Selección de atributos para todas las clases.
2. Selección de atributos por clase.

En la selección de atributos para todas las clases, un solo subconjunto de atributos es seleccionado para todas las clases de un problema de clasifica-

ción, se asume que los atributos seleccionados tienen el mismo poder para discriminar a cada clase de las demás.

En la selección de atributos por clase, como se mencionó anteriormente, un subconjunto de atributos, posiblemente diferente, es seleccionado para cada clase, asumiendo que diferentes grupos de atributos tienen diferentes capacidades de discriminación para cada clase.

Frecuentemente los objetos para un problema de clasificación son representados considerando sólo atributos numéricos o sólo atributos no numéricos. Sin embargo en los problemas de clasificación del mundo real se presentan ambos tipos de atributos simultáneamente, lo que se conoce como datos mezclados. Además en muchas ocasiones existen objetos en los que el valor de uno o más de sus atributos son desconocidos, a esto se le conoce como datos incompletos o ausencia de información. Un problema de clasificación puede contener datos mezclados e incompletos. Por lo tanto en esta tesis se propone un método de selección de atributos por clase para problemas de clasificación supervisada, que además de identificar los atributos relevantes para cada clase, permita trabajar con datos mezclados y ausencia de información.

El objetivo general de esta tesis consiste en desarrollar un método de selección de atributos por clase que permita trabajar con atributos numéricos y no numéricos, para lo cual se plantean dos objetivos específicos:

1. Desarrollar un método que permita seleccionar aquellos atributos que son relevantes para cada clase de un problema.

2. Definir una estrategia de clasificación que utilice los atributos seleccionados por clase y así permita aprovechar los resultados del método propuesto.

El segundo objetivo específico es importante ya que los clasificadores supervisados no permiten trabajar con una selección de atributos diferente para cada clase.

Esta tesis se encuentra organizada en 6 capítulos, en el capítulo 1 se introduce al tema de la selección de atributos por clase, se dan los objetivos de la tesis y la organización de la misma. En el capítulo 2 se presentan los conceptos y definiciones básicas necesarias para el desarrollo de este trabajo. En el capítulo 3 se hace un análisis de los trabajos previos que han sido propuestos en el campo de la selección de atributos, tanto en la selección de atributos para todas las clases como en la selección de atributos por clase. En el capítulo 4 se introduce el método de selección de atributos por clase propuesto, explicando cada una de las fases que lo conforman. En el capítulo 5 se muestran los resultados de los experimentos realizados con el método de selección de atributos por clase propuesto. Finalmente en el capítulo 6 se exponen las conclusiones y aportaciones del trabajo de tesis presentado, así como el trabajo futuro.

# Capítulo 2

# Conceptos preliminares

En este capítulo se abordarán los conceptos básicos necesarios para entender las técnicas utilizadas en el método de selección de atributos por clase propuesto en esta tesis. En el método propuesto en esta tésis, para realizar una selección de atributos por clase se aplica la tecnica de binarización $uno - contra - todos$. Debido a que al aplicar la técnica de binarización $uno - contra - todos$ se puede generar un problema desbalanceado, en esta tesis se propone realizar un balanceo de clases para la selección de atributos, con la finalidad de que los atributos seleccionados discriminen mejor a la clase para la que fueron seleccionados de las demás clases. Posteriormente para la clasificación de nuevos objetos se emplea un ensamble de clasificadores y se propone una estrategía de clasificación. Antes de mencionar como se aplica la binarización y balanceo de clases, así como el ensamble de clasificadores en el método propuesto, es necesario presentar algunas definiciones para comprender mejor estas técnicas.

## 2.1. Introducción

Dentro del reconocimiento de patrones, aprendizaje automático y minería de datos se estudian problemas relacionados con la clasificación de objetos, principalmente los problemas de clasificación supervisada. En la clasificación supervisada se cuenta con un conjunto de objetos previamente clasificados que sirven como base para clasificar nuevos objetos.

En la clasficación de objetos se utilizan además técnicas como la extracción de atributos y la selección de atributos, para mejorar la calidad de clasificación. En la extracción de atributos, nuevos atributos son derivados de los atributos originales para incrementar la eficiencia computacional y la exactitud en la clasificación [FW05]. Para realizar una extracción de atributos se utilizan combinaciones lineales, las cuales son útiles, simples de implementar y analíticamente tratables. Un método lineal proyecta datos de alta dimensión en un espacio de dimensión menor. Hay dos métodos comúnmente utilizados para encontrar trasformaciones lineales efectivas. Uno de ellos es el Análisis de Componentes Principales (PCA), el cual busca una proyección que mejor represente los datos. Otro método es conocido como Análisis de Discriminante Múltiple (MDA), el cual busca una proyección que separe mejor los datos [DHS00].

La selección de atributos puede definirse como un proceso que elige un subconjunto mínimo de atributos del conjunto original de atributos. En la clasificación supervisada la selección de atributos es a menudo vista como un problema de búsqueda en el espacio de atributos, de tal manera que el objetivo de la selección de atributos dentro de la clasificación supervisada es encontrar el subconjunto de atributos que proporcione la más alta calidad de clasificación [DB04].

## 2.2. Selección de atributos

La selección de atributos ha sido un campo de investigación y desarrollo desde los años 70's en el área del reconocimiento de patrones, aprendizaje automático y minería de datos [LY02]. Algunos beneficios de la selección de atributos son:

1. Aumentar la calidad de la clasificación.
2. Reducir los requerimientos de almacenamiento de datos (al reducir el tamaño de los datos).
3. Reducir los tiempos de ejecución.

### 2.2.1. Clasificación de los selectores de atributos

En general, los métodos de selección de atributos se distinguen por su forma de evaluar atributos y se pueden clasificar [GE03] en:

1. Filter: Utilizan un criterio de evaluación para seleccionar atributos que se basa únicamente en las propiedades de los datos, por lo que es independiente a un clasificador.
2. Wrapper: Utilizan el desempeño de algún clasificador para evaluar a los subconjuntos de atributos. El inconveniente que presentan los métodos Wrapper es que son más costosos computacionalmente que los Filter aunque suelen obtener mejores resultados.
3. Híbridos: Es una combinación de los dos anteriores, utilizan un criterio de evaluación y un clasificador para elegir un subconjunto de atributos.

Hay otra forma de clasificar a los métodos de selección de atributos [FW05] en:

1. Selección de atributos para todas las clases.
2. Selección de atributos por clase.

En la selección de atributos para todas las clases, un solo subconjunto de atributos es seleccionado para todas las clases, se asume que los atributos seleccionados tienen el mismo poder de discriminación para todas las clases de un problema de clasificación. La mayoría de trabajos desarrollados en selección de atributos son de este tipo. En la selección de atributos por clase un subconjunto de atributos, posiblemente diferente, es elegido para cada clase de un problema de clasificación.

## 2.3. Selección de atributos por clase

La selección de atributos por clase es un campo de estudio que surge en el reconocimiento de patrones para mejorar la calidad de clasificación. En [OLS99] afirman ser los primeros en utilizar el término de selección de atributos por clase, aunque en trabajos como [JR98] ya se había descrito una técnica para seleccionar subconjuntos de atributos específicos para discriminar entre clases.

La selección de atributos por clase se basa en la idea de que los atributos tienen diferente poder de discriminación para cada clase de un problema y el grado de discriminación de cada atributo depende de la naturaleza del mismo. Un atributo puede discriminar fuertemente o débilmente a una clase de las demás, es decir, un atributo con débil poder de discriminación permite distinguir un subconjunto de clases del resto de las clases de un problema de clasificación. Por ejemplo si se cuenta con 5 clases $w_1$, $w_2$, $w_3$, $w_4$, $w_5$; un atributo que nos ayuda a distinguir las clases $w_1$ y $w_2$ de las clases $w_3$, $w_4$ y $w_5$ es un atributo con débil poder de discriminación. Un atributo con fuerte poder de discriminación permite distinguir a una clase del resto de las clases de un problema de clasificación. Por ejemplo si se cuenta con 5 clases $w_1$, $w_2$, $w_3$,

$w_4$, $w_5$ un atributo que nos ayuda a distinguir la clase $w_1$ de las clases $w_2$, $w_3$, $w_4$ y $w_5$ es un atributo con fuerte poder de discriminación, en este caso para la clase $w_1$.

En [Bai02] se define a los atributos por clase como atributos con fuerte poder de discriminación para las clases que han sido seleccionados.

## 2.4. Binarización de clases

En problemas de clasificación del mundo real es común tener más de dos clases, algunos clasificadores son inherentemente binarios pues sólo son capaces de discriminar entre dos clases. Hay dos principales enfoques para aplicar tales clasificadores binarios a los problemas de clasificación multiclase, una es generalizar a los clasificadores binarios para trabajar con problemas de clasificación multiclase y otra es transformar los problemas de clasificación multiclase en una serie de problemas binarios. El segundo enfoque es conocido como binarización de clases (*class binarization*) la cual es definida en [Für02] como:

La binarización de clases es un mapeo de un problema de clasificación multiclase a una serie de problemas de clasificación binarios, de tal manera que permita la derivación de un solo resultado para el problema multiclase utilizando los resultados de los clasificadores binarios.

### 2.4.1. Técnicas de binarización de clases

Son dos las técnicas de binarización de clases más utilizadas: *uno-contra-todos (one-against-all) y uno-contra-uno (one vs one, round robin classification, parwise classification)*:

- *Uno-contra-todos*

La técnica *uno-contra-todos* es definida en [Für02] como:

La binarización de clase *uno-contra-todos* transforma un problema de $c$ clases en $c$ problemas de dos clases. Éstos son construidos usando los objetos de la clase $i$ como los ejemplos positivos y los objetos de las otras clases $j$ $(j = 1...c,\ j \neq i)$ como los ejemplos negativos. Para la clasificación se entrena un clasificador por cada problema binario, con lo que se cuenta con $c$ clasificadores binarios que requieren un criterio externo para combinar sus predicciones individuales en una predicción final. Por ejemplo tomando en cuenta la confianza de las predicciones, es decir, se elige el clasificador con el valor más alto en su salida.

- *Uno-contra-uno*

La técnica *uno-contra-uno* es definida en [Für02] como:

La técnica de binarización *uno-contra-uno* transforma un problema de $c$ clases en $c\,(c-1)\,/2$ problemas de dos clases $< i, j >$, uno para cada conjunto de clases $\{i, j\}$ , $i = 1...c-1,\ j = i+1...c$. El clasificador binario para el problema $< i, j >$ es entrenado con los objetos de las clases $i$ y $j$, donde los objetos de las demás clases $k \neq i, j$ son ignoradas. Para la clasificación se entrena un clasificador por cada problema binario, con lo que se cuenta con $c\,(c-1)\,/2$ clasificadores binarios cuyas predicciones se combinan mediante votación simple.

## 2.5. Clases desbalanceadas

El problema de las clases desbalanceadas se presenta cuando en un problema de clasificación supervisada se tiene una gran cantidad de ejemplos para una clase (clase mayoritaria) y sólo unos pocos para otra (clase minoritaria) [Jap00]. Este problema ha cobrado un gran interés ya que se presenta comúnmente en aplicaciones del mundo real. Algunos ejemplos en donde existen estos problemas [JS02] son:

1. Detección de fraudes, por ejemplo en llamadas telefónicas o transacciones bancarias.
2. Diagnóstico de fallas, por ejemplo en sistemas comerciales o de la milicia.
3. Diagnóstico médico, por ejemplo de enfermedades raras o que muy pocas personas padecen.
4. Percepción remota, por ejemplo detección de derrames de petróleo en imágenes satelitales.

Por ejemplo, si se tienen 1 000 objetos de los cuales sólo 50 pertenecen a la clase minoritaria, es común que un clasificador clasifique correctamente a todos los objetos de la clase mayoritaria teniendo una exactitud del 95 %, aunque clasifique mal a todos los objetos de la clase minoritaria. Esto representa un verdadero problema ya que a menudo es la clase minoritaria la que se desea clasificar correctamente.

En respuesta a este problema se han generado diferentes propuestas que se pueden dividir de la siguiente manera:

1. Mejores medidas para evaluar el desempeño de clasificadores ante clases desbalanceadas.
2. Técnicas que favorezcan la clasificación de la clase minoritaria.
3. Balanceo de clases.

Entre las técnicas que favorecen la clasificación de la clase minoritaria se encuentran los algoritmos que realizan aprendizaje sensible al costo, éste consiste en introducir un costo mayor a los objetos mal clasificados de la clase minoritaria que a los objetos de la clase mayoritaria. Un ejemplo muy común se encuentra en aplicaciones médicas donde se pueden cometer errores del tipo “falso positivo” (dar como enfermo a un paciente sano) y “falso negativo”

(dar como sano a un paciente enfermo), un error del tipo “falso positivo” tiene un costo menor ya que con estudios posteriores al paciente se le puede descartar cierta enfermedad, pero cometer un error del tipo “falso negativo” tiene un costo mayor ya que implicaría no hacer más pruebas a un paciente provocándole incluso hasta la muerte.

Existen dos formas de balancear las clases [JS02], ya sea haciendo sobre-muestro (over-sampling) que consiste en generar más objetos de la clase minoritaria o sub-muestreo (under-sampling) que consiste en eliminar objetos de la clase mayoritaria. Cualquiera de las dos formas se puede realizar de manera aleatoria o “inteligente”. La forma aleatoria, es la manera más sencilla de balancear las clases y ha presentado buenos resultados, aunque puede tener ciertos inconvenientes como eliminar objetos de la clase mayoritaria que pueden ser útiles o generar objetos de la clase minoritaria que pueden ser ruido, mientras que si se realiza de manera “inteligente” se esperaría generar únicamente objetos útiles de la clase minoritaria y no eliminar objetos importantes de la clase mayoritaria, aunque no siempre es así. Un método muy utilizado es el algoritmo de SMOTE el cual es descrito en la siguiente sección.

### 2.5.1. SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE es un algoritmo de sobre-muestreo de objetos de la clase minoritaria. Este algoritmo genera objetos sintéticos a partir de los $k$ vecinos más cercanos de objetos elegidos de manera aleatoria. SMOTE opera sobre el espacio de atributos en lugar del espacio de datos y fue propuesto con un $k = 5$ [CBHK02].

El algoritmo de SMOTE realiza los siguientes pasos:

- Recibe como parámetro el porcentaje de objetos a sobre-muestrear.
- Calcula el número de objetos a generar.
- Calcula los $k$ vecinos más cercanos de los objetos de la clase minoritaria.

- Genera los objetos sintéticos de la siguiente manera:
    - Para cada objeto de la clase minoritaria, elige aleatoriamente el vecino a utilizar para crear el nuevo objeto.
    - Calcula la diferencia entre cada uno de los atributos del objeto de la clase minoritaria y el vecino elegido.
    - Multiplica la diferencia obtenida por un número aleatorio entre 0 y 1.
    - Suma el valor obtenido en la multiplicación al valor del objeto de la clase minoritaria.
    - Regresa el conjunto de ejemplos sintéticos.

SMOTE puede presentar los siguientes inconvenientes:

1. Generar muchos objetos sintéticos cuyas semillas son objetos con ruido.
2. Al generar un nuevo objeto interpola entre dos objetos de la clase minoritaria, pero pueden existir muchos objetos cercanos de la clase mayoritaria generando modelos incorrectos.
3. No tiene una forma clara de determinar cuántos objetos generar.

## 2.6. Ensamble de clasificadores

Un método utilizado para incrementar la calidad de clasificación es entrenar varios clasificadores sobre un solo problema y entonces combinar los resultados de los diferentes clasificadores en un solo resultado. Hay varias razones para combinar múltiples clasificadores para resolver un problema de clasificación [JDM00]. Algunas de éstas son:

1. Se pueden tener diferentes clasificadores que trabajen en un contexto diferente y en una representación diferente del mismo problema. Por ejemplo identificar a una persona por su voz, rostro y escritura.

2. Algunas veces más de un solo conjunto de entrenamiento está disponible, colectado en un tiempo o ambiente diferente. Estos conjuntos de entrenamiento pueden usar diferentes atributos y pueden ser usados en clasificadores diferentes.

3. Diferentes clasificadores entrenados en un mismo conjunto de datos no sólo pueden diferir en su desempeño global, sino que cada clasificador puede tener su región en el espacio de atributos donde se desempeñe mejor.

4. Algunos clasificadores como las redes neuronales muestran diferentes resultados con diferentes inicializaciones en el proceso de entrenamiento. En lugar de seleccionar la mejor red y descartar las otras, se pueden combinar varias redes tomando ventaja de todos los intentos por aprender a partir de los datos.

Los ensambles de clasificadores son divididos en tres categorías de acuerdo a su arquitectura:

1. Paralelo. Todos los clasificadores son invocados de manera independiente y sus resultados combinados.

2. Cascada. Los clasificadores son invocados de manera consecutiva. El número de posibles clases para un objeto es reducido gradualmente conforme más clasificadores son invocados en la secuencia.

3. Jerárquico. Clasificadores individuales son combinados en una estructura similar a la de un clasificador de árboles de decisión, los nodos del árbol ahora son clasificadores complejos que utilizan una gran cantidad de atributos. La ventaja de esta arquitectura es su eficiencia y flexibilidad en explotar el poder de discriminación de diferentes tipos de atributos.

Cualquiera que sea la arquitectura a utilizar, para crear un ensamble de clasificadores se deben tomar en cuenta dos aspectos:

1. Qué clasificadores y cómo se utilizarán dentro del ensamble.
2. Cómo combinar los resultados individuales de los clasificadores en un solo resultado.

Respecto a la elección del clasificador a utilizar y cómo utilizarlo, se tienen diferentes enfoques que van desde utilizar el mismo tipo de clasificador en todo el ensamble o utilizar diferentes tipos de clasificadores y entrenar éstos sobre diferentes subconjuntos del conjunto de entrenamiento o entrenarlos con diferentes subconjuntos de atributos.

El segundo aspecto es cómo combinar el resultado de los clasificadores utilizados en uno solo, para lo cual se pueden emplear diferentes estrategias como por ejemplo votación, votación pesada o promedio.

### 2.6.1. Técnicas para ensamble de clasificadores

Recientemente se han desarrollado técnicas para construir conjuntos de clasificadores y combinarlos para clasificar nuevos objetos [OM99, Die00]. Entre las principales técnicas utilizadas para el ensamble de clasificadores se encuentran:

1. *Bagging*. Genera clasificadores de varias muestras de los objetos. Una muestra de objetos *bootstrap* se crea al muestrear uniformemente $m$ instancias del conjunto de entrenamiento con reemplazo. Se obtienen $T$ muestras, $B_1, ..., B_T$ y se construye un clasificador $C_i$ para cada muestra. Finalmente se combinan las salidas de cada clasificador usando alguna regla, por ejemplo el promedio.

2. *Boosting*. Genera un conjunto de clasificadores de manera secuencial. A todos los ejemplos les asigna inicialmente un peso igual, cada vez que se genera un clasificador, se cambian los pesos de los nuevos ejemplos usados para el siguiente clasificador. La idea es forzar al nuevo clasificador a minimizar el error esperado. Para esto se les asigna más peso a los ejemplos mal clasificados y menos a los bien clasificados. El clasificador final se forma usando un esquema de votación pesado que depende del desempeño de cada clasificador en su conjunto de entrenamiento.

3. *Stacking*. Se construyen $N$ clasificadores a partir de los datos usando algoritmos diferentes. Para producir una clasificación se utiliza un meta-algoritmo que emplea la salida de los clasificadores base como atributos de un nuevo clasificador.

Otras varientes de las técnicas utilizadas en el ensamble de clasificadores son [BK99]:

1. *Arcing*. Genera un conjunto secuencial de clasificadores. Como en *Boosting* asigna pesos a los ejemplos que es proporcional al número de errores que tuvo el último clasificador elevado a la cuarta potencia más uno.

2. Introducir aleatoriedad. Una forma de generar diversos clasificadores es introduciendo aleatoriedad en éstos. Por ejemplo, en redes neuronales los pesos iniciales pueden ser elegidos aleatoriamente. Si en un clasificador se entrena con el mismo conjunto de entrenamiento pero con diferentes pesos iniciales, el resultado del clasificador puede ser diferente.

3. *Error-correcting output code*s. Genera una codificación para cada clase de un problema de clasificación, mediante un código que permita fácilmente identificar errores en un clasificador.

4. *Wagging*. Es una variante de *Bagging* que introduce ruido gaussiano con media cero a cada peso de los objetos.

5. *Backfitting*. Es como *Bagging*, sólo que los clasificadores son entrenados con todos los ejemplos para obtener mejores estimaciones de su desempeño.

6. *Feature selection ensembles*. Surgen como una variante del ensamble de clasificadores que consiste en entrenar a un conjunto de clasificadores con diferentes subconjuntos de atributos del conjunto de atributos original [Gar07]. Así un ensamble de selección de atributos tiene como objetivo encontrar un conjunto de subconjuntos de atributos que no sólo proporcionen una alta calidad de clasificación, si no que además proporcionen diversidad entre los clasificadores base al permitir que cada clasificador trabaje con un subconjunto de atributos diferente [TPC05].

## 2.7. Resumen

En este capítulo se presentaron las definiciones de clasificación supervisada, selección de atributos (para todas las clases y por clase), balanceo de clases, ensamble de clasificadores y binarización de clases. Las definiciones proporcionadas anteriormente son necesarias debido a que en esta tesis se propone un método de selección de atributos por clase en el cual se aplica la técnica de binarización *uno-contra-todos*, por lo que se sugiere balancear los datos dado que al aplicar dicha técnica de binarización se producen problemas con clases desbalanceadas, se propone además un ensamble de clasificadores con una arquitectura en paralelo. La forma en la que se utiliza cada una de las técnicas descritas anteriormente en el método propuesto se explican con más detalle en el capítulo 4. En el siguiente capítulo se expondrán los métodos propuestos en la selección de atributos por clase, para poder observar las diferencias y ventajas que presenta el método propuesto.

# Capítulo 3

# Estado del arte

Como se mencionó en el capítulo 2, la selección de atributos se puede dividir en selección de atributos para todas las clases y en selección de atributos por clase. Por lo tanto en este capítulo se analizarán trabajos realizados en ambos tipos de selección de atributos, enfocándose más en la selección de atributos por clase.

## 3.1. Métodos de selección de atributos para todas las clases

La mayoría de los trabajos propuestos en selección de atributos realizan una selección de atributos para todas las clases, como se muestra en [DL97] en donde se presenta un estudio hasta 1997 de 32 métodos de selección de atributos. En estos métodos un solo subconjunto de atributos es obtenido para todas las clases de un problema de clasificación y sólo se manejan atributos numéricos. Muchos métodos de selección de atributos para todas las clases sólo son mejoras de métodos propuestos anteriormente, en donde por ejemplo diferentes métricas son aplicadas o están desarrolladas para un área de

investigación en específico, o bien se combinan con diferentes técnicas como máquinas de soporte vectorial, redes neuronales y algoritmos genéticos. Debido a que en los problemas del mundo real como en medicina se cuenta con atributos numéricos y no numéricos, es importante el desarrollo de métodos de selección de atributos para todas las clases que permitan utilizar datos mezclados. Para este tipo de problemas se han propuesto trabajos como en [TM07], donde aunque permite trabajar con datos mezclados, se utiliza un solo subconjunto de atributos para discriminar entre todas las clases.

La idea de seleccionar varios subconjuntos de atributos para un mismo problema de clasificación surge con los ensambles de clasificadores, a esta técnica se le conoce como ensamble de selección de atributos. Ejemplos de estos trabajos son los de [SF07] y [SD05], en donde aunque seleccionan atributos para cada clasificador, se elige para cada uno de ellos un solo subconjunto de atributos para todas las clases. En [SD05], se utiliza la combinación de múltiples clasificadores y cada uno de éstos trabaja con diferentes subconjuntos de atributos. Este método en un paso inicial elige un subconjunto de atributos y en cada iteración se seleccionan subconjuntos de atributos de los atributos que fueron descartados en pasos previos, hasta que todos los atributos son incluidos en algún subconjunto. Este trabajo tiene el inconveniente de que los atributos irrelevantes o redundantes son tomados en cuenta. En [SF07] se enfoca al mismo problema de múltiples clasificadores pero en este trabajo no usan todos los atributos y para este propósito se emplean diferentes selectores de atributos.

## 3.2. Métodos de selección de atributos por clase

Antes de que se utilizara la selección de atributos por clase, se propuso el término de atributos específicos de clase, pero esto dentro de la extracción de atributos. En trabajos como [Bag98], [Bag99] y [Bag02] se presenta un nuevo enfoque para obtener un subconjunto de atributos para cada clase de un problema de clasificación y de éstos se desprenden más trabajos sobre

extracción de atributos específicos de clase como [BN00], [Bag04], [BB04] y [Kay00]. En [BN00], se implementa una extensión del método de extracción de atributos específicos de clase utilizando modelos ocultos de Markov o solo prueban la existencia de los atributos específicos de la clase, como en [Kay00].

En [OHJ03] se utiliza el término atributos por clase en lugar de atributos específicos de clase, al igual que los anteriores se aplica para la extracción de atributos aplicando modelos ocultos de Markov en un sistema de reconocimiento automático de voz. Este trabajo analiza a los atributos por clase con fuerte poder de discriminación.

Aunque los trabajos mencionados anteriormente son aplicados a la extracción de atributos han servido de base para los trabajos en donde se aborda el problema de la selección de atributos por clase.

En [OLS99] se propone la selección de atributos por clase para mejorar el desempeño de los clasificadores en el reconocimiento de patrones, partiendo de la idea de que los atributos tienen diferente poder de discriminación en diferentes clases. La elección de atributos se realiza evaluando la separación de las clases con cada atributo, la cual es una medida que muestra que tan bien un atributo separa a las clases en el espacio de atributos.

En [OLS99] se considera un problema con $g$ clases, cada una representada por $w_i$, $i = 1, 2, ..., g$; un vector de atributos $X$ es un vector de dimensión $d$ tal que $X = (x_0, x_1, ..., x_{d-1})$. Además se propone a $S^{cg}(w_i, \Omega, X)$ como medida de separación entre la clase $w_i$ y un grupo de $g - 1$ clases donde $\Omega = \{w_k | 1 \leq k \leq g, k \neq i\}$ por un vector $X$. Para la selección del subconjunto de atributos para cada clase se calcula la separación de clase para cada atributo de manera independiente utilizando $S^{cg}$, posteriormente se ordenan los atributos de acuerdo a su valor de separación de clase de mayor a menor y se fija el número de atributos a seleccionar para cada subconjunto de atributos en $d/2$.

Debido a que se tienen $g$ diferentes subconjuntos de atributos (uno para cada clase), se propone un clasificador modular de redes neuronales. Este consiste

de $g$ subredes cada una correspondiente a una clase $w_i$, en donde se organizan los objetos de entrenamiento en dos grupos, los ejemplos positivos y lo ejemplos negativos, los ejemplos positivos corresponden a la clase $w_i$ y los negativos a $\Omega = \{w_k | 1 \leq k \leq g, k \neq i\}$. Así cada clase $w_i$ tiene su propio subconjunto de atributos que alimentan a una subred $M_{wi}$ que a su vez produce una sola salida $O_{wi}$. Un nuevo ejemplo es clasificado dentro de la clase $w_i$ con el valor máximo en su salida.

En [FW02] y [FW05] se seleccionan atributos por clase y se construye un solo clasificador de función de base radial (RBF por sus siglas en inglés), al contrario de trabajos como [OLS99], en éstos una sola red RBF es requerida para seleccionar atributos por clase en problemas de clasificación multiclase. Estos trabajos tienen como base la idea de que una red neuronal RBF cuenta con un conjunto de unidades ocultas, en donde cada unidad oculta sirve para identificar a una clase en particular y por lo tanto un subconjunto de unidades ocultas sirve para discriminar a una clase de las demás. Para buscar los subconjuntos de atributos para cada clase, se emplean algoritmos genéticos (GA). Los GA son usados para determinar una máscara de atributos para cada subconjunto de unidades ocultas que discriminan a la clase $i$ de las demás clases. Los parámetros utilizados para el algoritmo genético son individuos de longitud $nM$ bits, donde $n$ es la cantidad de atributos originales y $M$ el número de clases. Un individuo (cromosoma) $G$ es representado por:

$$G = \left\{ \left(g_1^1, ..., g_i^1, ..., g_n^1\right), ..., \left(g_1^k, ..., g_i^k, ..., g_n^k\right), ..., \left(g_1^M, ..., g_i^M, ..., g_n^M\right) \right\}$$

donde $g_i^k \in \{0, 1\}$, $i = 1, 2, ..., n \, y \, k = 1, 2, ..., M$.

Para las operaciones de cruza y mutación se utilizó la selección por ruleta y cruza de dos puntos, la tasa de mutación es dinámica y la función de aptitud para cada cromosoma fue :

$$F(G) = 1 - E_v(G)$$

donde $E_v(G)$ es el error en la clasificación al validar los datos con la selección representada por el cromosoma $G$. Una vez que se obtinene un subconjunto de atributos para cada clase por medio de los GA, para la clasificación los atributos seleccionados para la clase $i$ sirven como entradas al subconjunto de unidades ocultas que discriminan a la clase $i$ de las demás clases.

En otros trabajos como [Nan06] se introduce un método de selección de atributos únicamente para problemas de dos clases, en éste cada clase del problema es dividido en clusters de manera independiente y posteriormente se seleccionan atributos para cada cluster. El número de clusters a formar y el número de atributos a seleccionar para cada cluster son parámetros que deben ser proporcionados, se establece que para clusters de la misma clase se selecciona el mismo número de atributos, pero el número de atributos a seleccionar entre clases puede ser diferente. Para clasificar, se entrena a un clasificador por cada uno de los clusters con los atributos seleccionados para dicho cluster, por lo tanto un nuevo objeto debe reducir su dimensionalidad de acuerdo al clasificador en el que este siendo probado. Para asignar al nuevo objeto a una clase, cada clasificador mide la similitud del nuevo objeto con los objetos del cluster para el que fue entrenado y posteriormente se asigna al nuevo objeto la clase del cluster al que tuvo mayor similitud.

## 3.3. Discusión

Como se mencionó son muchos los métodos propuestos para la selección de atributos, la mayoría de los cuales se enfocan en el problema de la selección de atributos para todas las clases de un problema de clasificación. Este tipo de métodos han sido ampliamente estudiados y presentan soluciones para trabajar con problemas multiclase y datos mezclados. Por otra parte una línea de investigación poco estudiada es la selección de atributos por clase, la cual surge a partir de los trabajos propuestos para encontrar atributos específicos por clases pero en la extracción de atributos. De los pocos trabajos que se tienen para la selección de atributos por clase, en este capítulo se mencionaron

tres: el trabajo de [OLS99] fue propuesto para el reconocimiento de escritura y para atributos numéricos únicamente, aunque se eligen diferentes subconjuntos de atributos para cada clase es forzado a elegir el mismo número de atributos para todos los subconjuntos, por lo que para algunas clases se pueden escoger atributos que no son relevantes o dejar fuera atributos que si son relevantes.

Los trabajo de [FW02] y [FW05], proponen un método de selección de atributos por clase que no es para un dominio en específico y presentan una forma novedosa de utilizar una sola red de base radial para elegir atributos por clase, aunque en estos trabajos el número de atributos seleccionados para cada clase puede variar, sólo se realizaron experimentos sobre 3 bases de datos del repositorio de la universidad de California: Glass, Thyroid y Wine, las cuales son bases de datos que contienen atributos numéricos únicamente y no presentan ausencia de información.

Finalmente en el trabajo de [Nan06], se propone un método de selección de atributos por clase que únicamente permite trabajar con problemas de clasificación de 2 clases y además divide a cada clase en clusters, adaptar este trabajo a problemas multiclase sería demasiado complejo dada la forma en la que seleccionan los atributos para cada cluster de cada clase. Este trabajo además sólo permite trabajar con atributos numéricos y no permite ausencia de información. Como en este trabajo el número de clusters a formar y el número de atributos a seleccionar para cada cluster son parámetros desconocidos, se deben realizar una gran cantidad de experimentos variando los valores de estos dos parámetros para definir la combinación que permita encontrar una buena selección de atributos por clase.

En los 3 trabajos mencionados anteriormente se puede observar que aunque se encuentran dentro de la selección de atributos por clase, no utilizan datos mezclados y ausencia de información, además algunos no permiten que para cada clase sean seleccionados un número diferente de atributos. El método propuesto en esta tesis permite trabajar con datos mezclados, ausencia de información y no se establece un número fijo de atributos para cada clase.

En el método propuesto se presenta una manera de incorporar los selectores de atributos para todas las clases para abordar el problema de la selección de atributos por clase, esto es debido a que ya se han desarrollado una gran cantidad de selectores de atributos para todas las clases en los que se han resuelto los problemas de la presencia de atributos numéricos y no numéricos, la ausencia de información, así como la capacidad de determinar automáticamente el número de atributos presentes en el subconjunto seleccionado. Dado que se cuenta ya con estas ventajas se propone la utilización de éstos selectores de atributos para todas las clases para desarrollar un método que ayude a discriminar a una clase de las demás, partiendo de la idea de que los atributos tienen un poder de discriminación diferente para cada clase. Al realizar una selección de atibutos por clase, se propone además un ensamble de clasificadores y una estrategia de decisión para clasificar nuevos objetos, debido a que se cuenta con un subconjunto de atributos posiblemente diferente para cada clase.

## 3.4. Resumen

En este capítulo inicialmente se mencionó de manera breve los trabajos relacionados con la selección de atributos para todas las clases. Posteriormente se presentaron los trabajos relacionados a la selección de atributos por clase, en primer lugar la manera en la que estos trabajos abordan el problema de la selección de atributos por clase y posteriormente se brindó un análisis de cada uno. El análisis presentado en la sección de discusión es importante ya que menciona los puntos que se tratan de cubrir en esta tesis y que en algunos de los trabajos analizados no se abordan.

# Capítulo 4

# Método propuesto

En este capítulo se presenta el método de selección de atributos por clase propuesto, el cual permite trabajar con datos mezclados y/o ausencia de información. El método propuesto consiste de tres etapas que son: la binarización de clases, la selección de atributos por clase y la etapa de clasificación.

## 4.1. Solución propuesta

Como ya se mencionó, el método propuesto para la selección de atributos por clase en problemas con datos mezclados consta de tres etapas: una binarización de clases, la selección de atributos por clase y la clasificación. La etapa de clasificación no es por si misma parte de la selección de atributos, es necesario definir una etapa de clasificación dado que los clasificadores supervisados sólo permiten trabajar con un subconjunto de atributos para todas las clases de un problema de clasificación.

El método propuesto pretende analizar a cada clase de un problema de clasificación de manera independiente, para encontrar los atributos que mejor la discriminen de las demás clases. Para analizar a cada clase se emplea la técnica de binarización *uno-contra-todos*, ya que permite unir a varias clases en

una sola clase. La selección de atributos por clase se realiza sobre subproblemas binarios, en donde los atributos seleccionados favorecen únicamente a dos clases: la clase que se desea analizar y las demás clases que se encuentran unidas en una sola clase. Así los atributos seleccionados de cada subproblema binario se asignan a la clase a partir de la cual se construyó el subproblema binario. Para la etapa de clasificación se emplea un ensamble de clasificadores, es decir, se entrena un clasificador para cada clase con los atributos que se seleccionaron para ésta. Cada etapa del método propuesto es descrita con más detalle en las siguientes secciones.

### 4.1.1. Binarización de clases: *uno-contra-todos*

Dado un problema de clasificación con $c$ clases, éste es transformado en una serie de subproblemas binarios; tal que para cada $w_i \in W$, donde $i = 1, 2, ..., c$; se crea un subproblema de dos clases $< w_i, \Omega >$, siendo $\Omega$ un grupo de $c - 1$ clases tal que $\Omega = \{w_j | 1 \leq j \leq c,\ j \neq i\}$. Para cada subproblema binario los objetos de la clase $w_i$ son los ejemplos positivos y los objetos de las clases restantes son los ejemplos negativos (ver figura 4.1). En este trabajo se decidió aplicar la técnica de binarización de clase *uno-contra-todos*, debido a que nos permite analizar a cada clase de manera independiente a las demás clases, para elegir a los atributos que discriminen mejor a cada clase de las clases restantes y con lo cual se pueda realizar una selección de atributos por clase.

Si se utilizara la técnica de binarización *uno-contra-uno*, se generaría un mayor número de subproblemas binarios y por consiguiente se necesitaría un mayor número de selectores y clasificadores. Como en la técnica *uno-contra-uno* una clase puede estar en más de un subproblema binario, al momento de la selección de atributos por clase no se establecería de manera clara que atributos fueron seleccionados para cada clase del problema de clasificación.

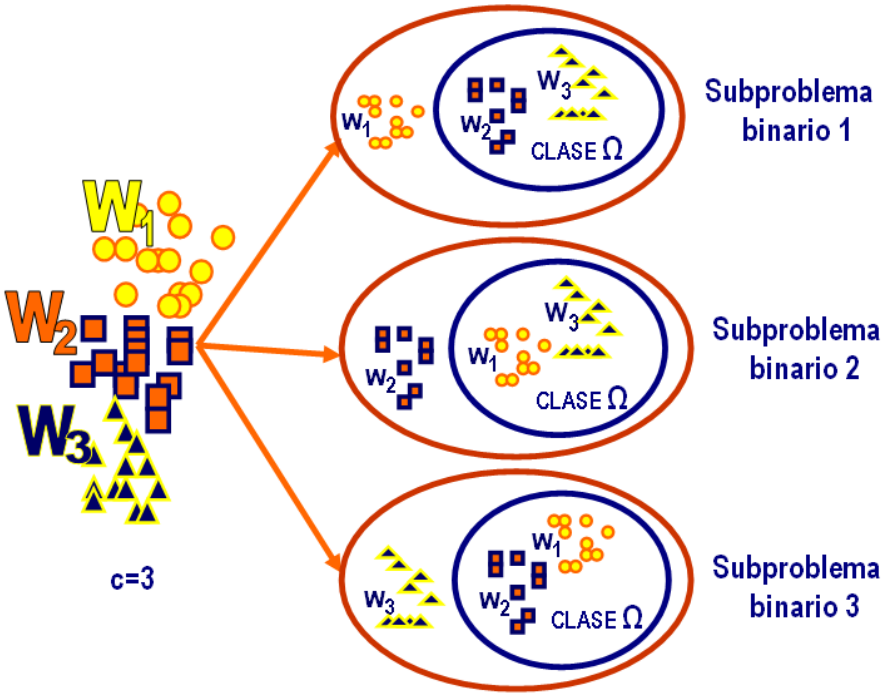


Figura 4.1: Binarización de un problema de 3 clases con la técnica *uno-contra-todos*.

En la figura 4.1 se puede observar como un problema de $c$ clases es transformado en $c$ subproblemas binarios, uno por cada clase del problema multiclase original. En cada subproblema binario los objetos de la clase $w_i$ conservan su clase y los objetos de las clases restantes cambian su clase original para poder ser integrados en $\Omega$ como una sola clase.

### 4.1.2. Selección de atributos por clase

Posteriormente a la binarización de clases se cuenta con un conjunto de subproblemas binarios, para cada subproblema binario se seleccionan atributos de manera independiente aplicando un selector de atributos para todas las clases, y los atributos seleccionados son asignados a la clase de la cual se formó el subproblema binario. Lo que se desea es mejorar la calidad de clasificación, por lo que al seleccionar atributos por clase se espera que estos

atributos discriminen lo mejor posible a cada clase $w_i$, donde $i = 1, ..., c$, de las demás clases. Después de realizar la selección de atributos en cada subproblema binario, se obtienen $c$ subconjuntos de atributos con dimensionalidad posiblemente diferente, uno para cada clase del problema multiclase original. El selector de atributos empleado para esta etapa puede ser el mismo o diferente para cada subproblema binario.

Es conocido que al aplicar la técnica de binarización *uno-contra-todos*, se puede generar un problema binario desbalanceo. Cuando se seleccionan atributos de un problema desbalanceado, los atributos seleccionados benefician en mayor grado la clasificación de los objetos de la clase mayoritaria. En el método propuesto se observó que en los subproblemas binarios $< w_i, \Omega >$, los atributos seleccionados beneficiaban en mayor grado a la clase $\Omega$ sobre la clase $w_i$, siendo que es la clase $w_i$ a la que se desea favorecer con la selección de atributos por clase. Para resolver el problema de las clases desbalanceadas, en este trabajo se aplicaron dos técnicas de balanceo de clases:

1. Repetición
2. SMOTE

Estas técnicas de balanceo se aplican sobre cada subproblema binario antes de seleccionar atributos, por lo tanto el balanceo se realiza exclusivamente para la etapa de selección de atributos por clase. En el balanceo de clases por repetición se calcula el valor $\beta_i$ para cada subproblema binario tal que:

$$\beta_i = |\Omega| - |w_i|$$

donde $i = 1, ..., c$; $|w_i|$ representa el número de objetos en la clase $w_i$ y $|\Omega|$ representa el número de objetos de las clases restantes. Si $\beta_i > 0$ se balancea la clase $w_i$ repitiendo los objetos de esta clase en el mismo orden en que aparecen en el problema de clasificación, hasta que el número de objetos en la clase $w_i$ iguale al número de objetos presentes en la clase $\Omega$.

SMOTE es un algoritmo de sobre-muestreo que se explicó anteriormente en el capítulo 2. En SMOTE se establece un porcentaje de sobre-muestro a partir del tamaño de la clase minoritaria:

$$\beta_i = ((|\Omega| - |w_i|) * 100) / |w_i|$$

donde $i = 1, ..., c$. Si $\beta_i > 0$ se aplica el algoritmo de SMOTE con el porcentaje $\beta_i$ de sobre-muestreo para la clase $w_i$.

Después de balancear los subproblemas binarios se realiza la selección de atributos sobre cada subproblema binario como se mencionó al inicio de esta sección. El proceso de selección de atributos por clase se ejemplifica en la figura 4.2.

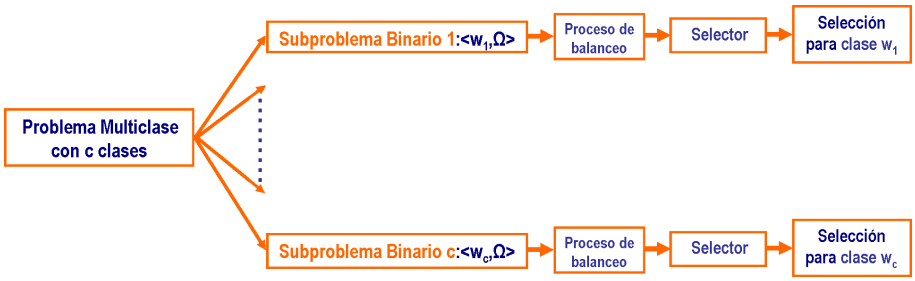


Figura 4.2: Selección de atributos por clase.

En la figura 4.2 se puede observar que para cada subproblema binario se realiza un balanceo de clases antes de seleccionar atributos de éste, posteriormente los atributos seleccionados son asignados a la clase $w_i$ para la que se formó el subproblema binario.

### 4.1.3. Clasificación con la selección de atributos por clase

Si en esta etapa se entrenara un clasificador por cada subproblema binario obtenido, se tendría una salida binaria de cada clasificador que nos indicaría si el nuevo objeto pertenece a la clase $w_i$ o no. Dado que para cada clase se

utiliza un subconjunto de atributos diferente, se puede presentar el caso de que todos los clasificadores indiquen que el nuevo objeto no pertenece a la clase $w_i$ para el que fue entrenado o que más de un clasificador indique que si y entonces se tendría un objeto sin clasificar.

Para evitar la situación anterior, en la clasificación se entrena a cada clasificador $e_i$ con el problema de clasificación multiclase original, pero solo considerando los atributos seleccionados para la clase $w_i$, es decir, los objetos que sirven de entrenamiento al clasificador conservan su clase original. Como para cada clase $w_i$ se entrena un clasificador, se forma un ensamble de clasificadores $E = \{e_1, ..., e_c\}$, así la salida de cada clasificador es una clase para el nuevo objeto. Cuando se quiere clasificar un nuevo objeto $O_y$, se debe reducir su dimensionalidad $d$ a la dimensionalidad $d_i$ utilizada por el clasificador $e_i$ en el que se este probando el nuevo objeto; como cada clasificador asigna una clase al nuevo objeto se aplica el siguiente criterio de decisión:

1. Mayor prioridad al clasificador $e_i$ si éste da como salida la clase $w_i$ para la cual se seleccionaron los atributos con los que fue entrenado. En caso de empate se aplica un voto mayoritario entre todos los clasificadores y si persistiera el empate, el objeto $O_y$ se asignaría a la clase mayoritaria de las clases empatadas.

2. En caso de que ningún clasificador dé como salida la clase para la que se realizó la selección de atributos con la que fue entrenado, se realiza un voto mayoritario entre todos los clasificadores. Si en el voto mayoritario persiste el empate entonces se asigna a la clase mayoritaria de las clases empatadas.

En la figura 4.3 se ilustra el proceso de clasificación para la selección de atributos por clase, en donde un nuevo objeto debe reducir su dimensionalidad a la utilizada por el clasificador en el que se este probando. Posteriormente se decide la clase a la que pertenece el nuevo objeto.

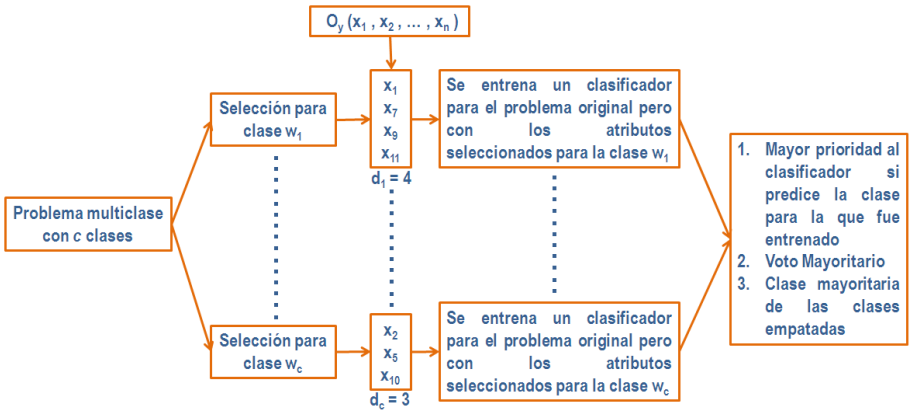


Figura 4.3: Proceso de clasificación utilizando la selección de atributos por clase.

En la tabla 4.1 se ejemplifican algunas de las posibles salidas de cada clasificador y las decisiones que se tomarían para asignar un nuevo objeto a una clase, suponiendo que se tiene un problema de clasificación de 3 clases.

| Salida de cada clasificador | | | Decisión |
|---|---|---|---|
| $e_1$ entrenado para $w_1$ | $e_2$ entrenado para $w_2$ | $e_3$ entrenado para $w_3$ | |
| $w_1$ | $w_3$ | $w_2$ | Se asigna a la clase $w_1$ dado que se le da mayor prioridad a $e_1$ debido a que predijo la clase $w_1$ para la que fue entrenado. |
| $w_1$ | $w_1$ | $w_3$ | Se asigna a la clase $w_1$ por voto mayoritario debido al empate entre $e_1$ y $e_3$. |
| $w_1$ | $w_2$ | $w_3$ | Se asigna a la clase $w_l$ que sea la mayoritaria dado el empate entre $e_1$, $e_2$ y $e_3$. |
| $w_2$ | $w_1$ | $w_1$ | Se asigna a la clase $w_1$ por voto mayoritario dado que ningún clasificador dio como resultado la clase para la que fue entrenado. |
| $w_2$ | $w_3$ | $w_1$ | Se asigna a la clase $w_l$ que sea la mayoritaria dado que ningún clasificador dio como resultado la clase para la que fue entrenado y persiste el empate con voto mayoritario. |

Tabla 4.1: Ejemplo del criterio de decisión propuesto para la clasificación.

## 4.2. Resumen

En este capítulo se describió cada una de las etapas del método propuesto en el cuál se aplicaron técnicas como binarización de clases *uno-contra-todos* y balanceo de clases, además se explicó la etapa de selección de atributos por clase en donde el método propuesto permite incorporar selectores de atributos para todas las clases. Debido a que los clasificadores convencionales sólo permiten trabajar con un mismo subconjunto de atributos para todas las clases de un problema de clasificación y en el método propuesto se seleccionan diferentes subconjuntos de atributos para cada clase, se propuso un ensamble de clasificadores para el que cualquier clasificador puede ser empleado y una estrategia de decisión para combinar los resultados de cada clasificador dentro del ensamble en un solo resultado para clasificar nuevos objetos.

# Capítulo 5

# Experimentación y resultados

En este capítulo se presentan los experimentos realizados sobre algunas bases de datos del repositorio de la universidad de California, aplicando sobre estas bases de datos el método de selección de atributos por clase propuesto sin balanceo, con balanceo por repetición y balanceo con SMOTE. Además se comparan los resultados obtenidos con los selectores de atributos implementados en la herramienta WEKA. Los selectores de atributos implementados en la herramienta WEKA se encuentran dentro del enfoque de selección de atributos para todas las clases y trabajan con problemas de 2 o más clases, con datos mezclados y/o ausencia de información. Así mismo para la etapa de clasificación del método propuesto, se utilizan los clasificadores: bayes (Naivebayes), vecinos más cercanos (IBk), C4.5(J48) y redes neuronales (Multilayerperceptron), estos clasificadores se encuentran implementados en la herramienta WEKA y los nombres entre paréntesis indican el nombre que utiliza la herramienta WEKA para referirse a cada clasificador.

## 5.1. Bases de datos

Los experimentos con el método propuesto se realizaron sobre 19 bases de datos del repositorio de la universidad de California en Irving [MM98]. Estas bases de datos se seleccionaron con características diferentes (número de atributos, número de clases, tipos de atributos y ausencia de información) para evaluar el desempeño del método propuesto.

En la tabla 5.1 se muestra un resumen de las características de cada una de las bases de datos utilizadas en los experimentos realizados.

| Base de Datos | No. De Clases | No. De Objetos | No. De Atributos | | Ausencia de Información |
|---|---|---|---|---|---|
| | | | Numéricos | No numéricos | |
| Annealing | 5 | 798 | 6 | 32 | Sí |
| Bridges | 7 | 106 | 2 | 9 | Sí |
| Credit-screening | 2 | 690 | 6 | 9 | Sí |
| Dermatology | 6 | 366 | 32 | 0 | Sí |
| Diabetes | 2 | 768 | 8 | 0 | No |
| Ecoli | 8 | 336 | 7 | 1 | No |
| Echocardiogram | 3 | 132 | 11 | 0 | Sí |
| Glass | 6 | 214 | 9 | 0 | No |
| Hepatitis | 2 | 155 | 19 | 0 | Sí |
| Ionosphere | 2 | 351 | 34 | 0 | No |
| Iris | 3 | 150 | 4 | 0 | No |
| Nursery | 5 | 12960 | 0 | 8 | No |
| Optdigits | 10 | 3823 | 64 | 0 | No |
| Page-blocks | 5 | 5473 | 10 | 0 | No |
| Postoperative | 3 | 90 | 1 | 7 | Sí |
| Segment | 7 | 210 | 19 | 0 | No |
| Thyroid gland | 3 | 215 | 5 | 0 | No |
| House-votes-84 (Votes) | 2 | 435 | 0 | 16 | Sí |
| Wine | 3 | 178 | 13 | 0 | No |

Tabla 5.1: Resumen de las características de las bases de datos utilizadas.

Otra característica que interesa analizar es el número y proporción de objetos existentes en cada una de las clases de las bases de datos que se utilizaron en los experimentos con el método propuesto. En la tabla 5.2 se presenta la

cantidad de objetos por clase y el porcentaje que representan respecto al total de objetos de la base de datos. En esta tabla se puede observar que algunas bases de datos ya se encuentran desbalanceadas y por lo tanto al aplicar la técnica de binarización *uno-contra-todos* se genera un desbalance mayor.

| Base de Datos | Objetos por clase |
|---|---|
| Annealing | 8(1%), 88(11.03%), 608(76.19%), 60(7.52%), 34(4.26%) |
| Bridges | 13(12.26%), 11(10.38%), 10(9.43%), 1(0.94%), 44(41.51%), 11(10.38%), 16(15.09%) |
| Credit-screening | 307(44.49%), 383(55.51%) |
| Dermatology | 112(30.60%), 61(16.67%), 72(19.67%), 49(13.39%), 52(14.21%), 20(5.46%) |
| Diabetes | 500(65.10%), 268(34.90%) |
| Ecoli | 143(42.56%), 77(22.92%), 52(15.48), 35(10.42%), 20(5.95%), 5(1.49%), 2(0.6%), 2(0.6%) |
| Echocardiogram | 50(37.88%), 24(18.18%), 58(43.94%) |
| Glass | 70(32.71%), 76(35.51%), 17(7.94%), 13(6.07%), 9(4.21%), 29(13.55%) |
| Hepatitis | 32(20.65%), 123(79.35%) |
| Ionosphere | 126(35.90%), 225(64.10%) |
| Iris | 50(33.33%), 50(33.33%), 50(33.33%) |
| Nursery | 4320(33.33%), 2(0.02%), 328(2.53%), 4266(32.92%), 4044(31.20%) |
| Optdigits | 376(9.84%), 389(10.18%), 380(9.94%), 389(10.18%), 387(10.12%), 376(9.84%), 377(9.86%), 387(10.12%), 380(9.94%), 382(9.99%) |
| Page-blocks | 4913(89.77%), 329(6.01%), 28(0.51%), 88(1.61%), 115(2.10%) |
| Postoperative | 64(71.11%), 2(2.22%), 24(26.67%) |
| Segment | 30(14.29%), 30(14.29%), 30(14.29%), 30(14.29%), 30(14.29%), 30(14.29%), 30(14.29%) |
| Thyroid gland | 150(69.77%), 35(16.28%), 30(13.95%) |
| Votes | 267(61.38%), 168(38.62%) |
| Wine | 59(33.15%), 71(39.89%), 48(26.97%) |

Tabla 5.2: Número y proporción de objetos por clase en las bases de datos utilizadas.

Existen bases de datos como iris y thyroid gland que contienen pocos atributos (4 y 5 respectivamente) y pocos objetos (150 y 215 respectivamente) y se

podría pensar que no es necesaria la selección de atributos en éstas, pero con el método propuesto se desea comprobar si seleccionando atributos por clase se puede aumentar la calidad de clasificación.

## 5.2. Estimación del error

Existen varios métodos para estimar el error en la clasificación. Para este trabajo se utilizó la validación cruzada (*k-fold cross validation*), este método es un punto intermedio entre el método de dejar uno fuera (*leave-one-out*) y el método de utilizar una parte del conjunto de datos para entrenamiento y la otra parte para prueba (*Holdout*) [JDM00]. La validación cruzada, consiste en dividir el conjunto de entrenamiento en $k$ subconjuntos disjuntos. De esta manera se utilizan $k-1$ subconjuntos para entrenar al clasificador y el subconjunto restante se utiliza para prueba. Esto se repite $k$ veces dejando a cada uno de los subconjuntos como prueba, finalmente se promedia el resultado de las $k$ pruebas.

Cabe mencionar que para efectos de comparación con los selectores de atributos implementados en la herramienta WEKA se utilizó la misma validación cruzada a 10 partes generada por WEKA, es decir, los subconjuntos generados por la herramienta WEKA para la validación cruzada son los mismos que se utilizaron para evaluar el método propuesto.

## 5.3. Selectores de atributos y clasificadores utilizados

Para la selección de atributos y para efectos de comparación se utilizaron los selectores de atributos tipo $filter$ implementados en la herramienta WEKA, estos selectores de atributos están dentro del enfoque de selección de atributos para todas las clases, dichos selectores buscan subconjuntos de atributos

en el espacio de atributos evaluando a los subconjuntos encontrados. Por tal motivo para realizar una selección de atributos con la herramienta WEKA, se debe especificar un evaluador y el método de búsqueda que se desee utilizar.

En el método propuesto se emplearon los dos evaluadores tipo $filter$ implementados en la herramienta WEKA, nombrados como CfsSubsetEval (e1) [Hal98] y ConsistencySubsetEval (e2) [LS96], éstos son evaluadores de subconjuntos de atributos los cuales toman un subconjunto de atributos y retornan una medida numérica, que es una evaluación de la calidad del subconjunto que sirve para guiar la búsqueda. Se utilizaron únicamente estos dos evaluadores ya que son tipo $filter$ y por lo tanto tienen un menor costo computacional. Además estos evaluadores determinan automáticamente el número de atributos seleccionados y por lo tanto se pueden combinar con la mayoría de métodos de búsqueda implementados igualmente en la herramienta WEKA. Como métodos de búsqueda se emplearon BestFirst (b1), ExhaustiveSearch (b2), GeneticSearch (b3), GreedyStepwise (b4) y RandomSearch (b5) [WF05]. Así para el método propuesto se probaron 10 selectores de atributos (ver tabla 5.3), que son las combinaciones de los 2 evaluadores con los 5 métodos de búsqueda.

| Selector de atributos | Evaluador | Método de búsqueda |
|---|---|---|
| e1b1 | CfsSubsetEval | BestFirst |
| e1b2 | CfsSubsetEval | ExhaustiveSearch |
| e1b3 | CfsSubsetEval | GeneticSearch |
| e1b4 | CfsSubsetEval | GreedyStepwise |
| e1b5 | CfsSubsetEval | RandomSearch |
| e2b1 | ConsistencySubsetEval | BestFirst |
| e2b2 | ConsistencySubsetEval | ExhaustiveSearch |
| e2b3 | ConsistencySubsetEval | GeneticSearch |
| e2b4 | ConsistencySubsetEval | GreedyStepwise |
| e2b5 | ConsistencySubsetEval | RandomSearch |

Tabla 5.3: Selectores de atributos para todas las clases utilizados en el método propuesto.

Para evaluar la selección de atributos por clase realizada por el método propuesto se utilizaron los clasificadores: NaiveBayes (Bayes) [JL95], vecinos más cercanos (IBk) [AKA91], este clasificador se probó con $k = 1$ y $k = 3$, se decidió variar el valor de $k$ únicamente en 3 primero para observar si ofrecia mejores resultados que $k = 1$, además de hacer rápido y simple el proceso de clasificación; C4.5 (J48) usado para generar árboles de decisión [Qui93] y MultilayerPerceptron (MLP) aplica redes neuronales y utiliza retropropagación. Estos clasificadores son rápidos, sencillos y ampliamente usados para resolver problemas de clasificación supervisada. En todos los experimentos realizados en esta tesis, estos clasificadores se utilizaron con los parámetros establecidos por defecto en la herramienta WEKA.

## 5.4. Experimentos realizados

En esta tesis se realizaron experimentos con tres variantes del método propuesto las cuales fueron:

1. Selección de atributos por clase sin balanceo de clases (Método propuesto sin balanceo).
2. Selección de atributos por clase con balanceo por repetición (Método propuesto con repetición).
3. Selección de atributos por clase con balanceo utilizando SMOTE (Método propuesto con SMOTE).

Para cada uno de los métodos, se realizó una selección de atributos por clase utilizando los 2 evaluadores y los 5 métodos de búsqueda implementados en la herramienta WEKA con los párametros establecidos por defecto, con lo que en total se probó a cada método con 10 selectores de atributos (combinación evaluador-método de búsqueda) distintos. Para evaluar la calidad de

la selección de atributos por clase realizada por cada método se utilizaron los clasificadores mencionados en la sección anterior.

Además para comparar los resultados obtenidos con el método propuesto se realizó:

1. Selección de atributos para todas las clases (WEKA con selección).
2. No hacer selección de atributos (WEKA sin selección).

Para la selección de atributos para todas las clases se usaron los mismos selectores de atributos y clasificadores implementados en la herramienta WEKA, que se utilizaron para los experimentos con el método propuesto. Además se obtuvo el porcentaje de error en la clasificación de cada una de las bases de datos sin hacer selección de atributos, utilizando los clasificadores implementados en la herramienta WEKA que se usaron en los experimentos con el método propuesto.

Los nombres entre parentésis tanto en las variantes del método propuesto como en la selección de atributos para todas las clases y no hacer selección, son con los que se hará referencia a estos métodos en el transcurso de la tesis. Por lo tanto en las siguientes secciones se reportan los resultados obtenidos con cada uno de los experimentos realizados con el método propuesto, así como los resultados obtenidos con la selección de atributos para todas las clases y sin hacer selección de atributos.

## 5.5. Comparación entre la selección de atributos para todas las clases y no hacer selección de atributos

Antes de comparar los resultados obtenidos del método propuesto contra la selección de atributos para todas las clases, se compara cuál es el porcentaje

de error obtenido con los 10 selectores de atributos empleados para realizar una selección de atributos para todas las clases y el porcentaje de error obtenido sin hacer selección de atributos sobre las 19 bases de datos del repositorio de la universidad de California. No en todos los selectores de atributos se probaron las 19 bases de datos, debido a que para algunas bases de datos las pruebas con los selectores de atributos usados eran muy tardadas[1] y en bases de datos como Postoperative, no se pudo hacer la selección de atributos para todas las clases[2]. Por lo tanto los selectores de atributos se encuentran agrupados de acuerdo a las bases de datos que se pudieron probar en éstos.

Así en las figuras 5.1 a 5.5 se muestra el promedio de los porcentajes de error de 5 selectores de atributos (e1b1, e1b3, e1b4, e2b1, e2b3), para los que se pudieron probar las 19 bases de datos y para medir la calidad del resultado de cada selector de atributos se utilizaron los clasificadores Bayes, IBK con k=1 y k=3, J48 y MLP. Estos mismos clasificadores se utilizaron para obtener el porcentaje de error sin hacer selección de atributos.

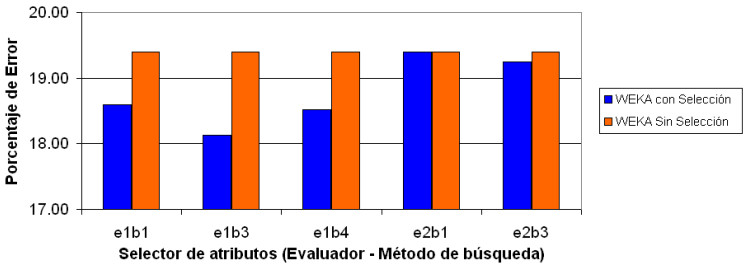


Figura 5.1: Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador Bayes.

[1]Las pruebas eran muy tardadas para las bases de datos que contiene más de 30 atributos, en donde haciendo una selección de atributos para todas las clases se podía tardar más de 1 día.

[2]Esto debido a que el selector de atributos para todas las clases implementado en la herramienta WEKA no retornaba algún atributo seleccionado como resultado.

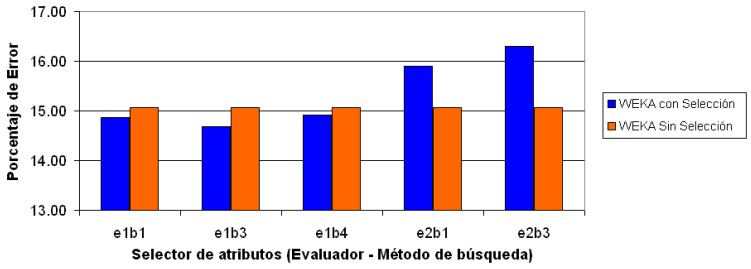


Figura 5.2: Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador J48.

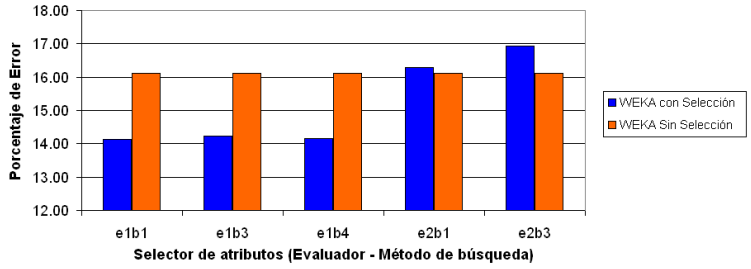


Figura 5.3: Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=1.

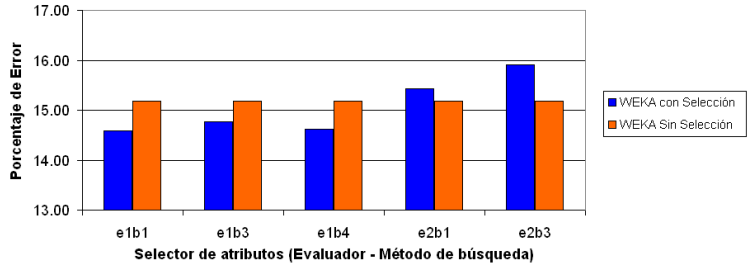


Figura 5.4: Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=3.

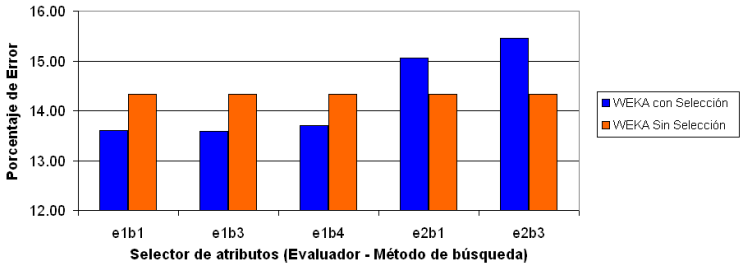


Figura 5.5: Comparación del promedio del porcentaje de error de 19 bases de datos entre WEKA con y sin selección de atributos para el clasificador MLP.

En las figuras 5.1 a 5.5 se puede observar que si se utiliza el evaluador ConsistencySubsetEval (e2) con BestFirst (b1) o GeneticSearch (b3) como método de búsqueda, el promedio del porcentaje de error es mayor al hacer una selección de atributos para todas las clases que no hacer selección de atributos, excepto con el clasificador Bayes para el cual con cualquier selector de atributos es menor el porcentaje de error al realizar una selección de atributos para todas las clases. Para el evaluador CfsSubsetEval (e1) aunque se utilizaron los mismos métodos de búsqueda que para e2 más GreedyStepwise (b4) el promedio en el porcentaje de error es menor al hacer una selección de atributos para todas las clases. Por lo tanto se puede observar que el tipo de evaluador influye en conseguir una buena selección de atributos.

En las figuras 5.6 a 5.10 se muestra el promedio de los porcentajes de error de 4 selectores (e1b2, e1b5, e2b2, e2b5), para los que solo se pudieron probar 15 bases de datos[3] y para medir la calidad del resultado de cada selector de atributos se utilizaron los clasificadores Bayes, IBK con k=1 y k=3, J48 y MLP. Estos mismos clasificadores se utilizaron para obtener el porcentaje de error sin hacer selección de atributos.

[3]Las bases de datos que no se pudieron probar fueron: Annealing, Dermatology, Ionosphere y Optdigits. Estas bases de datos contienen más de 30 atributos y las pruebas eran tardadas.

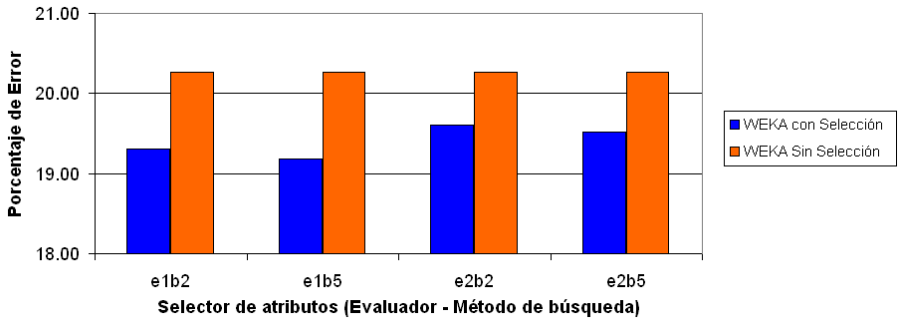


Figura 5.6: Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador Bayes.

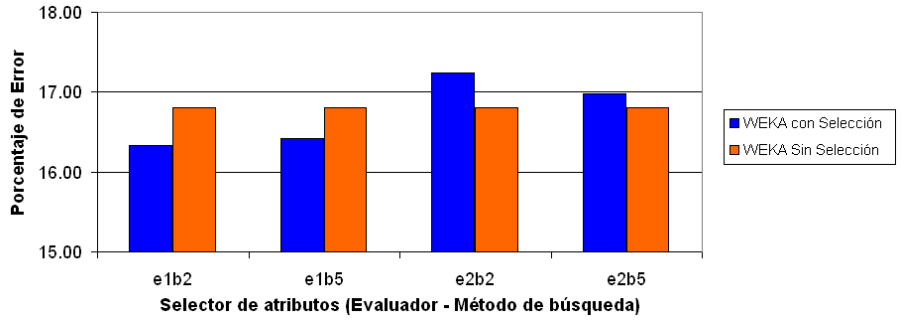


Figura 5.7: Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador J48.

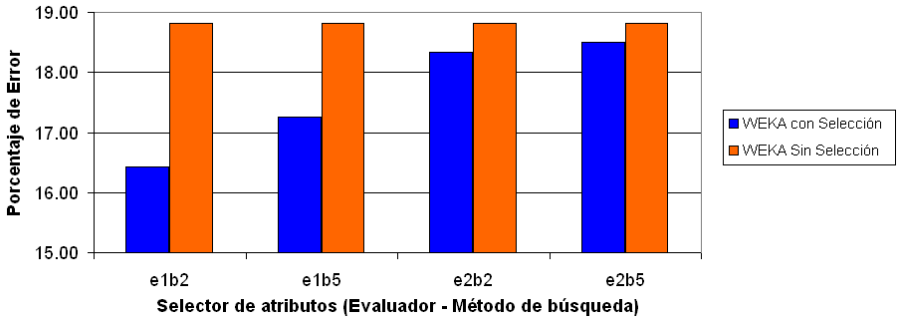


Figura 5.8: Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=1.

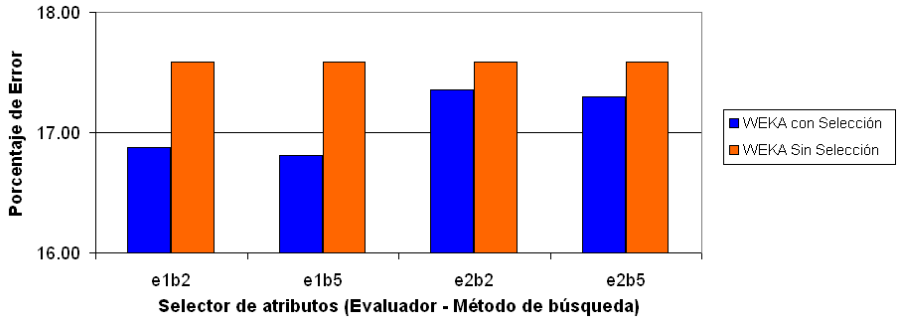


Figura 5.9: Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador IBk con k=3.

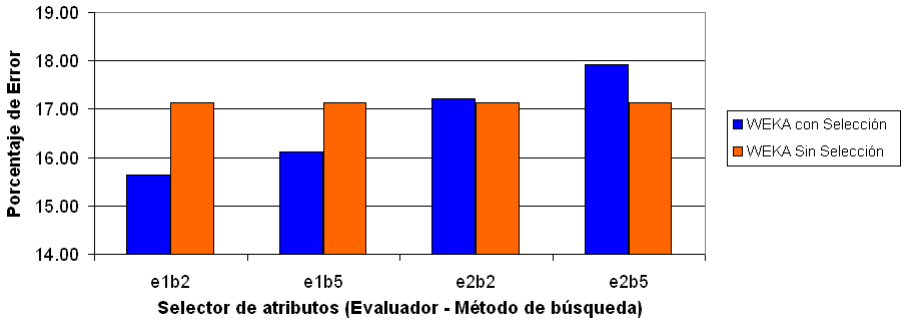


Figura 5.10: Comparación del promedio del porcentaje de error de 15 bases de datos entre WEKA con y sin selección de atributos para el clasificador MLP.

En las figuras 5.6 a 5.10 se puede observar que en clasificadores como Bayes e IBk con k=1 y k=3, el promedio en el porcentaje de error es menor con cualquier selector de atributos que se utilice, por lo tanto es mejor realizar una selección de atributos para todas las clases que no hacerla, pero para los clasificadores J48 y MLP si se utiliza el evaluador ConsistencySubsetEval (e2) ya sea con ExhaustiveSearch (b2) o RandomSearch (b5) como método de búsqueda, el promedio del porcentaje de error es mayor al hacer una selección de atributos para todas las clases que no hacer selección de atributos. El aumento en el promedio del porcentaje de error de los selectores de atributos

e2b2 y e2b5 puede deberse a que clasificadores como J48 y MLP realizan de manera interna una selección o pesado de atributos al momento de clasificar.

En la figura 5.11 se muestra el promedio de los porcentajes de error del selector de atributos e2b4, para el que se pudieron probar 18 bases de datos[4] y para medir la calidad del resultado del selector de atributos se utilizaron los clasificadores Bayes, IBK con k=1 y k=3, J48 y MLP. Estos mismos clasificadores se utilizaron para obtener el porcentaje de error sin hacer selección de atributos.

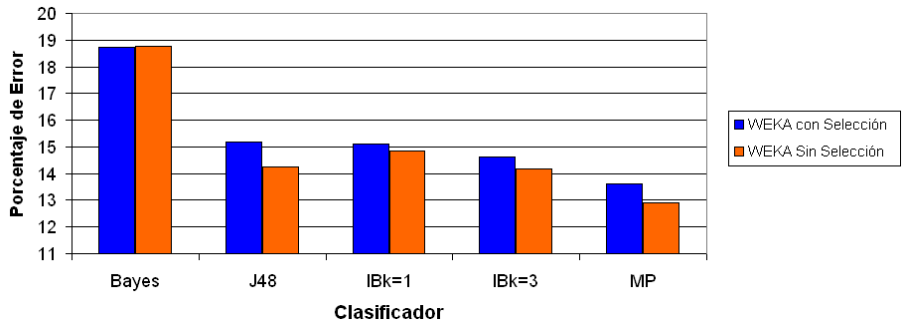


Figura 5.11: Comparación del promedio del porcentaje de error de 18 bases de datos entre WEKA con y sin selección de atributos para el selector de atributos e2b4.

En la figura 5.11 se puede observar que para el selector de atributos e2b4 se presenta un menor promedio en el porcentaje de error al no hacer una selección de atributos que al hacerla, excepto para el clasificador Bayes.

En general, en las figuras 5.1 a 5.11 se puede observar que en algunas ocasiones el error en la clasificación aumenta al hacer selección de atributos para todas las clases y que uno de los clasificadores que mejor resultado ofrece con los selectores de atributos utilizados es el clasificador Bayes. Con el método propuesto se pretende disminuir el promedio del porcentaje de error comparado con el que se obtiene haciendo una selección de atributos para todas las clases y sin hacer selección de atributos.

[4] La base de datos que no se pudo probar fue Postoperative

En las figuras 5.1 a 5.11 se presentaron 50 comparaciones de los resultados obtenidos haciendo una selección de atributos para todas las clases y sin hacer selección de atributos, de los cuales aproximadamente en el 68 % se disminuye el porcentaje de error en la clasificación al hacer una selección de atributos para todas las clases comparado con no hacer selección de atributos.

## 5.6. Resultados del método propuesto

En esta sección se presenta el promedio del porcentaje de error obtenido con el método propuesto en sus tres variantes sobre las bases de datos utilizadas y se comparan estos resultados contra los resultados obtenidos haciendo una selección de atributos para todas las clases y sin hacer selección de atributos. Se presentan los resultados de los 10 selectores de atributos con cada uno de los clasificadores mencionados anteriormente.

En las figuras 5.12 a 5.16 se reportan los promedios del porcentaje de error de utilizar los 10 selectores de atributos implementados en la herramienta WEKA sobre 14 bases de datos[5] que son las que se pudieron probar para todos los selectores de atributos. Se presentan los resultados haciendo selección de atributos para todas las clases, sin hacer selección de atributos y el método propuesto sin balanceo, con balanceo por repetición y SMOTE.

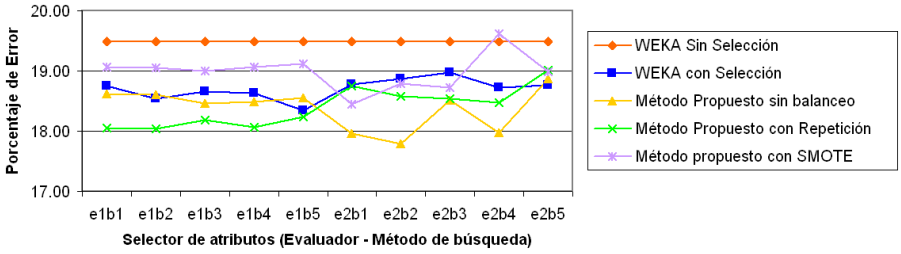


Figura 5.12: Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador Bayes.

[5]Las bases de datos que no se pudieron probar fueron: Annealing, Dermatology, Ionosphere, Optsdigits y Postoperative.

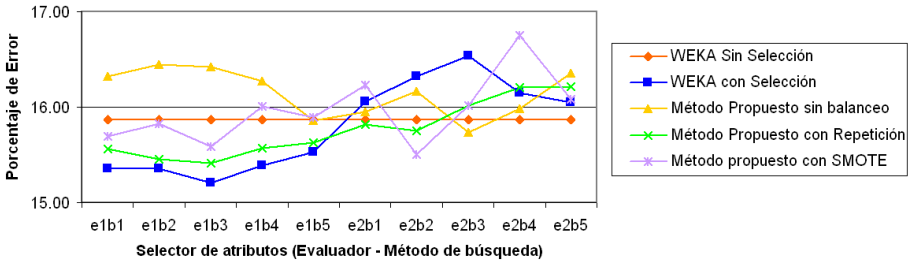


Figura 5.13: Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador J48.

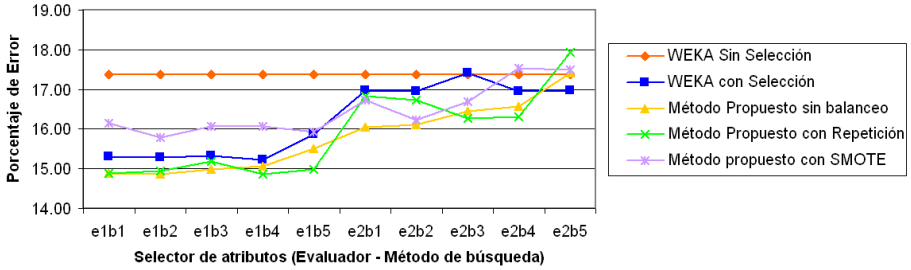


Figura 5.14: Comparación del promedio de porcentaje de error entre de 14 bases de datos el método propuesto y WEKA para el Clasificador IBk con k=1.

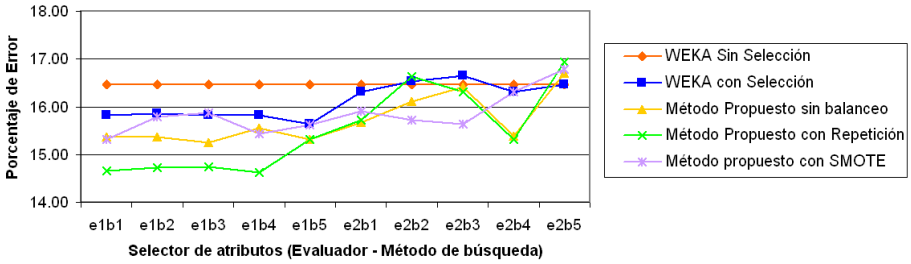


Figura 5.15: Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador IBk con k=3.

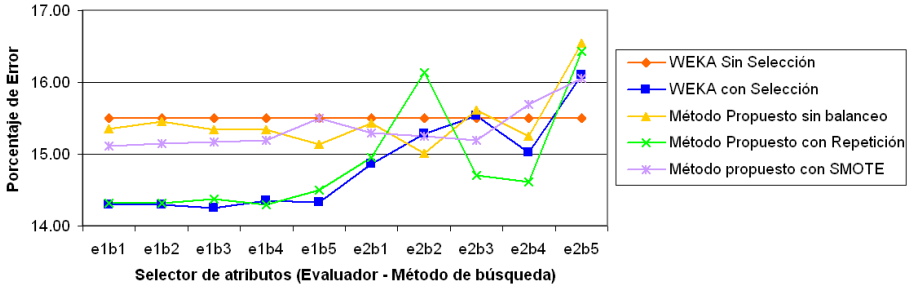


Figura 5.16: Comparación del promedio de porcentaje de error de 14 bases de datos entre el método propuesto y WEKA para el Clasificador MLP.

En las figuras 5.12 a 5.16 se puede observar que en total se probaron 10 selectores de atributos (evaluador – método de búsqueda) para cada uno de los 5 clasificadores (incluyendo la variación de IBk con k=3), por lo tanto en general se cuenta con 50 combinaciones selector – clasificador para las cuales en el 36.54 % de éstas, el método propuesto con balanceo por repetición presenta el menor promedio en el porcentaje de error, el método propuesto sin balanceo es mejor en el 25 % al igual que la selección de atributos para todas las clases, no hacer selección de atributos es mejor en 7.69 % de las veces y el método propuesto con balanceo utilizando SMOTE es solo mejor el 5.77 % de las veces. En resumen alguna variante del método propuesto es mejor en el 67.31 % de los casos, la selección de atributos para todas las clases es mejor en el 25 % y no hacer selección de atributos es mejor en el 7.69 %.

En las figuras 5.12 y 5.14 se puede observar que para el clasificador Bayes e IBk con k=1, el promedio del porcentaje de error obtenido con el método propuesto con balanceo por repetición en la mayoría de los selectores de atributos, es menor que hacer una selección de atributos para todas las clases a excepción del selector e2b5. Además para el clasificador Bayes el menor promedio del porcentaje de error se obtiene con e2b2 y para IBk con k=1 es e1b2, ambos con el método propuesto sin balanceo. En la figura 5.15 se puede observar que para el clasificador IBk con k=3 el método propuesto con balanceo por repetición es mejor en la mayoría de selectores a excepción de e2b2 y e2b5 y el menor promedio del porcentaje de error se obtiene con e1b4 para

el método propuesto con balanceo por repetición. Para los clasificadores J48 y MLP (figuras 5.13 y 5.16), se observa que para la mayoría de los selectores es mejor realizar una selección de atributos para todas las clases, especialmente con el selector e1b3, esto puede deberse a que estos clasificadores seleccionan atributos de manera interna para la clasificación y al realizar una selección de atributos por clase se reduce aun más la cantidad de atributos a utilizar para cada clase al momento de la clasificación.

En la tabla 5.4 se presenta la comparación de cada una de las variantes del método propuesto contra hacer selección de atributos para todas las clases y sin hacer selección de atributos, en esta tabla se muestra el porcentaje de veces en el que una variante del método propuesto obtuvo el menor promedio del porcentaje de error comparado con los otros métodos. Las 50 combinaciones (selector-clasificador) representan el 100 % de las pruebas que fueron ilustradas en las figuras 5.12 a 5.16.

| WEKA con selección o sin selección | % de ser el mejor | | Métodos propuestos | % de ser el mejor |
|---|---|---|---|---|
| WEKA con selección | 40% | Vs. | Método propuesto sin balanceo | 60% |
| WEKA con selección | 36% | Vs. | Método propuesto con balanceo por repetición | 64% |
| WEKA con selección | 64% | Vs. | Método propuesto con balanceo por SMOTE | 36% |
| WEKA sin selección | 24% | Vs. | Método propuesto sin balanceo | 76% |
| WEKA sin selección | 16% | Vs. | Método propuesto con balanceo por repetición | 84% |
| WEKA sin selección | 24% | Vs. | Método propuesto con balanceo por SMOTE | 76% |
| WEKA con selección | 28% | Vs. | Utilizar alguna variante del método propuesto | 72% |
| WEKA sin selección | 10% | Vs. | Utilizar alguna variante del método propuesto | 90% |

Tabla 5.4: Comparación entre los métodos probados por pares para 14 bases de datos.

En la tabla 5.4 se puede observar que el método propuesto sin balanceo y el método propuesto con balanceo por repetición, ofrece mejores resultados con la mayoría de selectores y clasificadores comparado con no hacer selección de atributos o hacer selección de atributos para todas las clases. El método propuesto con balanceo utilizando SMOTE es mejor que no hacer selección

de atributos, pero no es así haciendo una selección de atributos para todas las clases. Esto último se puede deber a que al aplicar el método de SMOTE se pueden generar muchos objetos sintéticos que tienen como semillas objetos con ruido; que a su vez provoca que los atributos seleccionados no presenten una buena discriminación de la clase para la que fueron elegidos respecto a las demás clases.

En las figuras 5.17 a 5.21 se muestra el promedio de los porcentajes de error de 5 selectores de atributos (e1b1, e1b3, e1b4, e2b1, e2b3), implementados en WEKA para los que se pudieron probar las 19 bases de datos con los clasificadores Bayes, IBK con k=1 y k=3, J48 y MLP. Estas figuras muestran los resultados obtenidos haciendo selección de atributos para todas las clases y no hacer selección de atributos, así como los resultados del método propuesto sin balanceo, con balanceo por repetición y SMOTE.

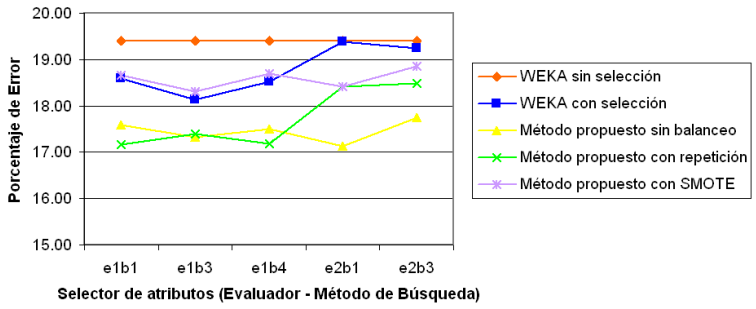


Figura 5.17: Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el Clasificador Bayes.

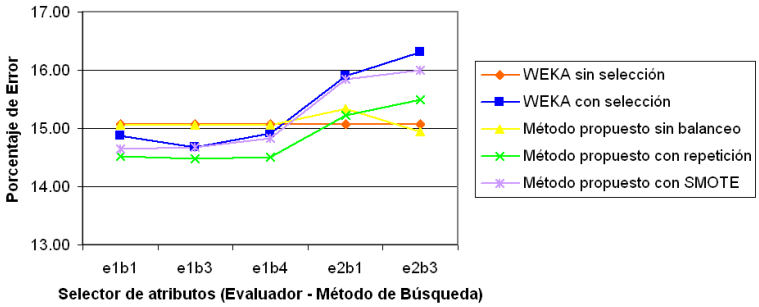


Figura 5.18: Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el Clasificador J48.

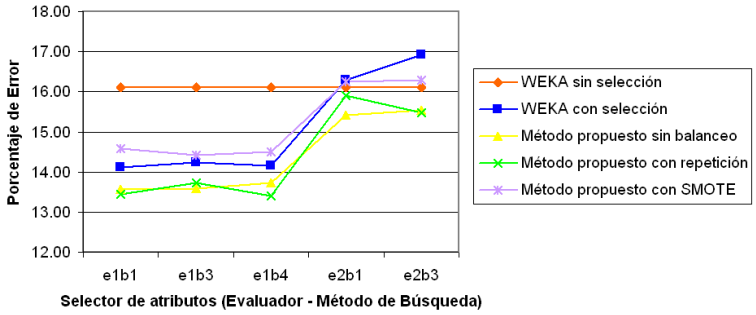


Figura 5.19: Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=1.

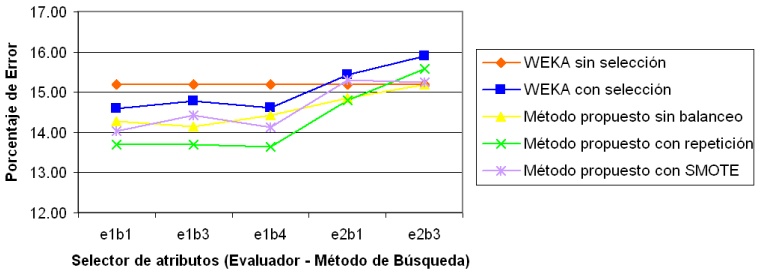


Figura 5.20: Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=3.

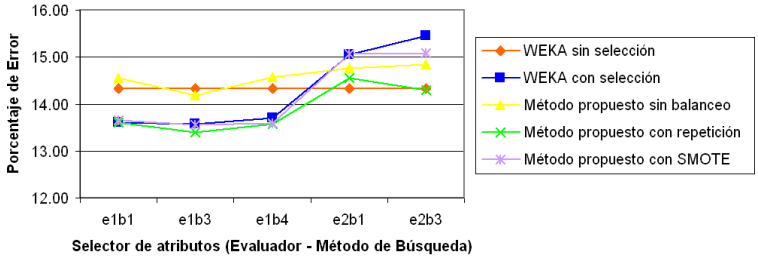


Figura 5.21: Comparación del promedio de porcentaje de error de 19 bases de datos entre el método propuesto y WEKA para el clasificador MLP.

En la figura 5.17 se puede observar que para el clasificador Bayes con las 19 bases de datos, el promedio del porcentaje de error se disminuye al hacer selección de atributos por clase utilizando cualquier selector de atributos implementado en WEKA, comparado con hacer selección de atributos para todas las clases y no hacer selección de atributos. En la figura 5.18 se puede observar que para el clasificador J48, aunque en los selectores e2b1 y e2b3 con el método propuesto con balanceo por repetición no se logra obtener un promedio del porcentaje de error menor al obtenido sin hacer selección de atributos, se logra disminuir el promedio del porcentaje de error para todos los selectores de atributos utilizados, comparado con la selección de atributos para todas las clases. En la figura 5.19 se puede observar que para el clasificador IBk con k=1, en los selectores e2b1 y e2b3 la selección de atributos para todas las clases proporciona un promedio del porcentaje de error mayor comparado con no hacer selección de atributos y con el método propuesto ya sea sin balanceo o con balanceo por repetición, el promedio del porcentaje de error que se obtiene es menor para cualquiera de los selectores de atributos utilizados comparado con no hacer selección de atributos. En la figura 5.20 para el clasificador IBk con k=3, sucede la misma situación que con k=1, excepto que en el selector e2b3 el promedio del porcentaje de error obtenido con el método propuesto con balanceo por repetición es mayor que no hacer selección de atributos. En la figura 5.21 se puede observar que para el clasificador MLP con el método propuesto con balanceo por repetición se logra

disminuir el promedio del porcentaje de error en todos los selectores de atributos utilizados, comparado con hacer selección de atributos para todas las clases, incluso para el selector de atributos e2b3 se logra un promedio menor comparado con no hacer selección de atributos.

En las figuras 5.22 a 5.26 se muestra el promedio de los porcentajes de error de 4 selectores de atributos (e1b2, e1b5, e2b2, e2b5) implementados en WEKA, para los que se pudieron probar 15 bases de datos[6] y para medir la calidad del resultado de cada selector de atributos se utilizaron los clasificadores Bayes, IBK con k=1 y k=3, J48 y MLP. Estas figuras muestran los resultados obtenidos haciendo selección de atributos para todas las clases y no hacer selección de atributos, así como los resultados del método propuesto sin balanceo, con balanceo por repetición y SMOTE.

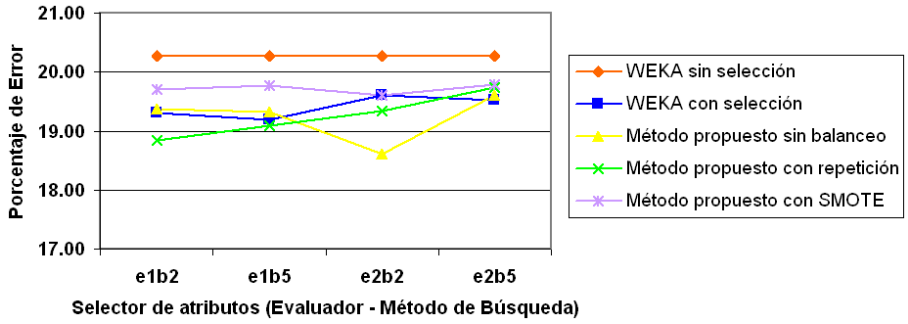


Figura 5.22: Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el Clasificador Bayes.

[6]Las bases de datos que no se pudieron probar fueron: Annealing, Dermatology, Ionosphere y Optdigits.

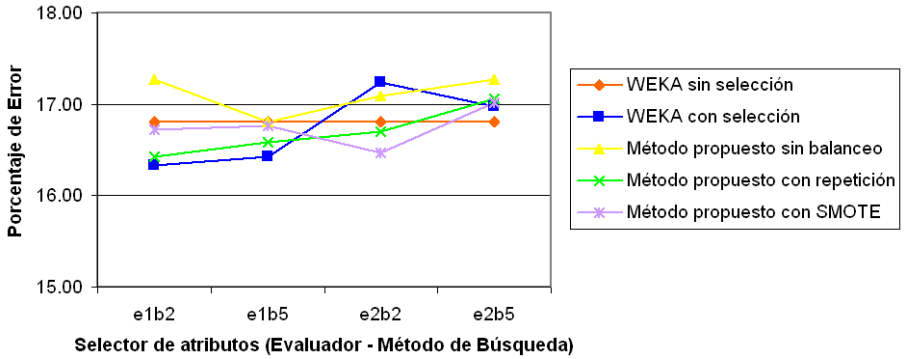


Figura 5.23: Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el Clasificador J48.

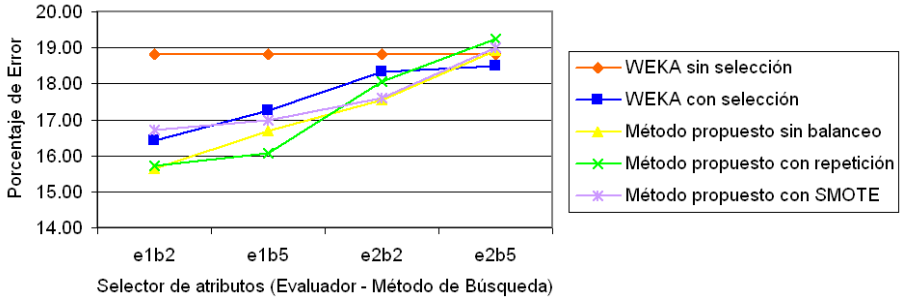


Figura 5.24: Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=1.

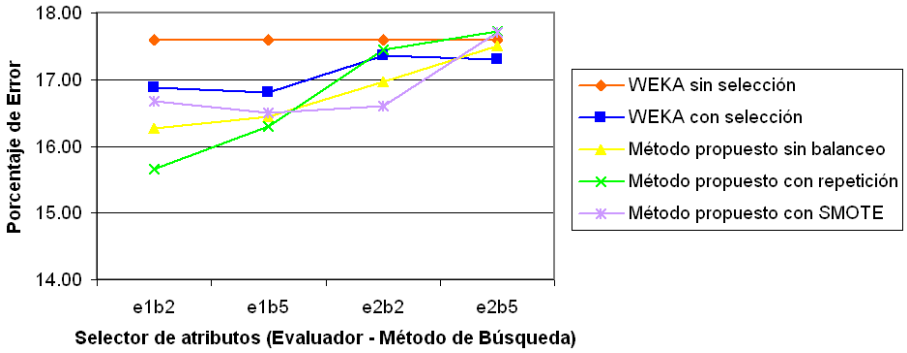


Figura 5.25: Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el clasificador IBk con k=3.

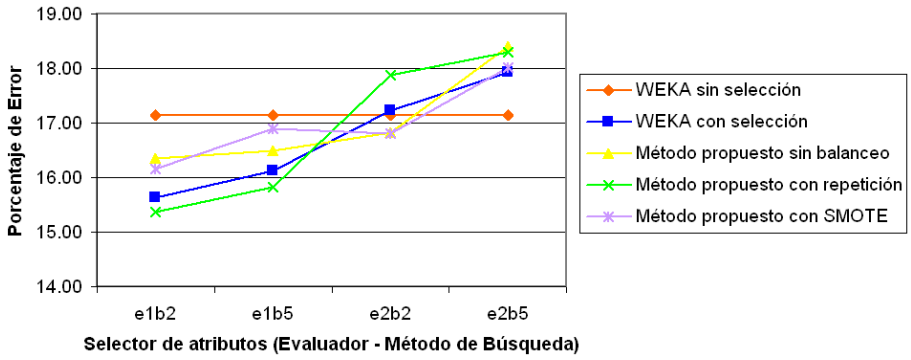


Figura 5.26: Comparación del promedio de porcentaje de error de 15 bases de datos entre el método propuesto y WEKA para el clasificador MLP.

En la figura 5.22 para el clasificador Bayes con el método propuesto con balanceo por repetición se obtiene un promedio del porcentaje de error menor, comparado con hacer una selección de atributos para todas las clases y no hacer selección de atributos, excepto con el selector de atributos e2b5 en donde el promedio obtenido con el método propuesto con balanceo por repetición es mayor, comparado con hacer selección de atributos para todas las clases. En la figura 5.23 para el clasificador J48 solo se logra disminuir el promedio del porcentaje de error en el selector e2b2 con el método propuesto con balanceo ya sea por repetición o con SMOTE. En la figura 5.24 para el clasificador IBK con k=1, el método propuesto con balanceo por repetición y sin balanceo presentan un promedio del porcentaje de error menor, comparado con hacer una selección de atributos para todas las clases y no hacer selección de atributos, excepto para el selector de atributos e2b5. Para el clasificador IBk con k=3 (ver figura 5.25), se presenta una situación similar a la de k=1. Excepto que para el selector de atributos e2b2, el método propuesto con balanceo por repetición presenta un promedio del porcentaje de error mayor comparado con hacer una selección de atributos para todas las clases. En la figura 5.26 para el clasificador MLP con los selectores de atributos e1b2 y e1b5 el promedio del porcentaje de error se logra disminuir con el método propuesto con balanceo por repetición comparado con hacer una selección de atributos para todas las clases y no hacer selección de atributos, con el método propuesto

sin balanceo y con balanceo con SMOTE para el selector de atributos e2b2 se logra disminuir el promedio del porcentaje de error, comparado con hacer una selección de atributos para todas las clases y no hacer selección de atributos.

En la figura 5.27 se muestra el promedio de los porcentajes de error del selector de atributos e2b4 para el que se pudieron probar 18 bases de datos[7] y para medir la calidad del resultado del selector de atributos se utilizaron los clasificadores Bayes, IBK con k=1 y k=3, J48 y MLP. La figura 30 muestra los resultados obtenidos haciendo selección de atributos para todas las clases y no hacer selección de atributos, así como los resultados del método propuesto sin balanceo, con balanceo por repetición y SMOTE.

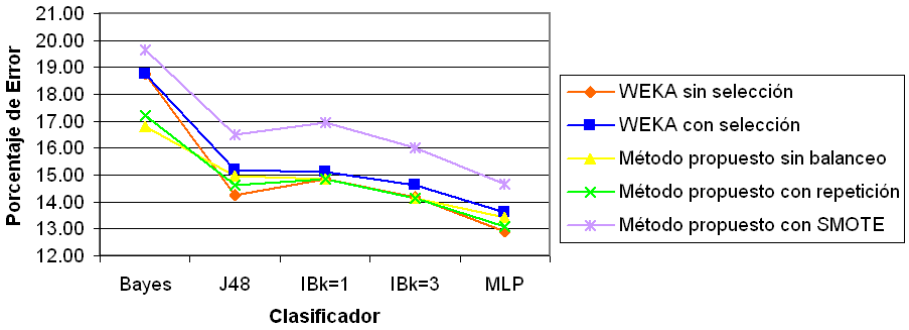


Figura 5.27: Comparación del promedio de porcentaje de error de 18 bases de datos entre el método propuesto y WEKA para el selector de atributos e2b4.

En la figura 5.27 se puede observar que con el método propuesto con balanceo por repetición y sin balanceo presenta un promedio del porcentaje de error menor, comparado con hacer una selección de atributos para todas las clases.

En general de las figuras 5.17 a 5.27 se puede observar que de las 50 combinaciones selector – clasificador presentadas, en el 50 % de las veces es mejor utilizar el método propuesto con balanceo por repetición, en el 20 % de las veces es mejor utilizar el método propuesto sin balanceo, en el 14 % de las veces es mejor no realizar selección de atributos, en el 10 % de las veces es

[7] La base de datos que no se pudo probar fue Postoperative

mejor realizar una selección de atributos para todas las clases y sólo en el 6 % es mejor el método propuesto con SMOTE. En resumen en el 76 % de los casos fue mejor alguna de las variantes del método propuesto, en el 14 % fue mejor no hacer selección de atributos y en el 10 % fue mejor hacer selección para todas las clases.

En la tabla 5.5 se presenta la comparación de cada una de las variantes del método propuesto contra hacer selección de atributos para todas las clases y sin hacer selección de atributos, en esta tabla se muestra el porcentaje de veces en el que una variante del método propuesto obtuvo el menor promedio del porcentaje de error comparado con los otros métodos. Las 50 combinaciones (selector-clasificador) representan el 100 % de las pruebas que fueron ilustradas en las figuras 5.17 a 5.27.

| WEKA con selección o sin selección | % de ser el mejor | | Métodos propuestos | % de ser el mejor |
|---|---|---|---|---|
| WEKA con selección | 34% | Vs. | Método propuesto sin balanceo | 66% |
| WEKA con selección | 18% | Vs. | Método propuesto con balanceo por repetición | 82% |
| WEKA con selección | 50% | Vs. | Método propuesto con balanceo por SMOTE | 50% |
| WEKA sin selección | 26% | Vs. | Método propuesto sin balanceo | 74% |
| WEKA sin selección | 22% | Vs. | Método propuesto con balanceo por repetición | 78% |
| WEKA sin selección | 34% | Vs. | Método propuesto con balanceo por SMOTE | 66% |
| WEKA con selección | 14% | Vs. | Utilizar alguna variante del método propuesto | 86% |
| WEKA sin selección | 16% | Vs. | Utilizar alguna variante del método propuesto | 84% |

Tabla 5.5: Comparación entre los métodos probados por pares.

Finalmente en la tabla 5.6 se muestra el promedio por clasificador de todos los métodos probados en este trabajo, es decir, el promedio de los 10 selectores de atributos por clasificador y se encuentra marcado cuál método de los probados es mejor para cada clasificador.

| Experimento | Bayes | J48 | IBk=1 | IBk=3 | MLP |
|---|---|---|---|---|---|
| WEKA sin selección | 19.68 | 15.68 | 17.07 | 16.04 | 15.31 |
| WEKA con selección | 19.02 | 15.88 | 16.14 | 15.83 | 15.19 |
| Método propuesto sin balanceo | 18.10 | 15.88 | 15.55 | 15.42 | 15.44 |
| Método propuesto con repetición | 18.29 | 15.56 | 15.59 | 15.27 | 14.99 |
| Método propuesto con SMOTE | 19.15 | 15.95 | 16.33 | 15.67 | 15.35 |

Tabla 5.6: Promedio por clasificador de los selectores de atributos utilizados en los métodos probados.

Como se puede observar en la tabla 5.6 el método propuesto ya sea sin balanceo o con balanceo por repetición, presenta un promedio en el porcentaje de error menor comparado con hacer selección de atributos para todas las clases o no hacer selección de atributos. Además se puede observar que el método propuesto con balanceo por repetición es el método que en general se comporta mejor para todos los clasificadores probados.

Para comprobar si los resultados obtenidos con el método propuesto son mejores significativamente, se realizó la prueba *paired t test* [Die98] con un nivel de confianza del 90 % sobre los mejores resultados obtenidos para cada base de datos con el método de selección de atributos por clase con balanceo por repetición comparados contra los resultados obtenidos en la selección de atributos para todas las clases utilizando el mismo selector de atributos y clasificador. En la tabla 5.7 se puede observar el porcentaje de error de cada base de datos con el método propuesto y con la selección de atributos para todas las clases, así como el selector de atributos y clasificador utilizados para ambos métodos. Además se puede apreciar que en 16 de las 19 bases de datos existe una diferencia significativa entre los resultados, es decir, en el 84.21 % de los casos los resultados obtenidos con el método propuesto fueron mejores significativamente.

| Base de Datos | Evaluador | Método de Búsqueda | Clasificador | % de Error método propuesto con repetición | % de Error selección para todas las clases WEKA | Diferencia | Mejor Significativamente |
|---|---|---|---|---|---|---|---|
| Annealing | e2 | b4 | Bayes | 13.41 | 37.34 | 23.93 | Si |
| Bridges | e2 | b3 | IBk1 | 39.62 | 52.83 | 13.21 | Si |
| Credit | e2 | b2 | J48 | 14.2 | 15.65 | 1.45 | Si |
| Dermatology | e1 | b1 | J48 | 3.55 | 6.56 | 3.01 | Si |
| Diabetes | e2 | b4 | IBk3 | 26.56 | 28.13 | 1.57 | No |
| Echocardiogram | e1 | b1 | IBk3 | 33.08 | 42.42 | 9.34 | Si |
| Ecoli | e1 | b1 | IBk1 | 17.86 | 19.64 | 1.78 | No |
| Glass | e1 | b5 | Bayes | 45.33 | 52.34 | 7.01 | Si |
| Hepatitis | e1 | b4 | IBk3 | 18.06 | 23.23 | 5.17 | Si |
| Ionosphere | e2 | b3 | IBk1 | 8.55 | 13.96 | 5.41 | Si |
| Iris | e2 | b5 | J48 | 4 | 5.33 | 1.33 | Si |
| Nursery | e1 | b5 | IBk1 | 2.1 | 6.14 | 4.04 | Si |
| Optdigits | e2 | b1 | J48 | 6.72 | 18.10 | 11.38 | Si |
| Page-Blocks | e2 | b1 | Bayes | 6.38 | 9.03 | 2.65 | Si |
| Postoperative | e1 | b5 | MP | 34.44 | 41.11 | 6.67 | Si |
| Segment | e2 | b5 | Bayes | 16.19 | 21.90 | 5.71 | Si |
| Thyroid | e2 | b3 | IBk3 | 4.14 | 6.51 | 2.37 | Si |
| Votes | e2 | b4 | MP | 4.37 | 5.29 | 0.92 | No |
| Wine | e2 | b1 | IBk1 | 3.37 | 7.87 | 4.50 | Si |

Tabla 5.7: Significancia estadística de los mejores resultados del método propuesto con repetición, para cada base de datos utilizada.

## 5.7. Resumen

En este capítulo se presentaron los experimentos realizados sobre algunas bases de datos del repositorio de la universidad de California, aplicando sobre estas bases de datos el método propuesto sin balanceo, el método propuesto con balanceo por repetición y con balanceo utilizando SMOTE. Además se compararon los resultados obtenidos con cada uno de los métodos propuestos con los obtenidos haciendo una selección de atributos para todas las clases y sin hacer selección de atributos.

En los resultados presentados en este capítulo se puede observar que con el método propuesto sin balanceo se obtienen buenos resultados en los experimentos realizados con los selectores de atributos que tienen como evaluador a $e_2$. El método de selección de atributos con balanceo por repetición es el método que presentó mejores resultados en todos los experimentos realizados. Además se presentaron los resultados obtenidos con el método propuesto con balanceo utilizando SMOTE, pero se mostró que este método dió resultados

que muy pocas veces fueron mejores que realizar un balanceo con repetición o no realizar balanceo en el método propuesto, esto se puede deber a los inconvenientes que se presentan al utilizar SMOTE como generación de ejemplos sintéticos apartir de objetos que pueden ser ruido.

# Capítulo 6

# Conclusiones y trabajo futuro

A pesar de que la selección de atributos es un área ampliamente estudiada, son pocos los trabajos en donde se realiza una selección de atributos por clase. La selección de atributos por clase permite observar la relación que hay entre los atributos y cada una de las clases de un problema, dicha relación no es distinguible cuando se realiza una selección de atributos para todas las clases. En problemas multiclase debido a que cada clase es diferente a otra, es razonable pensar que dichas clases exhiban propiedades diferentes y sean descritas mejor por un conjunto diferente de atributos.

## 6.1. Conclusiones

De los resultados obtenidos en este trabajo se puede concluir que efectivamente realizar una selección de atributos para todas las clases es en la mayoría de veces mejor que no realizar selección de atributos, no obstante en clasificadores como J48 y MLP la calidad de clasificación se ve afectada al realizar selección de atributos para todas las clases. Esto último se debe a que clasificadores como J48 y MLP usualmente seleccionan atributos de manera interna al momento de la clasificación. En la tabla 5.6 se puede observar

que el método propuesto ya sea sin balanceo o con balanceo por repetición tiene un promedio de porcentaje de error por clasificador menor en todos los clasificadores probados y esto se debe a que en la mayoría de bases de datos el método propuesto reduce el porcentaje de error. Además para algunas bases de datos se disminuye en mayor grado el porcentaje de error con el método propuesto, comparado con el aumento en el porcentaje de error que se presenta al aplicar el método propuesto en algunas bases de datos utilizadas.

A partir de los experimentos realizados, se puede concluir que el método propuesto con balanceo por repetición es el método que en general se comporta mejor para todos los clasificadores probados, éste mejora en promedio la calidad de clasificación incluso para los clasificadores J48 y MLP, comparado con la selección de atributos para todas las clases y sin hacer selección de atributos. El método propuesto con balanceo por repetición además es mejor que el método con balanceo utilizando SMOTE, esto se puede deber a que con el balanceo por repetición los objetos de la clase sólo se repiten consecutivamente y con SMOTE al generar objetos sintéticos, muchos de estos objetos sintéticos pudieron haber sido generados a partir de objetos con ruido. El método propuesto con balanceo por repetición también presenta mejores resultados que el método propuesto sin balanceo, esto se debe a que cuando se balancean las clases de los subproblemas binarios antes de la selección de atributos, los atributos elegidos discriminan mejor a una clase de las demás. Los atributos seleccionados con el método propuesto sin balanceo muchas veces favorecen en mayor medida a la clase mayoritaria que usualmente suele estar formada por las clases restantes que se unieron por la técnica de binarización *uno-contra-todos*. A pesar de que a veces los atributos seleccionados no favorezcan a la clase para la cual se realizó la selección de atributos por clase, el método propuesto sin balanceo presenta un menor promedio en el porcentaje de error para clasificadores como Bayes e IBk en algunos selectores de atributos probados. El método propuesto con balanceo utilizando SMOTE para la mayoría de los clasificadores solo es mejor que no realizar una selección de atributos.

Una de las principales aportaciones de este trabajo es proponer un método en donde los selectores de atributos para todas las clases son utilizados para analizar los atributos que describen mejor a cada clase de un problema de clasificación, además se presenta un ensamble de clasificadores y una estrategia de decisión para aprovechar los resultados de la selección de atributos por clase.

## 6.2. Trabajo futuro

A partir de los resultados obtenidos con este trabajo, como trabajo futuro se plantea identificar de acuerdo a las características de las bases de datos cuál de las variantes del método propuesto es mejor aplicar. Además en base a las características de cada clase de un problema de clasificación, identificar que selector de atributos para todas las clases y que clasificador se puede emplear para cada clase con la finalidad de aumentar la calidad en la clasificación. Como trabajo futuro también se pueden realizar más experimentos utilizando otros selectores de atributos dentro el enfoque de selección de atributos para todas las clases y/o probar la selección de atributos por clase obtenida con el método propuesto con otros clasificadores.

# Referencias


[AKA91] David W. Aha, Dennis F. Kibler y Marc K. Albert, “Instance-Based Learning Algorithms,” *Machine Learning*, Vol. 6:páginas 37–66, 1991. [Citada en p. 40]

[Bag98] Paul M. Baggenstoss, “Class-Specific Feature Sets in Classification,” *In proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC)–National Institute of Standards and Technology*, páginas 413–416, 1998. [Citada en p. 20]

[Bag99] Paul M. Baggenstoss, “Class-Specific Feature Sets in Classification,” *IEEE Transactions on Signal Processing*, Vol. 47:páginas 3428–3432, 1999. [Citada en p. 20]

[Bag02] Paul M. Baggenstoss, “Class-Specific Feature Sets in Classification,” *Office Naval Research–United States of America*, 2002. [Citada en p. 20]

[Bag04] Paul M. Baggenstoss, “Class-Specific Classifier: Avoiding the Curse of Dimensionality,” *IEEE Aerospace and Electronic Systems Magazine*, Vol. 19:páginas 37–52, 2004. [Citada en p. 20]

[Bai02] Alex Bailey, *Class-Dependent Features and multicategory classification*, Tesis Doctoral, University of SouthHampton, Department of Electronics and Computer Science, 2002. [Citada en p. 9]

[BB04] Paul M. Baggenstoss y Thomas Beierholm, “Speech music discrimination using class-specific features,” *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, Vol. 2:páginas 379–382, 2004. [Citada en p. 20]

[BK99] Eric Bauer y Ron Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Machine Learning*, Vol. 36(1,2):páginas 1–38, 1999. [Citada en p. 16]

[BN00] P. Baggenstoss y H. Niemann, “A Theoretically Optimal Probabilistic Classifier using Class-Specific Features,” en “In Proceedings of International Conference on Pattern Recognition (ICPR), Barcelona,” Vol. 2, páginas 767–772, 2000. [Citada en p. 20]

[CBHK02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall y W. Philip Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligent Research (JAIR)*, Vol. 16:páginas 321–357, 2002. [Citada en p. 12]

[DB04] Jennifer G. Dy y Carla E. Brodley, “Feature Selection for Unsupervised Learning,” *Journal of Machine Learning Research*, Vol. 5:páginas 845–889, 2004. [Citada en p. 6]

[DHS00] Richard O. Duda, Peter. E. Hart y David G. Stork, *Pattern Classification, Second Edition*, John Wiley and Sons, 2000. [Citada en p. 6]

[Die98] Thomas G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,” *Neural Computation*, Vol. 10(7):páginas 1895–1924, 1998. [Citada en p. 60]

[Die00] Thomas G. Dietterich, “Ensemble Methods in Machine Learning,” en “Proceedings of the First International Workshop on Multiple Classifier Systems,” páginas 1–15, Springer Verlag, Berlin, 2000. [Citada en p. 15]

[DL97] Manoranjan Dash y Huan Liu, “Feature Selection for Classification,” *Intelligent Data Analysis*, Vol. 1(1-4):páginas 131–156, 1997. [Citada en p. 19]

[Für02] Johannes Fürnkranz, “Round Robin Classification,” *Journal of Machine Learning Research*, Vol. 2:páginas 721–747, 2002. [Citada en p. 9, 10]

[FW02] Xiuju Fu y Lipo Wang, “A GA-Based Novel RBF Classifier with Class-Dependent Features,” en “Proceedings of the 2002 Congress on Evolutionary Computation CEC2002,” Vol. 2, páginas 1890–1894, 2002. [Citada en p. 22, 24]

[FW05] Xiuju Fu y Lipo Wang, “Data Mining with Computational Intelligence,” *Springer*, páginas 145–156, 2005. [Citada en p. 2, 6, 7, 22, 24]

[Gar07] Ramon Armengol Garganté, “A Genetic Algorithm to Ensemble Feature Selection,” *Vrije Universiteit Amsterdam, Faculty of Computer Sciences*, 2007. [Citada en p. 17]

[GE03] Isabelle Guyon y André Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, Vol. 3:páginas 1157–1182, 2003. [Citada en p. 7]

[Hal98] M. Hall, *Correlation-based Feature Selection for Machine Learning*, Tesis Doctoral, Waikato University, 1998. [Citada en p. 39]

[Jap00] Nathalie Japkowicz, “The Class Imbalance Problem: Significance and Strategies,” en “Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI’2000),” Vol. 1, páginas 111–117, junio 2000. [Citada en p. 10]

[JDM00] Anil K. Jain, Robert P. W. Duin y Jianchang Mao, “Statistical Pattern Recognition: A Review,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 22(1):páginas 4–37, 2000. [Citada en p. 13, 38]

[JL95] George H. John y Pat Langley, “Estimating continuous distributions in Bayesian classifiers,” en “Proceedings 11th Conference on Uncertainty in Artificial Intelligence,” páginas 338–345, 1995. [Citada en p. 40]

[JR98] Xiuping Jia y John A. Richards, “Progressive two-class decision classifier for optimization of class discriminations,” *Remote Sensing of Environment*, Vol. 63:páginas 289–297, 1998. [Citada en p. 8]

[JS02] Nathalie Japkowicz y Shaju Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, Vol. 6(5):páginas 429–449, 2002. [Citada en p. 10, 12]

[Kay00] Steven Kay, “Sufficiency, classification, and the class-specific feature theorem,” *IEEE Transactions on Information Theory*, Vol. 46(4):páginas 1654–1658, 2000. [Citada en p. 20, 21]

[LS96] Huan Liu y Rudy Setiono, “A probabilistic approach to feature selection – a filter solution,” en “Proceedings 13th International Conference on Machine Learning,” páginas 319–327, 1996. [Citada en p. 39]

[LY02] Huan Liu y Lei Yu, “Feature Selection for Data Mining,” *Technical report Arizona State–Department of computer science and Engineering*, 2002. [Citada en p. 7]

[MM98] C. J. Merz y P. M. Murphy, “UCI Repository of machine learning databases,” *University of California at Irvine, Department of Information and Computer Science, Irvine*, 1998. [Citada en p. 36]

[Nan06] Loris Nanni, “Cluster-based pattern discrimination: A novel technique for feature selection,” *Pattern Recognition Letters*, Vol. 27(6):páginas 682–687, 2006. [Citada en p. 2, 23, 24]

[OHJ03] Mohamed Kamal Omar y Mark Hasegawa-Johnson, “Strong-sense class-dependent features for statistical recognition,” *Statistical Sig-*

*nal Processing–2003 IEEE Workshop*, páginas 490–493, 2003. [Citada en p. 21]

[OLS99] Il-Seok Oh, Jin-Seon Lee y Ching Y. Suen, “Analysis of Class Separation and Combination of Class-Dependent Features for Handwriting Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21(10):páginas 1089–1094, 1999. [Citada en p. 2, 8, 21, 22, 23]

[OM99] David W. Opitz y Richard Maclin, “Popular Ensemble Methods: An Empirical Study,” *Journal of Artificial Intelligence Research(JAIR)*, Vol. 11:páginas 169–198, 1999. [Citada en p. 15]

[Qui93] Ross Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993. [Citada en p. 40]

[SD05] Marina Skurichina y Robert P. W. Duin, “Combining Feature Subsets in Feature Selection,” en “Multiple Classifier Systems,” Vol. 3541 de *Lecture Notes in Computer Science*, páginas 165–175, Springer, 2005. [Citada en p. 20]

[SF07] Hugo Silva y Ana L. N. Fred, “Feature Subspace Ensembles: A Parallel Classifier Combination Scheme Using Feature Selection,” en “Multiple Classifier Systems,” Vol. 4472 de *Lecture Notes in Computer Science*, páginas 261–270, Springer, 2007. [Citada en p. 20]

[TM07] Wenyin Tang y K. Z. Mao, “Feature selection algorithm for mixed data with both nominal and continuous features,” *Pattern Recognition Letters*, Vol. 28(5):páginas 563–571, 2007. [Citada en p. 20]

[TPC05] Alexey Tsymbal, Mykola Pechenizkiy y Padraig Cunningham, “Diversity in search strategies for ensemble feature selection,” *Information Fusion*, Vol. 6(1):páginas 83–98, 2005. [Citada en p. 17]

[WF05] Ian H. Witten y Eibe Frank, *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2005. [Citada en p. 39]

# Apéndice A

# Apéndice

En este apéndice se muestran las tablas de los resultados obtenidos con los experimentos realizados en esta tesis. La tabla A.2 muestra el promedio del porcentaje de error obtenido con los métodos probados sobre 14 bases de datos, de esta tabla se obtuvieron las gráficas mostradas en las figuras 5.12 a 5.16. La tabla A.4 muestra el porcentaje de error obtenido por cada selector de atributos con los métodos probados, de esta tabla se obtuvieron las gráficas mostradas en las figuras 5.17 a 5.27, en donde los selectores de atributos fueron agrupados de acuerdo a la cantidad de bases de datos que se pudieron probar.

En las tablas A.6 a A.24 se presentan los porcentajes de error por base de datos obtenidos con cada uno de los selectores de atributos utilizados, por lo tanto se tienen 10 tablas una por cada selector de atributos (combinación evaluador – método de búsqueda). En estas tablas se presenta más detalladamente los porcentajes de error obtenidos por base de datos con las variantes del método propuesto y haciendo selección de atributos para todas las clases. Si se analizan con detalle estas tablas se puede observar que en varias de las bases de datos el porcentaje de error se disminuye al utilizar el método propuesto.

| Experimento | | Clasificador | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Selector de atributos** | **Bayes** | **J48** | **IBk=1** | **IBk=3** | **MLP** |
| WEKA sin Selección | Cualquiera | 19.49 | 15.87 | 17.38 | 16.47 | 15.5 |
| WEKA con Selección | e1b1 | 18.75 | 15.36 | 15.3 | 15.82 | 14.29 |
| | e1b2 | 18.54 | 15.36 | 15.3 | 15.86 | 14.29 |
| | e1b3 | 18.66 | 15.21 | 15.34 | 15.82 | 14.25 |
| | e1b4 | 18.63 | 15.39 | 15.23 | 15.81 | 14.35 |
| | e1b5 | 18.34 | 15.53 | 15.87 | 15.63 | 14.33 |
| | e2b1 | 18.78 | 16.06 | 16.97 | 16.31 | 14.87 |
| | e2b2 | 18.86 | 16.33 | 16.96 | 16.54 | 15.28 |
| | e2b3 | 18.98 | 16.54 | 17.41 | 16.64 | 15.53 |
| | e2b4 | 18.73 | 16.15 | 16.96 | 16.31 | 15.02 |
| | e2b5 | 18.77 | 16.05 | 16.97 | 16.47 | 16.11 |
| Método sin Balanceo | e1b1 | 18.63 | 16.32 | 14.88 | 15.36 | 15.35 |
| | e1b2 | 18.61 | 16.44 | 14.86 | 15.37 | 15.45 |
| | e1b3 | 18.46 | 16.42 | 14.99 | 15.25 | 15.34 |
| | e1b4 | 18.48 | 16.27 | 15.06 | 15.56 | 15.34 |
| | e1b5 | 18.55 | 15.86 | 15.51 | 15.31 | 15.14 |
| | e2b1 | 17.96 | 15.95 | 16.04 | 15.67 | 15.43 |
| | e2b2 | 17.79 | 16.16 | 16.11 | 16.11 | 15.01 |
| | e2b3 | 18.51 | 15.74 | 16.44 | 16.42 | 15.61 |
| | e2b4 | 17.97 | 15.98 | 16.57 | 15.38 | 15.25 |
| | e2b5 | 18.87 | 16.36 | 17.41 | 16.69 | 16.55 |
| Método con Repetición | e1b1 | 18.05 | 15.56 | 14.89 | 14.66 | 14.32 |
| | e1b2 | 18.04 | 15.46 | 14.95 | 14.72 | 14.32 |
| | e1b3 | 18.18 | 15.42 | 15.18 | 14.74 | 14.38 |
| | e1b4 | 18.07 | 15.57 | 14.87 | 14.62 | 14.29 |
| | e1b5 | 18.23 | 15.63 | 14.99 | 15.31 | 14.5 |
| | e2b1 | 18.75 | 15.82 | 16.82 | 15.72 | 14.95 |
| | e2b2 | 18.58 | 15.75 | 16.74 | 16.64 | 16.13 |
| | e2b3 | 18.54 | 16.01 | 16.26 | 16.32 | 14.71 |
| | e2b4 | 18.47 | 16.21 | 16.3 | 15.31 | 14.62 |
| | e2b5 | 19.01 | 16.21 | 17.93 | 16.93 | 16.43 |
| Método con SMOTE | e1b1 | 19.07 | 15.69 | 16.15 | 15.32 | 15.12 |
| | e1b2 | 19.05 | 15.83 | 15.8 | 15.8 | 15.15 |
| | e1b3 | 19 | 15.59 | 16.06 | 15.87 | 15.18 |
| | e1b4 | 19.07 | 16 | 16.06 | 15.43 | 15.19 |
| | e1b5 | 19.12 | 15.9 | 15.92 | 15.62 | 15.5 |
| | e2b1 | 18.45 | 16.23 | 16.73 | 15.91 | 15.3 |
| | e2b2 | 18.79 | 15.5 | 16.23 | 15.72 | 15.25 |
| | e2b3 | 18.72 | 16.02 | 16.7 | 15.63 | 15.2 |
| | e2b4 | 19.61 | 16.75 | 17.54 | 16.32 | 15.69 |
| | e2b5 | 18.98 | 16.09 | 17.49 | 16.78 | 16.05 |

Tabla A.2: Promedio del porcentaje de error de 14 bases de datos con los métodos probados por clasificador.

| Experimento | | Clasificador | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Selector de atributos** | **Bayes** | **J48** | **IBk=1** | **IBk=3** | **MLP** |
| WEKA sin Selección | e1b1 | 19.4 | 15.07 | 16.11 | 15.18 | 14.32 |
| | e1b3 | 19.4 | 15.07 | 16.11 | 15.18 | 14.32 |
| | e1b4 | 19.4 | 15.07 | 16.11 | 15.18 | 14.32 |
| | e2b1 | 19.4 | 15.07 | 16.11 | 15.18 | 14.32 |
| | e2b3 | 19.4 | 15.07 | 16.11 | 15.18 | 14.32 |
| | e1b2 * | 20.27 | 16.81 | 18.81 | 17.59 | 17.14 |
| | e1b5 * | 20.27 | 16.81 | 18.81 | 17.59 | 17.14 |
| | e2b2 * | 20.27 | 16.81 | 18.81 | 17.59 | 17.14 |
| | e2b5 * | 20.27 | 16.81 | 18.81 | 17.59 | 17.14 |
| | e2b4 * | 18.75 | 14.24 | 14.85 | 14.17 | 12.9 |
| WEKA con Selección | e1b1 | 18.59 | 14.87 | 14.12 | 14.59 | 13.61 |
| | e1b3 | 18.13 | 14.68 | 14.24 | 14.78 | 13.58 |
| | e1b4 | 18.52 | 14.91 | 14.16 | 14.62 | 13.7 |
| | e2b1 | 19.39 | 15.9 | 16.3 | 15.44 | 15.06 |
| | e2b3 | 19.25 | 16.3 | 16.93 | 15.9 | 15.45 |
| | e1b2 * | 19.31 | 16.33 | 16.42 | 16.88 | 15.63 |
| | e1b5 * | 19.19 | 16.42 | 17.25 | 16.81 | 16.12 |
| | e2b2 * | 19.61 | 17.24 | 18.34 | 17.36 | 17.23 |
| | e2b5 * | 19.52 | 16.98 | 18.5 | 17.3 | 17.92 |
| | e2b4 * | 18.74 | 15.17 | 15.1 | 14.63 | 13.62 |
| Método sin Balanceo | e1b1 | 17.59 | 15.06 | 13.57 | 14.27 | 14.55 |
| | e1b3 | 17.32 | 15.05 | 13.58 | 14.15 | 14.19 |
| | e1b4 | 17.5 | 15.06 | 13.74 | 14.43 | 14.56 |
| | e2b1 | 17.13 | 15.34 | 15.42 | 14.85 | 14.77 |
| | e2b3 | 17.74 | 14.94 | 15.54 | 15.2 | 14.85 |
| | e1b2 * | 19.37 | 17.27 | 15.64 | 16.27 | 16.35 |
| | e1b5 * | 19.32 | 16.8 | 16.69 | 16.44 | 16.5 |
| | e2b2 * | 18.61 | 17.09 | 17.55 | 16.97 | 16.82 |
| | e2b5 * | 19.61 | 17.27 | 18.92 | 17.51 | 18.41 |
| | e2b4 * | 16.81 | 14.96 | 14.84 | 14.13 | 13.43 |
| Método con Repetición | e1b1 | 17.16 | 14.52 | 13.45 | 13.69 | 13.6 |
| | e1b3 | 17.4 | 14.48 | 13.73 | 13.7 | 13.4 |
| | e1b4 | 17.18 | 14.5 | 13.39 | 13.64 | 13.57 |
| | e2b1 | 18.42 | 15.22 | 15.91 | 14.8 | 14.56 |
| | e2b3 | 18.48 | 15.49 | 15.49 | 15.57 | 14.3 |
| | e1b2 * | 18.84 | 16.42 | 15.73 | 15.66 | 15.36 |
| | e1b5 * | 19.09 | 16.59 | 16.06 | 16.29 | 15.83 |
| | e2b2 * | 19.34 | 16.7 | 18.07 | 17.45 | 17.87 |
| | e2b5 * | 19.74 | 17.06 | 19.25 | 17.73 | 18.29 |
| | e2b4 * | 17.21 | 14.63 | 14.83 | 14.13 | 13.1 |
| Método con SMOTE | e1b1 | 18.66 | 14.65 | 14.57 | 14.04 | 13.65 |
| | e1b3 | 18.32 | 14.67 | 14.43 | 14.42 | 13.56 |
| | e1b4 | 18.69 | 14.83 | 14.51 | 14.13 | 13.6 |
| | e2b1 | 18.41 | 15.85 | 16.25 | 15.31 | 15.07 |
| | e2b3 | 18.85 | 16 | 16.3 | 15.25 | 15.08 |
| | e1b2 * | 19.71 | 16.72 | 16.72 | 16.67 | 16.16 |
| | e1b5 * | 19.77 | 16.76 | 16.99 | 16.51 | 16.89 |
| | e2b2 * | 19.61 | 16.47 | 17.61 | 16.6 | 16.8 |
| | e2b5 * | 19.79 | 17.01 | 18.99 | 17.71 | 18.02 |
| | e2b4 * | 19.65 | 16.51 | 16.93 | 16.02 | 14.65 |

Tabla A.4: Promedio del porcentaje de error de 19 bases de datos con los métodos probados por clasificador, los * señalan los selectores de atributos en los que se probaron menos de 19 bases de datos.

| CfsSubsetEval - BestFirst | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **IBk=1** | **IBk=3** | **MP** |
| WEKA con Selección | Echocardiogram | 45.45 | 34.85 | 34.09 | 42.42 | 34.85 |
| | Postoperative | 30 | 30 | 32.22 | 31.11 | 34.44 |
| | Iris | 4 | 4 | 4 | 5.33 | 4.67 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 19.35 | 20 | 18.71 | 21.94 | 14.84 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | 11.68 | 9.4 | 10.54 | 11.4 | 8.55 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.6 |
| | Wine | 2.81 | 6.18 | 3.93 | 2.81 | 0.56 |
| | Glass | 50.93 | 30.84 | 21.03 | 24.3 | 33.18 |
| | Segment | 17.62 | 11.9 | 10.48 | 10 | 10 |
| | Dermatology | 3.55 | 6.56 | 3.83 | 3.55 | 3.55 |
| | Annealing | 37.84 | 12.03 | 6.02 | 8.02 | 9.4 |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 5.43 | 2.89 | 4.44 | 4.29 | 3.87 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3.23 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.26 |
| | Bridges | 35.85 | 29.25 | 37.74 | 38.68 | 33.96 |
| | Optdigits | 7.77 | 9.63 | 1.62 | 1.49 | 2.49 |
| Método sin Balanceo | Echocardiogram | 42.58 | 40.71 | 31.76 | 32.53 | 37.12 |
| | Postoperative | 30 | 28.89 | 26.67 | 28.89 | 28.89 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 19.35 | 20 | 18.71 | 21.94 | 14.84 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | 11.68 | 9.4 | 10.54 | 11.4 | 8.55 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.6 |
| | Wine | 2.81 | 6.7 | 2.22 | 2.22 | 2.81 |
| | Glass | 47.71 | 33.64 | 27.94 | 30.32 | 35.05 |
| | Segment | 20 | 9.05 | 7.62 | 9.05 | 10 |
| | Dermatology | 3.82 | 5.47 | 3.27 | 3.81 | 4.37 |
| | Annealing | 19.67 | 9.03 | 6.39 | 8.9 | 14.66 |
| | Ecoli | 16.67 | 16.96 | 19.64 | 19.05 | 18.45 |
| | Page-Blocks | 5.97 | 2.85 | 3.76 | 3.95 | 4.11 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3 |
| | Thyroid | 3.26 | 8.84 | 3.26 | 6.98 | 5.12 |
| | Bridges | 34.91 | 34.91 | 32.08 | 33.02 | 35.85 |
| | Optdigits | 8.29 | 4.94 | 2.75 | 3.03 | 5.02 |
| Método con Repetición | Echocardiogram | 42.42 | 37.8 | 32.36 | 33.08 | 34.85 |
| | Postoperative | 30 | 30 | 26.67 | 28.89 | 30 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.3 | 26.56 | 29.94 | 26.96 | 23.96 |
| | Hepatitis | 16.87 | 20.04 | 17.37 | 18 | 18.06 |
| | Credit | 24.2 | 16.38 | 17.83 | 14.93 | 16.52 |
| | Ionosphere | 10.27 | 9.41 | 10.27 | 12.54 | 10.26 |
| | Votes | 5.04 | 4.58 | 4.58 | 4.81 | 4.14 |
| | Wine | 2.78 | 6.14 | 3.37 | 2.78 | 1.12 |
| | Glass | 47.21 | 28.03 | 26.65 | 29.44 | 31.31 |
| | Segment | 17.14 | 11.43 | 10.48 | 9.52 | 7.14 |
| | Dermatology | 2.18 | 3.55 | 2.46 | 2.74 | 1.91 |
| | Annealing | 22.17 | 10.02 | 5.51 | 8.4 | 10.78 |
| | Ecoli | 14.88 | 15.18 | 17.86 | 15.77 | 14.58 |
| | Page-Blocks | 6.47 | 2.94 | 3.78 | 3.93 | 5.12 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.84 | 3.26 | 7.44 | 6.05 |
| | Bridges | 35.85 | 33.02 | 34.91 | 31.13 | 33.02 |
| | Optdigits | 8.74 | 4.97 | 2.17 | 2.3 | 5.02 |
| Método con SMOTE | Echocardiogram | 47.09 | 35.86 | 35.1 | 31.06 | 34.85 |
| | Postoperative | 28.89 | 29.26 | 30 | 28.89 | 30.74 |
| | Iris | 3.55 | 4 | 4.67 | 4.89 | 4.67 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 16.52 | 17.42 | 22.58 | 21.08 | 21.51 |
| | Credit | 27.92 | 15.7 | 16.43 | 15.55 | 16.43 |
| | Ionosphere | 9.31 | 9.31 | 11.02 | 11.59 | 8.83 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.14 |
| | Wine | 3.37 | 5.06 | 4.49 | 3.56 | 1.31 |
| | Glass | 46.26 | 30.69 | 26.48 | 27.41 | 32.24 |
| | Segment | 20.95 | 13.02 | 11.43 | 10.48 | 10 |
| | Dermatology | 2.82 | 4.19 | 3.64 | 3.28 | 2.46 |
| | Annealing | 38.05 | 11.03 | 4.6 | 6.93 | 3.59 |
| | Ecoli | 15.18 | 17.16 | 20.64 | 16.96 | 16.97 |
| | Page-Blocks | 6.07 | 2.87 | 3.95 | 4.01 | 4.18 |
| | Nursery | 9.75 | 3.55 | 2.89 | 2.71 | 1.16 |
| | Thyroid | 3.26 | 8.34 | 2.64 | 6.36 | 3.72 |
| | Bridges | 38.68 | 34.91 | 40.25 | 38.68 | 36.48 |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **IBk=1** | **IBk=3** | **MP** |
| | Optdigits | 8.57 | 4.84 | 1.53 | 1.52 | 2.13 |

Tabla A.6: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda BestFirst (e1b1).

| **CfsSubsetEval - ExhaustiveSearch** | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 45.45 | 34.85 | 34.09 | 42.42 | 34.85 |
| | Postoperative | 30 | 30 | 32.22 | 31.11 | 34.44 |
| | Iris | 4 | 4 | 4 | 5.33 | 4.67 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 19.35 | 20 | 18.71 | 21.94 | 14.84 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.6 |
| | Wine | 2.81 | 6.18 | 3.93 | 2.81 | 0.56 |
| | Glass | 50.93 | 30.84 | 21.03 | 24.3 | 33.18 |
| | Segment | 14.76 | 11.9 | 10.48 | 10.48 | 10 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 5.43 | 2.89 | 4.44 | 4.29 | 3.87 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3.23 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.26 |
| | Bridges | 35.85 | 29.25 | 37.74 | 38.68 | 33.96 |
| | Optdigits | x | x | x | x | x |
| Método sin Balanceo | Echocardiogram | 42.42 | 40.91 | 31.82 | 32.58 | 37.12 |
| | Postoperative | 30 | 28.89 | 26.67 | 28.89 | 28.89 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 19.35 | 20 | 18.71 | 21.94 | 14.84 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.6 |
| | Wine | 2.81 | 7.3 | 2.25 | 2.25 | 2.81 |
| | Glass | 47.66 | 33.64 | 27.57 | 30.37 | 35.05 |
| | Segment | 20 | 10 | 7.62 | 9.05 | 11.43 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 16.67 | 16.96 | 19.64 | 19.05 | 18.45 |
| | Page-Blocks | 5.97 | 2.85 | 3.76 | 3.95 | 4.11 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3 |
| | Thyroid | 3.26 | 8.84 | 3.26 | 6.98 | 5.12 |
| | Bridges | 34.91 | 34.91 | 32.08 | 33.02 | 35.85 |
| | Optdigits | x | x | x | x | x |
| Método con Repetición | Echocardiogram | 42.42 | 37.88 | 32.58 | 33.33 | 34.85 |
| | Postoperative | 30 | 30 | 26.67 | 28.89 | 30 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 16.7 | 20 | 17.42 | 18.06 | 18.06 |
| | Credit | 24.2 | 16.38 | 17.83 | 14.93 | 16.52 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.14 |
| | Wine | 2.81 | 5.06 | 3.93 | 2.81 | 1.12 |
| | Glass | 47.2 | 28.04 | 26.64 | 29.44 | 31.31 |
| | Segment | 17.14 | 10.96 | 10.48 | 10 | 7.14 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.88 | 15.18 | 17.86 | 15.77 | 14.58 |
| | Page-Blocks | 6.47 | 2.94 | 3.78 | 3.93 | 5.12 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.84 | 3.26 | 7.44 | 6.05 |
| | Bridges | 35.85 | 33.02 | 34.91 | 31.13 | 33.02 |
| | Optdigits | x | x | x | x | x |

Continuación...

| Método | Base de Datos | Bayes | J48 | KNN1 | KNN3 | MP |
|---|---|---|---|---|---|---|
| Método con SMOTE | Echocardiogram | 49.24 | 40.91 | 35.86 | 34.85 | 35.35 |
| | Postoperative | 28.89 | 29.26 | 29.63 | 28.89 | 30.37 |
| | Iris | 3.78 | 4 | 4.45 | 4.89 | 3.55 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 15.91 | 17.42 | 21.29 | 21.08 | 21.29 |
| | Credit | 27.92 | 15.36 | 17.34 | 15.56 | 16.86 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.14 |
| | Wine | 3.37 | 4.87 | 4.68 | 2.62 | 2.44 |
| | Glass | 45.95 | 29.91 | 25.39 | 25.7 | 32.4 |
| | Segment | 19.84 | 12.7 | 10.79 | 10.95 | 9.21 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 15.08 | 17.26 | 20.04 | 17.06 | 16.27 |
| | Page-Blocks | 5.94 | 2.9 | 4.04 | 4.06 | 4.17 |
| | Nursery | 9.74 | 3.52 | 2.33 | 6.96 | 1.17 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 5.09 | 3.88 |
| | Bridges | 38.36 | 33.65 | 37.59 | 40.57 | 37.42 |
| | Optdigits | x | x | x | x | x |

Tabla A.8: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda ExhaustiveSearch (e1b2).

| CfsSubsetEval - GeneticSearch | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 45.45 | 34.85 | 34.09 | 42.42 | 34.85 |
| | Postoperative | 30 | 30 | 32.22 | 31.11 | 34.44 |
| | Iris | 4 | 4 | 4 | 5.33 | 4.67 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 18.71 | 20 | 19.35 | 20.65 | 15.48 |
| | Credit | 24.49 | 15.36 | 18.12 | 15.36 | 15.8 |
| | Ionosphere | 11.4 | 9.12 | 10.26 | 12.25 | 8.83 |
| | Votes | 5.06 | 3.68 | 4.37 | 4.83 | 4.37 |
| | Wine | 2.81 | 6.18 | 3.93 | 2.81 | 0.56 |
| | Glass | 50.93 | 30.84 | 21.03 | 24.3 | 33.18 |
| | Segment | 17.62 | 10.48 | 10.48 | 10.48 | 8.57 |
| | Dermatology | 1.64 | 4.92 | 3.83 | 3.28 | 3.55 |
| | Annealing | 31.45 | 12.03 | 7.52 | 10.78 | 8.65 |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 5.43 | 2.89 | 4.44 | 4.29 | 3.87 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3.23 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.26 |
| | Bridges | 35.85 | 29.25 | 37.74 | 38.68 | 33.96 |
| | Optdigits | 8.66 | 9.84 | 2.09 | 1.86 | 3.09 |
| Método sin Balanceo | Echocardiogram | 42.42 | 40.91 | 31.82 | 32.58 | 37.12 |
| | Postoperative | 30 | 28.89 | 26.67 | 28.89 | 28.89 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 18.71 | 20 | 19.35 | 20.65 | 15.48 |
| | Credit | 24.49 | 15.36 | 18.12 | 15.36 | 15.8 |
| | Ionosphere | 11.4 | 9.12 | 10.26 | 12.25 | 8.83 |
| | Votes | 5.06 | 3.68 | 4.37 | 4.83 | 4.37 |
| | Wine | 2.81 | 7.3 | 2.25 | 1.69 | 2.81 |
| | Glass | 47.71 | 32.71 | 26.17 | 28.5 | 34.11 |
| | Segment | 19.52 | 10 | 8.57 | 8.57 | 9.05 |
| | Dermatology | 3.01 | 3.28 | 2.73 | 3.01 | 2.73 |
| | Annealing | 17.92 | 10.28 | 6.64 | 8.9 | 11.03 |
| | Ecoli | 17.56 | 18.15 | 20.54 | 19.94 | 20.24 |
| | Page-Blocks | 3.76 | 2.87 | 3.69 | 3.93 | 4.15 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3 |
| | Thyroid | 3.26 | 8.84 | 3.26 | 6.98 | 5.12 |
| | Bridges | 35.85 | 34.91 | 33.02 | 33.96 | 34.91 |
| | Optdigits | 8.19 | 4.53 | 1.91 | 2.25 | 3.4 |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| Método con Repetición | Echocardiogram | 42.42 | 37.88 | 32.58 | 33.33 | 34.85 |
| | Postoperative | 30 | 30 | 26.67 | 28.89 | 30 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 16.77 | 20.65 | 20 | 18.06 | 18.71 |
| | Credit | 25.07 | 16.09 | 18.99 | 15.22 | 16.81 |
| | Ionosphere | 10.54 | 10.83 | 11.4 | 12.82 | 9.4 |
| | Votes | 5.06 | 3.68 | 4.37 | 4.83 | 4.37 |
| | Wine | 2.81 | 5.06 | 3.93 | 2.81 | 1.12 |
| | Glass | 47.2 | 28.5 | 26.64 | 28.97 | 30.37 |
| | Segment | 18.1 | 10.48 | 10 | 10.48 | 7.62 |
| | Dermatology | 1.64 | 3.55 | 2.19 | 2.19 | 2.73 |
| | Annealing | 25.81 | 10.28 | 6.39 | 8.15 | 8.52 |
| | Ecoli | 14.88 | 15.18 | 17.86 | 15.77 | 14.58 |
| | Page-Blocks | 6.54 | 2.96 | 3.87 | 3.96 | 5.21 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.84 | 3.26 | 7.44 | 6.05 |
| | Bridges | 35.85 | 33.02 | 34.91 | 31.13 | 33.02 |
| | Optdigits | 8.03 | 4.6 | 1.67 | 1.86 | 2.64 |
| Método con SMOTE | Echocardiogram | 51.26 | 39.39 | 35.36 | 34.09 | 37.63 |
| | Postoperative | 28.89 | 29.26 | 29.26 | 28.89 | 30 |
| | Iris | 3.33 | 4 | 4.22 | 5.11 | 3.78 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 16.77 | 16.77 | 23.44 | 22.37 | 20.86 |
| | Credit | 25.31 | 15.22 | 18.02 | 15.89 | 17.15 |
| | Ionosphere | 10.26 | 9.97 | 10.35 | 11.49 | 7.98 |
| | Votes | 5.06 | 3.68 | 4.37 | 4.83 | 4.37 |
| | Wine | 3 | 4.87 | 5.43 | 4.12 | 2.62 |
| | Glass | 45.64 | 29.91 | 25.08 | 28.19 | 31.93 |
| | Segment | 20 | 11.43 | 10.95 | 10.79 | 8.41 |
| | Dermatology | 2 | 3.56 | 3.19 | 2.92 | 2.1 |
| | Annealing | 32.12 | 12.49 | 4.97 | 7.06 | 3.26 |
| | Ecoli | 14.68 | 17.86 | 20.83 | 16.77 | 16.37 |
| | Page-Blocks | 6.1 | 2.84 | 4.01 | 4.09 | 4.06 |
| | Nursery | 9.77 | 3.51 | 2.85 | 2.85 | 1.12 |
| | Thyroid | 3.1 | 8.53 | 2.64 | 6.2 | 3.72 |
| | Bridges | 38.68 | 33.65 | 37.74 | 39.94 | 36.48 |
| | Optdigits | 8.72 | 5.28 | 1.46 | 1.45 | 1.94 |

Tabla A.10: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda GeneticSearch (e1b3).

| **CfsSubsetEval - GreedyStepwise** | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 45.45 | 34.85 | 34.09 | 42.42 | 34.85 |
| | Postoperative | 30 | 30 | 32.22 | 31.11 | 34.44 |
| | Iris | 4 | 4 | 4 | 5.33 | 4.67 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 18.71 | 20 | 18.71 | 23.23 | 16.13 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | 11.68 | 9.97 | 10.83 | 11.68 | 8.83 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.6 |
| | Wine | 2.81 | 6.18 | 3.93 | 2.81 | 0.56 |
| | Glass | 50.93 | 30.84 | 21.03 | 24.3 | 33.18 |
| | Segment | 16.67 | 12.38 | 9.52 | 8.57 | 9.52 |
| | Dermatology | 1.91 | 5.46 | 4.37 | 3.28 | 3.28 |
| | Annealing | 39.6 | 12.78 | 6.77 | 8.77 | 10.28 |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 5.43 | 2.89 | 4.44 | 4.29 | 3.87 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3.23 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.26 |
| | Bridges | 35.85 | 29.25 | 37.74 | 38.68 | 33.96 |
| | Optdigits | 7.77 | 9.63 | 1.62 | 1.49 | 2.49 |

Continuación...

| Método | Base de Datos | Bayes | J48 | KNN1 | KNN3 | MP |
|---|---|---|---|---|---|---|
| Método sin Balanceo | Echocardiogram | 42.42 | 40.91 | 31.82 | 32.58 | 37.12 |
| | Postoperative | 30 | 28.89 | 26.67 | 28.89 | 28.89 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 24.61 |
| | Hepatitis | 18.71 | 20 | 18.71 | 23.23 | 16.13 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | 11.68 | 9.97 | 10.83 | 11.68 | 8.83 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.6 |
| | Wine | 2.81 | 6.18 | 2.25 | 2.25 | 2.25 |
| | Glass | 47.66 | 32.24 | 28.5 | 29.91 | 34.58 |
| | Segment | 19.05 | 9.05 | 7.62 | 8.57 | 9.52 |
| | Dermatology | 3.83 | 5.46 | 3.28 | 3.83 | 4.37 |
| | Annealing | 19.67 | 9.02 | 6.39 | 8.9 | 14.66 |
| | Ecoli | 16.37 | 16.96 | 19.64 | 19.35 | 18.45 |
| | Page-Blocks | 6.07 | 2.92 | 3.84 | 4 | 4.13 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3 |
| | Thyroid | 3.26 | 8.84 | 3.26 | 6.98 | 5.12 |
| | Bridges | 34.91 | 35.85 | 33.96 | 34.91 | 35.85 |
| | Optdigits | 8.58 | 5.02 | 2.96 | 3.17 | 5.23 |
| Método con Repetición | Echocardiogram | 42.42 | 37.88 | 32.58 | 33.33 | 34.85 |
| | Postoperative | 30 | 30 | 26.67 | 28.89 | 30 |
| | Iris | 4 | 4 | 4 | 5.33 | 4 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 16.77 | 20 | 17.42 | 18.06 | 18.06 |
| | Credit | 24.2 | 16.38 | 17.83 | 14.93 | 16.52 |
| | Ionosphere | 10.26 | 9.4 | 9.4 | 12.25 | 10.54 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.14 |
| | Wine | 2.81 | 6.18 | 2.81 | 2.81 | 1.69 |
| | Glass | 47.66 | 28.5 | 26.64 | 28.5 | 29.91 |
| | Segment | 16.67 | 10.95 | 10.48 | 9.52 | 7.62 |
| | Dermatology | 2.19 | 3.28 | 2.46 | 2.73 | 2.19 |
| | Annealing | 22.31 | 10.03 | 5.51 | 8.4 | 10.78 |
| | Ecoli | 15.18 | 15.18 | 17.86 | 15.77 | 14.58 |
| | Page-Blocks | 6.47 | 2.94 | 3.78 | 3.93 | 5.12 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.84 | 3.26 | 7.44 | 6.05 |
| | Bridges | 35.85 | 33.02 | 34.91 | 31.13 | 33.02 |
| | Optdigits | 8.71 | 4.92 | 2.22 | 2.33 | 4.26 |
| Método con SMOTE | Echocardiogram | 46.77 | 40.66 | 34.6 | 32.58 | 34.85 |
| | Postoperative | 28.89 | 28.89 | 30 | 28.69 | 30 |
| | Iris | 3.78 | 4 | 4.22 | 4.67 | 3.78 |
| | Diabetes | 23.31 | 26.56 | 29.95 | 26.95 | 23.96 |
| | Hepatitis | 16.77 | 18.06 | 22.58 | 20.65 | 23.01 |
| | Credit | 27.44 | 15.7 | 17.29 | 15.46 | 16.76 |
| | Ionosphere | 9.31 | 8.83 | 10.83 | 11.78 | 7.12 |
| | Votes | 5.06 | 4.6 | 4.6 | 4.83 | 4.23 |
| | Wine | 3.37 | 4.87 | 4.31 | 3.18 | 2.06 |
| | Glass | 46.26 | 29.91 | 25.55 | 27.73 | 32.24 |
| | Segment | 21.9 | 13.02 | 11.11 | 10.79 | 8.89 |
| | Dermatology | 2.19 | 4.01 | 3.83 | 3.1 | 2.55 |
| | Annealing | 39.31 | 11.07 | 4.72 | 7.27 | 3.51 |
| | Ecoli | 14.88 | 17.66 | 20.93 | 16.67 | 16.47 |
| | Page-Blocks | 6.23 | 2.91 | 3.89 | 3.97 | 4.3 |
| | Nursery | 9.73 | 3.62 | 2.75 | 2.7 | 1.18 |
| | Thyroid | 3.1 | 7.91 | 2.79 | 6.2 | 3.88 |
| | Bridges | 38.37 | 34.59 | 40.25 | 39.62 | 37.11 |
| | Optdigits | 8.5 | 4.94 | 1.49 | 1.58 | 2.47 |

Tabla A.12: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda GreedyStepwise (e1b4).

| CfsSubsetEval - RandomSearch | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 39.39 | 34.85 | 33.33 | 37.12 | 35.61 |
| | Postoperative | 31.11 | 28.89 | 36.67 | 33.33 | 41.11 |
| | Iris | 4.67 | 4 | 5.33 | 4.67 | 3.33 |
| | Diabetes | 22.92 | 26.82 | 30.34 | 28.13 | 24.74 |
| | Hepatitis | 18.71 | 20 | 18.71 | 18.71 | 14.19 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.75 | 4.6 | 5.06 | 5.06 | 5.06 |
| | Wine | 2.81 | 6.18 | 3.93 | 2.81 | 0.56 |
| | Glass | 52.34 | 31.31 | 25.23 | 26.64 | 33.18 |
| | Segment | 14.29 | 11.9 | 10.48 | 10.48 | 9.05 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 6.83 | 2.98 | 5.43 | 5.04 | 5.39 |
| | Nursery | 10.54 | 5.26 | 6.14 | 4.86 | 3.9 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.26 |
| | Bridges | 35.85 | 30.19 | 37.74 | 39.62 | 33.96 |
| | Optdigits | x | x | x | x | x |
| Método sin Balanceo | Echocardiogram | 44.7 | 39.39 | 38.64 | 37.12 | 39.39 |
| | Postoperative | 30 | 30 | 33.33 | 32.22 | 35.56 |
| | Iris | 4.67 | 4 | 5.33 | 4.67 | 4.67 |
| | Diabetes | 22.92 | 26.82 | 30.34 | 28.13 | 24.74 |
| | Hepatitis | 18.71 | 20 | 18.71 | 18.71 | 14.19 |
| | Credit | 25.07 | 15.07 | 17.97 | 14.64 | 15.36 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.75 | 4.6 | 5.06 | 5.06 | 5.06 |
| | Wine | 2.25 | 6.18 | 2.25 | 2.25 | 2.25 |
| | Glass | 48.6 | 29.91 | 26.17 | 29.91 | 34.58 |
| | Segment | 18.57 | 10.48 | 7.62 | 9.05 | 9.52 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.29 | 14.88 | 16.96 | 15.18 | 13.99 |
| | Page-Blocks | 5.99 | 2.83 | 3.69 | 3.98 | 3.71 |
| | Nursery | 10.05 | 4.58 | 4.77 | 4.26 | 3 |
| | Thyroid | 3.26 | 8.37 | 2.79 | 6.51 | 3.72 |
| | Bridges | 34.91 | 34.91 | 36.79 | 34.91 | 37.74 |
| | Optdigits | x | x | x | x | x |
| Método con Repetición | Echocardiogram | 43.94 | 37.12 | 31.82 | 35.61 | 37.12 |
| | Postoperative | 31.11 | 30 | 31.11 | 30 | 34.44 |
| | Iris | 4.67 | 4 | 5.33 | 4.67 | 4.67 |
| | Diabetes | 22.92 | 26.62 | 30.34 | 28.12 | 23.96 |
| | Hepatitis | 16.77 | 20 | 17.42 | 20 | 17.42 |
| | Credit | 24.2 | 16.38 | 17.83 | 14.93 | 16.52 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.75 | 4.6 | 5.06 | 5.06 | 4.83 |
| | Wine | 2.81 | 5.06 | 3.93 | 2.81 | 1.12 |
| | Glass | 45.33 | 29.91 | 26.64 | 28.5 | 33.64 |
| | Segment | 16.19 | 10.48 | 10.48 | 10 | 7.62 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.29 | 15.18 | 18.45 | 14.58 | 13.1 |
| | Page-Blocks | 6.69 | 2.78 | 3.71 | 3.75 | 4.71 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.72 |
| | Bridges | 38.68 | 35.85 | 33.96 | 37.74 | 33.96 |
| | Optdigits | x | x | x | x | x |
| Método con SMOTE | Echocardiogram | 48.99 | 36.36 | 35.86 | 30.3 | 40.66 |
| | Postoperative | 28.89 | 28.89 | 31.85 | 28.89 | 36.3 |
| | Iris | 4.67 | 4 | 5.33 | 5.11 | 4.45 |
| | Diabetes | 22.92 | 26.77 | 30.34 | 28.12 | 23.96 |
| | Hepatitis | 15.91 | 17.2 | 23.87 | 20.65 | 21.94 |
| | Credit | 27.06 | 15.8 | 16.62 | 15.65 | 16.38 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 5.75 | 4.6 | 5.06 | 5.06 | 4.83 |
| | Wine | 3 | 5.62 | 4.31 | 3.56 | 2.62 |
| | Glass | 46.42 | 31.31 | 23.83 | 30.06 | 33.49 |
| | Segment | 20.79 | 14.38 | 10.79 | 10 | 8.41 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.98 | 18.05 | 20.83 | 16.96 | 16.67 |
| | Page-Blocks | 5.82 | 2.75 | 3.64 | 3.89 | 3.55 |
| | Nursery | 9.73 | 3.5 | 2.86 | 2.71 | 1.1 |
| | Thyroid | 3.26 | 7.91 | 2.79 | 6.51 | 3.72 |
| | Bridges | 38.37 | 34.28 | 36.79 | 40.11 | 35.22 |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| | Optdigits | x | x | x | x | x |

Tabla A.14: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador CfsSubsetEval y Método de búsqueda RandomSearch (e1b5).

| **ConsistencySubsetEval - BestFirst** | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 43.18 | 40.15 | 37.88 | 40.15 | 32.58 |
| | Postoperative | 30 | 30 | 37.78 | 28.89 | 44.44 |
| | Iris | 4.67 | 4 | 4.67 | 4.67 | 3.33 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 17.42 | 21.94 | 21.94 | 18.71 | 19.35 |
| | Credit | 21.16 | 15.07 | 19.71 | 15.65 | 16.67 |
| | Ionosphere | 15.67 | 9.69 | 11.97 | 13.39 | 11.4 |
| | Votes | 8.05 | 3.22 | 6.9 | 6.44 | 6.67 |
| | Wine | 7.3 | 5.62 | 7.87 | 7.87 | 6.74 |
| | Glass | 49.07 | 31.31 | 23.36 | 27.57 | 32.24 |
| | Segment | 18.1 | 10.48 | 10.95 | 9.52 | 7.14 |
| | Dermatology | 3.55 | 5.74 | 6.28 | 6.01 | 4.1 |
| | Annealing | 37.34 | 13.78 | 2.51 | 4.26 | 1.63 |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 9.03 | 2.98 | 4.07 | 4 | 3.73 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 1.31 |
| | Thyroid | 3.26 | 9.77 | 2.79 | 6.51 | 6.05 |
| | Bridges | 33.02 | 33.02 | 46.23 | 42.45 | 34.91 |
| | Optdigits | 19.02 | 18.1 | 13.6 | 12.4 | 16.32 |
| Método sin Balanceo | Echocardiogram | 37.88 | 35.61 | 31.06 | 37.12 | 34.85 |
| | Postoperative | 30 | 30 | 37.78 | 28.89 | 42.22 |
| | Iris | 5.33 | 5.33 | 6.67 | 4 | 5.33 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 17.42 | 21.94 | 21.94 | 18.71 | 19.35 |
| | Credit | 21.16 | 15.07 | 19.71 | 15.65 | 16.67 |
| | Ionosphere | 15.67 | 9.69 | 11.97 | 13.39 | 11.4 |
| | Votes | 8.05 | 3.22 | 6.9 | 6.44 | 6.67 |
| | Wine | 5.62 | 3.93 | 3.37 | 3.37 | 4.49 |
| | Glass | 49.53 | 32.24 | 28.89 | 28.04 | 32.24 |
| | Segment | 19.05 | 13.81 | 10.95 | 10.95 | 13.33 |
| | Dermatology | 5.46 | 7.92 | 8.48 | 8.47 | 5.19 |
| | Annealing | 14.04 | 15.29 | 7.14 | 9.02 | 1.25 |
| | Ecoli | 13.69 | 16.37 | 18.75 | 15.48 | 15.48 |
| | Page-Blocks | 6.83 | 3.07 | 4.33 | 4.17 | 3.58 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 3.26 | 9.77 | 2.79 | 5.58 | 5.12 |
| | Bridges | 29.25 | 32.08 | 37.74 | 39.62 | 33.96 |
| | Optdigits | 8.87 | 5.26 | 3.01 | 2.98 | 4.47 |
| Método con Repetición | Echocardiogram | 40.15 | 38.64 | 41.67 | 34.85 | 28.79 |
| | Postoperative | 30 | 30 | 36.67 | 28.89 | 42.22 |
| | Iris | 4 | 4 | 5.33 | 5.33 | 4 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 19.35 | 20 | 23.23 | 20 | 18.06 |
| | Credit | 24.06 | 15.07 | 20 | 17.1 | 19.13 |
| | Ionosphere | 12.82 | 11.4 | 9.12 | 10.83 | 9.69 |
| | Votes | 8.05 | 3.22 | 6.9 | 6.44 | 6.21 |
| | Wine | 5.62 | 4.49 | 3.37 | 3.93 | 6.74 |
| | Glass | 45.78 | 29.91 | 28.5 | 28.97 | 36.45 |
| | Segment | 20.95 | 12.81 | 12.86 | 12.86 | 9.52 |
| | Dermatology | 3.55 | 5.46 | 8.47 | 7.65 | 5.74 |
| | Annealing | 29.7 | 14.16 | 6.39 | 8.15 | 1.88 |
| | Ecoli | 14.29 | 15.77 | 19.64 | 15.48 | 13.39 |
| | Page-Blocks | 6.38 | 2.94 | 4.33 | 4.33 | 4.17 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.84 | 2.33 | 4.65 | 4.19 |
| | Bridges | 36.79 | 34.91 | 35.85 | 35.85 | 31.13 |
| | Optdigits | 11.38 | 6.72 | 6.04 | 5.73 | 7.69 |

| Contuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| Método con SMOTE | Echocardiogram | 42.93 | 40.66 | 36.11 | 37.12 | 35.61 |
| | Postoperative | 31.11 | 30 | 37.41 | 28.89 | 40 |
| | Iris | 4.45 | 4.67 | 6.89 | 4.89 | 4.44 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 17.63 | 22.37 | 25.59 | 20 | 20.22 |
| | Credit | 23.33 | 13.96 | 19.28 | 15.12 | 17.49 |
| | Ionosphere | 11.49 | 11.59 | 9.4 | 12.06 | 10.45 |
| | Votes | 8.05 | 3.22 | 6.9 | 6.44 | 6.21 |
| | Wine | 3.56 | 5.43 | 3.56 | 3 | 4.5 |
| | Glass | 49.69 | 29.13 | 31.78 | 29.75 | 31 |
| | Segment | 17.62 | 11.43 | 7.94 | 10.48 | 10.48 |
| | Dermatology | 6.01 | 7.83 | 9.02 | 7.19 | 7.56 |
| | Annealing | 26.52 | 13.41 | 6.89 | 8.52 | 2.59 |
| | Ecoli | 14.78 | 16.27 | 19.15 | 14.88 | 13.79 |
| | Page-Blocks | 6.7 | 2.9 | 3.86 | 3.99 | 3.88 |
| | Nursery | 9.85 | 4.08 | 2.84 | 2.84 | 2.08 |
| | Thyroid | 2.33 | 9.3 | 2.48 | 5.58 | 5.12 |
| | Bridges | 32.71 | 35.85 | 38.37 | 40.57 | 32.39 |
| | Optdigits | 16.5 | 11.1 | 11.78 | 11.45 | 11.57 |

Tabla A.16: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda BestFirst (e2b1).

| **ConsistencySubsetEval - ExhaustiveSearch** | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 42.42 | 37.12 | 37.88 | 38.64 | 33.33 |
| | Postoperative | 30 | 30 | 37.78 | 28.89 | 44.44 |
| | Iris | 4.67 | 4 | 4.67 | 4.67 | 3.33 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 15.48 | 20 | 20 | 16.77 | 14.84 |
| | Credit | 21.45 | 15.65 | 20 | 16.52 | 17.83 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 6.44 | 3.22 | 4.6 | 5.75 | 3.91 |
| | Wine | 7.3 | 6.18 | 5.06 | 6.74 | 8.43 |
| | Glass | 49.07 | 31.31 | 23.36 | 27.57 | 32.24 |
| | Segment | 21.43 | 12.38 | 10.95 | 9.52 | 8.1 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 9.03 | 2.98 | 4.07 | 4 | 3.73 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 1.31 |
| | Thyroid | 3.26 | 9.77 | 2.79 | 6.51 | 6.05 |
| | Bridges | 34.91 | 38.68 | 52.83 | 50 | 43.4 |
| | Optdigits | x | x | x | x | x |
| Método sin Balanceo | Echocardiogram | 37.88 | 38.64 | 35.61 | 39.39 | 31.06 |
| | Postoperative | 30 | 30 | 37.78 | 28.89 | 42.22 |
| | Iris | 5.33 | 5.33 | 6.67 | 5.33 | 6 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 15.48 | 20 | 20 | 16.77 | 14.84 |
| | Credit | 21.45 | 15.65 | 20 | 16.52 | 17.83 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 6.44 | 3.22 | 4.6 | 5.75 | 3.91 |
| | Wine | 5.62 | 5.06 | 2.81 | 3.37 | 2.81 |
| | Glass | 48.13 | 30.84 | 26.17 | 28.97 | 33.64 |
| | Segment | 18.1 | 10.95 | 9.05 | 10 | 11.9 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.29 | 16.07 | 19.05 | 15.18 | 14.88 |
| | Page-Blocks | 6.71 | 3.07 | 4.28 | 4.18 | 3.58 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 3.26 | 9.77 | 2.33 | 5.58 | 5.12 |
| | Bridges | 32.08 | 36.79 | 43.4 | 44.34 | 39.62 |
| | Optdigits | x | x | x | x | x |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| Método con Repetición | Echocardiogram | 40.91 | 38.64 | 37.88 | 37.12 | 36.36 |
| | Postoperative | 30 | 30 | 36.67 | 28.89 | 42.22 |
| | Iris | 4 | 4 | 4.67 | 5.33 | 4.67 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 15.48 | 23.23 | 21.94 | 20 | 18.71 |
| | Credit | 22.61 | 14.2 | 19.71 | 17.1 | 17.83 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 6.44 | 3.22 | 4.6 | 5.75 | 4.37 |
| | Wine | 5.62 | 2.25 | 4.49 | 6.18 | 5.06 |
| | Glass | 43.93 | 29.44 | 28.04 | 31.31 | 34.58 |
| | Segment | 24.29 | 13.33 | 13.33 | 12.86 | 15.71 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.58 | 16.67 | 19.64 | 15.18 | 14.29 |
| | Page-Blocks | 6.4 | 2.76 | 4.18 | 4.28 | 3.91 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.84 | 2.79 | 4.19 | 6.05 |
| | Bridges | 38.68 | 33.02 | 41.51 | 43.4 | 36.79 |
| | Optdigits | x | x | x | x | x |
| Método con SMOTE | Echocardiogram | 43.69 | 38.13 | 36.11 | 35.1 | 37.88 |
| | Postoperative | 31.11 | 30 | 37.04 | 28.89 | 38.52 |
| | Iris | 4.89 | 4.45 | 6.22 | 4.89 | 4.45 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 18.5 | 20 | 21.94 | 18.49 | 19.14 |
| | Credit | 22.8 | 14.35 | 18.07 | 15.41 | 17.3 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 6.44 | 3.22 | 4.6 | 5.75 | 4.37 |
| | Wine | 3.93 | 4.87 | 3.18 | 2.44 | 4.68 |
| | Glass | 50.78 | 27.73 | 31 | 29.28 | 32.71 |
| | Segment | 19.21 | 11.43 | 10.16 | 10.16 | 11.11 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.39 | 16.17 | 19.05 | 14.68 | 13.39 |
| | Page-Blocks | 7.04 | 2.85 | 3.84 | 4.04 | 4.08 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 8.68 | 2.48 | 5.58 | 5.43 |
| | Bridges | 34.28 | 34.28 | 38.99 | 44.03 | 31.45 |
| | Optdigits | x | x | x | x | x |

Tabla A.18: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda ExhaustiveSearch (e2b2).

| **ConsistencySubsetEval - GeneticSearch** | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 44.7 | 37.88 | 40.91 | 37.12 | 36.36 |
| | Postoperative | 30 | 30 | 36.67 | 28.89 | 43.33 |
| | Iris | 4.67 | 4 | 4.67 | 4.67 | 3.33 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 16.13 | 20.65 | 18.71 | 17.42 | 17.42 |
| | Credit | 22.17 | 14.78 | 19.71 | 16.67 | 17.68 |
| | Ionosphere | 15.95 | 9.69 | 13.96 | 13.11 | 11.4 |
| | Votes | 6.67 | 3.22 | 6.44 | 5.98 | 5.29 |
| | Wine | 5.06 | 6.74 | 6.18 | 7.3 | 5.06 |
| | Glass | 49.07 | 31.31 | 23.36 | 27.57 | 32.24 |
| | Segment | 21.43 | 13.33 | 12.86 | 11.9 | 9.05 |
| | Dermatology | 4.64 | 5.19 | 6.28 | 5.74 | 5.46 |
| | Annealing | 28.32 | 14.91 | 8.77 | 9.9 | 1.88 |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 9.03 | 2.98 | 4.07 | 4 | 3.73 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 1.31 |
| | Thyroid | 3.26 | 9.77 | 2.79 | 6.51 | 6.05 |
| | Bridges | 34.91 | 39.62 | 52.83 | 49.06 | 42.45 |
| | Optdigits | 21.06 | 18.41 | 12.29 | 11.54 | 14.13 |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| Método sin Balanceo | Echocardiogram | 43.18 | 34.09 | 35.61 | 40.15 | 31.65 |
| | Postoperative | 30 | 30 | 37.78 | 28.89 | 41.11 |
| | Iris | 5.33 | 5.33 | 6.67 | 5.33 | 6 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 16.13 | 20.65 | 18.71 | 17.42 | 17.42 |
| | Credit | 22.17 | 14.78 | 19.71 | 16.67 | 17.68 |
| | Ionosphere | 15.95 | 9.69 | 13.96 | 13.11 | 11.4 |
| | Votes | 6.67 | 3.22 | 6.44 | 5.98 | 5.29 |
| | Wine | 5.06 | 3.93 | 2.81 | 1.69 | 4.49 |
| | Glass | 50.47 | 33.64 | 33.18 | 35.05 | 35.51 |
| | Segment | 18.57 | 10 | 8.1 | 10.95 | 11.43 |
| | Dermatology | 2.46 | 3.01 | 3.01 | 3.83 | 3.83 |
| | Annealing | 17.92 | 14.16 | 6.64 | 9.15 | 2.01 |
| | Ecoli | 13.99 | 16.07 | 19.35 | 16.07 | 14.88 |
| | Page-Blocks | 6.96 | 3.03 | 4.28 | 4.17 | 3.58 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 3.26 | 9.77 | 2.33 | 5.58 | 5.12 |
| | Bridges | 33.02 | 34.91 | 41.51 | 40.57 | 40.57 |
| | Optdigits | 11.61 | 6.64 | 3.74 | 3.92 | 5.21 |
| Método con Repetición | Echocardiogram | 41.67 | 36.36 | 35.61 | 38.64 | 30.3 |
| | Postoperative | 31.11 | 30 | 36.67 | 28.89 | 41.11 |
| | Iris | 4 | 4 | 5.33 | 5.33 | 4 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 13.55 | 21.29 | 20 | 17.42 | 15.48 |
| | Credit | 23.91 | 14.49 | 20 | 16.38 | 18.26 |
| | Ionosphere | 17.95 | 9.69 | 8.55 | 13.68 | 8.55 |
| | Votes | 6.67 | 3.22 | 6.44 | 5.98 | 5.52 |
| | Wine | 5.06 | 6.18 | 5.62 | 5.06 | 5.06 |
| | Glass | 47.66 | 33.18 | 27.1 | 28.5 | 32.71 |
| | Segment | 21.43 | 9.52 | 11.43 | 11.9 | 11.43 |
| | Dermatology | 3.55 | 3.55 | 4.1 | 4.37 | 3.28 |
| | Annealing | 21.18 | 15.16 | 8.65 | 11.78 | 3.01 |
| | Ecoli | 13.99 | 16.96 | 18.45 | 14.88 | 14.29 |
| | Page-Blocks | 7.24 | 2.94 | 3.73 | 3.78 | 4.53 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.33 | 8.37 | 2.79 | 4.14 | 3.72 |
| | Bridges | 37.74 | 36.79 | 39.62 | 46.23 | 33.02 |
| | Optdigits | 17.82 | 11.8 | 8.63 | 8.74 | 9.97 |
| Método con SMOTE | Echocardiogram | 43.94 | 37.63 | 37.37 | 36.36 | 38.13 |
| | Postoperative | 31.11 | 30 | 37.04 | 28.89 | 38.52 |
| | Iris | 4.67 | 4.89 | 6.67 | 5.34 | 4.67 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 17.85 | 20.86 | 19.36 | 17.85 | 18.93 |
| | Credit | 23.33 | 14.2 | 19.13 | 14.69 | 17.29 |
| | Ionosphere | 16.9 | 11.68 | 11.11 | 13.01 | 8.74 |
| | Votes | 6.67 | 3.22 | 6.44 | 5.98 | 5.52 |
| | Wine | 3.18 | 6.56 | 4.31 | 3.37 | 4.68 |
| | Glass | 48.13 | 29.13 | 31.16 | 30.53 | 31.93 |
| | Segment | 20 | 10.47 | 10 | 9.84 | 9.21 |
| | Dermatology | 2.91 | 4.92 | 4.19 | 3.37 | 3.37 |
| | Annealing | 18.96 | 12.74 | 5.93 | 9.23 | 2.59 |
| | Ecoli | 14.19 | 16.47 | 18.65 | 14.78 | 13.59 |
| | Page-Blocks | 7.43 | 2.91 | 3.87 | 3.88 | 4.07 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 9.3 | 2.79 | 5.12 | 4.19 |
| | Bridges | 35.54 | 37.74 | 42.45 | 40.88 | 33.02 |
| | Optdigits | 26.24 | 20.43 | 17.71 | 16.43 | 20.61 |

Tabla A.20: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda GeneticSearch (e2b3).

| ConsistencySubsetEval - GreedyStepwise | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 43.18 | 37.88 | 37.12 | 38.64 | 34.09 |
| | Postoperative | x | x | x | x | x |
| | Iris | 5.33 | 7.33 | 7.33 | 6.67 | 6.67 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 16.77 | 21.94 | 20.65 | 18.71 | 18.06 |
| | Credit | 21.16 | 15.07 | 19.71 | 15.65 | 16.67 |
| | Ionosphere | 15.67 | 9.69 | 11.97 | 13.39 | 11.4 |
| | Votes | 7.36 | 3.45 | 6.21 | 5.98 | 5.29 |
| | Wine | 7.3 | 5.62 | 7.87 | 7.87 | 6.74 |
| | Glass | 49.07 | 31.31 | 23.36 | 27.57 | 32.24 |
| | Segment | 18.1 | 10.48 | 10.95 | 9.52 | 7.14 |
| | Dermatology | 3.01 | 5.46 | 6.28 | 4.92 | 5.46 |
| | Annealing | 37.34 | 13.78 | 2.51 | 4.26 | 1.63 |
| | Ecoli | 14.29 | 16.37 | 19.64 | 14.58 | 13.1 |
| | Page-Blocks | 9.03 | 2.98 | 4.07 | 4 | 3.73 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 1.31 |
| | Thyroid | 3.26 | 9.77 | 2.79 | 6.51 | 6.05 |
| | Bridges | 33.02 | 33.02 | 46.23 | 42.45 | 34.91 |
| | Optdigits | 19.02 | 18.1 | 13.6 | 12.4 | 16.32 |
| Método sin Balanceo | Echocardiogram | 40.91 | 36.36 | 39.39 | 37.12 | 34.09 |
| | Postoperative | x | x | x | x | x |
| | Iris | 5.33 | 7.33 | 7.33 | 6.67 | 6.67 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.13 | 24.35 |
| | Hepatitis | 16.77 | 21.94 | 20.65 | 18.71 | 18.06 |
| | Credit | 21.16 | 15.07 | 19.71 | 15.65 | 16.67 |
| | Ionosphere | 15.67 | 9.69 | 11.97 | 13.39 | 11.4 |
| | Votes | 7.36 | 3.45 | 6.21 | 5.98 | 5.29 |
| | Wine | 5.62 | 4.49 | 3.37 | 3.37 | 5.06 |
| | Glass | 47.66 | 32.24 | 26.64 | 27.1 | 31.31 |
| | Segment | 17.62 | 14.29 | 11.9 | 9.05 | 14.29 |
| | Dermatology | 13.66 | 13.39 | 14.48 | 14.75 | 13.11 |
| | Annealing | 13.41 | 14.79 | 7.14 | 9.02 | 1.25 |
| | Ecoli | 13.99 | 16.07 | 19.05 | 14.88 | 16.37 |
| | Page-Blocks | 6.96 | 3.07 | 4.28 | 4.2 | 3.53 |
| | Nursery | 9.74 | 2.91 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 3.72 | 9.3 | 3.26 | 5.58 | 5.12 |
| | Bridges | 30.19 | 29.25 | 38.68 | 36.79 | 32.08 |
| | Optdigits | 8.24 | 7.72 | 1.62 | 1.78 | 2.43 |
| Método con Repetición | Echocardiogram | 40.91 | 36.36 | 39.39 | 37.12 | 34.09 |
| | Postoperative | x | x | x | x | x |
| | Iris | 5.33 | 7.33 | 7.33 | 6.67 | 6.67 |
| | Diabetes | 24.22 | 28.52 | 28.52 | 26.56 | 24.61 |
| | Hepatitis | 14.19 | 19.35 | 19.35 | 19.35 | 13.55 |
| | Credit | 23.77 | 14.78 | 20.14 | 15.8 | 17.39 |
| | Ionosphere | 12.82 | 13.39 | 11.11 | 10.54 | 9.12 |
| | Votes | 8.97 | 3.22 | 6.9 | 6.44 | 4.37 |
| | Wine | 5.62 | 4.49 | 3.37 | 3.37 | 5.06 |
| | Glass | 47.66 | 32.24 | 26.64 | 27.1 | 31.31 |
| | Segment | 17.62 | 14.29 | 11.9 | 9.05 | 14.29 |
| | Dermatology | 13.66 | 6.01 | 14.48 | 14.75 | 13.11 |
| | Annealing | 13.41 | 10.28 | 7.14 | 9.02 | 1.25 |
| | Ecoli | 14.58 | 15.77 | 19.64 | 15.48 | 13.69 |
| | Page-Blocks | 6.38 | 2.94 | 4.33 | 4.33 | 4.17 |
| | Nursery | 9.77 | 2.9 | 2.1 | 2.1 | 0.6 |
| | Thyroid | 2.79 | 9.77 | 2.79 | 5.12 | 3.72 |
| | Bridges | 36.79 | 34.91 | 35.85 | 35.85 | 31.13 |
| | Optdigits | 11.38 | 6.72 | 6.04 | 5.73 | 7.69 |
| Método con SMOTE | Echocardiogram | 42.93 | 40.66 | 37.12 | 34.34 | 31.06 |
| | Postoperative | x | x | x | x | x |
| | Iris | 4.67 | 4.22 | 6 | 4.89 | 4.45 |
| | Diabetes | 24.61 | 27.99 | 29.43 | 28.12 | 26.95 |
| | Hepatitis | 18.71 | 23.87 | 25.59 | 21.51 | 19.14 |
| | Credit | 22.85 | 14.01 | 19.76 | 14.78 | 17.15 |
| | Ionosphere | 12.25 | 10.83 | 10.64 | 11.4 | 9.12 |
| | Votes | 7.36 | 3.45 | 6.21 | 5.98 | 5.75 |
| | Wine | 3.74 | 4.3 | 3.37 | 2.44 | 4.87 |
| | Glass | 59.19 | 34.89 | 38.63 | 35.2 | 37.07 |
| | Segment | 22.7 | 12.06 | 12.38 | 12.38 | 16.98 |
| | Dermatology | 6.28 | 8.74 | 9.65 | 8.65 | 7.37 |
| | Annealing | 27.28 | 13.49 | 6.89 | 8.6 | 3.01 |
| | Ecoli | 16.27 | 17.26 | 19.25 | 15.68 | 15.38 |
| | Page-Blocks | 6.86 | 2.92 | 3.86 | 4 | 3.78 |
| | Nursery | 9.85 | 4.08 | 2.84 | 2.84 | 2.08 |
| | Thyroid | 2.48 | 8.99 | 2.79 | 5.12 | 4.81 |
| | Bridges | 32.39 | 35.85 | 38.37 | 41.19 | 30.19 |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| | Optdigits | 33.27 | 29.48 | 31.97 | 31.28 | 24.62 |

Tabla A.22: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda GreedyStepwise (e2b4).

| **ConsistencySubsetEval - RandomSearch** | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| WEKA con Selección | Echocardiogram | 43.94 | 40.15 | 39.39 | 37.12 | 37.88 |
| | Postoperative | 30 | 30 | 40 | 28.89 | 43.33 |
| | Iris | 4.67 | 5.33 | 5.33 | 5.33 | 4.67 |
| | Diabetes | 24.48 | 27.34 | 30.99 | 26.04 | 26.04 |
| | Hepatitis | 12.26 | 20 | 17.42 | 17.42 | 18.71 |
| | Credit | 23.04 | 14.78 | 20 | 15.51 | 18.12 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 8.51 | 3.45 | 6.67 | 5.98 | 6.21 |
| | Wine | 2.81 | 7.3 | 4.49 | 5.06 | 3.93 |
| | Glass | 52.34 | 30.37 | 24.3 | 27.1 | 30.84 |
| | Segment | 21.9 | 10.48 | 9.52 | 10 | 9.52 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 15.18 | 17.56 | 21.13 | 15.77 | 16.67 |
| | Page-Blocks | 6.23 | 3 | 4.44 | 4.26 | 3.69 |
| | Nursery | 10.17 | 4.97 | 6.75 | 5.44 | 4.5 |
| | Thyroid | 3.26 | 9.77 | 2.79 | 6.51 | 6.05 |
| | Bridges | 33.96 | 30.19 | 44.34 | 49.06 | 38.68 |
| | Optdigits | x | x | x | x | x |
| Método sin Balanceo | Echocardiogram | 44.7 | 36.36 | 38.64 | 37.88 | 33.33 |
| | Postoperative | 30 | 30 | 40 | 28.89 | 44.44 |
| | Iris | 6.67 | 7.33 | 5.33 | 5.33 | 5.33 |
| | Diabetes | 24.48 | 27.34 | 30.99 | 26.04 | 26.04 |
| | Hepatitis | 12.26 | 20 | 17.42 | 17.42 | 18.71 |
| | Credit | 23.04 | 14.78 | 20 | 15.51 | 18.12 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 8.51 | 3.45 | 6.67 | 5.98 | 6.21 |
| | Wine | 3.93 | 8.43 | 7.87 | 4.49 | 8.99 |
| | Glass | 48.6 | 31.78 | 28.5 | 28.5 | 32.24 |
| | Segment | 19.05 | 12.38 | 10.95 | 12.86 | 9.05 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 15.48 | 15.48 | 20.54 | 16.37 | 15.77 |
| | Page-Blocks | 6.23 | 2.98 | 4.4 | 4.22 | 3.71 |
| | Nursery | 10.17 | 4.97 | 6.75 | 5.44 | 4.73 |
| | Thyroid | 3.26 | 10.7 | 2.33 | 6.51 | 7.91 |
| | Bridges | 37.74 | 33.02 | 43.4 | 47.17 | 41.51 |
| | Optdigits | x | x | x | x | x |
| Método con Repetición | Echocardiogram | 46.21 | 37.12 | 40.15 | 40.15 | 34.85 |
| | Postoperative | 30 | 28.89 | 37.78 | 28.89 | 44.44 |
| | Iris | 6 | 4 | 6 | 5.33 | 4 |
| | Diabetes | 24.48 | 27.34 | 30.99 | 26.04 | 26.3 |
| | Hepatitis | 14.84 | 20.65 | 20 | 16.67 | 18.06 |
| | Credit | 23.23 | 14.78 | 20.14 | 15.36 | 17.52 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 8.51 | 3.45 | 6.67 | 5.98 | 6.21 |
| | Wine | 5.06 | 7.87 | 8.43 | 5.62 | 8.43 |
| | Glass | 46.73 | 31.78 | 30.37 | 28.5 | 32.24 |
| | Segment | 16.19 | 11.43 | 10.95 | 13.33 | 9.52 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 14.88 | 17.86 | 21.02 | 16.07 | 15.77 |
| | Page-Blocks | 6.92 | 2.96 | 3.84 | 3.95 | 4.13 |
| | Nursery | 10.17 | 4.97 | 6.75 | 5.44 | 4.5 |
| | Thyroid | 2.33 | 10.7 | 2.33 | 6.51 | 9.77 |
| | Bridges | 40.57 | 32.08 | 43.4 | 48.11 | 38.68 |
| | Optdigits | x | x | x | x | x |

| Continuación... | | | | | | |
|---|---|---|---|---|---|---|
| **Método** | **Base de Datos** | **Bayes** | **J48** | **KNN1** | **KNN3** | **MP** |
| Método con SMOTE | Echocardiogram | 45.71 | 38.63 | 40.4 | 36.36 | 35.35 |
| | Postoperative | 31.11 | 30 | 40 | 30.74 | 45.56 |
| | Iris | 5.11 | 5.11 | 5.33 | 4.89 | 4.89 |
| | Diabetes | 24.48 | 27.34 | 30.99 | 26.04 | 26.3 |
| | Hepatitis | 17.63 | 20.21 | 18.06 | 15.7 | 18.93 |
| | Credit | 22.95 | 14.06 | 19.33 | 15.07 | 17.44 |
| | Ionosphere | x | x | x | x | x |
| | Votes | 8.51 | 3.45 | 6.67 | 5.98 | 6.21 |
| | Wine | 2.25 | 8.06 | 3.74 | 4.12 | 2.62 |
| | Glass | 51.4 | 31.78 | 31 | 29.91 | 30.68 |
| | Segment | 19.84 | 11.43 | 10.32 | 15.87 | 9.68 |
| | Dermatology | x | x | x | x | x |
| | Annealing | x | x | x | x | x |
| | Ecoli | 15.08 | 17.26 | 21.23 | 15.77 | 14.68 |
| | Page-Blocks | 6.17 | 2.96 | 4.43 | 4.2 | 3.48 |
| | Nursery | 10.17 | 4.97 | 6.75 | 5.44 | 4.73 |
| | Thyroid | 2.48 | 10.7 | 2.33 | 6.51 | 8.84 |
| | Bridges | 33.96 | 29.25 | 44.34 | 49.06 | 40.88 |
| | Optdigits | x | x | x | x | x |

Tabla A.24: Porcentaje de error de las 19 bases de datos con los métodos probados por clasificador para el evaluador ConsistencySubsetEval y Método de búsqueda RandomSearch (e2b5).