VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN ENSEÑANZA DE LA MATEMÁTICA

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN ENSEÑANZA DE LA MATEMÁTICA

TEMA: ESTUDIO COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA DEVELAR PATRONES DE DESEMPEÑO ACADÉMICO EN ENSEÑANZA MEDIA

AUTORA: CHAMORRO SANGOQUIZA, DIANA CRISTINA

DIRECTOR: PROF. PELUFFO ORDÓÑEZ, DIEGO HERNÁN

SANGOLQUÍ

2019

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

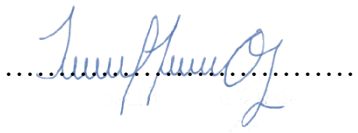
CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación: "***ESTUDIO COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS PARA DEVELAR PATRONES DE DESEMPEÑO ACADÉMICO EN ENSEÑANZA MEDIA***" fue realizado por la señorita Chamorro Sangoquiza, Diana Cristina el mismo que ha sido revisado en totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 20 de junio de 2019

Firma:



Prof. Peluffo Ordóñez, Diego Hernán

C.C: 1757278443

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

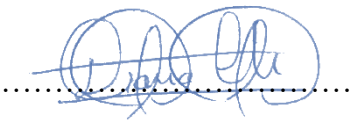
AUTORÍA DE RESPONSABILIDAD

Yo, **Chamorro Sangoquiza, Diana Cristina** con cédula de ciudadanía n° 1003220942, declaro que el contenido, ideas y criterios del trabajo de titulación: **Estudio comparativo de técnicas de minería de datos para develar patrones de desempeño académico en enseñanza media** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 20 de junio de 2019

Firma:



……

Chamorro Sangoquiza, Diana Cristina

C.C.: 1003220942

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

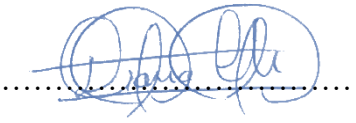
CENTRO DE POSGRADOS

AUTORIZACIÓN

Yo, **Chamorro Sangoquiza, Diana Cristina** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Estudio comparativo de técnicas de minería de datos para develar patrones de desempeño académico en enseñanza media** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 20 de junio de 2019

Firma:



……

Chamorro Sangoquiza, Diana Cristina

C.C.: 1003220942

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

DEDICATORIA

A mi padre, por permitirme conocer la importancia del mundo de las letras y números. A mi madre por toda la dedicación y apoyo incondicional en todas las etapas de mi vida.

A mi compañero de vida, Andrés Vargas por todo su apoyo y compresión, en el caminar juntos en la vida.

A mis hermanas, Gina, Mayra, Viviana y mi hermano Daniel, por las palabras de aliento, y por confiar en mí.

A todos mis hermosos y tiernos sobrinos, (Matheo, Sebastián, Alejandra, Nicolás, Daniela, Erik, Isaac), que con su inocencia y ternura ayudan a que este camino sea más llevadero.

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AGRADECIMIENTOS

Agradezco a Dios que nos brinda la oportunidad de aprender de nuestros errores y aciertos día a día.

Doy mis más sinceros agradecimientos a Andrés Vargas, por su invaluable apoyo y compañía.

Agradezco al profesor Diego Peluffo por su asesoría y apoyo en el desarrollo de este trabajo.

Agradezco al colegio Nacional Ibarra por permitirme realizar este estudio, facilitándome la base de datos.

Y a todas las personas que, de alguna manera, contribuyeron al desarrollo de este trabajo, gracias.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ECUACIONES

RESUMEN

La minería de datos es ampliamente utilizada en diversos campos: educación, computación móvil, minería web, análisis financiero, análisis de delitos, ingeniería, gestión, medicina, etc. Naturalmente, en la dinámica de la sociedad actual, se ha dado, cada vez, mayor importancia a la educación y la investigación. Adicionalmente, las instituciones de educación generan y almacenan datos sobre los estudiantes, que con un procesado subsecuente adecuado pueden resultar útiles para tomar decisiones estratégicas en pro de todos los procesos académicos internos de las mismas instituciones. En este sentido, las técnicas computarizadas, particularmente las técnicas de minería de datos, han tomado importancia dado que permiten comprender mejor a los estudiantes y los entornos en los que aprenden, orientando a las instituciones en cómo proceder para brindar mejoras continuas en la calidad de educación. El presente trabajo presenta un estudio de técnicas de minería de datos, aplicada a datos de estudiantes de educación media, para proponer una base conceptual y algorítmica para diseñar una herramienta que tenga un margen mayor de análisis de la información, a través de los datos generados por la aplicación, apoyándose de igual manera en algoritmos de inteligencia artificial, los cuales son: método ponderado de vecinos cercanos, árboles de decisión, máquinas de vectores de soporte, y enfoques de multi-clasificadores.

PALABRAS CLAVE:

- **MINERÍA DE DATOS**
- **PATRONES DE DESEMPEÑO ACADEMICO**
- **SELECCIÓN DE CARACTERÍSTICAS**
- **CLASIFICADORES**
- **MULTI-CLASIFICADORES**
- **MATLAB**

ABSTRACT

Data mining is widely used in several fields, such as: education, mobile computing, web mining, financial analysis, crime analysis, engineering, management, and medicine, among others. Naturally, in the dynamics of today's society, education and research have become increasingly important. Furthermore, educational institutions are generating and storing students' data, which – undergoing an appropriate, subsequent processing- can be useful to make strategic decisions in favor of all the internal academic processes thereof. In this sense, computerized techniques, particularly data mining techniques, have become a key tool since they allow students for a better understanding, as well as identifying the environments in which they learn, by guiding the institutions on how to proceed to provide continuous improvements in reaching the high quality in education. This master's thesis presents a study of data mining techniques -applied to data of students from secondary education- to propose a conceptual and algorithmic basis to design a tool having a greater margin of analysis of the information. Specifically, the explored methods are: weighted nearest neighbors, decision trees, support vector machines, and multi-classifier approaches.

KEY WORDS:

- **DATA MINING**
- **ACADEMIC PERFORMANCE PATTERNS**
- **FEATURE SELECTION**
- **CLASSIFIERS**
- **MULTIPLE CLASIFIERS**
- **MATLAB**

CAPÍTULO I

1. INTRODUCCIÓN

1.1. Planteamiento del problema

En la actualidad, se vive en un mundo digitalizado, razón por la cual se ha dado, naturalmente, un explosivo crecimiento en el volumen de los datos. Con esto surge la necesidad imperiosa de crear técnicas avanzadas y metodologías capaces de integrar estos datos que puedan, de manera inteligente y de forma automatizada, procesarlos y transformarlos en información útil; que generen conocimiento y brinden soporte en su interpretación de una forma robusta y eficiente para realizar la toma de decisiones con respecto a diversos ámbitos. Estas técnicas y herramientas son el objeto del campo de Descubrimiento de Conocimiento en Bases de Datos – DCBD (Bolaños Ramírez, 2017).

La expansión de datos no siempre supone un aumento de conocimiento, como naturalmente podría esperarse, puesto que procesarlos con los métodos clásicos resulta ser en muchos casos imposible, sumamente tedioso y con resultados superficiales e insatisfactorios (U. M. Fayyad, 1996).

En 2003, (Vesonder & Wright, 2003) sostienen que “es bastante común que las bases de datos tengan del 60% al 90% de problemas de calidad en los datos”, razón por la cual se plantea un estudio comparativo de los diferentes métodos de selección y clasificación de características, para luego aplicar la técnica de minería de datos.

Una de las áreas más sobresalientes en los últimos años es el Aprendizaje Automático (Machine Learning, ML), donde se dispone de diferentes clases de algoritmos de clasificación que han servido de soporte al momento de toma de decisiones; desarrollando algoritmos robustos, fiables y no necesariamente con menor coste computacional. Si embargo, no es una tarea fácil de encontrar

una técnica o algoritmo únicos capaz de solventar todos los problemas en el ámbito educacional; llevando a tener problemas en la correcta clasificación o al incremento en el coste computacional. De acuerdo con lo anterior, se aprecia que la selección de un método de minería de datos no es una tarea trivial y que, por tanto, es necesario realizar una comparación entre varios algoritmos. En este caso, se busca algoritmos de selección y clasificación de características que mejor se adapten al tipo de datos educativos de interés (información de estudiantes de enseñanza media), de manera que se logre un balance entre precisión e interpretación en contexto, además de ofrecer un mejor soporte en el descubrimiento de patrones académicos.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar una metodología de comparación de técnicas recientes de minería de datos, usando criterios de precisión e interpretación en contexto, para descubrir patrones de desempeño estudiantil en educación media.

1.2.2. Objetivos específicos

- Seleccionar diferentes técnicas de selección de atributos con el fin de obtener un adecuado subconjunto de características de datos sociodemográficos que mejor representen a los individuos de estudio.
- Diseñar un estudio comparativo de varias técnicas de clasificación de datos recomendadas por el estado del arte con el fin de identificar patrones académicos.
- Proponer un esquema para la adecuada interpretación de los resultados obtenidos en un contexto académico y sociodemográfico.

1.3. Contribuciones de esta tesis

En lo referente a tratar gran cantidad de datos, en el ámbito de la educación, es claro que el desarrollo y las tecnologías ha ganado espacio en este ámbito, aplicando al estudio del

comportamiento de los estudiantes, causas de deserción, etc. Esta investigación aporta al estudio de descubrimiento de patrones de desempeño académico en la educación media, mediante el desarrollo de una aplicación, usando técnicas de aprendizaje automático. El aporte puede referirse en tres partes:

- Seleccionar las características de las bases de datos, con el fin de obtener un adecuado subconjunto de datos, que mejor representen la totalidad de los individuos de estudio.
- Diseñar un estudio comparativo de varias técnicas de clasificación de datos recomendadas por el estado del arte con el fin de identificar patrones académicos.
- Proponer un esquema para la adecuada interpretación de los resultados obtenidos en un contexto académico y sociodemográfico.

1.4. Organización del documento

Este estudio está dividido en cinco capítulos, estructurados así:

El Capítulo 1, esta detallado la problemática que conlleva a plantearse el estudio, así también los objetivos que se plantea en la investigación y sus contribuciones.

El Capítulo 2, corresponde al marco conceptual, que corresponde a todas las definiciones y conceptos necesarios para llevar a cabo esta investigación.

El Capítulo 3, se realiza la metodología de comparación entre los métodos de selección y clasificación de datos, así también la propuesta e interpretación en contexto

El Capítulo 4, se expone la base de datos a utilizar, se describe los experimentos y sus resultados, también se expone la interfaz desarrollada.

El Capítulo 5, concluye el trabajo, destacando las conclusiones de la investigación, se expone las recomendaciones y las propuestas de trabajo futuro.

CAPÍTULO II

2. MINERÍA DE DATOS

2.1. Introducción

La cantidad de datos a nivel mundial va aumentando, conforme aumenta la digitalización de la información, llegando a ser cada vez más difícil para el ser humano obtener información relevante; los datos se vuelven menos entendibles, debido a que en contadas ocasiones se hace explícita o se aprovecha esta información. Por tanto, surge la necesidad de usar métodos automatizados de procesamiento de datos que permitan dilucidar los patrones que lo subyacen.

La minería de datos es ampliamente usada para descubrir patrones en bases de datos extensas, las cuales ayudan a resolver problemas de la vida real (Khube, Islam, & Ashad, 2017); su análisis se lo realiza sobre datos almacenados electrónicamente y mediante un análisis automatizado. La minería de datos ha sido aplicada en numerosas áreas, incluyendo el sector educacional (Mhetre & Nagar, 2017), denominada Minería de Datos Educativos (EDM), centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales; el uso de esos métodos ayudan a una mejor comprensión de los estudiantes, así también, el entorno en el que aprenden.

La disciplina EDM, emerge como un paradigma orientado a diseñar modelos, tareas, métodos y algoritmos con el objetivo de explorar los datos del entorno educativo (Peña-Ayala, 2014). El objetivo del EDM, es analizar datos para extraer conocimiento (Mhetre & Nagar, 2017) mediante la búsqueda de patrones; además, realizar predicciones que caractericen el comportamiento y desempeño de los estudiantes o para predecir su rendimiento, deserción, entre muchos otros (Luan, 2002).

¿Cómo se expresan los patrones? Los patrones útiles permiten hacer predicciones no triviales sobre nuevos datos. Hay dos extremos para la expresión de un patrón: como una caja negra cuyo contenido es efectivamente incomprensibles y como una caja transparente cuya construcción revela la estructura del patrón. (Data Mining, 2011).

2.2. Esquema general del descubrimiento de conocimiento en bases de datos (KDD)

Los autores (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996) definen el KDD como, "*El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos*".

En la ***Figura 1*** se puede observar las fases del proceso KDD (Knowledge Discovery in Databases), cuya intervención del usuario en la toma de decisiones hace que el proceso sea interactivo e iterativo.

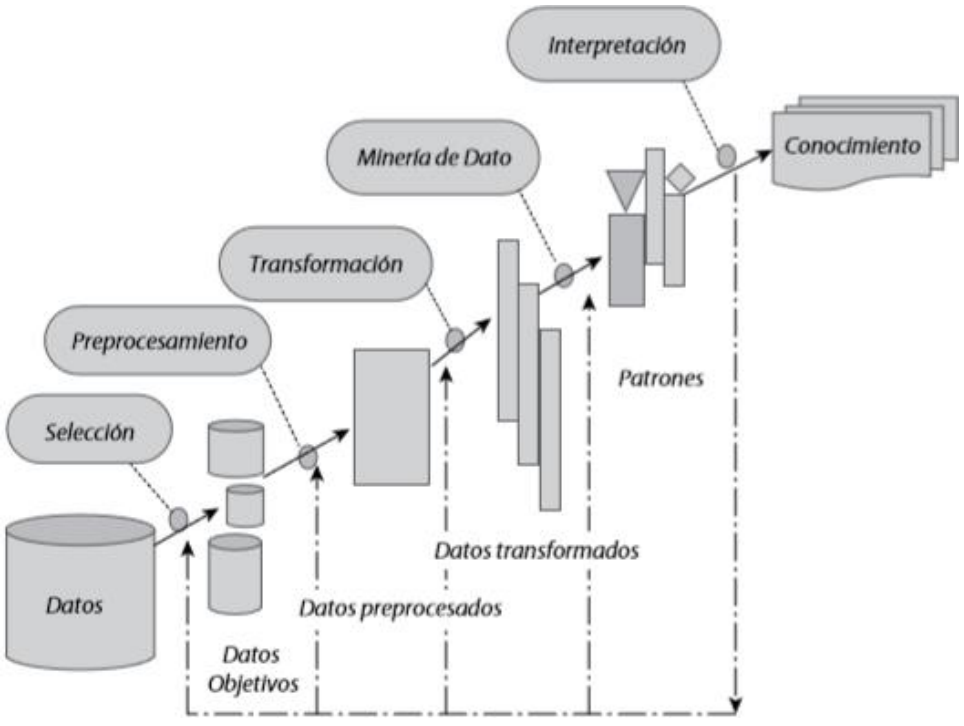


Figura 1. Etapas del proceso KDD

Fuente: (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016)

El proceso KDD está en constante evolución, desde la intersección de la investigación en campos como bases de datos, aprendizaje automático, reconocimiento de patrones, estadísticas, inteligencia artificial y razonamiento con incertidumbre, adquisición de conocimiento para sistemas expertos, visualización de datos, descubrimiento de máquinas (Langley & Simon, 1995).

2.2.1. Etapas del proceso KDD

En esta sección se expondrán cada una de las etapas de proceso KDD, tomando en cuenta que no es un proceso lineal.

Selección. En esta etapa, se seleccionan los atributos más relevantes, que ofrezcan información no redundante, permitiendo obtener un conjunto o una muestra de los datos representativos para el modelo.

Preprocesamiento/limpieza.- Los datos incompletos (sin valores de atributo o ciertos atributos de interés), ruidosos (que contienen errores o valores atípicos) e inconsistentes, son propiedades comunes de las grandes bases de datos del mundo real (Bhardwaj, 2015).

El preprocesamiento de datos consiste en la limpieza, transformación, reducción de dimensionalidad, selección de subconjuntos de características y manejo de datos desconocidos, nulos y duplicados mediante uso de métricas estadísticas para corregirlos.

En el proceso de limpieza de datos se completa valores faltantes, se suaviza los datos ruidosos, se resuelve inconsistencias y se elimina valores atípicos.

A continuación se detalla varios métodos de limpieza de datos recomendados para obtener un mejor resultado en el proceso de minería (Bhardwaj, 2015).

- Eliminar las instancias que tienen etiquetas de clase perdidas.
- Rellenar manualmente los datos que tienen información perdida, no es una estrategia factible para bases de datos con grandes cantidades de instancias.

- Usar una constante global para todos los datos perdidos, aunque es un método simple de limpieza, no es recomendado debido a que el programa de minería puede tomar estas etiquetas como un valor común de agrupación.
- Usar la media de los todos atributos para rellenar los datos perdidos.
- Usar la media de los atributos pertenecientes a la misma clase para rellenar los datos perdidos.
- Usar los valores más probables para rellenar los datos perdidos, estos pueden ser determinados usando técnicas Bayesianas o inducción mediante árboles de decisión.

Transformación/reducción en esta etapa de la minería de datos se aplican técnicas que permiten cambiar la representación de los datos sin comprometer la integridad de la calidad de conocimiento original. Dentro de estos procesos se aplican técnicas de reducción de dimensiones, compresión de datos, selección de atributos, entre otros.

Minería de datos La Minería de Datos es la extracción de información implícita, no conocida previamente y potencialmente útil de los datos recolectados (*Data Mining*, 2011) mediante la aplicación de diferentes tipos de algoritmos. Analizando varios modelos y eligiendo el que tiene el mejor desempeño de pronóstico - evaluación competitiva de modelos.

Interpretación/evaluación En esta etapa se realiza el análisis de los patrones encontrados en los datos, permitiendo la posibilidad de retomar pasos anteriores para mejorar los resultados y aplicar el modelo a nuevos datos para producir pronósticos y estimaciones correctas para los problemas investigados.

2.3. Pre-proceso de datos

Los datos sin procesar son altamente susceptibles al ruido, conteniendo valores perdidos e inconsistencias. El preprocesamiento de datos es una etapa esencial del proceso de descubrimiento

de información o KDD (Han & Kamber, 2001) (Zaki & Meira, Jr, 2014). Esta etapa se encarga de la limpieza de datos (data cleaning), en la cual se analiza la calidad de datos y se aplica operaciones básicas como integración, transformación y reducción para la siguiente fase de minería de datos (Bhardwaj, 2015).

2.4. Representación de datos

2.4.1. Reducción de dimensión

Las bases de datos contienen gran volumen de información (alta dimensionalidad), siendo un problema en el descubrimiento de patrones en la minería de datos (Hernández, Ramírez, & Ferri, 2004). La mayoría de los datos, pueden contener información irrelevante o redundante con respecto al problema a resolver (Timarán Pereira et al., 2016), por lo cual, se aplican técnicas de reducción de dimensiones que permiten obtener un subconjunto de datos representativos del conjunto original, pero de una dimensión más baja, sin perder su especificidad y representatividad. Estos procesos ayudan a evitar varios problemas en general, como:

- La existencia de atributos irrelevantes.
- La existencia de atributos redundantes.
- La alta dimensionalidad de los datos.

Los conjuntos basados en subconjuntos de características facilitan potencialmente la creación de un clasificador para conjuntos de datos de alta dimensionalidad, buscando características útiles para representar dependiendo del objetivo de la investigación. Una reducción se define como el subconjunto de características más pequeño que tiene igual o mayor poder predictivo que el conjunto de características completo (Rokach & Maimon, 2015). Además, estos métodos pueden usarse para mejorar el rendimiento de la clasificación debido a la reducida correlación entre los clasificadores (Han & Kamber, 2001); tener una reducción del tamaño del conjunto de datos

implica una inducción más rápida de los clasificadores. Entre las técnicas más utilizadas de reducción se pueden mencionar: agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras (Han & Kamber, 2001).

Unos de los objetivos principales del método de la selección de características es ahorrar espacio de almacenamiento o reducir el costo computacional, para así poder enfrentar problemas de alta dimensionalidad o de recursos limitados (Xu, King, Lyu, & Jin, 2010).

2.4.2. Selección de características

Es importante realizar selección de características previo al aprendizaje, debido al efecto negativo de atributos irrelevantes en la mayoría de esquemas de aprendizaje automático (Liu & Motoda, 1998) (Liu & Motoda, 2012). Por lo cual, un objetivo de la minería es la selección de características en la base de datos, este proceso permite eliminar datos irrelevantes, redundantes y con ruido; permite mejorar la calidad de los datos, y por ende reducir el tiempo de procesamiento del algoritmo de aprendizaje automático, obteniendo una mejor interpretación del modelo (Phyu, 2009). En la Figura 2 se puede observar el proceso básico de selección de características, el cual se repite hasta que se cumpla el criterio planteado.

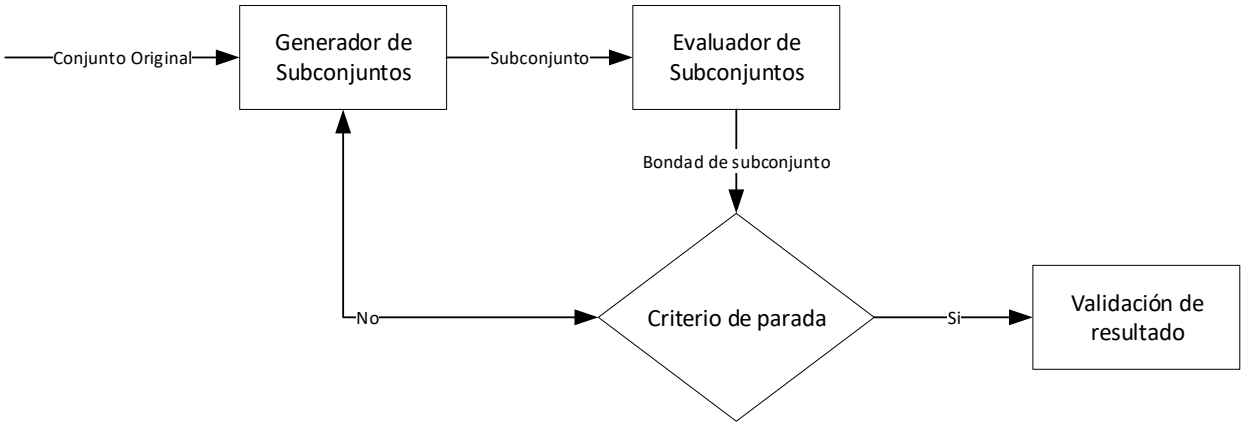


Figura 2. Proceso de selección de características
Fuente: (Herrera, 2006)

Existen varios algoritmos de aprendizaje automático diseñados para encontrar los atributos más apropiados para la toma de decisiones (Herrera, 2006), permitiendo reducir el tamaño de los datos mediante la eliminación de características redundantes; La ***Figura 3*** muestra la clasificación de los métodos de selección de características.

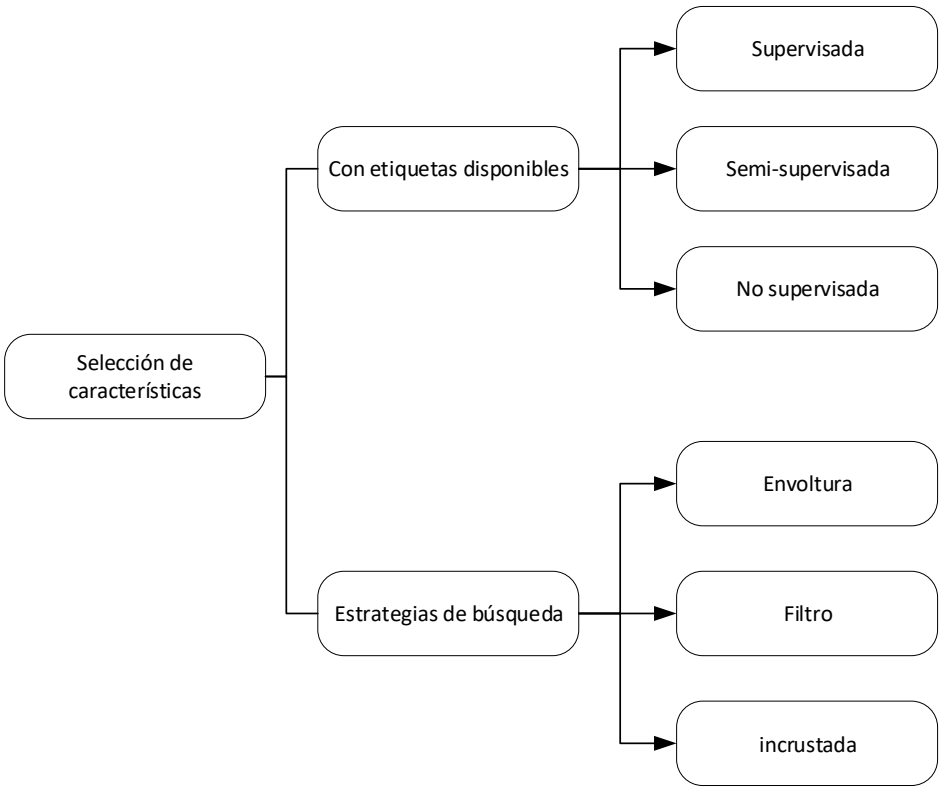


Figura 3. Categorías de los métodos de selección de características
Fuente: (Bolaños Ramírez, 2017)

A continuación, se presenta diferentes técnicas empleadas para efectuar la selección de características. Para ello, previamente se introduce diferentes aproximaciones a la clasificación de las mismas. Existen varios algoritmos de búsqueda de características relevantes, entre los cuales se menciona los más aplicados en investigaciones anteriores.

2.4.2.1.Best-First

Permite buscar un subconjunto de atributos mediante la expansión del nodo más prometedor de acuerdo a ciertas reglas especificadas. Existen varias configuraciones iniciales en el algoritmo de búsqueda, las cuales pueden comenzar con un conjunto vacío de atributos y buscar hacia adelante, o comenzar con un conjunto completo de atributos y buscar hacia atrás, o comenzar en cualquier punto y buscar en ambas direcciones («BestFirst», 2019).

2.4.2.2.Ranker

El método Ranker, realiza evaluaciones individuales para la clasificación de atributos (Chetty, Vaisla, & Sudarsan, 2015), proporcionando una lista ordenada de características de acuerdo a su evaluación. Este método se evalúa en conjunto con los evaluadores de atributos (ReliefF, Gain Ratio, Entropy, etc.) con el fin de generar un parámetro de ranking ordenados por su puntaje al ser evaluado. Algunas de las ventajas de este método de selección están orientadas a las selecciones de un número de características, también, establece que características son importantes y cuáles no, teniendo en cuenta el orden de relevancia entre ellas (Pulgarín, 2012). Se puede configurar un intervalo, en el cual los atributos son descartados o especificar cuantos atributos mantener, incluso mantener atributos independientemente de su ranking (*Data Mining*, 2011).

2.4.2.3.PCA

El análisis de componentes principales (PCA) se usa a lo largo de la ciencia y la ingeniería para ayudar a resumir, representar y mostrar datos medidos en muchas variables en términos de un número menor de variables derivadas (Johnstone & Paul, 2018).

CorrelationAtributeEval

Evalúa el valor de un atributo midiendo la correlación (Pearson) entre este y la clase. Los atributos nominales se consideran valor por valor al tratar cada valor como un indicador. Se llega a una correlación global para un atributo nominal a través de un promedio ponderado («CorrelationAttributeEval», 2019).

El evaluador de atributos es la técnica por la cual cada atributo en su conjunto de datos (también llamado columna o entidad) se evalúa en el contexto de la variable de salida (por ejemplo, la clase). El método de búsqueda es la técnica mediante la cual se puede probar o navegar por diferentes combinaciones de atributos en el conjunto de datos para llegar a una breve lista de características elegidas. Algunas técnicas de evaluación de atributos requieren el uso de métodos de búsqueda específicos. Por ejemplo, la técnica CorrelationAttributeEval utilizada en la siguiente sección solo se puede usar con un método de búsqueda de clasificación, que evalúa cada atributo y enumera los resultados en un orden de clasificación (Brownlee, 2016).

CFS: Correlation-based Feature Selection:

Como la mayoría de los programas de selección de características, CFS utiliza un algoritmo de búsqueda junto con una función para evaluar el mérito de los subconjuntos de características. La heurística mediante la cual la CSF mide la bondad de los subconjuntos de características que tiene en cuenta la utilidad de las características individuales para predecir la etiqueta de clase junto con el nivel de intercorrelación entre ellas. La hipótesis en la que se basa la heurística se puede afirmar: Los buenos subconjuntos de características contienen características altamente correlacionadas con la clase, pero no correlacionadas entre sí.

2.4.2.4.Selección de instancias

La selección de instancias (SII) consiste en escoger las muestras más representativas de un conjunto determinado (Kibler & Aha, 1987) (Brighton & Mellish, 2002) (Liu & Motoda, 2002).

El disminuir la cantidad de atributos no tiene que ser sinónimo de perdida de información, debido a que la información eliminada podría ser datos repetidos o con ruido; este proceso permite reducir tanto la complejidad computacional, como los recursos de almacenamiento. Cada ejemplo presenta un cierto grado de libertad suficiente, tal que, si se reduce su número de atributos puede en ciertos casos superar situaciones de sobre aprendizaje. La (SII) se puede llevar a cabo siguiendo diferentes vías, como se puede ver en la Figura 4.

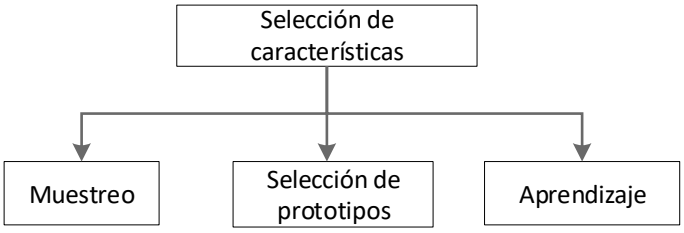


Figura 4. Estrategia de selección de instancias
Fuente: (Herrera, 2006)

2.4.2.5.Algoritmo de Kennard – Stone (KS)

La técnica de algoritmo KS se aplicó originalmente para generar un conjunto de entrenamiento cuando no se puede implementar un diseño experimental estándar. El algoritmo inicia considerando candidatos a todos los objetos para el conjunto de entrenamiento, los candidatos seleccionados en el proceso son elegidos secuencialmente. El algoritmo KS se puede resumir de la siguiente manera: primero, el algoritmo KS toma el par de muestras con la mayor distancia euclidiana de vectores x (predictores) y luego selecciona una muestra para maximizar la distancia euclidiana entre

vectores x de muestras ya seleccionadas y las muestras restantes. Este proceso se repite hasta que se alcanza el número requerido de muestras (Saptoro, Tadé, & Vuthaluru, 2012).

2.5. Métodos de clasificación de datos

En esta sección se menciona los clasificadores utilizados con mayor frecuencia en el proceso de minería de datos educativos.

2.5.1. Clasificador de los *k* vecinos más cercanos (K-NN)

El algoritmo de los *k* vecinos más cercanos (k-NN Nearest Neighbour), es un sistema de clasificación supervisado basado en criterios de vecindad (Aha, Kibler, & Albert, 1991) (Sancho, 2019). Su funcionamiento se basa en determinar las instancias de un conjunto de datos que son más parecidas a una nueva observación.

El resultado de la clasificación por medio de este algoritmo puede ser discreto o continuo. En el caso discreto, el resultado de la clasificación es la clase más común de los k vecinos (Phyu, 2009) (Bolaños Ramírez, 2017).

En particular, la clasificación k-NN se basa en la búsqueda de la mayor cantidad de vecinos más cercanos dentro de un conjunto de entrenamiento, basándose simplemente en "recordar" todos los ejemplos que se vieron en la etapa de entrenamiento. El algoritmo k-NN es conocido como un algoritmo perezoso, es decir, una técnica que espera hasta que llega la consulta para generalizar más allá de los datos de entrenamiento. Cuando un nuevo dato se presenta al sistema de aprendizaje, este se clasifica según el comportamiento del dato más cercano (Aha et al., 1991).

Para disminuir el tiempo empleado en escanear la totalidad de los datos, se emplea la regla de los *k* vecinos más cercanos (k-NN), donde se busca los *k* ejemplos del conjunto de entrenamiento más cercanos. Hay varios elementos clave de este enfoque: (Wu & Kumar, 2009, p. 10).

- El conjunto de objetos etiquetados que se usarán para evaluar la clase de un objeto de prueba.

- Una métrica de distancia o similitud que se puede usar para calcular la cercanía de los objetos.
- El valor de k, el número de vecinos más cercanos
- El método utilizado para determinar la clase del objeto de destino en función de las clases y distancias de los k vecinos más cercanos.

A continuación, se expone las métricas de distancia aplicables en el algoritmo k-NN. La notación es la siguiente: La matriz de datos es $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{n}}$, está formada por vectores N de dimensión n, representado los datos, de forma que el valor k del i-ésimo vector se denota por x_{ik}, donde $k \in \{1, \dots, n\}$ e $i \in \{1, \dots, N\}$.

- **Distancia euclídea**: Esta distancia se mide entre dos observaciones de un espacio, la cual se deduce a partir de la generalización del teorema de Pitágoras;

$$\boldsymbol{d_{ij}} = \sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} (\boldsymbol{x_{ik}} - \boldsymbol{x_{jk}})^{\boldsymbol{2}}} \qquad \text{Ecuación 1}$$

- **Distancia *cityblock*:** Suma de diferencias absolutas:

$$\boldsymbol{d_{ij}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \left|\boldsymbol{x_{ik}} - \boldsymbol{x_{jk}}\right| \qquad \text{Ecuación 2}$$

- **Distancia coseno:** Es una medida del coseno del ángulo entre las observaciones que son tratadas como vectores:

$$\boldsymbol{cos\theta} = \frac{\boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_j}}{|\boldsymbol{x_i}||\boldsymbol{x_j}|} \qquad \text{Ecuación 3}$$

- **Distancia de correlación:** Correlación entre los elementos. Cada centroide es el promedio de las observaciones que hacen parte de un determinado grupo, después de haber centrado y normalizado dichas observaciones.

2.5.2. Clasificadores árboles de decisión (Ruggieri, 2002)

En la investigación planteada por (Domingo et al., 2015), se presenta los resultados obtenidos en la competencia genérica de Lectura Crítica, cuyo objetivo fue detectar patrones de desempeño académico en las competencias genéricas de los estudiantes. La investigación se lleva a cabo con datos sociodemográficos, económicos, académicos e institucionales de los alumnos, llegando a descubrir patrones asociados al desempeño académico de los estudiantes; se utiliza un modelo de clasificación basado en árboles de decisión y la metodología CRISP-DM. Entre los patrones descubiertos, se destacan la acreditación institucional y la modalidad de estudio presencial, como dos factores importantes asociados al buen desempeño académico. Así también en el estudio planteado por (Spositto, Etcheverry, Ryckeboer, & Bossero, 2009), se presenta resultados sobre el nivel de deserción de los estudiantes; se utiliza el árbol de decisión implementados en Weka (algoritmo C4.5) como algoritmo de minería de datos. Sin embargo, este método no logra encontrar un clasificador con un alto grado de precisión y compresibilidad. Por ende, se plantea realizar futuros estudios implementando redes neuronales. Por otro lado, el estudio planteado por (Pereira & de Pasto, 2013), tuvo el objetivo de determinar perfiles de bajo rendimiento académico y deserción estudiantil en la comunidad universitaria, aplicando técnicas de descubrimiento de conocimiento, este estudio lo realizan con la herramienta TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios de DCBD del Departamento de Ingeniería e implementando el algoritmo C4.5.

Los árboles son uno de los algoritmos más sencillos y de fácil implementación, además de existir múltiple variaciones, por lo que son ampliamente utilizados (Rodallegas Ramos et al., 2010). Las aplicaciones son muy diversas como: el diagnóstico médico, juegos, predicciones meteorológicas y control de calidad, datos educativos, etc. Los árboles pueden ser representado

mediante un grafo, en el que cualquiera de los dos nodos están conectados exactamente a un camino (Ornella, 2010) (Solarte & Ocampo, 2009). El modelo más utilizado es el aprendizaje inductivo supervisado no paramétrico. Es decir, éste método se basa en reglas de forma recursiva, considera el criterio de la mayor proporción de ganancia de información, eligiendo atributos que mayor clasifiquen a los datos (Phyu, 2009) (Bolaños Ramírez, 2017).

El método de clasificación se basa en que la nueva instancia se parte de la raíz y recorre los nodos del árbol de acuerdo a los valores de los atributos, el proceso termina cuando se alcanza una hoja del árbol, asignando a la instancia el valor de clase de dicha hoja (Hall et al., 2009). Teóricamente el proceso se detiene cuando todos los nodos hojas contienen casos de una misma clase, como no siempre se desea llegar a este extremo. En la ***Figura 5*** se observa la estructura de un árbol de decisión, donde los nodos internos indican los diferentes atributos, las ramas los valores que estos pueden asumir y las hojas o nodos terminales presentan las etiquetas de clase.

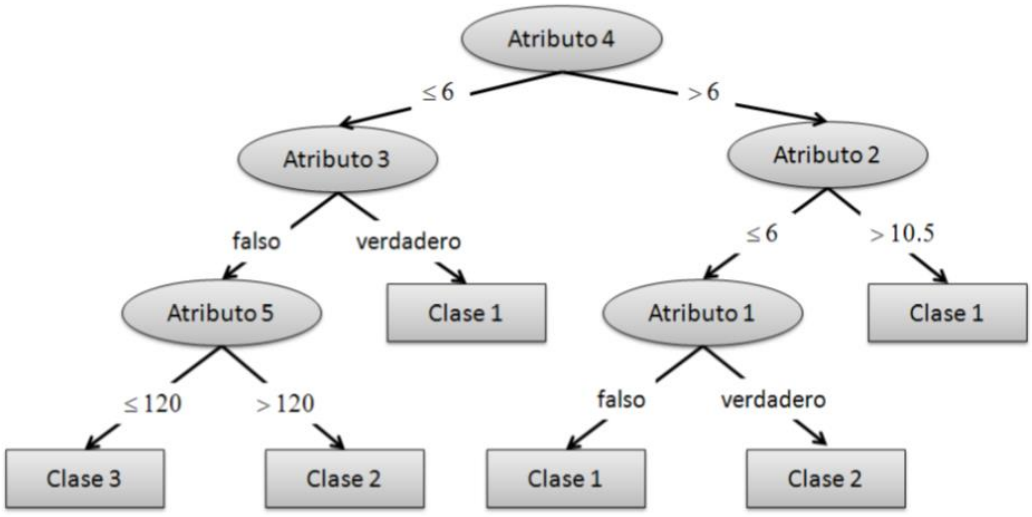


Figura 5. Estructura de un árbol de decisión
Fuente: (Ornella, 2010)

2.5.3. Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial (Support Vector Machines, SVM), se fundamentan en la estadística de aprendizaje desarrollada por el grupo de Vladimir Vapnik (Cortes & Vapnik, 1995), tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los años 90.

Inicialmente los SVM se plantearon para resolver problemas de clasificación binarios, actualmente se ocupan para (regresión, agrupamiento, multi-clasicación) (Suárez, 2014). Son diversos los campos en los que han sido utilizados los SVM con éxito, tales como visión artificial, reconocimiento de caracteres, categorización de texto e hipertexto, clasificación de proteínas, procesamiento de lenguaje natural, análisis de series temporales. Gracias a sus sólidos fundamentos desde su introducción, han ido ganando un merecido reconocimiento (Suárez, 2014).

Los clasificadores basados en SVM, realizan la búsqueda de un hiperplano que separe de forma óptima a los elementos de cada clase, que han sido previamente proyectados a un espacio de dimensión superior. La característica fundamental de los SVM es la "separación óptima", por eso también a veces se les conoce como clasificadores de margen máximo; este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. Logrando que los puntos del vector etiquetados de una clase estarán a un lado del hiperplano y los casos que se encuentren en la otra clase estarán al otro lado. Las máquinas de vectores de soporte pertenecen a la familia de los clasificadores lineales. Al vector formado por los puntos más cercanos al hiperplano se le llama vector de soporte (Pulgarín, 2012) (L. H. R. González, 2015).

2.5.4. Mezcla de clasificadores

Una mezcla de clasificadores es un conjunto de clasificadores cuyas predicciones individuales se combinan de alguna forma para así obtener una predicción final conjunta. Las técnicas de combinación se pueden agrupar y analizar de diferentes maneras, según el criterio de clasificación principal adoptado por (Valentini & Masulli, 2002), identifican dos grandes grupos dependiendo del grupo de entrada como criterio principal, uno que usa el mismo y otro que usa diferentes representaciones de los datos.

El multi-clasificador Bagging es un método para generar múltiples versiones de un predictor y usarlas para obtener un predictor agregado. La agregación promedia las versiones cuando predice un resultado numérico y hace un voto de pluralidad cuando predice una clase. Las múltiples versiones se forman haciendo repeticiones de arranque del conjunto de aprendizaje y usándolas como nuevos conjuntos de aprendizaje. Las pruebas muestran que el empaquetamiento puede proporcionar ganancias sustanciales en la precisión, en conjuntos de datos reales y simulados, utilizando árboles de clasificación y regresión (Friedman, 2001). En la Figura 6 se observa el diagrama del multi-clasificador Bagging.

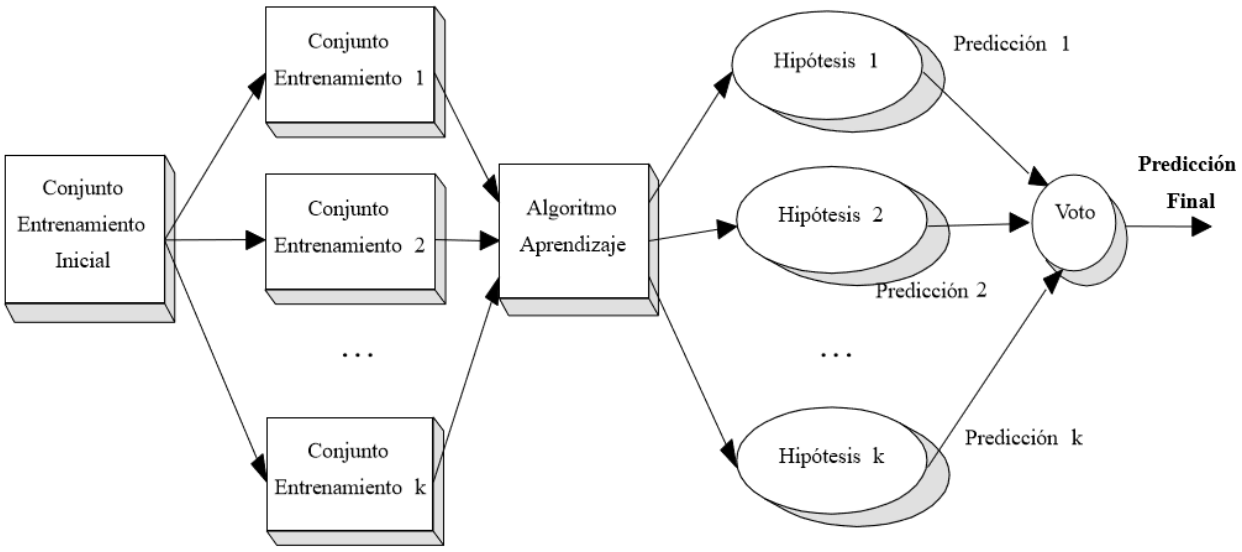


Figura 6. Diagrama de Bagging
Fuente: Fuente: (C. J. A. González, 2019)

El multi-clasificador Boosted es un tipo de técnica iterativa que persigue la minimización de los errores mediante la introducción de nuevos modelos. Se basa en errores de iteraciones anteriores. Su aplicación más común es cuando los clasificadores de forma individual presentan bajo porcentaje de clasificación (Campillo, Ibáñez, & Vargas, 2004.). Este algoritmo se desarrolla en la década del siglo XX (Kuhn & Johnson, 2013), en vista que los clasificadores por sí solos en algunos casos no presentan un buen resultados; se realiza la combinación con el objetivo de producir un clasificador de ensamble que reduzca el error de clasificación incorrecta en comparación con el obtenido de los modelos individuales. En la ***Figura 7*** se observa el diagrama del multi-clasificador Boosting.

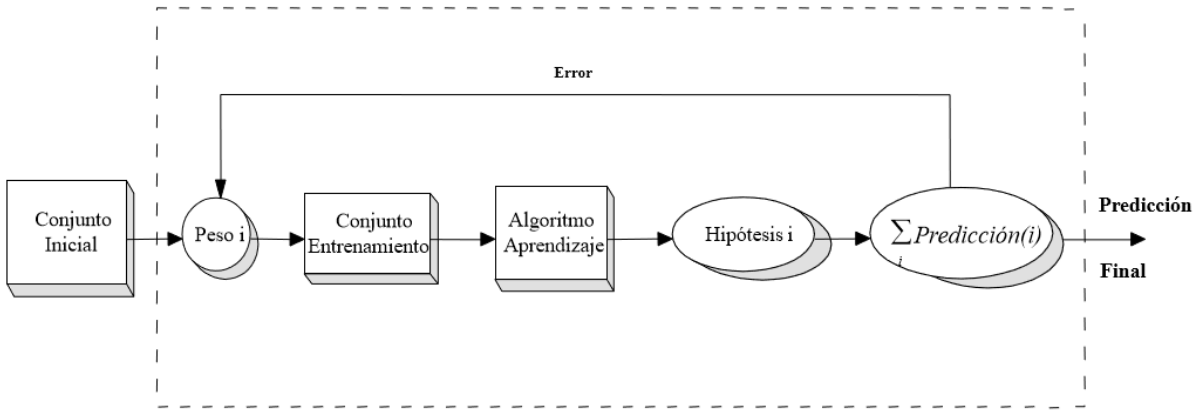


Figura 7. Esquema Boosting
Fuente: (C. J. A. González, 2019)

Para (Valentini & Masulli, 2002) existen cuatro niveles de actuación en la construcción de multi-clasificadores:

1. El nivel de combinación de los clasificadores base. A este nivel existen distintos modos de combinación de las predicciones individuales.
2. El nivel de los clasificadores base, seleccionando qué combinación de clasificadores base se van a implementar.
3. El nivel de las características. Se basa en tener distintos clasificadores base ya sea quitando, añadiendo y/o modificando las características del conjunto de datos, de una manera distinta para cada clasificador base, aun cuando los clasificadores base sean del mismo tipo.
4. En nivel del conjunto de datos. Haciendo que con algún criterio los conjuntos de datos de cada clasificador base sean distintos, aun cuando los clasificadores base sean del mismo tipo.

2.6. Medidas de desempeño

Las medidas de desempeño son una herramienta de evaluación y apoyo a la decisión, en esta investigación se utilizan las siguientes medidas de desempeño: sensibilidad (Se), especificidad (Sp) y porcentaje de clasificación (CP). Las ecuaciones de estas medidas de desempeño se expresan en la Tabla 1.

Tabla 1

Ecuaciones de las medidas de desempeño

Descripción	Definición	Nombre de la medida
Mide la proporción de muestras positivas correctamente clasificadas	$\frac{V_n}{V_n + F_p} \times 100$	Sensibilidad
Mide la proporción de muestras negativas correctamente clasificadas	$\frac{V_p}{V_p + F_n} \times 100$	Especificidad
Entrega una relación de los datos correctamente clasificados en base al número total de datos del conjunto de prueba	$\frac{V_n + V_p}{V_n + F_p + V_p + F_n} \times 100$	Exactitud

Donde:

- V_p Son los verdaderos positivos. Un alumno de la clase de interés es clasificado correctamente.
- V_n Son los verdaderos negativos. Un alumno diferente a la clase de interés es clasificado correctamente.
- F_p Son los falsos positivos. Un alumno que no pertenece a la clase de interés es clasificado como clase de interés.
- F_n Son los falsos negativos. Un alumno que no pertenece a la clase de interés es clasificado como no clase de interés.

2.6.1. Matriz De Confusión

La matriz de confusión es una herramienta fundamental para evaluar el desempeño de un algoritmo de clasificación, empleado en el aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa las instancias en la clase real. En la Tabla 2 se puede observar el caso de una matriz bi-clase.

Tabla 2

Matriz de confusión de una matriz bi-clase

		Clasificación	
		Clase estimada	
		Verdadero	**Negativo**
Clase real	**Verdadero**	Verdaderos positivos	Falsos positivos
	Negativo	Falsos negativos	Verdaderos negativo

2.7. Visualización de datos

La Visualización de Información, utiliza interfaces interactivas entre el usuario y la computadora, con el objetivo principal de representar con mínima entropía visual una serie de datos (Keim, Mansmann, Schneidewind, & Ziegler, 2006) (Mazza, 2009), (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008). Según (Córdoba Cely & Alatriste Martínez, 2012) (Roweis & Saul, 2000) la visualización de información se caracteriza por:

- Ser inter-relacional,
- Transformar datos "crudos" en información relevante,
- Buscar la mínima pérdida de información en dicha transformación,
- Dirigirse a usuarios que interactúan, transforman e interpretan esta información.

La manera en que las personas perciben la información es fundamental en el campo de la visualización, por lo cual, la comprensión individual del usuario puede mejorar notablemente.

2.7.1. Diagrama de dispersión

Un diagrama de dispersión ayuda a identificar patrones de respuesta en entornos naturales; se puede ver una aproximación más formal a la visualización de la dependencia espacial, basada en el concepto del Scatter plot (Moreno & Vayá, 2004).

El diagrama de dispersión es un gráfico que tiene en las abscisas las observaciones de la variable de estudio y en el de ordenadas el retardo espacial de la misma (Moran, 1948), en ambos casos las variables están normalizadas. De este modo, los cuatro cuadrantes reproducen diferentes tipos de dependencia espacial. Si la representación por puntos está dispersa en los cuatro cuadrantes es indicio de ausencia de correlación espacial. Por el contrario, los valores se encuentran concentrados sobre la diagonal que cruza los cuadrantes I (derecha superior) y III (izquierda inferior), existe una elevada correlación espacial positiva de la variable, coincidiendo su pendiente con el valor de la I

de Moran. La dependencia será negativa si los valores se concentran en los dos cuadrantes restantes (Moreno & Vayá, 2004).

2.7.2. Diagrama de caja

El diagrama de caja (boxplot) es un tipo de gráfico que resume información útil de 5 medidas estadísticas (Minnaard, Condesse, & Rabino, 2010); las medidas representadas en el gráfico de caja son: el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo.

Los cuartiles permiten dividir la muestra de datos en cuatro partes iguales, lo cual permite evaluar la dispersión y la tendencia central de los mismos.

En el diagrama de caja se representa el rango intercuartil de los datos en una caja, se tiene como extremos el percentil 75 (cuartil superior) y el percentil 25 (cuartil inferior), mientras que el percentil 50 corresponde a la mediana.

Un percentil es una medida de posición que se usa en estadística para indicar el valor de las variables por debajo del cual se encuentra un porcentaje, luego de ordenar los datos de menor a mayor. Además de la caja se prolonga extensiones, que muestran las observaciones externas en la muestra (Walpole & Myers, 2012).

La información visual representada en el diagrama de caja y su extensión sirve como una herramienta de diagnóstico, en la ***Figura 8*** se pueden observar las características que contiene un diagrama de caja.

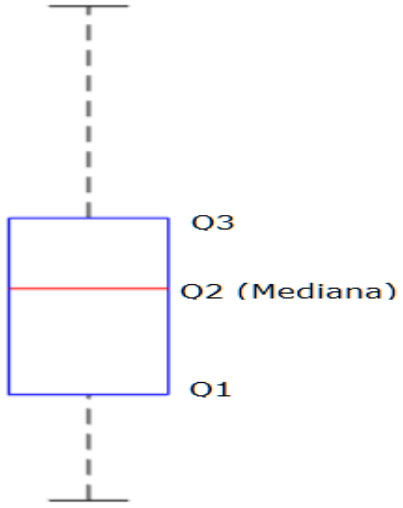


Figura 8. Representación gráfica de un diagrama de caja

2.7.3. Diagrama de barra

El diagrama de barras representa gráficamente un conjunto de datos o valores que pueden estar orientados horizontal o verticalmente, está conformado por barras rectangulares de longitudes proporcionales a los valores representados. Los gráficos de barras ***Figura 9*** son usados para comparar dos o más valores (Der & Everitt, 2014).

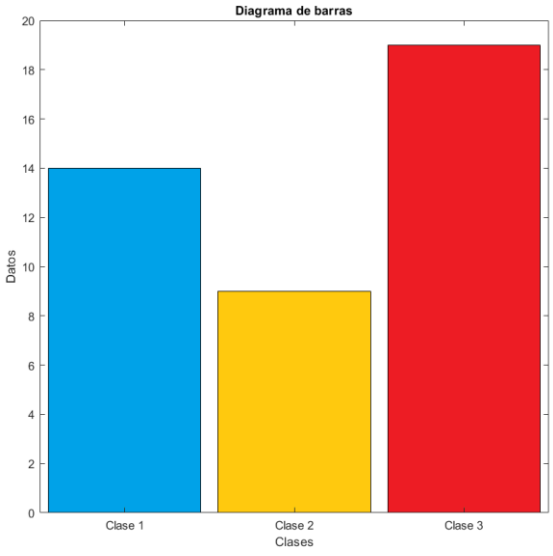


Figura 9. Diagrama de barras

CAPÍTULO III

3. METODOLOGÍA DE COMPARACIÓN PROPUESTA

3.1. Introducción

En esta sección se detalla la metodología utilizada, resumida en la ***Figura 10***, esta contiene etapas de pre-procesamiento de datos, caracterización, selección de los mejores atributos, entrenamiento de modelos individuales y de múltiples clasificadores, continuando con las medidas de desempeño, que sirven para la evaluación de los algoritmos implementados, y concluir con una interpretación en contexto de los resultados.

3.2. Selección de métodos a comparar

La investigación se realiza con la ayuda del software Weka para la selección de las características y Matlab para la selección de mejores prototipos, clasificación, implementación de los clasificadores, multi-clasificadores y los resultados de la investigación, los cuales se detallan en la Tabla 3.

Tabla 3

Métodos y herramientas de la metodología seleccionada

Detalle	Métodos	Herramienta
Selección de características	BestFirst – SubsetEaval (Sección 4.2.1) Ranker – CorrelationAttribute (Sección 4.2.1) Ranker – PCA (Sección 4.2.1)	Weka
Selección de instancias	Kennard – Stone (Sección 2.4.2.5)	Matlab
Clasificación	Árbol de decisión (Sección 3.3.3) kNN – Ponderado (Sección 3.3.4) SVM (Sección 3.3.5)	Matlab
Multi-clasificadores	Bagged_Tree (Sección 3.3.7) Boosted (Sección 3.38)	Matlab
Presentación de resultados	Diagrama de caja (Sección 2.10.2) Diagrama de dispersión (Sección 2.10.1) Diagrama de barras (Sección 2.10.3)	Matlab

3.3. Desarrollo de la metodología

En la siguiente sección se explica los pasos a seguir en cada etapa del proceso, los cuales son: Selección de la base de datos, limpieza de datos, selección de características, selección de mejores atributos y balanceo, evaluación de clasificadores, diseño de modelo clasificador, medición de desempeño e interpretación en contexto.

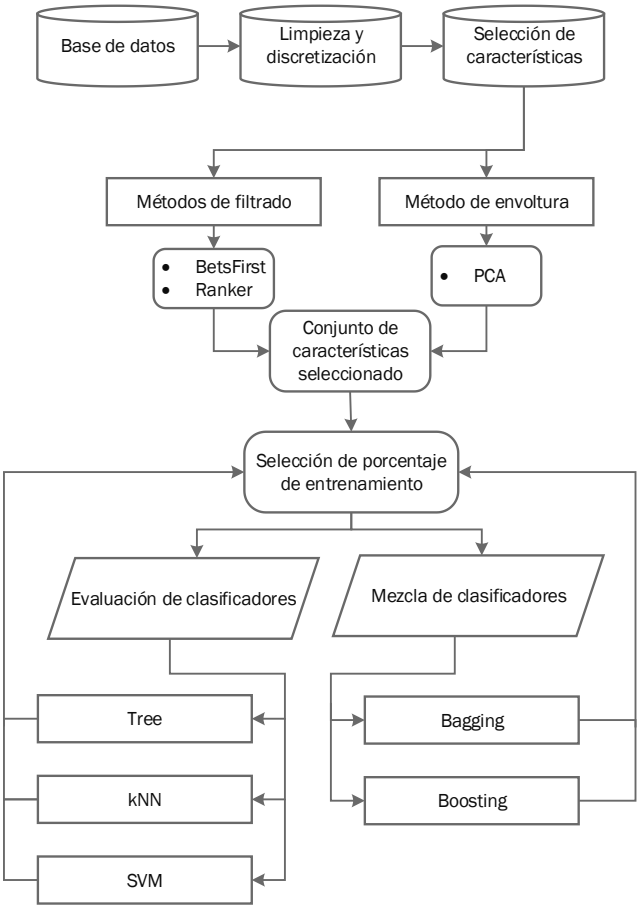


Figura 10. Metodología propuesta

3.3.1. Selección de características

El proceso de selección de características se realiza con la ayuda de la herramienta de software libre Weka, esta etapa tiene como objetivo determinar un subconjunto de características relevantes a partir de la totalidad de los datos; permitiendo mejorar el rendimiento de un clasificador,

disminuyendo el espacio de almacenamiento y el tiempo de procesamiento. El proceso de selección de características se puede observar a detalle en la sección 2.4.2. En primer lugar, se selecciona un subconjunto de datos candidatos desde el conjunto inicial. Posteriormente, con una función, se evalúa la calidad de los datos seleccionados. El proceso continúa hasta que un criterio de parada determine si continuar con la búsqueda de otro subconjunto o no (Espinoza, 2017).

Selección de atributos

En esta sección se detalla las diferentes estrategias en el desarrollo de la presente investigación para la selección de un subconjunto de datos representativos.

Cfs-SubsetEval, Este método extrae un subconjunto de atributos de forma jerárquica basada en correlaciones estadísticas, usando una función de evaluación heurística. De tal manera, busca atributos que tengan una buena correlación con la clase y que estén poco correlacionados entre sí, ignorando las características irrelevantes que mantienen baja o nula correlación con la clase. La información redundante por otra parte será penalizada, ya que el atributo redundante tendrá una alta correlación con una o varias de las características restantes. La inclusión de una característica por tanto depende de si esta es capaz de explicar la clase en fragmentos del espacio de instancias que no han sido ya explicadas por otros atributos. La función de evaluación utilizada es la siguiente (Bolaños Ramírez, 2017):

$$\boldsymbol{Ms} = \frac{\boldsymbol{k}\,\overline{\boldsymbol{ref}}}{\sqrt{\boldsymbol{k} + \boldsymbol{k}(\boldsymbol{k}-\boldsymbol{1})\overline{\boldsymbol{r}_{ff}}}} \qquad \text{Ecuación 4}$$

Donde:

Ms es el mérito heurístico del subconjunto S conteniendo

k representa las características

$\overline{ref}$ es el valor de la correlación media entre la clase y la característica

$f(f \subset S)$y $\overline{r_{ff}}$ es la mejor correlación entre dos características del conjunto *S*.

En la presente investigación se utiliza como algoritmo de búsqueda *BestFirst*.

Algoritmo BestFirst

Es un algoritmo de búsqueda heurística general (en profundidad) (Korf, 1993). Mantiene una lista abierta que contiene los nodos de la frontera del árbol que se han generado pero que aún no se han expandido, y una lista cerrada que contiene los nodos interiores o expandidos. Cada nodo tiene un valor de costo asociado. En cada ciclo, un nodo abierto de costo mínimo se expande, generando todos sus hijos. Los hijos son evaluados por la función de costo, se insertan en la lista abierta y el nodo principal se coloca en la lista cerrada. Inicialmente, la lista abierta contiene solo el nodo inicial, y el algoritmo termina cuando se elige un nodo objetivo para la expansión**.**

Para la búsqueda se emplea un árbol, que consiste en ir eliminando atributos hasta llegar a un numero de atributos, los cuales son seleccionados por el usuario, el subconjunto resultante es evaluado usando métricas monotónicas y el resultado es guardado como una cota. Se continúa eliminando atributos del total de datos de una manera organizada, cada subconjunto obtenido debe ser evaluado. Si se observa que el subconjunto tiene una evaluación igual o peor de la cota, se detienen la búsqueda y se elimina (se poda), debido a que la solución no conduce a una mejor solución de la actual. Sin embargo, si los subconjuntos nuevos tienen mejor resultado, se remplaza la cota con este nuevo valor, continua la clasificación hasta terminar de explorar todas las ramas. De este modo, se garantiza el procedimiento (usando la métrica correcta) ahorrando el tiempo de procesamiento (Gutiérrez García, 2016) (Bolaños Ramírez, 2017).

Componentes principales (PCA)

Para el presente estudio se usó el análisis de componentes principales (PCA). Esta técnica se basa en el análisis exploratorio de datos, con el objetivo de sintetizar la información o reducir la dimensión (número de variables). Es decir, iniciar con una base de datos de alta dimensión, para reducir a un menor número de variables latentes perdiendo la menor cantidad de información (Terrádez-Gurrea, 2006).

En la Figura 11 se puede observar la reducción de dimensión partiendo de la tabla de datos (Tabla izquierda) hacia las nuevos componentes reducidos, en las cuales se concentra lo esencial de la información (Tabla derecha); el cálculo se basa en aspectos matemáticos independiente de su interpretabilidad (Ramírez-Anormaliza et al., 2017) (Vásquez & Ramírez, 2012).

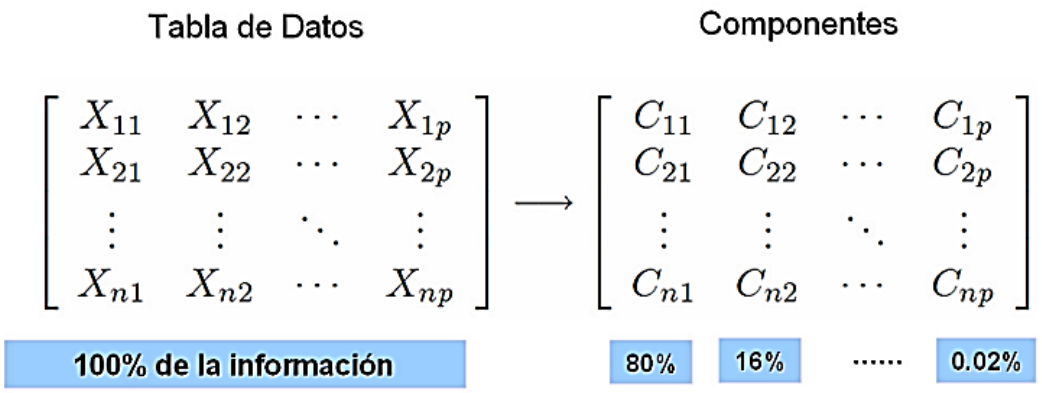


Figura 11. Transformación de las variables originales en componentes

Fuente: (Rodríguez Rojas, 2009)

3.3.2. Algoritmo Kennard – Stone

Después de realizar la selección de características por los diferentes métodos, se puede observar en la Tabla 4 un desequilibrio en el número de muestras por clase, lo que muestra la necesidad de aplicar una técnica de balanceo de datos. El balanceo es llevado a cabo con el algoritmo planteado

por (Kennard & Stone, 1969), permite extraer un subconjunto de muestras de mayor representatividad del conjunto de datos original, además, ayuda a disminuir los datos y de esta manera equilibrar las clases.

Tabla 4

Muestras por cada clase en base a los métodos de selección de características

Base de datos	Clase 1	Clase 2
Todos los datos	310	3416
Cfs-BestFirst	115	2377
Ranker	96	993
PCA	228	3208

El método de Kennard – Stone se basa es buscar los datos más dispersos dentro de cada clase; se aplica la distancia euclídea para cumplir este objetivo. Para cada par de muestras i, j la distancia euclídea en el espacio x se define como (Saptoro et al., 2012).

$$\boldsymbol{d_x(i,j) = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{M} (x_{ik} - x_{jk})^2}} \qquad \text{Ecuación 5}$$

Selección de métodos a comparar

Luego de realizar el proceso de limpieza y reducción de datos, se procede con el proceso de evaluación de los algoritmos de clasificación. Según la descripción de la literatura analizada, los clasificadores que más destacan son: Árboles de decisión, kNN ponderado y Máquina de soporte vectorial (SVM), algoritmos seleccionados para esta investigación; también se expone la mezcla de clasificadores Boosted_Tree y Bagged_Tree. Esto se lo realiza con la ayuda del Toolbox de Matlab, el cual proporciona algoritmos de aprendizaje automático supervisados y no supervisados. Muchas de las técnicas y los algoritmos de aprendizaje automático pueden usarse para cálculos en conjuntos de datos que son demasiado grandes para ser almacenados en la memoria.

3.3.3. Árbol de decisión

Para el método de árbol de decisión se ejecuta el modelo implementado en Matlab con el nombre *Fitctree*, el cual selecciona automáticamente el subconjunto óptimo de algoritmos para cada división utilizando el número conocido de clases y niveles de un predictor categórico. Se ejecuta el método con todos los datos y con las categorías obtenidas mediante los algoritmos BestFirst, Ranker y PCA mostrados de a detalle en el Anexo A.

3.3.4. kNN

El algoritmo kNN ponderado implementado en esta investigación mediante el uso del Toolbox de Matlab se lo explica a detalle en el Anexo B.

La categorización de los puntos de prueba, basando en su distancia a los puntos en un conjunto de datos de entrenamiento puede ser una forma simple pero efectiva de clasificar nuevos puntos. Puede usar varias métricas para determinar la distancia, que se describe a continuación.

3.3.5. SVM

Para el modelo de SVM se ejecuta el modelo implementado en Matlab con el nombre *templateSVM*, este devuelve una plantilla de aprendizaje adecuada para entrenar modelos multiclase de código de salida con corrección de errores (ECOC). El algoritmo usa un Kernel polinomial de grado 2.

3.3.6. Mezcla de clasificadores

Debido a la existencia de señales difíciles de clasificar, se requiere mejorar la exactitud y precisión (Dietterich, 1997), con el fin de mejorar el rendimiento de un clasificador individual se aplica una mezcla de clasificadores. Existen varios estudios que comprueban que la mezcla de clasificadores, denominadas multi-clasificadores, mejoran la precisión en comparación con el

rendimiento de cada clasificador de forma individual (Bauer, Kohavi, Chan, Stolfo, & Wolpert, 1999; Breiman, 1996; Dietterich, 2000).

Existen varios métodos de mezcla de clasificadores, entre los cuales se tiene:

- Métodos basados en generación de ensamblados. Se encargan de generar los clasificadores que van a conformar el ensamblado Estos sistemas fijan un esquema de combinación como por ejemplo el voto mayoritario. En general consiste en conformar clasificadores independientes en cuanto a sus respuestas. Ejemplos de estos métodos se tiene el Bagging, Boosting, Bosques Aleatorios, etc.
- Métodos basados en selección de clasificadores. Estos métodos tratan de seleccionar cuál de los *L* clasificadores es el más apropiado para asignar una clase a un objeto. Donde *L* es un conjunto de los clasificadores ya entrenados.
- Métodos basados en combinación de clasificadores. Dado un conjunto de *L* clasificadores, estos métodos combinan o fusionan los *L* resultados de sus miembros para retornar una respuesta final.
- Métodos híbridos. Aquí se agrupan los métodos que combinan varias o todas las estrategias descritas anteriormente.

En esta investigación se implementa dos mezclas de clasificadores detalladas a continuación.

Bagged: es un modelo predictivo compuesto por una combinación ponderada de árboles de regresión múltiple. Los árboles de decisión individuales tienden a sobre ajustarse. Los árboles de decisión agregados (empaquetados) combinan los resultados de muchos árboles de decisión, lo que reduce los efectos de sobreajuste y mejora la generalización. En general, la combinación de árboles de aumenta el rendimiento predictivo. Además, *Bagged* selecciona un subconjunto aleatorio de

predictores para usar en cada división de decisión como en el algoritmo de bosque aleatorio (Segal, 2004).

Boosted: Construye iterativamente los clasificadores base, y en cada iteración da mayor importancia a instancias que han sido clasificadas de forma errada. El resultado final de la clasificación se realiza mediante ponderaciones de los clasificadores base, dando mayor pesos a los clasificadores con mayor tasa de acierto (Toca, 2016). Este método es una combinación de árboles de regresión reforzados y empaquetado. Un conjunto de árbol de regresión es un modelo predictivo compuesto por una combinación ponderada de árboles de regresión múltiple. En general, la combinación de árboles de regresión múltiple aumenta el rendimiento predictivo (Segal, 2004).

3.4. Propuesta de la interpretación en contexto

Se propone en la interfaz de usuario, una sección de interpretación en contexto de los resultados con el uso de diagramas de pasteles y conclusiones de la investigación. Emitiendo un informe de los resultados obtenidos en la investigación para las autoridades de la Unidad Educativa Ibarra, de manera que sirva como sustento de información y permita plantear y ejecutar estrategias para disminuir el grado de alumnos que presenten bajas calificaciones.

CAPÍTULO IV

4. MARCO EXPERIMENTAL

En el desarrollo de los experimentos de esta sección, se utiliza la base de datos facilitada por la Unidad Educativa Ibarra de los terceros de bachillerato del periodo 2017-2018. Se evalúa el desempeño de los diferentes métodos de selección de características, balanceo y las combinaciones de clasificadores, las medidas de desempeño utilizadas son: matriz de confusión y gráficos de cajas, con el objetivo de identificar los mejores métodos que mejor se adapten al presente caso de estudio.

4.1. Base de datos a utilizar

En esta investigación se utiliza datos de los alumnos provenientes de los tercer año de bachillerato de la Unidad Educativa Ibarra del periodo 2017 – 2018. La adquisición de los datos se lo realiza mediante una encuesta, la cual es llenada por los alumnos al iniciar cada periodo académico, además, de sus datos de identificación personal; la encuesta contiene los principales antecedentes sociodemográficos, tales como edad, genero, estado civil, personas con las que viven, el nivel educativo alcanzado por los padres y el tipo de trabajo y categoría ocupacional de los mismos. También se dispone de los antecedentes académicos tales como: pérdida de año escolar, deseo de abandono de los estudios, la relación que tiene entre sus compañeros, tipo de aprendizaje, y en la sección económica comentan preguntas como su situación alimenticia, condición de su vivienda, entre otros. Este formulario es único para todo el colegio, el cual se completa, registra y almacena en la unidad DECE (Departamento de Consejería Estudiantil) de forma manual. Sin embargo, los datos son recolectados o facilitados por los alumnos, quienes en general completan la ficha sin ningún tipo de asesoramiento especializado, y esta situación deja libre a su comprensión las consignas o códigos establecidos en dicho formulario. Por lo tanto, las inconsistencias de los

datos en muchas veces, son de libre comprensión de cada estudiante con lo que hace meritorio un preprocesamiento de datos para continuar con el proceso de minería.

La base de datos de la Unidad Educativa Ibarra de los terceros de bachillerato del periodo se encuentra detallada en la Tabla 5.

Tabla 5

Base de datos de los estudiantes del tercero de bachillerato periodo 2017 – 2018 Unidad Educativa Ibarra

Número	Atributo	Descripción	Valores
1	genero	Género	Femenino Masculino
2	edad	Edad del estudiante en el momento de llenar la encuesta	Continuo
3	vive_madre	Si el estudiante vive con la madre	Si (1) No (0)
4	vive_padre	Si el estudiante vive con el padre	Si (1) No (0)
5	vive_hermanos	Si el estudiante vive con hermanos	Si (1) No (0)
6	vive_otros	Si el estudiante vive con otros familiares	Si (1) No (0)
7	viven_juntos	Si sus padres viven juntos	Juntos (1) Separados (2) Con nueva pareja (3) Fallecido (4)
8	relacion_con_padres	La relación entre padres y estudiante	Muy buena (1) Buena (2) Regular (3) Conflictiva (4)
9	relacion_entre_padres	La relación entre padres del estudiante	Muy buena (1) Buena (2) Regular (3) Conflictiva (4)
10	numero_hermanos	Cuántos hermanos tiene	Continua (0-8)
11	hermanos_estudiando	Cuántos hermanos tiene que están estudiando	Continua (0-7)
12	hermanos_trabajando	Cuántos hermanos tiene que están trabajando	Continua (0-6)
13	educacion_madre	Nivel de educación de la madre	Ve Tabla 6
14	educacion_padre	Nivel de educación del padre	Ver Tabla 7
15	trabajo_madre	Que trabajo tiene la madre	Ver Tabla 8
16	trabajo_padre	Que trabajo tiene el padre	Ver Tabla 9

CONTINÚA

17	situacion_economica	Situación económica del estudiante	Buena (1) Regular(2) Mala (3) Deficiente (4)
18	situacion_alimenticia	Situación alimenticia del estudiante	Buena (1) Regular(2) Deficiente (3)
19	condicion_vivienda	Condición de la vivienda	Propia (1) Prestada (2) Arrendada (3)
20	servicio_telefono	Si el estudiante cuenta con servicio telefónico	Si (1) No(0)
21	servicio_internet	Si el estudiante cuenta con servicio de internet	Si (1) No(0)
22	horas_dormir	Cuántas horas dedica a dormir	Continuo (5-9)
23	revisan_tareas	Si los representantes le revisan los deberes	Siempre (1) A veces (2) Nunca (3)
24	perdio_ano_escolar	Si el estudiante ha perdido algún año escolar	Si (1) No(0)
25	deseo_abandono	Si el estudiante ha deseado abonadora los estudios	A veces (1) Nunca (2)
26	ritmo_aprendizaje	Que ritmo de aprendizaje considera que tiene	Rápido (1) Lento (2) Muy lento (3)
27	aprendizaje_auditivo	Si el aprendizaje es auditivo	Si (1) No(0)
28	aprendizaje_visual	Si el aprendizaje es visual	Si (1) No(0)
29	aprendizaje_kinestesico	Si el aprendizaje es kinestesico	Si (1) No(0)
30	dislexia	Si el estudiante padece de dislexia	Si (1) No(0)
31	sentir_colegio	Como se siente en el colegio	Mucho (1) Poco (2) Quiero cambio de colegio (3)
32	relacion_companeros	Qué tipo de relación mantiene entre compañeros	Siempre (1) A veces (2) Nunca (3)
33	peleas_compañeros	Si ha tenido peleas entre compañeros	Siempre (1) A veces (2) Nunca (3)
34	Tipo de Bachillerato	Tipo de bachillerato cruza el estudiante	Ver Tabla 10
35	Nota 1	Nota que obtiene en las diferentes materias en la primera parcial	Continuo
36	Materia	Materia del estudiante	Ver Tabla 11
37	Resultado	Aprueba la materia sin supletorio	Si (1) No (0)

En la Tabla 6 se observa de menor a mayor grado el nivel de educación de los padres de los estudiantes, desde que no presenta educación hasta educación superior.

Tabla 6

Educación de la madre-padre

Código	Descripción
0	NA
1	Ninguna
2	Primaria
3	Secundaria
4	Superior

A continuación, en la Tabla 7 se observa los diferentes trabajos que registra la base de datos respecto a la educación de la madre.

Tabla 7

Trabajo de la madre

Código	Descripción
0	NA
1	Abogada
2	Administradora
3	Ama de casa
4	Artesana
5	Aux. odontología
6	Aux. Servicio
7	Cajera
8	Camarera
9	Carpintera
10	Chofer
11	Cocinera
12	Comerciante
13	Contadora
14	Costurera
15	Diseño gráfico
16	Empelada domestica
17	Empleada publica
18	Enfermera
19	Estilista
20	Modista
21	Parvulario
22	Peluquera
23	Policía
24	Profesora
25	Psicóloga
26	Secretaria
27	Supervisora
28	Trabajadora privada

En la Tabla 8 se detalla los diferentes tipos de trabajo de los padres.

Tabla 8

Trabajo del padre

Código	Descripción
0	NA
1	Abogado
2	Administrador
3	Agricultor
4	Albañil
5	Arbitro
6	Arquitecto
7	Artesano
8	Bombero
9	Carpintero
10	Chofer
11	Comerciante
12	Conserje
13	Consultor privado
14	Contratista
15	Controlador
16	Costurero
17	Eléctrico
18	Empleado privado
19	Empleado publico
20	Estilista
21	Guardia de seguridad
22	Ingeniero
23	Jardinero
24	Jubilado
25	Laboratorista
26	Mecánico
27	Militar
28	Nutricionista
29	Obrero
30	Panadero
31	Profesor
32	Relojero
33	Tejedor
34	Topógrafo
35	Trabajador publico
36	Tramitador de aduana

En la Tabla 9 se detalla los diferentes tipos de bachilleratos de la base de datos

Tabla 9

Tipo de bachillerato

Código	Bachillerato	Descripción
0	BAGED	Bachillerato unificado diurno
1	BAGEN	Bachillerato unificado general nocturno

CONTINÚA

2	BASCD	Bachillerato ciencias sociales diurno
3	CONTD	Bachillerato contabilidad diurno
4	INFOD	Bachillerato informática diurno
5	PRBID	Bachillerato internacional diurno

A continuación, en la Tabla 10 se enlista todas las materias que se dictan en los diferentes cursos de bachillerato.

Tabla 10

Materias de los diferentes cursos de bachillerato

Código	Descripción
0	Anatomía
1	Biología
2	Ciencias naturales
3	Contabilidad bancaria
4	Contabilidad de costos
5	Contabilidad general
6	Corrientes filosóficas
7	Desarrollo de funciones
8	Dibujo técnico aplicado
9	Educación artística
10	Educación física
11	Emprendimiento y gestión
12	Empresa y gestión
13	Estudios matemáticos
14	Estudios sociales
15	Física
16	Formación y orientación laboral
17	Geografía del ecuador
18	Gestión del talento humano
19	Historia
20	Implantación de aplicaciones
21	Informática aplicada a la educación
22	Ingles
23	Lengua y literatura
24	Matemática
25	Números complejos
26	Paquetes contable y tributarios
27	Problemas del mundo contemporáneo
28	Proyectos escolares
29	Química
30	Razonamiento numérico y abstracto
31	Razonamiento verbal
32	Redes de área local
33	Relaciones en el entorno de trabajo
34	Sistemas informáticos
35	Sociología
36	Teoría del conocimiento (TOC)

De acuerdo a la revisión bibliográfica sobre los factores que intervienen el desempeño académico y los antecedentes teóricos, los 36 atributos del conjunto de datos detallados en la Tabla 5, se agruparon en: sociodemográfica, económica, académica e institucional, en la Tabla 11 se detalla su agrupación.

Tabla 11

Clasificación de atributos en dimensiones

Dimensión	Atributo
Sociodemográfica	género, edad, vive_madre, vive_padre, vive_otros, viven_juntos, relacion_con_padres, relacion_entre_padres, numero_hermanos, servicio_internet
Económica	situacion_economica, situacion_alimenticia, condicion_vivienda, servicio_telefono
Académica	perdio_ano_escolar, deseo_abandono, ritmo_aprendizaje, aprendizaje_auditivo, aprendizaje_visual, aprendizaje_kinestesico, dislexia, Tipo de Bachillerato, Nota 1, Materia
Institucional	sentir_colegio, relacion_companeros, peleas_compañeros

4.2. Descripción de los experimentos

En esta sección se describe el desarrollo del trabajo acorde a la metodología y el marco experimental, en base a los datos económicos, sociodemográficos, académicos e institucionales, de la base de datos de la Unidad Educativa Ibarra. La metodología a seguir se encuesta detallada en la ***Figura 10***, la cual contiene los siguientes pasos: limpieza, selección de características, métodos de selección de características, modelos de entrenamiento y visualización de medidas de desempeño (matriz de confusión, sensibilidad, predicción, especificidad) también la combinación de métodos de clasificación.

En primera instancia se describe y compara los resultados de los diferentes métodos de selección de características aplicados sobre la base de datos. Posteriormente, se aplica y compara los diferentes algoritmos de clasificación seleccionados para este estudio. También, se aplica y compara los diferentes métodos de mezcla de clasificadores, finalmente, se calcula las medidas de desempeño.

4.2.1. Selección de características

En primer lugar, se realiza la selección de atributos en función a la revisión bibliográfica realizada, para esta investigación se decidió utilizar la combinación de los siguientes métodos; *CFSSubsetEval + Bestfirst*, *correlaccionAttributeEval + Ranker* y *PCA + Ranker*, implementados en la herramienta Weka 3.9.3. A continuación se detalla cada método, con los resultados obtenidos.

BestFirst

La Figura 12 muestra las características seleccionadas por la combinación del evaluador de atributos *BestFirst* y el método de búsqueda *CfsSubsetEval* implementados en Weka.

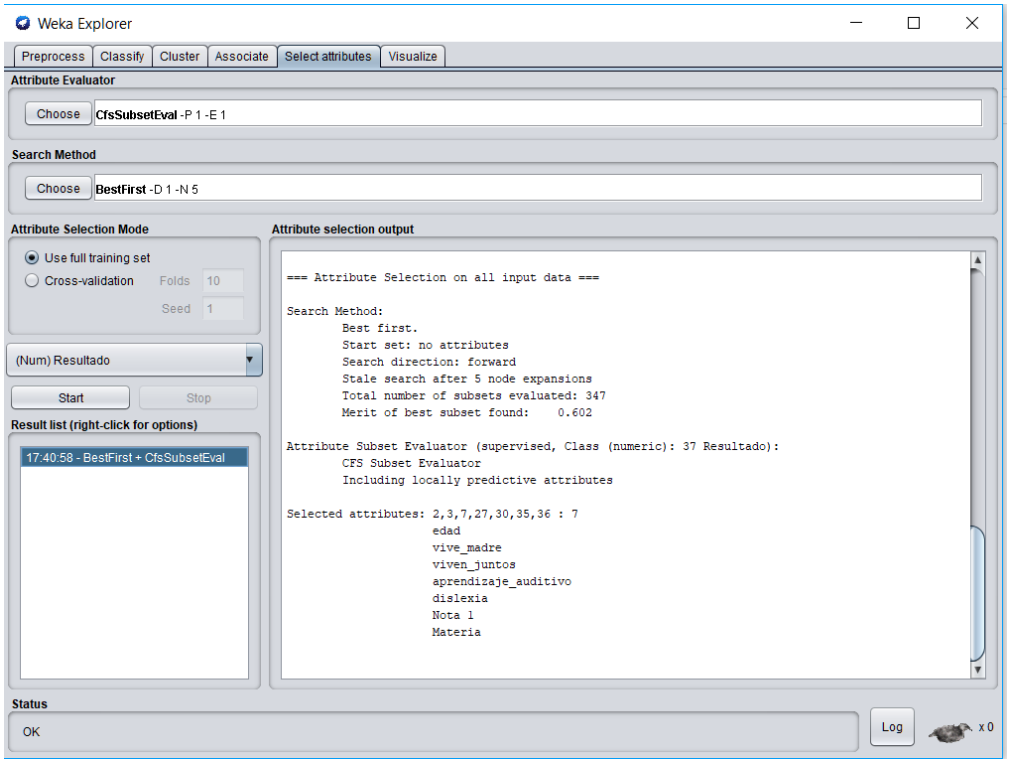


Figura 12. Selección de atributos método BesrtFirst y evaluador CFS

Este método busca a través de toda la combinación de las características en el conjunto de datos y concluye con un subconjunto que incluye características que tienen buenas capacidades de predicción. En el experimento, el número de variables se redujo a 6 de un total de 36. El método

de búsqueda utilizado en las pruebas fue *BestFirst*, que inicia la búsqueda sin atributos mientras va agregando uno a uno los atributos que mejor valor de predicción tengan sobre la clase objetivo. El método se detiene cuando los atributos que se agregan ya no generan un mejor desempeño. El mérito del subconjunto final fue 0,602. La Tabla 12 enlista los atributos seleccionados por este método.

Tabla 12

Atributos seleccionados por usando el método de búsqueda BestFirst

Evaluador / Método de búsqueda	Modo de selección de atributos	No. De atributos seleccionados	Orden de atributos seleccionados	Mérito
CFS/BestFirst	CfsSubsetEval	6	Edad Vive_madre Vivien_juntos Aprendizaje_auditivo Dilexia Materia Nota	0,602

Selección de atributos con el método de búsqueda Ranker y evaluador de atributos CorrelationAtributrEval

En la ***Figura 13*** se observa el resultado de Weka al implementar el método de búsqueda *Ranker* con el selector de atributos *CorrelationAtributeEval*.

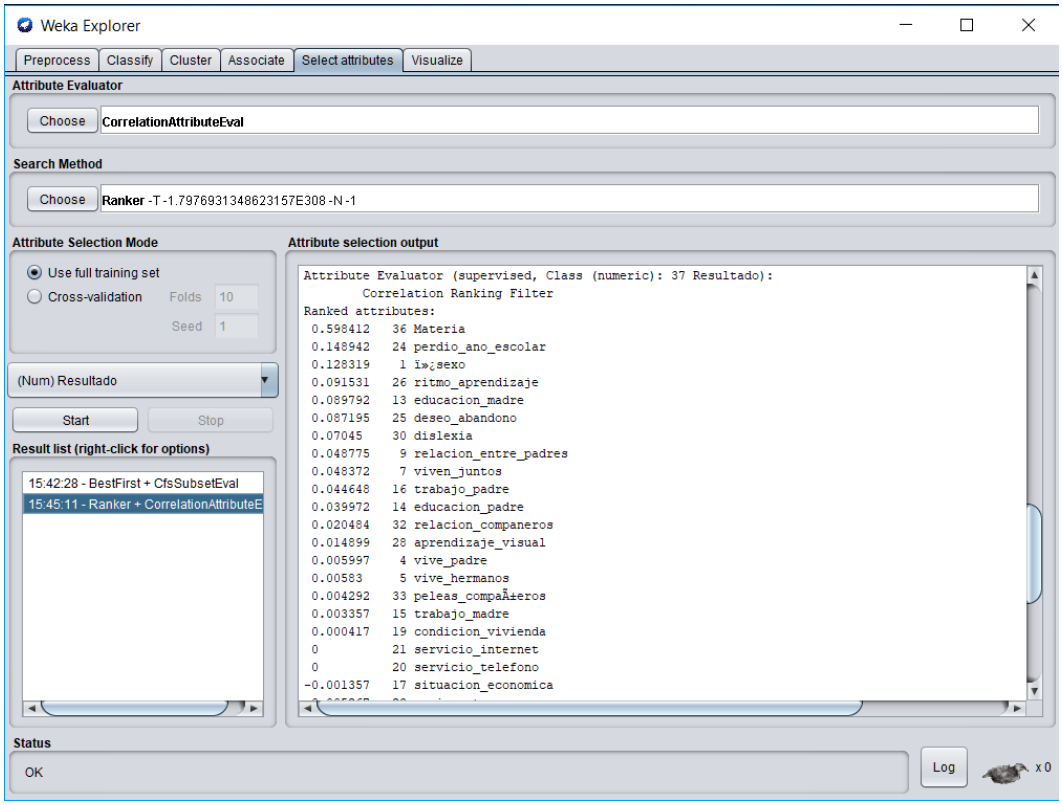


Figura 13. Selección de atributos método Ranker y evaluador CorrelationAtrributeEval

La Tabla 13 muestra los resultados de ejecutar el algoritmo de selección de atributos de correlación con el método de búsqueda *Ranker* sobre la base de datos. Este método selecciona las características evaluando el valor de un atributo midiendo la correlación (Pearson), entre este y la clase. Usando el método *CorrelationAtributeEval* se observa que la categoría *Materia* se clasifica en primer lugar con un valor de 0.598, mientras que la categoría *perdio_año_escolar* se clasifica en segundo lugar con un valor de 0.1489, algo que con el anterior método no sucedía, debido a que este atributo no quedó dentro de los 6 atributos seleccionados. También se puede observar que la categoría *edad* quedó fuera de la clasificación, pero ingresó la categoría *perdió_ano_escolar*, variables que tiene cierta relación.

Tabla 13

Atributos seleccionados usando el método de búsqueda Ranker (CorrelationAtrributeEval)

Nº	Rank	No. Atributo	Descripción
1	0.59841	36	materia
2	0.14894	24	perdio_ano_escolar
3	0.12831	1	genero
4	0.09153	26	ritmo_aprendizaje
5	0.08979	13	educacion_madre
6	0.08719	25	deseo_abandono

Otro método de selección de características aplicado es el método basado en envoltura de Análisis de Componentes Principales (PCA), con el método de búsqueda *Ranker*. En la ***Figura 14*** se observa los resultados que brinda Weka.

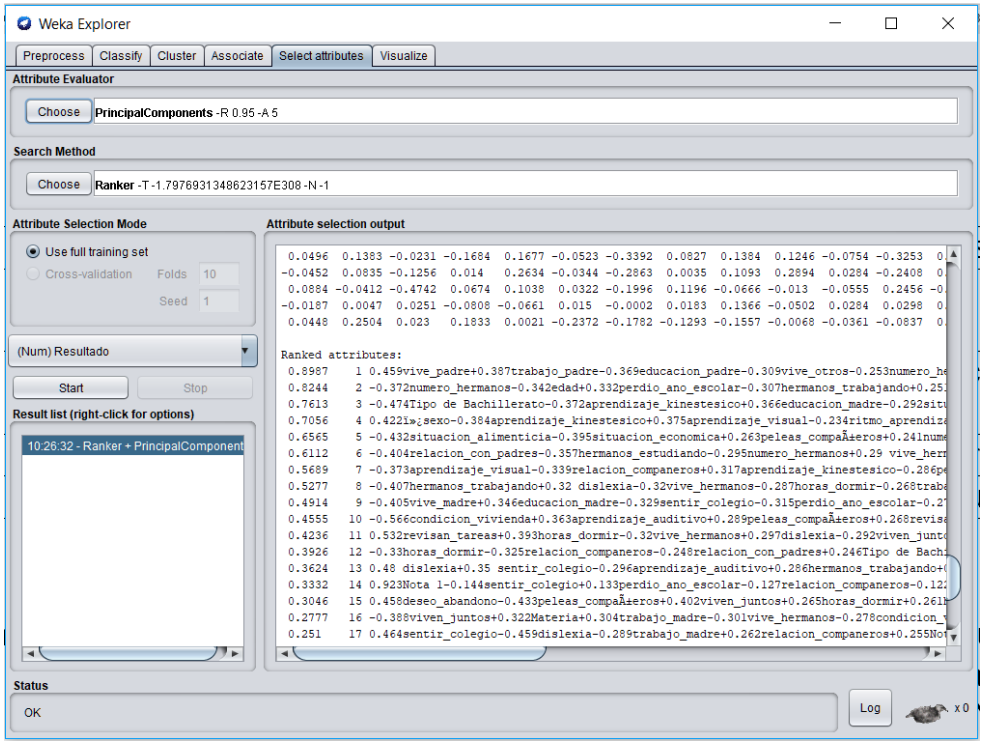


Figura 14. Selección de atributos método Ranker y evaluador PrincipalComponents

En la Tabla 14, se enlista desde el grupo de mayor a menor porcentaje de explicación los datos que tiene cada componente generado por el método Ranke, este método retorna un subconjunto de combinaciones lineales basado en los datos originales. El primer grupo obtiene un 0.8987 lo cual

es un porcentaje alto de explicación, en el proceso de clasificación se observará el rendimiento de estos resultados de acuerdo al método empleado.

Tabla 14

Atributos seleccionados del método de búsqueda Ranker (PrincipalComponents)

Ranking de atributos		
0.8987	1	0.459vive_padre+0.387trabajo_padre-0.369educacion_padre-0.309vive_otros-0.253numero_hermanos
0.8244	2	0.372numero_hermanos-0.342edad+0.332perdio_ano_escolar-0.307hermanos_trabajando+0.251educacion_madre
0.3356	3	-0.474Tipo de Bachillerato-0.372aprendizaje_kinestesico+0.366educacion_madre-0.292situacion_economica+0.29 aprendizaje_visual
0.2162	4	0.422genero-0.384aprendizaje_kinestesico+0.375aprendizaje_visual-0.234ritmo_aprendizaje+0.225situacion_economica
0.1261	5	-0.432situacion_alimenticia-0.395situacion_economica+0.263peleas_compaÃ±eros+0.241numero_hermanos-0.222sentir_colegio...
0.076	6	-0.404relacion_con_padres-0.357hermanos_estudiando-0.295numero_hermanos+0.29 vive_hermanos-0.265relacion_entre_padres
0.0371	7	-0.373aprendizaje_visual-0.339relacion_companeros+0.317aprendizaje_kinestesico-0.286peleas_compaÃ±eros-0.27ritmo_aprendizaje

4.3. Medidas de desempeño utilizadas

Para medir el desempeño de los clasificadores se usa la matriz de confusión y el cálculo de los parámetros evaluadores de los algoritmos como son sensibilidad Se, especificidad Sp, y porcentaje de clasificación Cp.

4.4. Resultados

En esta sección se detalla todos los experimentos realizados con la base de datos de la Unidad Educativa Ibarra tanto con los clasificadores y multi-clasificadores.

4.4.1. Resultados por cada experimento

Una vez realizada la selección de características por los diferentes métodos detallados en el apartado 4.2.1 y continuando con el segundo objetivo, se aplica los diferentes métodos de

clasificación y mezclas de clasificadores, y para medir su desempeño se utiliza diferentes medidas de desempeño y la matiz de confusión.

En primer lugar, se aplica el clasificador de árboles de decisión, con todas las características del conjunto de datos original, así como las seleccionadas por los métodos *Cfs-BestFirst, Ranker y Componentes Principales (PCA).* La ejecución se realiza utilizando la configuración predeterminada en Matlab, que incluye un máximo de 4 ramificaciones, ver **¡Error! No se encuentra el origen de la referencia.**. En todos los clasificadores se realiza una iteración de cien veces para comprobar la estabilidad del modelo. Los resultados al ejecutar el método *árbol_de_decisión* se muestra en la ***Figura 15*** mediante un diagrama de caja.

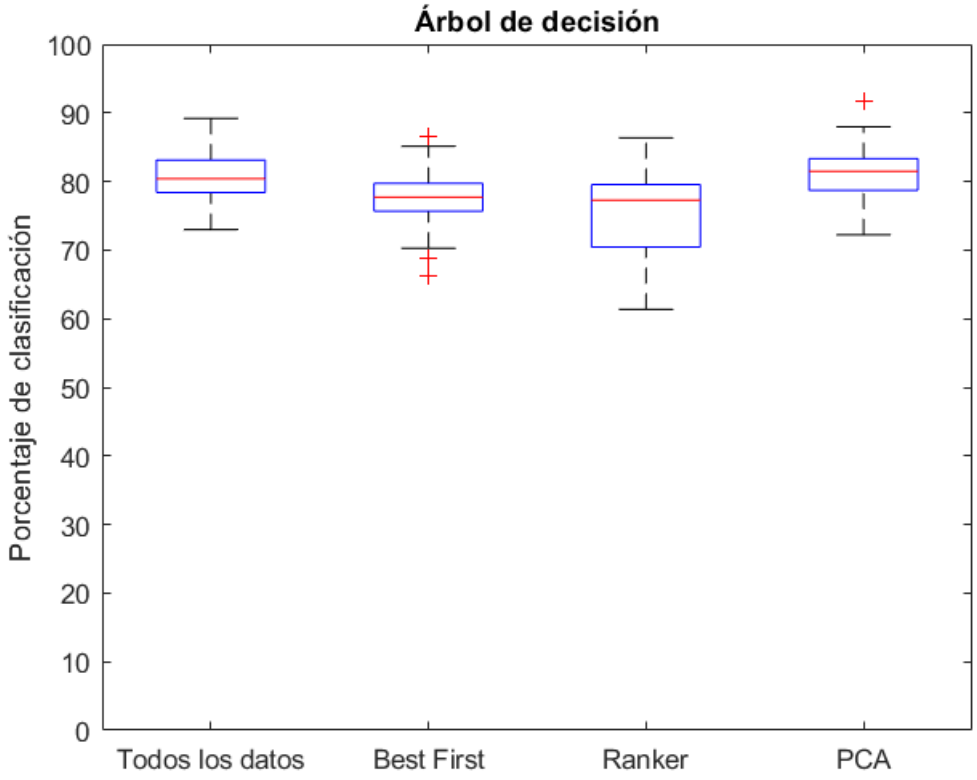


Figura 15. Cuartiles del clasificador árbol de decisión con los diferentes grupos de datos

En la ***Figura 15*** se observa que presenta mayor estabilidad el método si se ejecuta con todo el grupo de datos. Se prosigue a calcular las medidas de desempeño del clasificador que se resumen en la Tabla 15.

Tabla 15

Rendimiento del clasificador Árbol de decisión sobre la base de datos, sensibilidad (Se), especificidad (Sp), porcentaje de clasificación (Cp) con 76% de entrenamiento.

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
Árbol_de_decisión_Cjto_todos	**99.63**	77.03	**97.76**
Árbol_de_decisión_CfsBestFirst	83.78	97.30	90.54
Árbol_de_decisión _Ranker	73.91	**100**	86.96
Árbol_de_decisión_PCA	76.36	98.18	87.27

En la Tabla 15 se puede observar que la reducción de características en la base de datos totales, se logra obtener un 100% en porcentaje de *especificidad* con los datos seleccionados con el método Ranker. Sin embargo, la tasa de sensibilidad y porcentaje de clasificación disminuyó al momento de disminuir las características en los tres conjuntos de datos. No obstante, el método *BestFirts* tiene un equilibrio entre especificidad y porcentaje de clasificación. A continuación, se expone la matriz de confusión, aplicando el modelo de *árboles de decisión* con los diferentes conjuntos de datos.

La Tabla 16 se detalla la matriz de confusión al aplicar a todos los datos usando el clasificador árboles de decisión, se puede observar que las clases están desbalanceadas.

Tabla 16

Matriz de confusión con el clasificador árboles de decisión con todo el conjunto de datos.

Clase	Negativo	Positivo
Negativo	57	17
Positivo	3	817

En la Tabla 17 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *BestFirst* y las clases balanceadas.

Tabla 17

Matriz de confusión del clasificador árboles de decisión con los atributos seleccionados por CFS/BestFirst

Clase	Negativo	Positivo
Negativo	36	1
Positivo	6	31

En la Tabla 18 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método Ranker y las clases balanceadas.

Tabla 18

Matriz de confusión del clasificador árboles de decisión con atributos seleccionados por Ranker (CorrelationAtrributeEval)

Clase	Negativo	Positivo
Negativo	23	0
Positivo	6	17

En la Tabla 19 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método PCA y las clases balanceadas.

Tabla 19

Matriz de confusión del clasificador árboles de decisión con los atributos seleccionados por Ranker PCA

Clase	Negativo	Positivo
Negativo	54	1
Positivo	13	42

Resultados con el clasificador SVM Cuadrático

Siguiendo el mismo procedimiento, se aplica el clasificador *SVM Cuadrático* con todas las características del conjunto de datos original, así como los atributos seleccionados por los métodos *Cfs-BestFirst*, *Correlation AtrributeEval* y *Componentes Principales*. La ejecución se realiza utilizando la configuración predeterminada en MATLAB, ver **¡Error! No se encuentra el origen de la referencia.**.

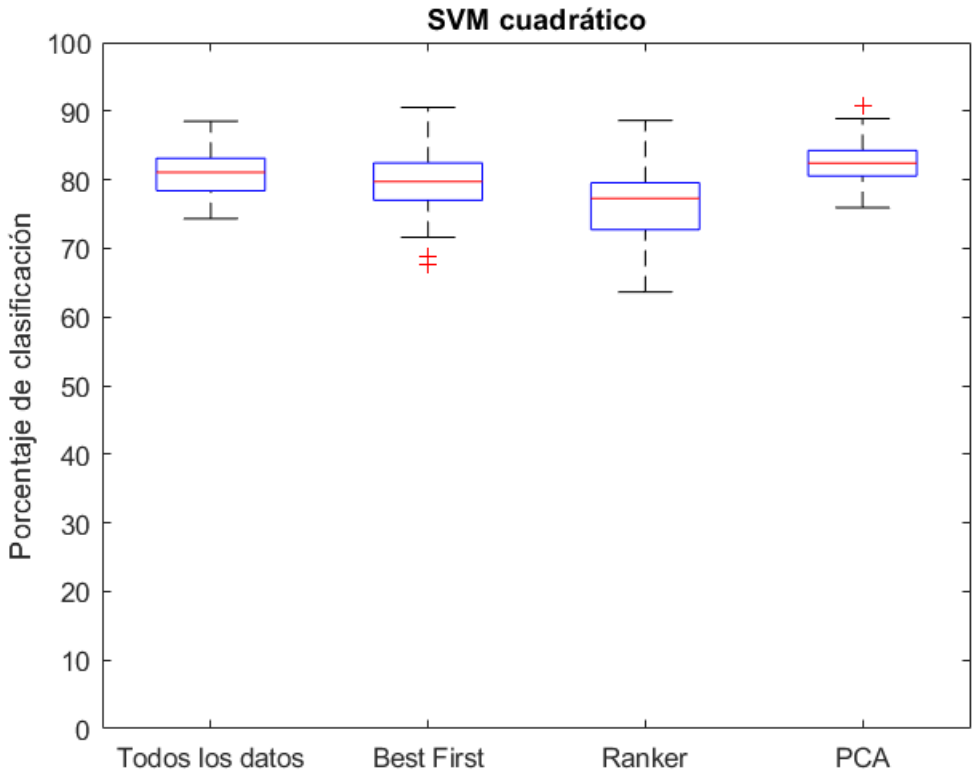


Figura 16. Cuartiles del clasificador SVM cuadrático con los diferentes grupos de datos

En la Figura 16 se puede observar los diagramas de caja después de ejecutar cien veces el método *SVM_Cuadrático* con los diferentes grupos de datos, se observa que presenta mayor estabilidad el grupo seleccionados mediante *PCA*. A continuación, se detallas la matriz de confusión con los diferentes datos y el método *SVM_Cuadrático*.

Tabla 20

Rendimiento del clasificador SVM cuadrático, sobre la base de datos en términos de sensibilidad (Se), especificidad (Sp) y porcentaje de clasificación (Cp) con un porcentaje de 76% de entrenamiento.

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
SVM_Cuadrático_Cjto_todos	74.32	**98.65**	86.49
SVM_Cuadrático_CfsBestFirst	**89.19**	97.30	**93.24**
SVM_Cuadrático_Ranker	73.91	95.65	84.78
SVM_Cuadrático_PCA	81.82	98.18	90.00

Como se puede observar en la Tabla 20, la reducción del conjunto de características en el conjunto de datos mejora las puntuaciones del porcentaje de sensibilidad llegando a tener 89.19% y en porcentaje de clasificación 93.24 % con las características seleccionadas por el método *BestFirst*. Sin embargo, la tasa de especificidad disminuyó en los tres conjuntos de datos.

La Tabla 21 se detalla la matriz de confusión al aplicar a todos los datos usando el clasificador SVM cuadrático, con las clases desbalanceadas.

Tabla 21

Matriz de confusión con el clasificador SVM cuadrático con todo el conjunto de datos.

Clase	Negativo	Positivo
Negativo	73	1
Positivo	19	55

En la Tabla 22 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método BestFirst y las clases balanceadas.

Tabla 22

Matriz de confusión con el clasificador SVM cuadrático con los atributos seleccionados por CFS/BestFirst

Clase	Negativo	Positivo
Negativo	36	1
Positivo	4	33

En la Tabla 23 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método Ranker y las clases balanceadas.

Tabla 23

Matriz de confusión con el clasificador SVM cuadrático con los atributos seleccionados por Ranker (CorrelationAtrributeEval)

Clase	Negativo	Positivo
Negativo	22	1
Positivo	6	17

En la Tabla 24 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método PCA y las clases balanceadas.

Tabla 24

Matriz de confusión con el clasificador SVM cuadrático con los atributos seleccionados por PCA

Clase	Negativo	Positivo
Negativo	54	1
Positivo	10	45

Resultados con el clasificador kNN Ponderado

Siguiendo el mismo procedimiento, se aplica el clasificador kNN ponderado con todas las características del conjunto de datos original, así como las seleccionadas por los métodos *Cfs-BestFirst, CorrelationAtrributeEval* y *componentes principales PCA*. La ejecución se realizó utilizando la configuración predeterminada en Matlab, **¡Error! No se encuentra el origen de la referencia.**. Como método de validación se ejecuta cien veces el algoritmo con los diferentes grupos de datos, los resultados están expresado en diagrama de caja como se puede observar en la Figura 17.

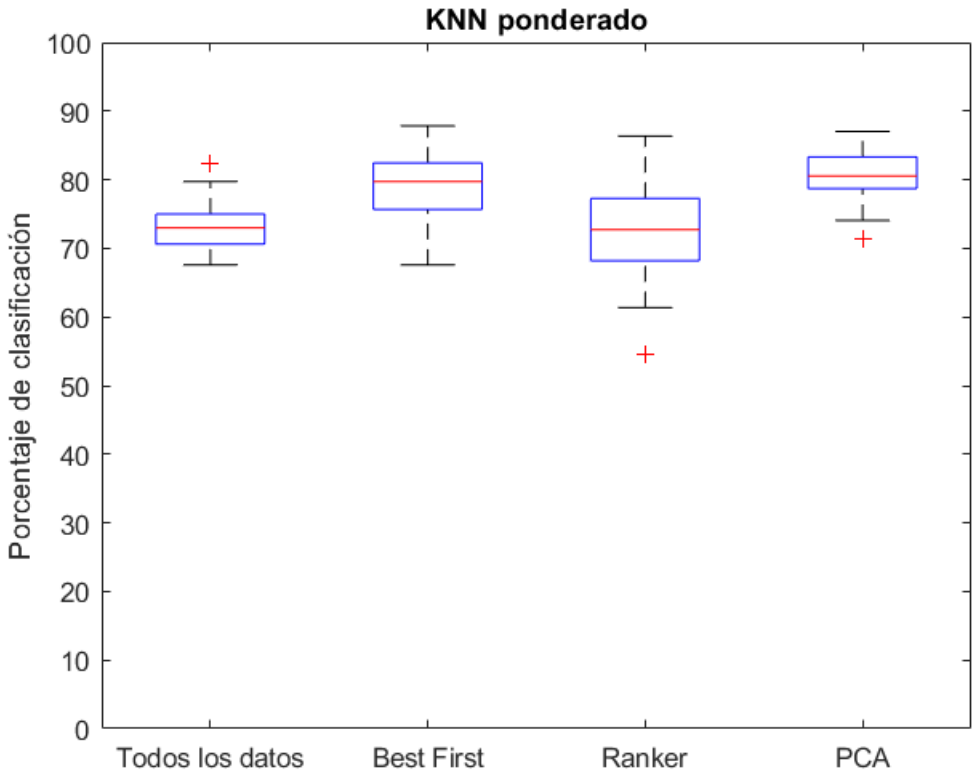


Figura 17. Cuartiles del clasificador kNN Ponderado con los diferentes grupos de datos

En la Figura 17 se observa que presenta mayor estabilidad el grupo seleccionados mediante *PCA*.

A continuación, se detalla la matriz de confusión con los diferentes datos.

Tabla 25

Rendimiento del clasificador kNN Ponderado sobre la base de datos, en términos de sensibilidad (Se), especificidad (Sp), porcentaje de clasificación (Cp) con un porcentaje de 76% de entrenamiento.

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
kNN_Ponderado_todos	**100**	58.11	**96.53**
kNN_Ponderado_Cfs-BestFirst	81.08	91.89	86.49
kNN_ Ponderado_Ranker	82.61	91.30	86.96
kNN_ Ponderado_PCA	60.00	**96.36**	78.18

En la Tabla 25, se puede observar que el mejor porcentaje de especificidad se obtiene con el método de selección de características *PCA* con un porcentaje de 96.36 %. Mientras que el mejor

porcentaje de clasificación se tiene con todos los datos, pero hay que tomar en cuenta que el porcentaje de especificidad es el más bajo con el 58.11%.

La Tabla 26 se detalla la matriz de confusión al aplicar a todos los datos usando el *kNN Ponderado*, con las clases desbalanceadas.

Tabla 26

Matriz de confusión con el clasificador kNN Ponderado con todo el conjunto de datos.

Clase	Negativo	Positivo
Negativo	42	32
Positivo	0	820

En la Tabla 27 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *BestFirst* y las clases balanceadas, con el método *kNN Ponderado*.

Tabla 27

Matriz de confusión con el clasificador kNN Ponderado con los atributos seleccionados por CFS/BestFirst

Clase	Negativo	Positivo
Negativo	34	3
Positivo	7	30

En la Tabla 28 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *Ranker* y las clases balanceadas, con el método *kNN Ponderado*.

Tabla 28

Matriz de confusión con el clasificador kNN Ponderado con los atributos seleccionados por Ranker (CorrelationAtrributeEval)

Clase	Negativo	Positivo
Negativo	21	22
Positivo	4	19

En la Tabla 29 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método PCA y las clases balanceadas, con el método k*NN Ponderado*.

Tabla 29

Matriz de confusión con el clasificador kNN Ponderado con los atributos seleccionados por CPA

Clase	Negativo	Positivo
Negativo	53	2

Positivo	22	33

Resultados con el multi-clasificador Bagged_Tree

Tabla 30

Rendimiento del multi-clasificador Bagged_Tree en términos de sensibilidad (Se), especificidad (Sp), porcentaje de clasificación (Cp) con un porcentaje de 76% de entrenamiento.

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
Bagged_Tree_todos	**100**	72.97	**97.76**
Bagged_Tree_Cfs-BestFirst	86.49	**100**	93.24
Bagged_Tree_Ranker	78.26	**100**	89.13
Bagged_Tree_PCA	76.36	98.18	87.27

La Tabla 30, muestra que el método de mezcla de clasificadores basado en Bagged_*Tree*, que se ajusta a los clasificadores básicos en subconjuntos aleatorios del conjunto de datos original y luego agrega sus predicciones individuales (ya sea votando o promediando), para formar una predicción final utilizando todo el conjunto de atributos. Se puede observar que el porcentaje de especificidad del 100% se tiene con los dos métodos de selección de características *Ranker* y *BestFirst.* Sin embargo, la sensibilidad disminuye en los grupos seleccionados con los diferentes métodos de clasificación, cabe aclarar que el grupo de todos los datos están desbalanceados por lo que al tener el 100% de sensibilidad no sería representativo como se puede observar en la matriz de confusión de la Tabla 31. En la Figura 18 se puede observar los diagramas de caja después de ejecutar cien veces el multi-clasificador *Bagged_Tree*, ver configuración **¡Error! No se encuentra el origen de la referencia.**.

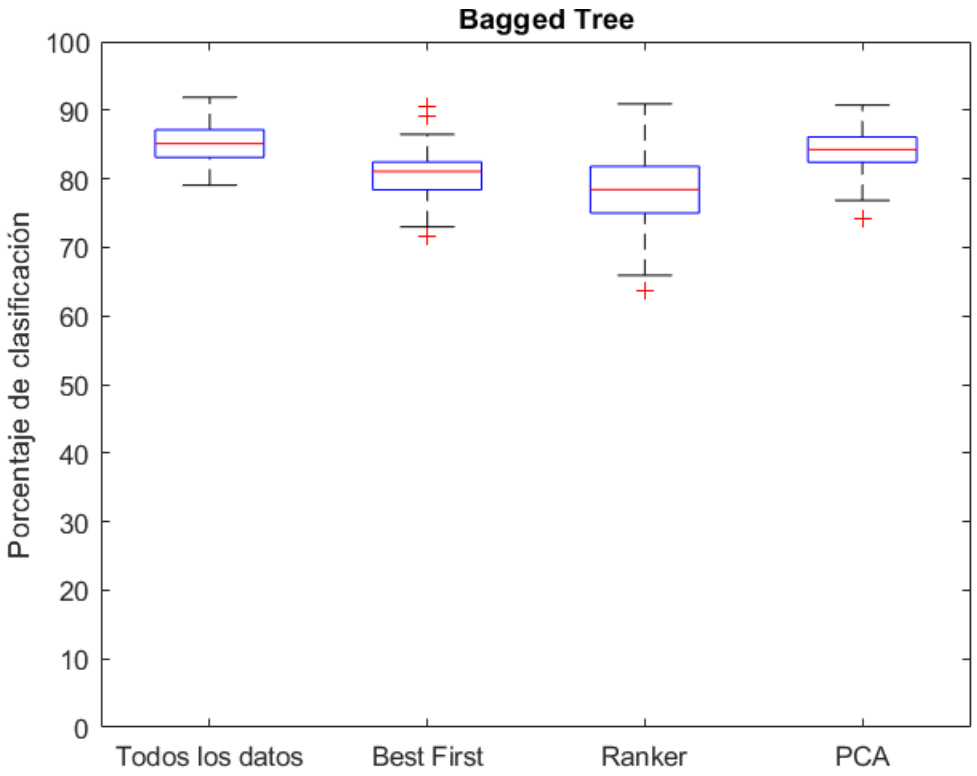


Figura 18. Cuartiles del multi-clasificador Bagged_Tree con los diferentes grupos de datos

En la Figura 18 se observa que presenta mayor estabilidad el multi-clasificador *Bagged_Tree* si se trabaja con todos los datos, sin realizar selección de características. A continuación, se detalla la matriz de confusión son los diferentes datos.

La Tabla 31 detalla la matriz de confusión al aplicar a todos los datos usando el multi-clasificador Bagged_tree, con las clases desbalanceadas.

Tabla 31

Matriz de confusión con el multi-clasificador Bagged_tree y todo el conjunto de datos.

Clase	Negativo	Positivo
Negativo	54	20
Positivo	0	820

En la Tabla 32 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *BestFirst* y las clases balanceadas, aplicado el *multi-clasificador Bagged_tree*.

Tabla 32
Matriz de confusión con el multi-clasificador Bagged_tree con los atributos seleccionados por CFS/BestFirst

Clase	Negativo	Positivo
Negativo	37	0
Positivo	5	32

En la Tabla 33 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *Ranker* y las clases balanceadas, aplicado el *multi-clasificador Bagged_tree*.

Tabla 33
Matriz de confusión con el multi-clasificador Bagged_tree con los atributos seleccionados por Ranker

Clase	Negativo	Positivo
Negativo	23	0
Positivo	5	18

En la Tabla 34 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *PCA* y las clases balanceadas, aplicado el *multi-clasificador Bagged_tree*.

Tabla 34
Matriz de confusión con el multi-clasificador Bagged_tree con los atributos seleccionados por PCA

Clase	Negativo	Positivo
Negativo	51	1
Positivo	13	42

Resultados con el multi-clasificador Boosted_Tree

Finalmente, se concluye la etapa experimental realizando la combinando los clasificadores a través del método *Boosted_Tree* con la configuración establecida en Matlab, ver Anexo E. Para el experimento se utiliza todas las características del conjunto de datos original, así como las seleccionadas por los métodos *Cfs-BestFirst, Ranker y Componentes Principales (PCA)* detalladas

en la Tabla 35. En la Figura 19, se puede el rendimiento del multi-clasificador en un diagrama de caja después de ejecutar el modelo cien veces.

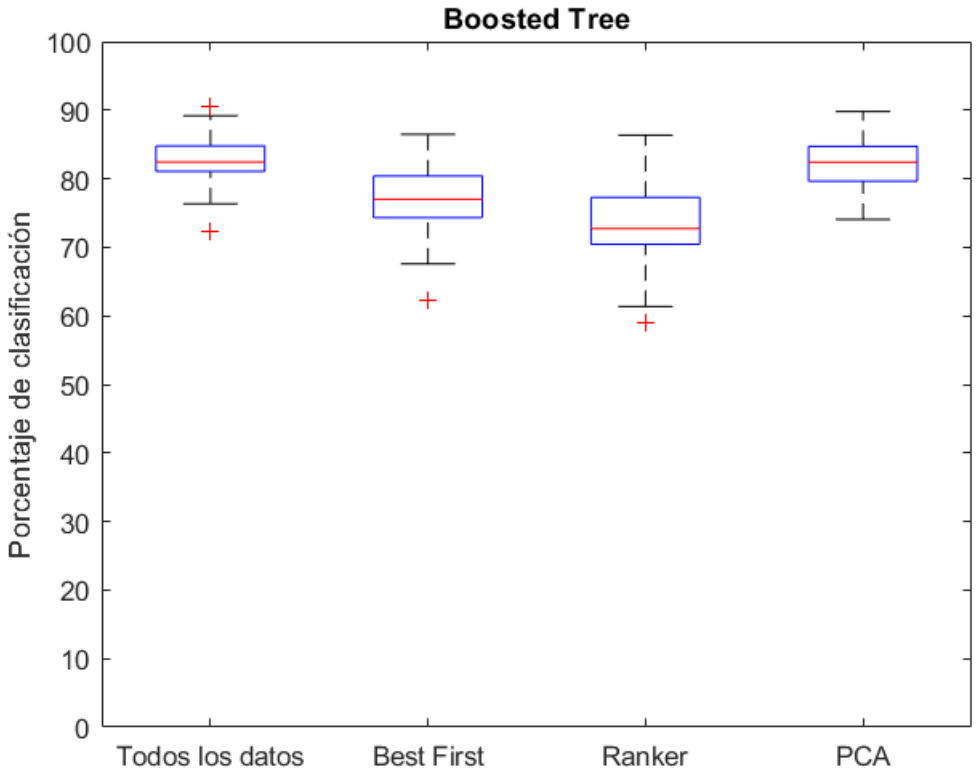


Figura 19. Cuartiles del multi-clasificador Boosted_Tree con los diferentes grupos de datos

En la Figura 19 se observa que presenta mayor estabilidad el multi-clasificador Boosted_Tree si se trabaja con todos los datos. A continuación, se detalla la matriz de confusión del multi-clasificador *Boosted_Tree* con los diferentes grupos de datos.

Tabla 35

Rendimiento del multi-clasificador Boosted_Tree, en términos de sensibilidad (Se), especificidad (Sp), porcentaje de clasificación (Cp) con un porcentaje de 76% de entrenamiento.

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
Boosted_Tree _todos	**99.51**	89.19	**98.66**
Boosted_Tree_Cfs-BestFirst	86.49	**100**	93.24
Boosted_Tree_Ranker	65.22	86.96	76.09
Boosted_Tree_PCA	76.36	98.18	87.27

La Tabla 35, muestra que el método de mezcla de clasificadores basado en *Boosted_Tree*. Se tiene que utilizando los datos seleccionados por el método *BestFirst*, la especificidad tiene un

porcentaje del 100%, lo cual es la prioridad en nuestro caso de estudio, que permite detectar a los alumnos que sea propensos a tener un bajo rendimiento. Por otro lado, el porcentaje de clasificación con todo el conjunto de atributos mejora a un 98.66%, pero reduce en especificidad.

La Tabla 36 se detalla la matriz de confusión al aplicar a todos los datos usando el multi-clasificador Boosted_Tree, con las clases desbalanceadas aplicando el *multi-clasificador Boosted_Tree*.

Tabla 36

Matriz de confusión aplicando el multi-clasificador Boosted_Tree con todo el conjunto de datos

Clase	Negativo	Positivo
Negativo	66	8
Positivo	4	816

En la Tabla 37 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método *BestFirst* y las clases balanceadas aplicando el *multi-clasificador Boosted_Tree*.

Tabla 37

Matriz de confusión aplicando el multi-clasificador Boosted_Tree con el conjunto de datos Cfs-BestFirst

Clase	Negativo	Positivo
Negativo	37	0
Positivo	5	32

En la Tabla 38 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método Ranker y las clases balanceadas, aplicando el *multi-clasificador Boosted_Tree*.

Tabla 38

Matriz de confusión aplicando el multi-clasificador Boosted_Tree con el conjunto de datos Ranker

Clase	Negativo	Positivo
Negativo	20	3
Positivo	8	15

En la Tabla 39 se observa la matriz de confusión con el grupo de datos seleccionado mediante el método PCA y las clases balanceadas.

Tabla 39

Matriz de confusión aplicando el multi-clasificador Boosted_Tree con el conjunto de datos PCA

Clase	Negativo	Positivo
Negativo	54	1
Positivo	13	42

4.4.2. Interfaz desarrollada

Este apartado contiene la descripción detallada de la interfaz creada para la interacción entre el sistema de búsqueda de patrones académicos y el usuario, el cual, permite obtener un mejor entendimiento de la interfaz y dar un mejor uso y rendimiento de la misma. Las características principales de esta interfaz son:

- Carga de la base de datos (encuesta Unidad Educativa Ibarra)
- Entrenamiento de los clasificadores
  Árboles de decisión (Tree),
  Máquinas de soporte vectorial SVM Cuadrático
  kNN Ponderado
- Entrenamiento de multi-clasificadores
  Bagged Tree
  Boosted Tree
- Visualización de la información de la base de datos
- Selección del porcentaje de entrenamiento y prueba
- Visualización de medias de desempeño de los entrenadores.
- Matriz de confusión
- Visualización de datos,
- Interpretación y puesta en contexto

En la Figura 20 se muestra la interfaz propuesta, donde se visualiza las diferentes secciones: información de los datos y visualización, configuración para los entrenadores, y por último visualización de las medidas de desempeño (resultados). Sin embargo, se puede observar que las secciones están deshabilitadas hasta que el usuario cargue la base de datos.

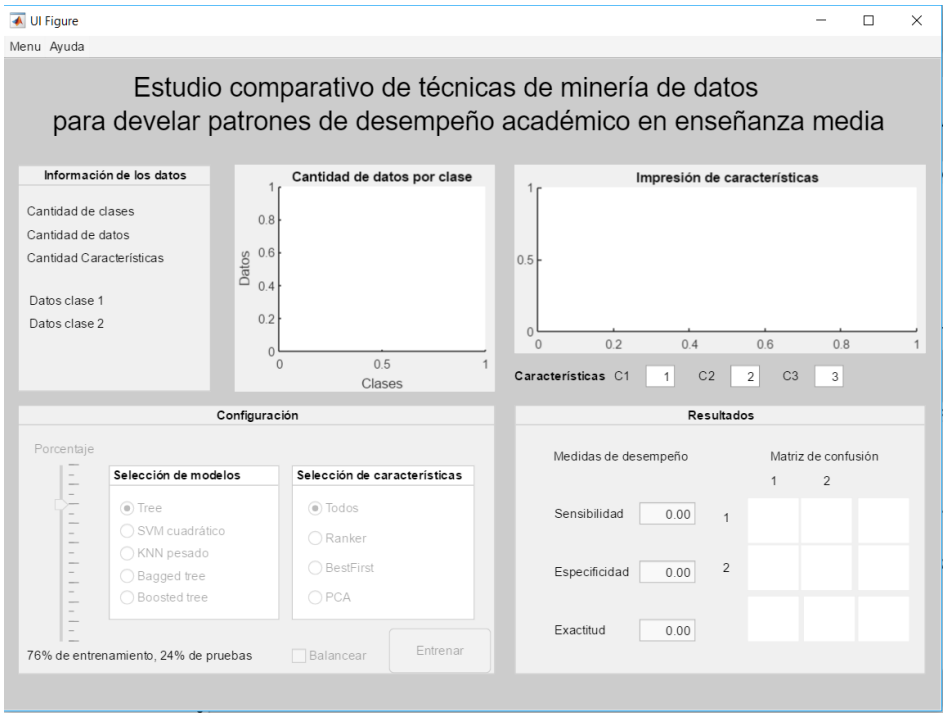


Figura 20. Interfaz propuesta

Para iniciar el análisis, se debe seleccionar la base de datos que vamos a trabajar, para ello se pulsa en “Menú” y se selecciona “Cargar Archivo”, tal como se detalla en la ***Figura 20***.

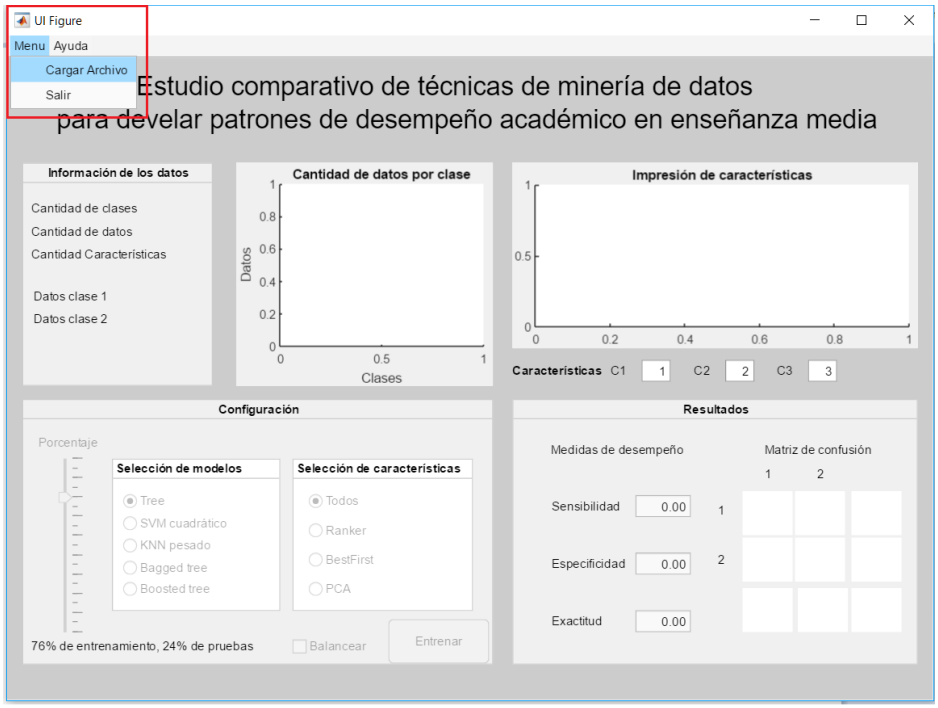


Figura 21. Menú para cargar la base de datos

Una vez cargado el archivo se habilitan las opciones de configuración y resultados como se muestra en la ***Figura 21***.

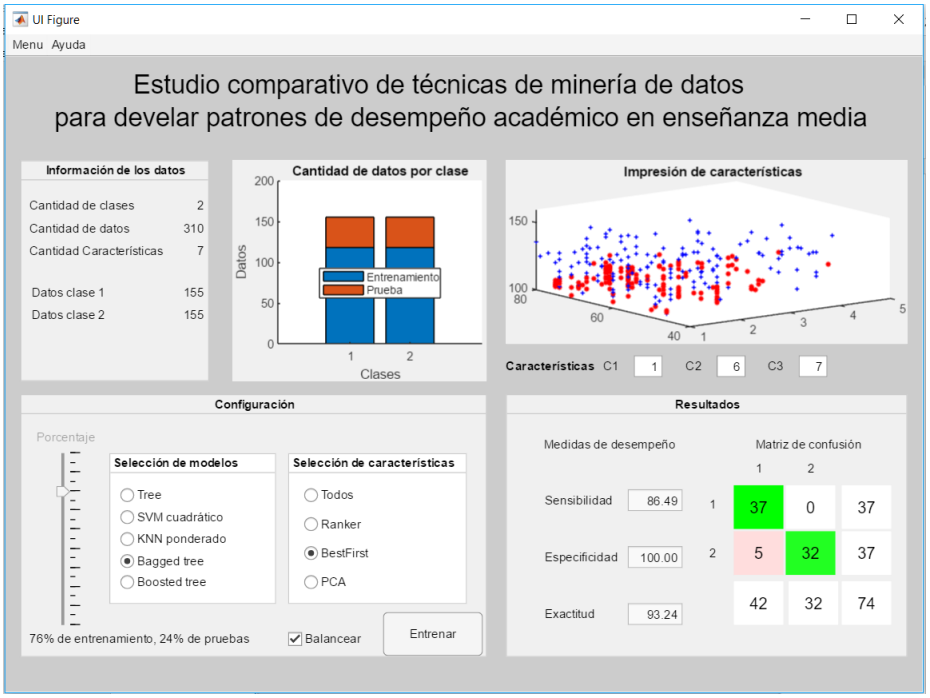


Figura 22. Interfaz con la base de datos seleccionada

Figura 23. Detalle de las secciones en la interfaz

La interfaz principal está dividida en seis secciones. En la sección de información datos (sección 1), se puede visualizar la información general de la base de datos, la cual contiene:

- Cantidad de clases
- Cantidad de datos
- Cantidad de características
- Cantidad de datos por cada clase

A continuación, se puede apreciar la sección información de la interfaz ***Figura 24***.

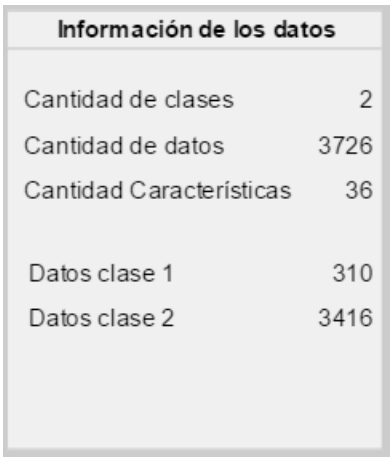


Figura 24. Sección de información de la interfaz

En la ***Figura 25*** aprecia un gráfico de barras, con la cantidad de registros por cada clase. Se muestra el porcentaje de entrenamiento y prueba de los datos, además, se muestra los datos balanceados o no balanceados.

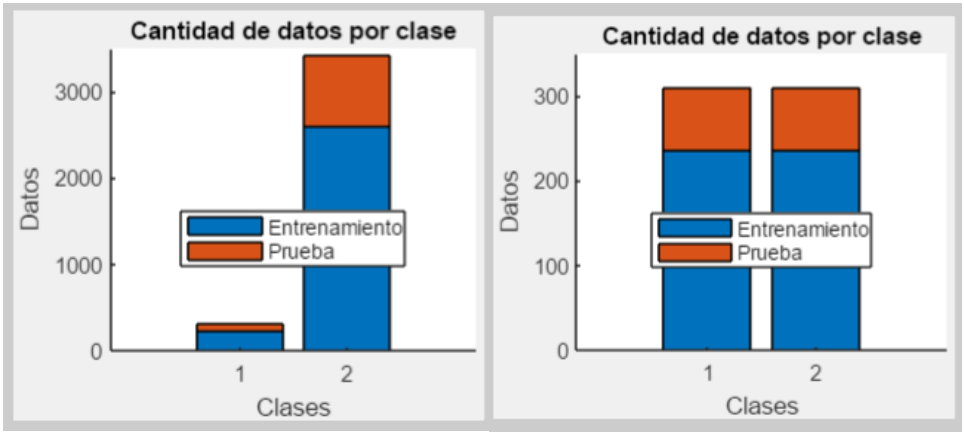


Figura 25. Visualización base de datos desbalanceados y balanceados con el porcentaje de entrenamiento y prueba

Mediante un cuadro de selección (Balancear), se puede elegir si los datos para el entrenamiento se deben balancear con respecto a la clase minoritaria, o se entrena con los datos desbalanceados. En la ***Figura 26*** se puede observar las dos opciones.

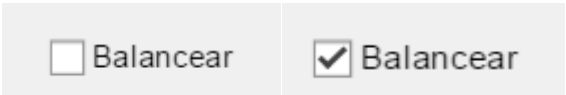


Figura 26. Selección de balanceo de datos

En la Figura 27, se visualiza mediante un diagrama de barras los datos de entrenamiento y prueba para el clasificador de forma balaceada.

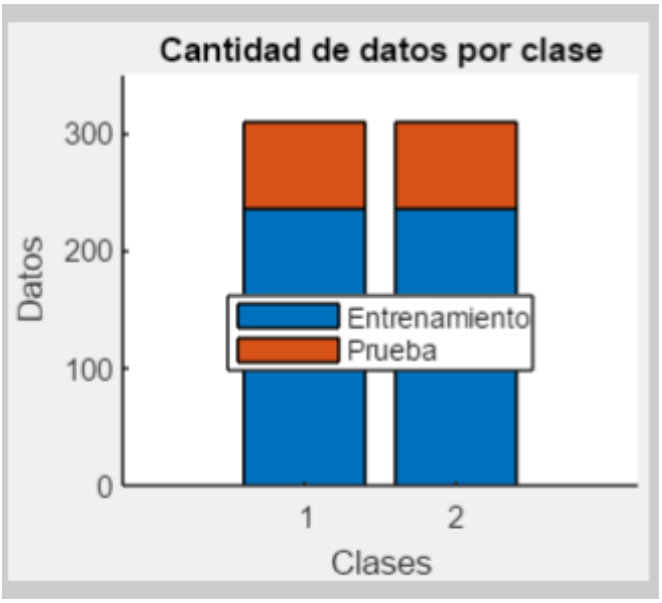


Figura 27. Cantidad de información a usar para entrenamiento y prueba del clasificador

Otra opción de configuración (sección 3) de la Figura 23, es la selección del porcentaje de entrenamiento y prueba, mediante el uso de las barras de movimiento mostrada en la Figura 28, por defecto se tiene el 76% de entrenamiento y 24% de prueba, cabe recalcar que no puede elegirse el 100% de entrenamiento ni el 0% tanto de entrenamiento como de prueba.

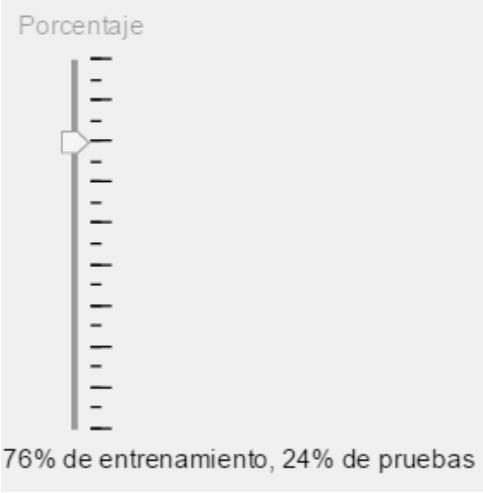


Figura 28. Cantidad de información a usar para entrenamiento y prueba del clasificador

En la sección de configuración existe la opción de elegir entre diferentes modelos de aprendizaje automático programados, así también, elegir si se desea trabajar con todos los datos o los que están seleccionados con los diferentes métodos de selección de características, como se puede visualizar en la ***Figura 29***.

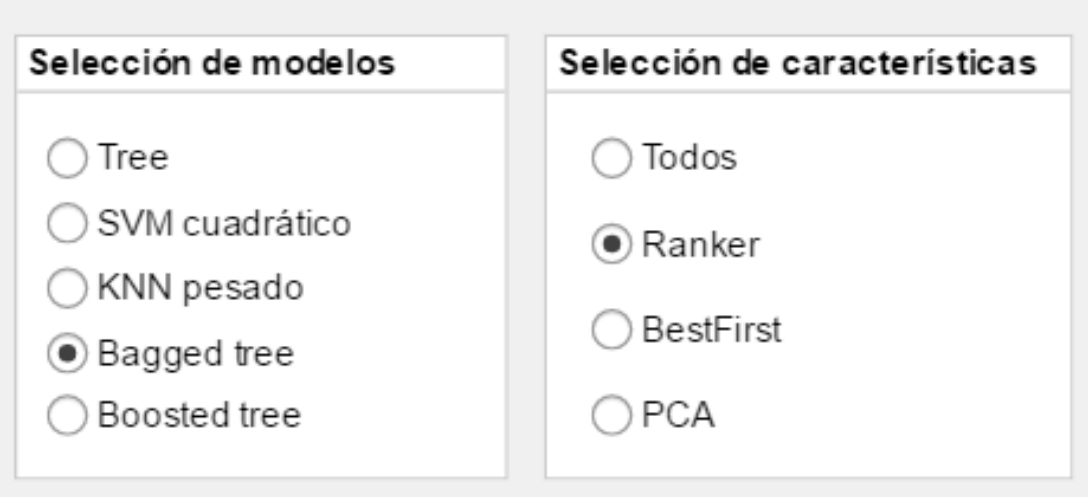


Figura 29. Seleción de métodos y caracteristicas para entrenar

Una vez terminado de configurar las opciones de entrenamiento, se procede a presionar el botón "Entrenar", ver ***Figura 30***.

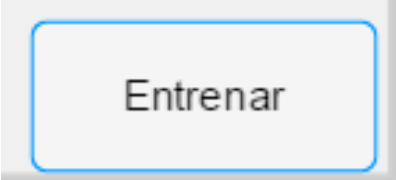


Figura 30. Botón entrenar

La evaluación del método entrenado se puede visualizar en la sección resultados. En esta sección se tienen las métricas de desempeño sensibilidad, especificidad, porcentaje de clasificación, y la matriz de confusión, ver ***Figura 31***.

Figura 31. Resultados del clasificador

Por último, para comprender los datos de una forma dinámica se puede visualizar en "Impresión de características", con la ayuda de un diagrama de dispersión, para ello se selecciona la combinación de máximo tres características que se dese representar gráficamente y observar su relación ***Figura 32***.

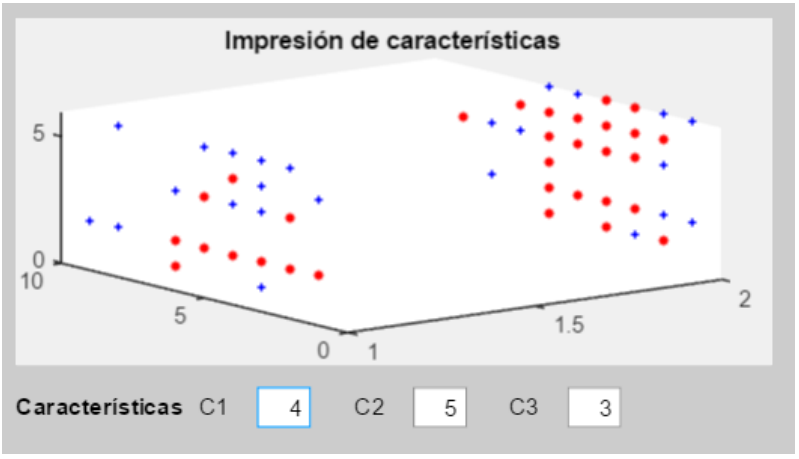


Figura 32. Resultados del clasificador

Resultados de la interfaz

- El número de clases, atributos y cantidad de datos total de la base de datos
- Visualización de porcentajes de entrenamiento y prueba
- Métricas de desempeño de los clasificadores una vez entrenados

- Matriz de confusión.
- Diagrama de dispersión

La interfaz permite:

- Visualización de la cantidad de porcentaje asignados para el proceso de entrenamiento y prueba
- Visualizar las métricas de desempeño de los clasificadores una vez entrenados.
- Visualizar la matriz de confusión.
- Visualizar de forma dinámica la relación entre máximo tres atributos (diagrama de dispersión).

Uso de la interfaz:

- Iniciamos con la selección de la base de datos a usar en el entrenamiento de los clasificadores. Para ellos vamos a la sección "menú" y luego "Carga archivo" ***Figura 21***, la base de datos debe introducirse inicialmente un Excel extensión ".csv" este archivo contiene la encuesta y las notas por cada materia del estudiante ya categorizado.
- En la Figura 24 se mostrarán la cantidad de atributos por clase la cantidad de datos, así también las clases.
- Se seleccionará el porcentaje, tanto de entrenamiento como de prueba, por default se tiene seleccionado 76% para entrenar y 24% para pruebas.
- Mediante un cuadro de selección, se elegirá si desea balancear o no los datos
- Se pulsará el botón "entrenar"
- Se visualizará las meticas de desempeño del modelo entrenado
- Los resultados se mostrarán en una matriz de confusión

4.5. Discusión

Este apartado presenta los resultados del desarrollo de la propuesta del trabajo de investigación según la metodología escogida sobre las bases de datos facilitada por la Unidad educativa Nacional Ibarra.

4.5.1. Resultados y análisis base de datos

Una vez realizado la categorización de los datos, el siguiente paso, es la reducción y selección de características, continuando con la clasificación, resultados e interpretación en contexto.

Para la selección de características, se aplica experimentos con varias técnicas, la cuales se realizan con la ayuda de los algoritmos implementados en Weka. Se utiliza las técnicas CFS con el método de búsqueda *Bestfirst, CorrelaciónAtributosEval* y el análisis de *componentes principales* (*PCA*) con el método de búsqueda *Ranker*. Se obtuvieron los resultados mostrados en la en la Tabla 40.

Tabla 40

Resultados de selección de características sobre base de datos

Técnica	Atributos seleccionados	Observación
CFS-BestFirst	6	Se reducen los atributos a 6 de una totalidad de 36, con un mérito de 0,602
CorrelaciónAtributosEval-Ranker	6	Se reducen los atributos de 36 a 6. Este método despliega un ranking de atributos que mejor predicen a la clase.
PCA-Ranker	7	Se reducen los atributos de 36 a 7. Aclarando que los atributos seleccionados son 7 compontes principales que en este caso han utilizado todos los atributos para las componentes logrando explicar con ellos un 89% los valores de la clase.

Al completar la selección de características, se realizaron los experimentos con los clasificadores y multi-clasificadores seleccionados para el presente estudio. Los cuales se analizaron con métricas de evaluación como son: porcentaje de clasificación, sensibilidad y especificidad. Los resultados se describen en la Tabla 41.

Tabla 41

Rendimiento de los clasificadores y multi-clasificadores sobre la base de datos, en términos de sensibilidad (Se), especificidad (Sp), porcentaje de clasificación (Cp) con un porcentaje de 76% de entrenamiento.

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
Bagged_Tree_Cjto_todos	**100**	72.97	97.76
Boosted_Tree _Cjto_todos	99.51	89.19	98.66
Árbol_de_decisión _todos	99.63	77.03	97.76

CONTINÚA

SVM_Cuadratico _Cjto_todos	74.32	**98.65**	86.49
kNN_ Ponderado_Cjto_todos	99.88	77.03	**97.99**
Bagged_Tree _ Cfs-BestFirst	86.49	**100**	**93.24**
Boosted_Tree_Cfs-BestFirst	86.49	**100**	**93.24**
Árbol_de_decisión_Cfs-BestFirst	83.78	97.30	90.54
SVM_Cuadratico_Cfs-BestFirst	**89.19**	97.30	**93.24**
kNN_ Ponderado_ Cfs-BestFirst	81.08	91.89	86.49
Bagged_Tree _Ranker	78.26	**100**	**89.13**
Boosted_Tree_Ranker	65.22	86.96	76.09
Árbol_de_decisión_Ranker	73.91	100	86.96
SVM_Cuadratico_Ranker	73.91	95.65	84.78
kNN_ Ponderado_ Ranker	**82.61**	91.30	86.96
Bagged_Tree _PCA	76.36	98.18	87.27
Boosted_Tree _PCA	76.36	98.18	87.27
Árbol_de_decisión_PCA	76.36	98.18	87.27
SVM_Cuadratico_PCA	**81.82**	98.18	**90.00**
kNN_ Ponderado_PCA	60.00	**96.36**	78.18

En la Tabla 41 se puede observar, un resumen de resultados de los clasificadores y multi-clasificadores implementados en este estudio en porcentajes de medidas de despeño. En lo que se refiere al porcentaje de sensibilidad se tiene 100 %, implementado los multi-clasificadores *Bagged_*Tree con todos los atributos, Por otro lado, se tiene que en porcentaje de especificidad que es la prioridad en este estudio detectar alumnos que sean propensos a quedarse a supletorio, se tiene 100% con las dos mezclas de clasificadores *Bagged* y *Boosted* y con el método de selección de atributos *BestFirst* determinó un mejor desempeño, así también se obtiene el 100 % de especificidad con el multi-clasificador *Bagged_Tree* y las características seleccionadas con el método *Ranker*. En lo que se refiere al porcentaje de clasificación se tiene 97.99 % con el método *kNN_Ponderado* y todos los datos.

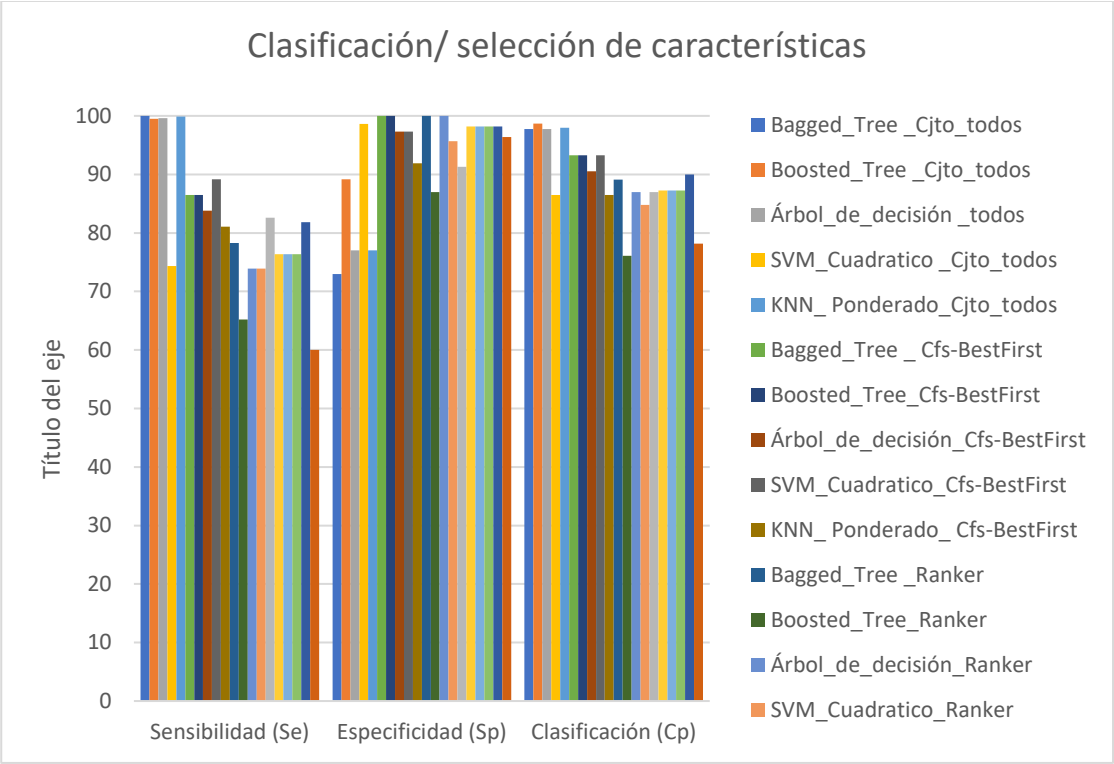


Figura 33. Medidas de desempeño de los clasificadores y multi-clasificadores sobre la base de datos

En la Figura 33, se puede observar las medidas de desempeño de los clasificadores y multi-clasificadores, como es sensibilidad, especificidad y porcentaje de clasificación.

En la Tabla 42, se enlista los resultados de los algoritmos implementados en este estudio. Y se realiza el análisis de forma individual con los diferentes grupos de estudio.

Tabla 42

Resultados rendimiento de los clasificadores y multi-clasificadores para los conjuntos de datos con selección de características

Algoritmo	Sensibilidad (Se)	Especificidad (Sp)	Clasificación (Cp)
Árbol_de_decisión_CfsBestFirst	83.78	97.30	90.54
Árbol_de_decisión _Ranker	73.91	100	86.96
Árbol_de_decisión_PCA	76.36	98.18	87.27
SVM_Cuadrático_CfsBestFirst	89.19	97.30	93.24

CONTINÚA

SVM_Cuadrático_Ranker	73.91	95.65	84.78
SVM_Cuadrático_PCA	81.82	98.18	90.00
kNN_Ponderado_Cfs_BestFirst	81.08	91.89	86.49
kNN_Ponderado_Ranker	82.61	91.30	86.96
kNN_Ponderado_PCA	60.00	96.36	78.18
Bagged_Tree _ Cfs_BestFirst	86.49	100	93.24
Bagged_Tree_Ranker	78.26	100	89.13
Bagged_Tree_PCA	76.36	98.18	87.27
Boosted_Tree _Cfs_BestFirst	86.49	100	93.24
Boosted_Tree_Ranker	65.22	86.96	76.09
Boosted_Tree_PCA	76.36	98.18	87.27

En la Tabla 42 a nivel individual para el algoritmo *Árbol_de_decisión* determino un mejor porcentaje de clasificación con un porcentaje de 90.54 % con el método de selección de atributos *BestFirts*. En lo que se refiere al método *SVM_Cuadrático* el método de selección de atributos *CfsBestFirst* determinó un mejor desempeño del clasificador en porcentaje de clasificación 93.24%, mientras que el porcentaje de clasificación con el método *kNN_Ponderado* se obtuvo (86.96%) con las características seleccionadas con *Ranker*. Con el multi-clasificador *Bagged_Tree* se obtuvo el 93.24 % en porcentaje de clasificación con el método de selección de características *BestFirst* y el 100% en especificidad, También se logró el 100% de especificidad con los datos obtenidos con el método *Ranker*. Por otro lado, el Multi-clasificador *Boosted* presentó mejores porcentajes en todas las medidas de desempeño con los datos seleccionados con *BestFirst*, sensibilidad 86.49%, especificidad 100% y porcentaje de clasificación 93.24%.

CAPÍTULO V

5. CONCLUSIONES Y TRABAJO FUTURO

5.1. Conclusiones

- Los resultados permiten, por una parte, observar la contribución al propósito de brindar más información respecto de la problemática objetivo del proyecto. Dicha información podrá orientar para toma de decisiones y acciones por parte las autoridades, para reducir la cantidad de alumnos con problemas de aprobación de la materia.

- Por otra parte, se pretende incrementar el conocimiento sobre las distintas técnicas de preprocesamiento de datos, dada la importancia que tiene esta etapa en la aplicación de minería de datos o en cualquier otro tipo de análisis de información.

- Los resultados obtenidos con los diferentes métodos de clasificación y los multi-clasificadores que se implementó en esta investigación, indican que son capaces de generar modelos con el respaldo teórico, basado en los datos recopilados por el DECE de la Unidad Educativa Ibarra.

- Los patrones de desempeño académico descubiertos para los alumnos de los terceros de bachillerato del periodo 2017-2018 son, edad que tiene relación si el alumno ha perdido un año escolar, si vive o no con su madres, si los padres viven juntos, género, su ritmo de aprendizaje, el nivel de educación de los padres, si ha tendido deseos de abandono, entre otros aspectos se encuentran asociados en el desempeño académico por encima de la media como por debajo de ella.

5.2. Trabajo futuro

Sobre la base de los datos, en esta investigación y luego de llegar a las conclusiones expuestas, se recomiendan los siguientes temas para investigación adicional.

- En esta investigación se selecciona los atributos con tres métodos, es importante seguir explorando otros métodos alternativos o metodologías que combinen los utilizados.

- Esta investigación se realizó la implementación de tres clasificadores que según la literatura y las pruebas realizadas, son los más utilizados en el ámbito de estudio, tomando en cuenta que nuestra prioridad es un gran grado de sensibilidad. Se plantea explorar la opción de aplicar otros algoritmos clasificadores a la base de datos.

- Finalmente, la implementación de nuevas mezclas de clasificadores para mejorar el índice de predicción, siempre que muestre resultados estables y aceptables para la bases de datos. Esto quedaría planteado para trabajos futuros.

- El sistema debería ser actualizado cada periodo académico, en vista que se vive un mundo cambiante, y variables que no se han tomado en cuenta para esta investigación podría ser útil para mejorar la predicción, y al actuar en función de las nuevas necesidades, se iría adaptando a las nuevas realidades de los estudiantes. En esta investigación se realiza usando las base de datos de los terceros de bachillerato del periodo 2017 - 2018, sería recomendable implementar este procesamiento de datos en diferentes niveles de la Unidad Educativa.

- Entre las dificultades presentadas en el desarrollo de la investigación, están la mala calidad de las bases de datos, en vista que no todos los estudiantes de este periodo habían llenado la encuesta (alumnos faltantes), además de la falta de datos y especificidad en las preguntas,

se tuvieron que descartar ciertos atributos por la imposibilidad de obtener sus valores, y que de alguna manera, podrían influir en el descubrimiento de los patrones objeto de este estudio.

Bibliografía

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37-66. https://doi.org/10.1007/BF00153759

Bauer, E., Kohavi, R., Chan, P., Stolfo, S., & Wolpert, D. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, *36*(August), 105-139. https://doi.org/10.1023/a:1007515423169

BestFirst. (2019). Recuperado 30 de mayo de 2019, de http://weka.sourceforge.net/doc.dev/weka/attributeSelection/BestFirst.html

Bhardwaj, A. (2015). *Data Preprocessing Techniques for Data Mining*. Recuperado de http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

Bolaños Ramírez, H. F. (2017). *Modelo multiclasificador aplicando minería de datos para el diagnóstico médico utilizando datos abiertos*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/10.1007/BF00058655

Brighton, H., & Mellish, C. (2002). Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Mining and Knowledge Discovery*, *6*(2), 153-172. https://doi.org/10.1023/A:1014043630878

Brownlee, J. (2016, julio 12). How to Perform Feature Selection With Machine Learning Data in Weka. Recuperado 30 de mayo de 2019, de Machine Learning Mastery website: https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

Campillo, J. P., Ibáñez, P. C., & Vargas, J. M. (2015). *la predicción del fracaso empresarial mediante modelos basados en la técnica boosted regression trees (brt)*. 24.

Chetty, N., Vaisla, K. S., & Sudarsan, S. D. (2015). Role of attributes selection in classification of Chronic Kidney Disease patients. *2015 International Conference on Computing, Communication and Security (ICCCS)*, 1-6. https://doi.org/10.1109/CCCS.2015.7374193

Córdoba Cely, C., & Alatriste Martínez, Y. (2012). Hacia una taxonomía de investigación entre Visualización de Información y Diseño. *No Solo Usabilidad*, (11). Recuperado de http://www.nosolousabilidad.com/articulos/taxonomia_visualizacion.htm

CorrelationAttributeEval. (2019). Recuperado 30 de mayo de 2019, de http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Data Mining: Practical Machine Learning Tools and Techniques. (2011). https://doi.org/10.1016/C2009-0-19715-5

Der, G., & Everitt, B. (2014). A Handbook of Statistical Graphics Using SAS ODS. Recuperado 7 de junio de 2019, de CRC Press website: https://www.crcpress.com/A-Handbook-of-Statistical-Graphics-Using-SAS-ODS/Der-Everitt/p/book/9781466599031

Dietterich, T. G. (1997). Machine Learning Research: Four Current Directions. *AI Magazine*, *18*, 97-136.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and ranomization. *Machine Learning*, *40*(2), 139-157. https://doi.org/10.1023/A:1007607513941

Domingo, S., Timaran Pereira, R., Hidalgo Troya, A., Caicedo Zambrano, J., Hernández Arteaga, I., & Alvarado Pérez, J. (2015, julio 29). *Descubrimiento de Patrones de Desempeño Académico en la Competencia de Lectura Crítica*.

Espinoza, J. R. (2017). *Desarrollo de algoritmos para la extracción de características y la Clasificación automática de la obesidad en registros médicos electrónicos con un enfoque jerárquico multiclase*.

Fayyad, U. M. (Ed.). (1996). *Advances in knowledge discovery and data mining*. Menlo Park, Calif.: AAAI Press [u.a.].

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, *39*, 27–34.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189-1232.

González, C. J. A. (2019). *Introducción a los sistemas Multiclasificadores*. 28.

González, L. H. R. (2015). *Sistema de clasificación y reconocimiento de imágenes*. 181.

Gutiérrez García, J. A. (2016). *Comenzando con Weka: Filtrado y selección de subconjuntos de atributos basada en su relevancia descriptiva para la clase*. Málaga.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18. https://doi.org/10.1145/1656274.1656278

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco, {CA}, {USA}: Morgan Kaufmann Publishers Inc.

Hernández, O. J., Ramírez, Q. M., & Ferri, FC. (2004). *Introducción a la Minería de Datos* (Pearson Prentice Hall, Ed.). Madrid.

Herrera, F. (2006). *Técnicas de reducción de datos en KDD. El uso de Algoritmos Evolutivos para la Selección de Instancias*. 19.

Johnstone, I. M., & Paul, D. (2018). PCA in High Dimensions: An Orientation. *Proceedings of the IEEE*, *106*(8), 1277-1292. https://doi.org/10.1109/JPROC.2018.2846730

Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in Visual Data Analysis. *Tenth International Conference on Information Visualisation (IV'06)*, 9-16. https://doi.org/10.1109/IV.2006.31

Kennard, R. W., & Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics*, *11*(1), 137. https://doi.org/10.2307/1266770

Khube, M., Islam, Z., & Ashad, M. (2017). *Analyzing Performance of Classification Techniques in Detecting Epileptic Seizure*. 396-408.

Kibler, D., & Aha, D. W. (1987). Learning Representative Exemplars of Concepts: An Initial Case Study. En P. Langley (Ed.), *Proceedings of the Fourth International Workshop on MACHINE LEARNING* (pp. 24-30). https://doi.org/10.1016/B978-0-934613-41-5.50006-4

Korf, R. E. (1993). Linear-space best-first search. *Artificial Intelligence*, *62*(1), 41-78. https://doi.org/10.1016/0004-3702(93)90045-D

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Liu, H., & Motoda, H. (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Springer Science & Business Media.

Liu, H., & Motoda, H. (2002). On Issues of Instance Selection. *Data Mining and Knowledge Discovery*, *6*(2), 115-130. https://doi.org/10.1023/A:1014056429969

Liu, H., & Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media.

Luan, J. (2002). Data Mining and Its Applications in Higher Education. *New Directions for Institutional Research*, *2002*(113), 17-36. https://doi.org/10.1002/ir.35

Mazza, R. (2009). *Introducción a la visualización de la información, Texto original*. Recuperado de https://www.springer.com/la/book/9781848002180

Mhetre, V., & Nagar, M. (2017). Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA. *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, 475-479. https://doi.org/10.1109/ICCMC.2017.8282735

Minnaard, C., Condesse, V., & Rabino, C. (2010). *Los Gráficos de Caja: Un Recurso Innovador*.

Moran. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 243-251.

Moreno, R., & Vayá, E. (2004). Econometría espacial: Nuevas técnicas para el análisis regional. Una aplicación a las regiones europeas. Recuperado 7 de junio de 2019, de https://ebuah.uah.es/dspace/bitstream/handle/10017/32600/econometria_moreno_IR_2002_N1.pdf?sequence=1&isAllowed=y

Ojeda, I. L. R. (2007). *PROBABILIDAD Y ESTADÍSTICA BÁSICA PARA INGENIEROS*. 336.

Ornella, L. A. (2010). Códigos Correctores de Error en Problemas de Clasificación Multiclase de Datos de Marcadores Moleculares. Recuperado 5 de junio de 2019, de https://docplayer.es/2970314-Tesis-doctoral-codigos-correctores-de-error-en-problemas-de-clasificacion-multiclase-de-datos-de-marcadores-moleculares-leonardo-alfredo-ornella.html

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Pereira, R. T., & de Pasto, S. J. (2013). *Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos*. 5.

Phyu, T. N. (2009). Survey of Classification Techniques in Data Mining. *International MUlticonference of Engineers and Computer Scientists*, *I*, 18-20.

Pulgarín, C. A. T. (2012). *Clasificación basada en la estimación de Parzen en espacios generalizados de disimilitudes*. 195.

Purchase, H. C., Andrienko, N., Jankun-Kelly, T. J., & Ward, M. (2008). Theoretical Foundations of Information Visualization. En A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization* (Vol. 4950, pp. 46-64). https://doi.org/10.1007/978-3-540-70956-5_3

Ramirez-Anormaliza, R., Guevara-Viejo, F., Regnault, M. D., Pena-Holguin, R., Farias-Lema, R., Bravo-Duarte, F., … Castelo-Gonzalez, J. (2017). *Análisis Multivariante: Teoría y práctica de las principales técnicas*. 212.

Rodallegas Ramos, E., Torres González, A., Gaona Couto, B. B., Gastelloú Hernández, E., Lezama Morale, R. As., & Valero Orea, S. (2010). Recursos digitales para la educación y la cultura. En U. T. Metropolitana (Ed.), *Recursos Digitales*. Cádiz, España.

Rodriguez Rojas, O. (2009). *Análisis de componentes principales* (p. 22). p. 22. San José, {COSTA RICA}.

Rokach, L., & Maimon, O. (2015). *Data mining with decision trees: Theory and applications* (Second edition). Hackensack, New Jersey: World Scientific.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323-2326. https://doi.org/10.1126/science.290.5500.2323

Ruggieri, S. (2002). Efficient C4.5 [classification algorithm]. *IEEE Transactions on Knowledge and Data Engineering*, *14*(2), 438-444. https://doi.org/10.1109/69.991727

Saptoro, A., Tadé, M. O., & Vuthaluru, H. (2012). A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models. *Chemical Product and Process Modeling*, *7*(1). https://doi.org/10.1515/1934-2659.1645

Segal, M. R. (2004). *Machine Learning Benchmarks and Random Forest Regression*. 15.

Solarte, G., & Ocampo, C. (2009). *Técnicas de clasificación y análisis de representación del conocimiento para problemas de diagnóstico*.

Spositto, O. M., Etcheverry, M. E., Ryckeboer, H. L., & Bossero, J. (2009). *Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil*. 5.

Suárez, E. J. C. (2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. 25.

Terrádez-Gurrea, M. (2006). Analisis De Componentes Principales. *Revista chilena de obstetricia y ginecolog*, *71*(1), 1-11. https://doi.org/10.4067/S0717-75262006000100004

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*. https://doi.org/10.16925/9789587600490

Toca, C. E. S. (2016). *Predicción de mapas de contactos de proteínas mediante multiclasificadores*.

Valentini, G., & Masulli, F. (2002). Ensembles of Learning Machines. En M. Marinaro & R. Tagliaferri (Eds.), *Neural Nets* (Vol. 2486, pp. 3-20). https://doi.org/10.1007/3-540-45808-5_1

Vásquez, M., & Ramírez, G. (2012). *Análisis de Datos*. Recuperado de http://saber.ucv.ve/bitstream/123456789/7374/1/Capitulo1AnadatPregrado_1.pdf

Vesonder, G., & Wright, J. (2003). *Data Quality through Knowledge Engineering*. 6.

Walpole, R., & Myers, R. (2012). *Probabilidad y Estadistica para Ingenieros*.

Wu, & Kumar. (2009). *The Top Ten Algorithms in Data Mining*.

Xu, Z., King, I., Lyu, T., & Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization, Neural Networks. *IEEE Transactions on Knowledgeand Data Engineering*, 1033–1047.

Zaki, M. J., & Meira, Jr, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms* (1.ª ed.). https://doi.org/10.1017/CBO9780511810114