Universidad Autónoma de San Luis Potosí
Facultad de Ingeniería
Centro de Investigación y Estudios de Posgrado

Predicción de la Deserción Escolar de Estudiantes de Ingeniería mediante Minería de Datos

T E S I S

Que para obtener el grado de:

Maestra en Ingeniería de la Computación

Presenta:

Lic. en Fís. Ana Estela Pérez Mejía

Asesor:

Dr. José Ignacio Núñez Varela

San Luis Potosí, S. L. P. Agosto de 2023

UASLP
Universidad Autónoma de San Luis Potosí

FACULTAD DE INGENIERÍA

18 de mayo de 2023

LIC. en FÍS. ANA ESTELA PÉREZ MEJÍA
P R E S E N T E.

En atención a su solicitud de Temario, presentada por el **Dr. José Ignacio Núñez Varela,** Asesor de la Tesis que desarrollará Usted, con el objeto de obtener el Grado de **Maestra en Ingeniería de la Computación**, me es grato comunicarle que en la sesión del H. Consejo Técnico Consultivo celebrada el día 18 de mayo del presente, fue aprobado el Temario propuesto:

TEMARIO:

"Predicción de la Deserción Escolar de Estudiantes de Ingeniería mediante Minería de Datos"

1. Introducción.
2. Minería de Datos Educativa.
3. Preprocesamiento de Datos Académicos.
4. Predicción de la Deserción Escolar.
5. Conclusiones.
   Referencias.

"MODOS ET CUNCTARUM RERUM MENSURAS AUDEBO"

A T E N T A M E N T E

DR. EMILIO JORGE GONZÁLEZ GALVÁN
DIRECTOR.

UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSI
FACULTAD DE INGENIERIA
DIRECCION

Copia. Archivo.
*etn.

ANIVERSARIO DE LA AUTONOMÍA UASLP 2023

www.uaslp.mx

Av. Manuel Nava 8
Zona Universitaria • CP 78290
San Luis Potosí, S.L.P.
tel. (444) 826 2330 al39
fax (444) 826 2336

Resumen

La deserción escolar es un tema de particular interés para las coordinaciones académicas de las instituciones educativas. Poder modelar el rendimiento académico de los estudiantes desde etapas tempranas del programa sería una herramienta de gran utilidad que podría evitar futuras deserciones o prevenir rezagos estudiantiles. La interacción del estudiante con el entorno académico genera grandes cantidades de información que puede ser usada para extraer conocimiento acerca de su comportamiento, desempeño, etc. La minería de datos es una herramienta que permite analizar extensas bases de datos y obtener información oculta para descubrir patrones y/o relaciones entre sus elementos. La aplicación de la minería de datos al entorno educativo es conocida como minería de datos educativa, y se ha vuelto muy popular en la última década, consiste en aplicar las técnicas básicas de la minería de datos para estudiar sistemas educativos. El objetivo de este trabajo es aplicar técnicas de minería de datos a los historiales académicos de los estudiantes del Área de Ciencias de la Computación, de la Facultad de Ingeniería de la UASLP, para construir modelos predictivos que permitan detectar a estudiantes con alto riesgo de deserción. Los resultados obtenidos demostraron que los datos procedentes de los exámenes del proceso de admisión no proveen información suficiente para identificar a los estudiantes en situación de riesgo de deserción. También, a partir de las predicciones realizadas, se logró reconocer que el rendimiento de los estudiantes, utilizado por la Facultad de Ingeniería como métrica para las inscripciones, es el mejor parámetro para predecir si los estudiantes terminarán la carrera o no.

A mis padres

Agradecimientos

Una vez más y siempre, mi agradecimiento principal es para mis padres, que han estado conmigo para apoyarme a lo largo de toda mi vida. Cada logro mío es de ellos también, porque sin su esfuerzo y su ejemplo nada hubiera sido posible. A mis hermanos, por estar siempre presentes y por ser un respaldo en mi vida.

Agradezco de forma muy especial a mi asesor, el Dr. José Ignacio Núñez Varela, por haberme permitido realizar este trabajo de investigación bajo su supervisión. Le agradezco por todo el conocimiento compartido durante este tiempo, por la confianza, y especialmente, le agradezco por su paciencia infinita, sin ella, no podríamos haber llegado al final de esto.

A mi compañero Mauricio González González, por haber compartido conmigo los *gajes del oficio* de esta maestría. Gracias por las risas, por las tareas y trabajos compartidos, por el apoyo, por ser cómplice de tantos momentos; definitivamente, esta maestría no hubiera sido lo mismo sin ti.

Al Dr. Jonathan Sánchez Muñoz de la Escuela de Ingeniería y Ciencias del ITESM SLP, por haberme apoyado, orientado y enseñado a utilizar las herramientas que fueron fundamentales para el desarrollo de este trabajo.

Al Dr. César Augusto Puente Montejano y al Dr. Juan Carlos Cuevas Tello por haber aceptado ser miembros de mi comité de tesis, por sus aportaciones y su disposición. A la Dra. Sandra Edith Nava Muñoz por sus comentarios y sugerencias siempre tan pertinentes. También agradezco a todos los profesores que contribuyeron a mi formación académica en este posgrado.

Finalmente, quiero agradecer a la Facultad de Ingeniería de la UASLP por haberme permitido realizar este posgrado y también por habernos brindado la confianza al compartir la información académica de los estudiantes, imprescindible para este trabajo de investigación.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El desarrollo de nuevas tecnologías y el aumento en el número de usuarios ha provocado un crecimiento dramático en la cantidad de información que se genera y se almacena continuamente en las bases de datos. La implementación de nuevas herramientas computacionales que ayuden a la extracción de información de estos grandes volúmenes de datos se volvió una necesidad, dando como resultado la aparición de la Minería de Datos (Data Mining, DM, por sus siglas en inglés) [1, 2].

La aplicación del DM en diversas áreas se ha vuelto tendencia, como por ejemplo, en *marketing*, medicina, finanzas, educación, etc. En el caso específico de la educación, la información académica almacenada en las instituciones puede servir para buscar solución a problemas como: estudiantes con bajo desempeño escolar, rezago estudiantil, deserción, etc. La Minería de Datos Educativa (Educational Data Mining, EDM, por sus siglas en inglés) surge como resultado de la aplicación de las técnicas de DM a la educación [3, 4]. El objetivo principal de la EDM es encontrar patrones en datos académicos y hacer predicciones sobre el comportamiento de los estudiantes, con la finalidad de proponer estrategias que permitan mejorar la calidad académica de los estudiantes.

1.1. Propuesta de Trabajo

Existen muchos factores que impiden a los estudiantes tener un desempeño académico satisfactorio, sin embargo, en la mayoría de los casos, los recursos y el tiempo que tienen los tutores no es suficiente para analizar a profundidad el historial de cada estudiante. Esto provoca que las situaciones de estudiantes en riesgo sean detectadas hasta etapas muy avanzadas de la carrera. Si se pudieran identificar los casos latentes en etapas tempranas se podrían buscar estrategias para disminuir la reprobación y los casos de abandono. Nuestra propuesta de trabajo es aplicar DM para apoyar a los coordinadores académicos y tutores de la Facultad de Ingeniería de la UASLP, particularmente al Área de Ciencias de la Computación, a identificar a estudiantes en situaciones de riesgo, a través de un sistema que permita predecir el desempeño académico de los estudiantes con base en su historial académico.

1.2. Objetivo General

Generar y analizar modelos predictivos computacionales que permitan identificar posibles casos de estudiantes en riesgo de deserción analizando los datos académicos históricos de los estudiantes de dos de los programas del Área de Ciencias de la Computación (Ingeniería en Computación e Ingeniería en Informática).

1.2.1. Objetivos Particulares

Para lograr el objetivo principal se plantean también los siguientes objetivos particulares:

- Analizar el conjunto de datos disponible con el fin de seleccionar los datos adecuados para nuestro estudio.
- Realizar el preprocesamiento del conjunto de datos académicos de los estudiantes.
- Llevar a cabo un análisis exploratorio de los datos para la identificación de atributos representativos.
- Evaluar las características de diversos algoritmos de aprendizaje de máquina con el fin de aplicar los más adecuados al conjunto de datos.
- Obtener una interpretación del conocimiento obtenido que sea útil para la coordinación académica de las carreras mencionadas.

1.3. Metodología de Investigación y de Trabajo

Después de haber planteado los objetivos del trabajo, es importante definir la metodología de investigación adecuada para el proyecto. En este trabajo se consideraron 6 etapas en el proceso de investigación: definición del problema, revisión bibliográfica, selección de herramientas y métodos, aplicación de métodos, evaluación y análisis de resultados, y finalmente, la conclusión del trabajo. La Figura 1.1 muestra el esquema de la metodología de investigación planteada.

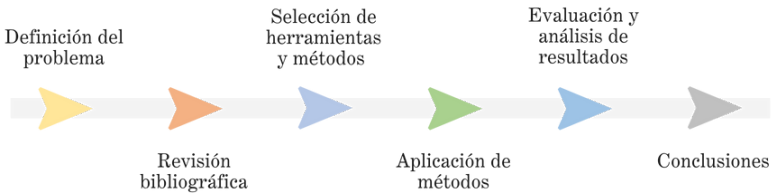

Figura 1.1. Metodología de investigación planteada.

Adicional a la metodología de investigación presentada, es conveniente proponer una metodología de trabajo que establezca etapas específicas para el desarrollo del proyecto. La metodología de trabajo planteada en este estudio tiene 5 etapas principales basadas en estudio relacionados [5, 6], estas etapas están orientadas a cumplir cada uno de los objetivos particulares establecidos. La Figura 1.2 muestra un esquema de dicha metodología. A continuación, se describe brevemente cada una de las 5 etapas.

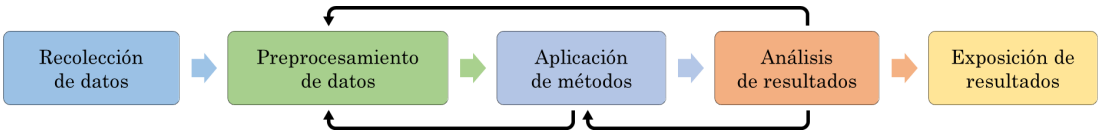

Figura 1.2. Metodología de trabajo basada en estudios relacionados [5, 6].

- **Recolección de datos:** Consiste en la selección del conjunto de datos a utilizar. En este caso, se seleccionará la información académica de los estudiantes de las carreras de Ingeniería en Computación e Ingeniería en Informática, del Área de Ciencias de la Computación de la Facultad, de las generaciones 2008 a 2013.
- **Preprocesamiento de datos:** Incluye diversas tareas como realizar la limpieza de los datos, llevar a cabo un análisis exploratorio del conjunto datos para conocer la información que aporta y elegir los atributos y muestras a evaluar.

- **Aplicación de métodos:** Conlleva la elección y aplicación de diversas técnicas supervisadas de DM a los datos seleccionados y el ajuste de los hiperparámetros correspondientes.
- **Análisis de resultados:** Constituye la comprensión de los resultados obtenidos con los diferentes métodos aplicados.
- **Exposición de resultados:** Implica la interpretación de la información derivada de los resultados y su representación en forma útil.

Es importante mencionar que este proceso no es unidireccional si no que puede ser cíclico, esto permite la repetición de algunas etapas del proceso como se muestra en la Figura 1.2. Poder regresar a etapas previas es fundamental, ya que en ocasiones es necesario modificar algunos datos sin tener que empezar desde cero, por ejemplo, modificar algunos atributos, cambiar la selección de las muestras, ajustar los hiperparámetros de los modelos, etc.

1.4. Contribuciones de la Tesis

De acuerdo con los objetivos planteados, la principal contribución de la tesis es la generación y el análisis de modelos predictivos que permitan la identificación de estudiantes en riesgo de deserción. Otra contribución importante es la transformación del conjunto de datos, ya que la estructura final obtenida sintetiza la información académica de los estudiantes para su posterior análisis.

1.5. Estructura de la Tesis

El resto de la tesis se organiza de la siguiente manera:

Capítulo 2: Presenta el fundamento teórico del DM y EDM, la descripción de las técnicas de predicción más comunes, las métricas de evaluación y las herramientas de desarrollo utilizadas. En este capítulo se incluye también el trabajo relacionado.

Capítulo 3: Describe la selección y el preprocesamiento del conjunto de datos. Incluye la limpieza, transformación, reducción e integración de los datos. Adicionalmente, se presenta

un análisis exploratorio del conjunto de datos.

Capítulo 4: Explica los experimentos realizados, los métodos empleados, los hiperparámetros ajustados y presenta el análisis de los resultados.

Capítulo 5: Incluye un resumen de los principales resultados obtenidos, la discusión de las limitaciones, las conclusiones más relevantes y el posible trabajo futuro.

Capítulo 2

Minería de Datos Educativa

2.1. Descubrimiento de Conocimiento en Bases de Datos

Un problema derivado de la era digital en la que nos encontramos, es que la sobrecarga de datos generados y almacenados ha superado nuestra capacidad de procesarlos. Las bases de datos son una fuente de información potencial, y en las últimas décadas se han destacado por el valor incalculable de lo que ahí se oculta. Ante esta necesidad, surgió un campo de estudio llamado *Descubrimiento de conocimiento en bases de datos* (Knowledge Discovery in Databases, KDD, por sus siglas en inglés), y la aplicación del KDD se ha vuelto esencial para la manipulación de dicha información [1, 6].

El objetivo del KDD es el desarrollo de métodos y técnicas que permitan analizar datos almacenados para obtener información relevante, es decir, transformar los grandes volúmenes de información en formas de representación más compactas y fáciles de analizar, que permitan convertir la información en algo útil.

Las etapas del KDD se pueden agrupar en cinco principales: selección de un conjunto de datos, preprocesamiento de la información, transformación de los datos, aplicación de algoritmos de minería de datos (Data Mining, DM, por sus siglas en inglés) y evaluación de resultados. El proceso completo del KDD (Figura 2.1) incluye también la interpretación y exposición de los resultados, esta etapa es fundamental ya que en ella se decide si el conocimiento descubierto es relevante para la investigación o no. La etapa de DM consiste en la aplicación de herramientas para el descubrimiento de patrones y extracción de información oculta dentro de los datos [6].

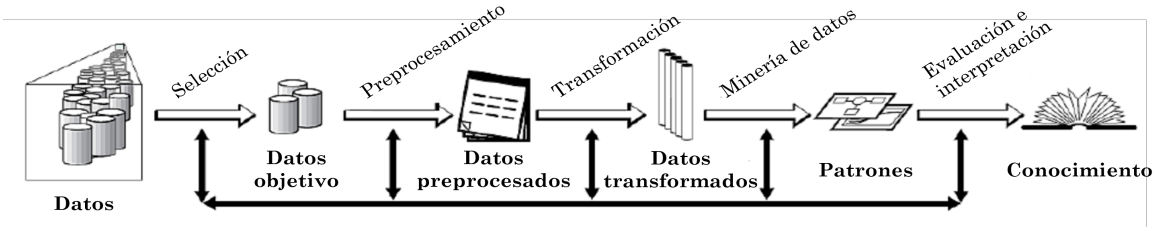

Figura 2.1. Etapas en el proceso del KDD, adaptada de [1].

El proceso del KDD es interactivo e iterativo, las decisiones que el analista tome a lo largo del proceso influyen directamente en los resultados y en su interpretación. Las etapas del proceso se describen brevemente a continuación [7].

Etapa de Selección

Cuando se conoce la problemática a resolver y se tiene claro el objetivo del estudio, se lleva a cabo la etapa de selección. En esta etapa se selecciona el conjunto de datos objetivo sobre el cual se realizará el proceso del KDD.

Etapa de Preprocesamiento

En la etapa de preprocesamiento y limpieza de los datos, se analizan los datos para identificar datos incompletos, vacíos o con formatos no permitidos, para evitar conflictos en etapas posteriores. Estos datos pueden ser eliminados o tratados estadísticamente para ser reemplazados, el analista toma la decisión de acuerdo a su criterio. Además, se busca eliminar el ruido en los datos, es decir, identificar valores que se encuentren significativamente lejos del resto de los valores del conjunto de datos. Cuando los datos provienen de distintas fuentes es más probable que sean conjuntos heterogéneos, por lo que se debe dedicar más tiempo a la limpieza y preprocesamiento, ya que resulta fundamental para las etapas posteriores.

Etapa de Transformación

En la etapa de transformación y reducción de datos, únicamente se conservan los datos que aporten información relevante para el estudio, los datos son analizados en busca de características útiles. Cuando se han identificado los datos/atributos representativos, se elimina el resto y el conjunto de datos se reduce, puede ser reducción horizontal o vertical.

La reducción horizontal se presenta cuando se eliminan registros idénticos que se generan como resultado de la creación de atributos de mayor nivel, por agrupamiento de datos, etc. Mientras que la reducción vertical implica la eliminación de columnas completas que representan atributos sin información trascendente para el estudio.

Etapa de Minería de Datos

El objetivo de esta etapa es buscar y encontrar patrones dentro del conjunto de datos reducido, para lograr este objetivo se emplean diversos algoritmos y métodos. Existe una gran variedad de técnicas que pueden ser aplicadas dependiendo del objetivo del estudio, pueden ser técnicas predictivas, o incluso, únicamente descriptivas. La elección del algoritmo incluye la selección del método, así como el ajuste de los parámetros.

Etapa de Evaluación e Interpretación

En la última etapa del proceso se lleva a cabo la interpretación de los patrones obtenidos. Los patrones obtenidos son analizados y son transformados por el analista para poder ser representados de una forma comprensible y útil. El conocimiento generado se documenta y se reporta.

El proceso del KDD puede volverse iterativo, es decir, es posible caer en ciclos repetidos entre dos o más etapas, sin embargo, es importante concluir con cada una de ellas correctamente antes de continuar con la siguiente.

2.2. Minería de Datos

El DM es únicamente un paso dentro de todo el proceso del KDD, consiste en explorar grandes bases de datos y crear modelos que representen la información de manera útil y más simplificada. El DM es un método de investigación que permite generar información y crear conocimiento basado en características ocultas de los datos. El término de "minería" hace referencia al concepto básico de extracción de minerales, la diferencia, es que en el DM se extrae conocimiento. El objetivo del DM es encontrar patrones de datos, organizar la información basándose en relaciones

escondidas, determinar reglas de asociación, clasificar objetos, agrupar datos, etc [8].

Decidir cuáles modelos resultan útiles y cuáles no, es parte del trabajo interactivo, queda a juicio del analista esta decisión, por eso es importante conocer el contexto del problema y las variables del entorno. Existe una gran variedad de algoritmos de DM descritos hasta el día de hoy, la mayor parte de ellos se basan en estadística, reconocimiento de patrones, aprendizaje supervisado y bases de datos [1].

2.2.1. Descubrimiento y Predicción

El estudio de un sistema puede tener dos objetivos: verificar o descubrir información. En la verificación, el sistema está limitado a verificar la hipótesis establecida por el analista; mientras que en el descubrimiento, el sistema encuentra nuevos patrones de manera autónoma. El descubrimiento, a su vez, se subdivide en dos, predicción y descripción. El estudio descriptivo encuentra similitudes y/o asociaciones en subconjuntos del conjunto de datos objetivo y crea patrones para representar la información de manera comprensible. Por su parte, el análisis predictivo se usa para clasificar y predecir comportamientos o eventos futuros. La Figura 2.2 presenta un esquema de la taxonomía de los métodos del DM descritos previamente, basada en [9].

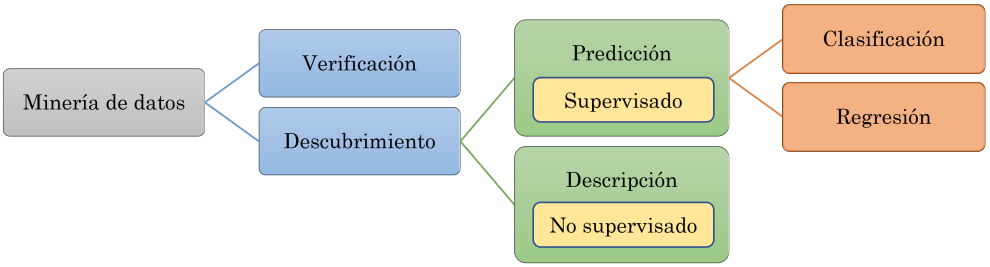

Figura 2.2. Taxonomía de los métodos del DM, basada en [9].

Los métodos de descubrimiento también son conocidos como métodos de aprendizaje supervisados (predicción) y no supervisados (descripción), como podemos observar en la Figura 2.2. Los métodos de aprendizaje supervisados son aquellos en los cuales los datos son etiquetados y sirven como datos de entrada para entrenar a los algoritmos. Los algoritmos utilizan estos datos para "aprender" y posteriormente, usan ese conocimiento para predecir el comportamiento de algún evento objetivo.

Los métodos de aprendizaje no supervisados son aquellos que no utilizan etiquetas y tampoco se tiene un conjunto de entrenamiento, este tipo de métodos se utilizan principalmente para analizar y agrupar los datos. Los algoritmos no supervisados identifican similitudes o diferencias entre grupos de datos, y con esta información pueden crear patrones y agrupaciones. La Figura 2.3 muestra un breve esquema de las diferencias entre el aprendizaje supervisado y el no supervisado, donde se observa que el aprendizaje supervisado clasifica de acuerdo a etiquetas definidas, mientras que el aprendizaje no supervisado clasifica de acuerdo a las características que el método considere relevantes.

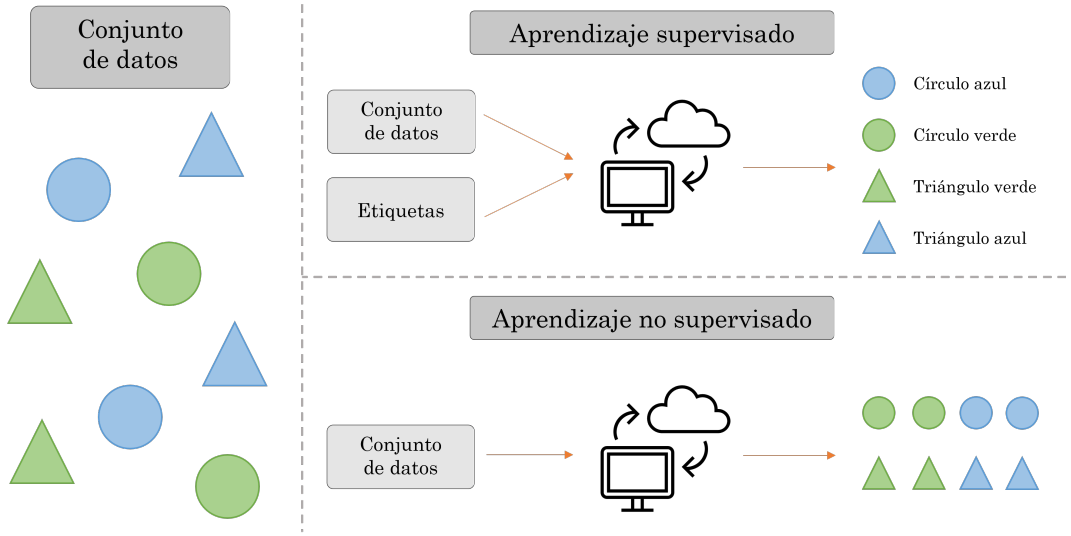

Figura 2.3. Esquema del aprendizaje supervisado y no supervisado.

La clasificación, por ser una subcategoría de la predicción en el DM, es considerada también un método de aprendizaje supervisado, es decir, será capaz de definir un valor de salida utilizando un conjunto de entrada para entrenarse y aprender. Este método se utiliza para clasificar a los datos en distintas clases cuando la variable objetivo tiene valores discretos y finitos.

La clasificación puede ser binaria o multicategórica. Si los datos se desean clasificar en dos tipos de clases únicamente, entonces es una clasificación binaria; por ejemplo, si se desea predecir si un individuo es hombre o mujer, los datos serán clasificado en dos grupos solamente. Sin embargo, se pueden tener más de dos clases objetivo, por ejemplo, si se desea que el algoritmo clasifique a las personas en bebés, niños, jóvenes, adultos y adultos mayores; esta sería una clasificación multicategórica, ya que se tienen cinco clases objetivo.

Existen diversos métodos de clasificación, algunos de los más comunes son: regresión lineal, regresión logística, máquina de soporte vectorial, árboles de decisión, bosques aleatorios y redes neuronales [11]. Lo que hace especial a cada una de estas técnicas es que cada una de ellas tiene diferentes fundamentos, algunas provienen de la inteligencia artificial y otras de la estadística. Debido a esto, es posible que algún método se adapte mejor a un conjunto de datos específico, mientras que para otro, podría no funcionar adecuadamente; por eso, es importante analizar las características de diferentes métodos con la intención de encontrar el que se adapte mejor a los datos y al estudio en cuestión.

2.2.2. Algoritmos de Predicción

En la actualidad existe una gran variedad de algoritmos de predicción los cuales han sido constantemente objeto de estudio. Un tema de particular interés es la clasificación de dichos algoritmos, un ejemplo es el estudio publicado en 2020 por Prenkaj y colaboradores, donde se presenta la clasificación de algunas de las principales estrategias utilizadas en estudios de predicción [10]. La Figura 2.4 muestra la clasificación presentada en dicho trabajo, donde se puede observar que las técnicas fueron clasificadas de acuerdo con el principio bajo el cual operan.

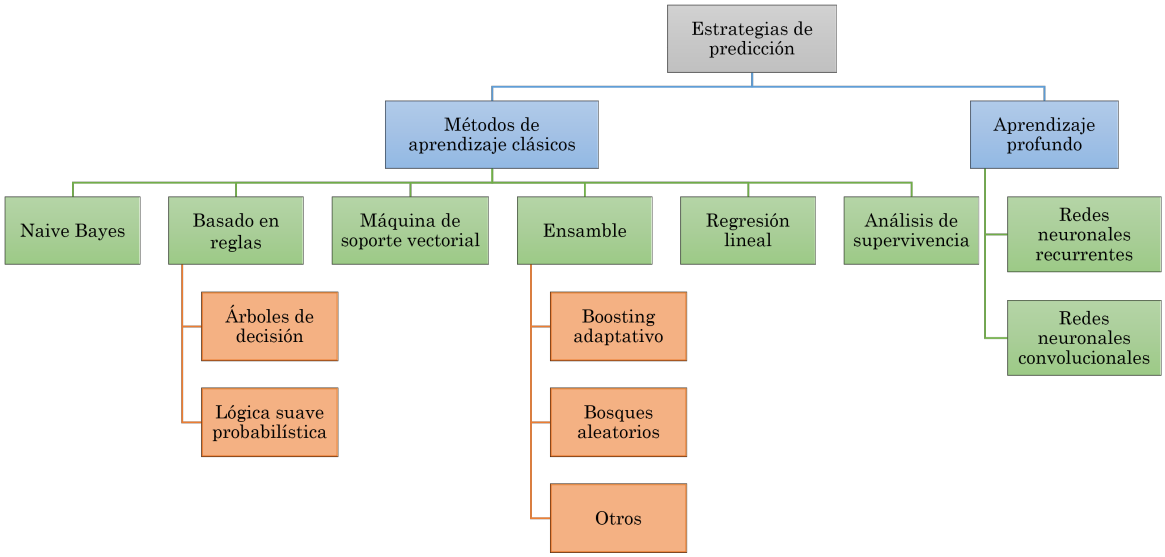

Figura 2.4. Clasificación de algunas técnicas de predicción de DM, traducida de [10].

La clasificación de la Figura 2.4 tiene dos ramas principales, estas ramas dividen a los méto-

dos en aquellos que se basan en el aprendizaje automático clásico y los que usan el aprendizaje profundo. Dentro de los métodos de aprendizaje automático clásico se encuentran: Naïve Bayes, los métodos basados en reglas, máquina de soporte vectorial, los métodos de ensamble, regresión lineal y análisis de supervivencia. Como parte de los métodos de aprendizaje profundo se encuentran las redes neuronales recurrentes y las redes neuronales convolucionales.

Para este trabajo se eligieron cinco métodos de clasificación: Naïve Bayes, árboles de decisión, máquina de soporte vectorial, bosque aleatorio y redes neuronales artificiales. Estos cinco métodos fueron elegidos debido a la diferencia que existe entre sus principios de operación, por ejemplo, los árboles de decisión son basados en reglas, el Naïve Bayes es probabilístico, el bosque aleatorio es un método de ensamble, etc. Además, durante la revisión bibliográfica se encontró que estos métodos son comúnmente utilizados en estudios similares. A continuación, se presenta un breve resumen de cada una de las cinco técnicas predictivas aplicadas a nuestro conjunto de datos.

Árbol de Decisión

Los árboles de decisión son algoritmos de aprendizaje supervisado basados en reglas. Las reglas de decisión pueden ser combinadas para formar estructuras similares a las de un árbol. Este método supervisado es muy utilizado, ya que al ser una representación gráfica de una decisión o condición, es muy fácil de interpretar [12]. Los árboles de decisión tienen dos objetivos: comprensión de los datos y predicción de nuevos elementos. Un árbol de decisión puede servir para representar al conjunto de datos de manera compacta y comprensible, o también es posible predecir el valor objetivo de un nuevo elemento a partir de esta representación.

Estos árboles tienen un arreglo jerárquico, el primer nodo es llamado raíz, después hay ramas, nodos internos (o nodos de decisión) y nodos hoja; la Figura 2.5 presenta los distintos nodos de un árbol de decisión. Cada nodo representa un atributo, los atributos se descomponen en ramas, las ramas pueden representar los valores de cada atributo o pueden indicar si la condición planteada es verdadera o falsa [13].

Los árboles usan la estrategia de “divide y vencerás” para identificar los puntos óptimos de división dentro del conjunto de datos. Este proceso de división continúa (de arriba hacia abajo) hasta que cada uno de los elementos del conjunto ha sido clasificado y cada rama llegue

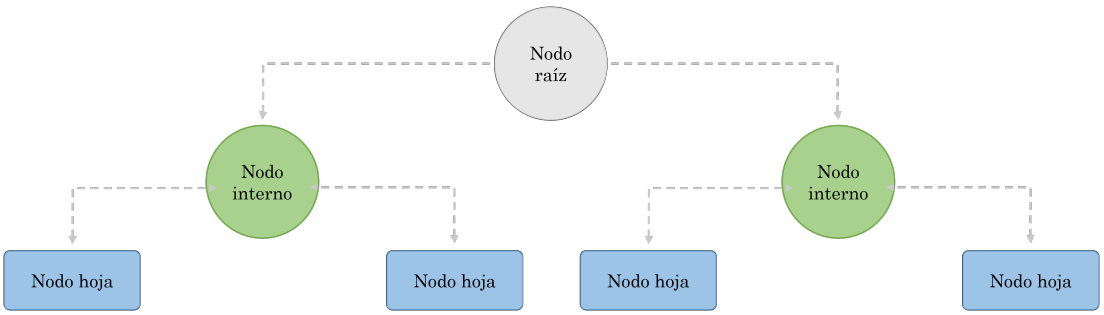

Figura 2.5. Nodos en un árbol de decisión.

a representar una solución única.

Bosque Aleatorio

El algoritmo de bosque aleatorio es tipo de ensamble que combina el producto de diferentes árboles de decisión para lograr un resultado de mayor calidad. El conjunto de datos original es fragmentado aleatoriamente y cada uno de los subconjuntos generados es utilizado para construir un árbol de decisión diferente. De cada árbol se obtiene una salida diferente, la salida que tenga mayor frecuencia de aparición será el resultado arrojado por el bosque aleatorio, esto reduce el riesgo de sesgo y sobreajuste en el resultado final [13].

Máquina de Soporte Vectorial

La técnica de máquina de soporte vectorial (Support Vector Machine, SVM, por sus siglas en inglés) es un algoritmo de clasificación que separa de manera eficiente a los datos en categorías. El principio del SVM consiste en la búsqueda de un separador que sea capaz de dividir correctamente a los elementos del conjunto de datos de acuerdo a sus caracterísitcas y a la clase a la que pertenecen, a este separador se le conoce como hiperplano; si fuera en dos dimensiones sería una línea, sin embargo, la mayoría de los problemas de clasificación tienen muchas dimensiones [14]. Cuando las clases han sido separadas por el hiperplano, se puede llevar a cabo la clasificación de nuevos elementos, analizando sus características es posible identificar la clase a la que pertenecen.

Los vectores de soporte son aquellos puntos que delimitan el espacio entre el hiperplano y los elementos de cada clase, cuando la separación entre las categorías es mayor la calidad del

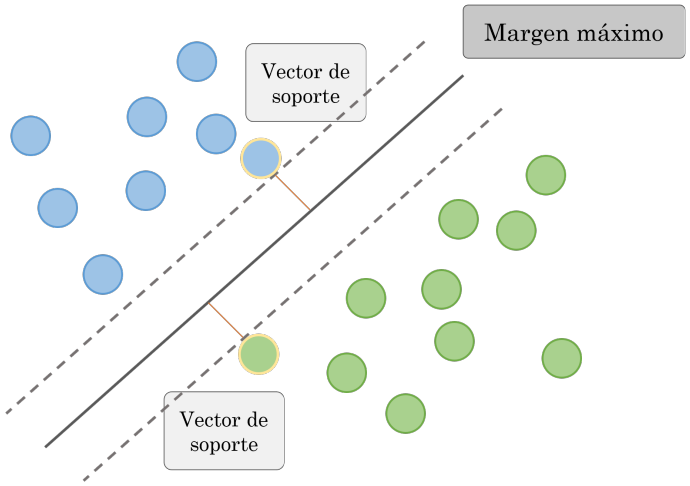

Figura 2.6. Vectores de soporte en SVM, adaptada de [14].

modelo aumenta. Muchas veces es necesario permitir que la clasificación de algunos elementos sea errónea para poder ampliar el espacio entre el hiperplano y los vectores de soporte. Una parte fundamental de este método consiste en encontrar la relación óptima entre el margen máximo y el número de elementos clasificados erróneamente. La Figura 2.6 representa una clasificación en dos dimensiones usando SVM, donde se puede observar el papel que juegan los vectores de soporte y el margen máximo.

En muchas ocasiones no es posible encontrar un hiperplano que separe fácilmente las clases, para esto, se aplica la función kernel. Esta función permite transformar el espacio y crear nuevas dimensiones que permitan separar las clases adecuadamente. Existen diferentes funciones kernel, tales como kernel lineal, kernel polinomial, kernel gaussiano, etc., elegir el adecuado es tarea del analista.

Naïve Bayes

El clasificador Naïve Bayes es un método de clasificación probabilístico basado en el Teorema de Bayes de probabilidad condicional. Este clasificador supone independencia entre las características de los elementos, es decir, la presencia de alguna característica no está relacionada con la presencia o ausencia de alguna otra característica; esta suposición simplifica los cálculos significativamente [15]. Es un algoritmo sencillo pero eficiente, pueden obtenerse buenos modelos rápidamente. Este tipo de algoritmo es muy útil cuando la dimensión del espacio de caracte-

rísticas es muy grande, ya que sus cálculos requieren mucho menos tiempo que el resto de los algoritmos de aprendizaje de máquina automáticos.

El objetivo del clasificador Naïve Bayes es calcular la probabilidad de que un elemento pertenezca a una clase especifica utilizando como base el Teorema de Bayes. El Teorema de Bayes calcula la probabilidad condicional de un evento aleatorio conociendo los valores de otro conjunto de eventos aleatorios [14]. Supongamos dos eventos, A y B, la probabilidad de que ocurra el evento A se denota por P(A) y la probabilidad de que ocurra B, es P(B). El Teorema de Bayes permite calcular la probabilidad de que ocurra el evento A dado un evento previo B, P(A|B); para esto utiliza las probabilidades P(A), P(B) y P(B|A) (probabilidad de que ocurra B dado que ocurrió A), que resultan mucho más fáciles de calcular. Este teorema está determinado por la Ecuación 2.1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

Un concepto que se debe tener claro es la diferencia entre P(A|B) y P(B|A), el Teorema de Bayes se utiliza para calcular P(A|B) usando P(B|A), ya que es más fácil conocer el valor de P(B|A) que el de P(A|B). El evento B normalmente corresponde a la combinación de diferentes variables o características, mientras que el evento A es único, esto hace que calcular P(A|B) sea mucho más complicado.

Redes Neuronales Artificiales

Las redes neuronales artificiales (Artificial Neural Network, ANN, por sus siglas en inglés) se basan en el funcionamiento de las neuronas del cerebro humano, imitan la forma en que las neuronas transmiten la información. Las ANN se componen de capas y nodos, incluyen una capa de entrada, una o varias capas ocultas y una capa de salida; cada nodo representa a una neurona. La Figura 2.7 muestra un esquema básico de la estructura de una ANN [12].

La capa de entrada recibe la información, cada neurona recibe datos específicos, los procesa y transmite la información a otra neurona; cada una de las neuronas se conecta a las neuronas de la siguiente capa. Cada neurona procesa los datos de forma diferente, los cálculos se realizan utilizando los pesos asignados a cada conexión de entrada, los pesos pueden entenderse como la

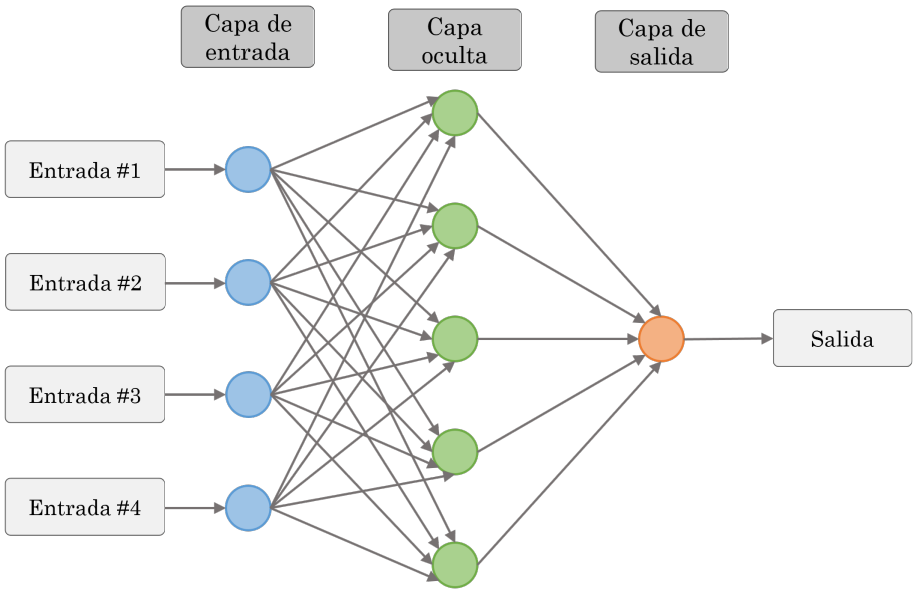

Figura 2.7. Esquema de una red neuronal artificial, adaptada de [12].

fuerza de una conexión sináptica [14]. El valor de salida de una neurona será una combinación lineal de los valores de entrada y sus pesos, pasando por una función de activación, por lo tanto, es fundamental ajustar los pesos adecuadamente. Las funciones de activación son las responsables de transformar la información y de propagarla de una neurona a otra, gracias a ellas la neuronas pueden aprender relaciones no lineales.

El entrenamiento de una ANN consiste en ajustar el valor de los pesos, de tal manera que las predicciones mejoren en cada iteración. La ANN modifica ligeramente el valor de los pesos en cada cálculo y analiza la calidad de la predicción, compara los resultados y así, es capaz de identificar cuáles variables tienen mayor influencia y reducir los errores.

2.2.3. Métricas de Evaluación

Como hemos visto hasta ahora, existen diversos algoritmos de clasificación que pueden ser aplicados sobre un mismo conjunto de datos, cada uno con un fundamento diferente, y una las tareas del analista consiste en elegir el método que mejor se adapte al conjunto de datos objetivo. Uno de los parámetros decisivos al momento de tomar esta decisión es la calidad de los modelos obtenidos, es decir, que tan buenos o malos son para clasificar nuevos elementos. Existen diferentes métricas que sirven para evaluar el desempeño de los modelos de clasificación, algunas de ellas

son: matriz de confusión, exactitud, precisión, sensibilidad y valor-F; todas estas evalúan algo diferente, por eso es importante conocerlas y saber cuándo aplicar cada una [12, 14]. Además de evaluar el desempeño de los modelos a través de las métricas mencionadas, es posible evaluar también los atributos considerados en cada modelo mediante una métrica llamada importancia. A continuación, se presenta una breve descripción de cada una de estas métricas.

Matriz de confusión

La matriz de confusión es una métrica de evaluación muy común que se usa para medir el rendimiento de los modelos de clasificación, si la clasificación es binaria, la matriz de confusión corresponde a una matriz de 2x2, una fila y una columna por cada clase. Es una matriz relativamente sencilla, que representa los resultados obtenidos tras una clasificación, para una clasificación binaria, cada elemento clasificado tiene cuatro posibles etiquetas dentro de la matriz de confusión, dependiendo de su coincidencia con el valor real: verdadero positivo, verdadero negativo, falso positivo o falso negativo (Figura 2.8) [12].

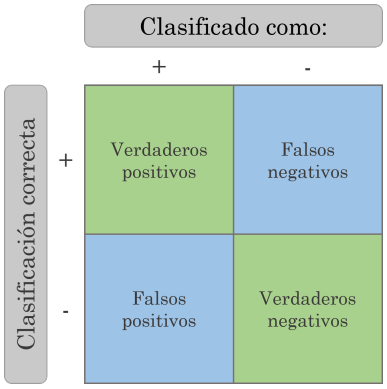

Figura 2.8. Matriz de confusión, adaptada de [12].

- *Verdadero positivo (VP):* El valor real es positivo y el modelo lo clasificó como positivo.
- *Verdadero negativo (VN):* El valor real es negativo y el modelo lo clasificó como negativo.
- *Falso positivo (FP):* El valor real es negativo y el modelo lo clasificó como positivo.
- *Falso negativo (FN):* El valor real es positivo y el modelo lo clasificó como negativo.

A partir de estos cuatro valores se obtienen el resto de las métricas de evaluación.

Exactitud

La exactitud (*accuracy*, en inglés) es una métrica que indica la fracción de casos que fueron clasificados correctamente, por lo tanto, tiene un valor mínimo de 0 y un máximo de 1. Si el valor obtenido es cercano a 1, significa que la mayoría de los elementos fueron clasificados correctamente. Este valor se calcula dividiendo la suma de los elementos verdaderos (positivos y negativos) entre la suma de todos los elementos (verdaderos y falsos) (Ecuación 2.2).

$$Exactitud = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{2.2}$$

Con la exactitud se calcula qué fracción de los elementos se encuentra en la diagonal de la matriz (Figura 2.9). Esta métrica es recomendada cuando el conjunto de datos está balanceado, es decir, el número de elementos de cada clase es similar. Sin embargo, una observación importante es que este valor no es una buena métrica cuando se tienen clases desbalanceadas, ya que el modelo podría basarse únicamente en la clase más frecuente y clasificar a todos los elementos como parte de esta clase. La exactitud no reflejaría este error, ya que seguiría siendo alta debido a la frecuencia de los casos verdaderos.

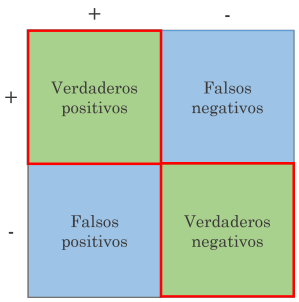

Figura 2.9. Elementos de la matriz de confusión analizados en la exactitud.

Precisión

La precisión (*precision*, en inglés) indica qué parte de los elementos clasificados como positivos son realmente positivos, es decir, nos sirve para saber qué tanto podemos confiar en los elementos clasificados como verdaderos. Se calcula dividiendo los valores verdaderos positivos entre la suma de todos los valores positivos (Ecuación 2.3).

$$Precisión = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.3}$$

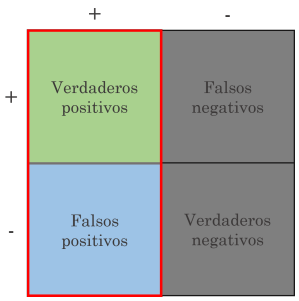

Figura 2.10. Elementos de la matriz de confusión que definen la precisión.

La Figura 2.10 enmarca los elementos que definen la precisión de un modelo. Como se puede observar, para esta métrica se consideran únicamente los elementos que fueron clasificados como positivos, columna de la izquierda. Dado que esta métrica solamente indica la fracción de elementos clasificados como positivos correctamente, no es recomendable evaluar la calidad de los modelos utilizando solo esta métrica.

Sensibilidad

La sensibilidad (*recall* o *sensitivity*, en inglés) mide la relación entre los verdaderos positivos y los positivos reales, es decir, esta métrica nos sirve para saber que tan bueno es el modelo para identificar elementos positivos, cuántos puede identificar del total. La sensibilidad se calcula como el cociente de los verdaderos positivos y los positivos reales (verdaderos positivos y falsos negativos) (Ecuación 2.4).

$$Sensibilidad = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.4}$$

En una matriz de confusión los elementos que definen la sensibilidad son los que conforman el renglón de los positivos reales. En la Figura 2.11 se enmarcan los valores implicados, primer renglón. Como esta métrica identifica a los elementos positivos clasificados correctamente no es una métrica que por sí sola nos permita evaluar la calidad de los modelos.

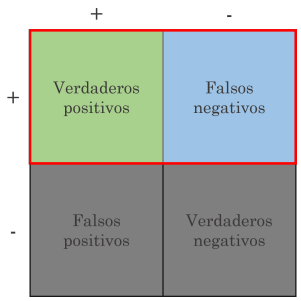

Figura 2.11. Elementos de la matriz de confusión que definen la sensibilidad.

Valor F1

El valor F1 es una métrica que combina los valores de la precisión y la sensibilidad, esto resulta muy útil ya que se pueden evaluar ambas métricas en una sola. El valor F1 se obtiene calculando la media armónica de estos dos valores (Ecuación 2.5).

$$F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad} \tag{2.5}$$

El valor F1 puede parecerse mucho a la media aritmética cuando la precisión y la sensibilidad son muy similares, sin embargo, esto cambia cuando existe una diferencia notable entre ambas métricas. Cuando uno de estos valores es muy pequeño en comparación con el otro, el valor F1 que se obtiene también es bajo, haciendo saber al analista que el modelo es deficiente. Esta métrica es muy recomendada cuando se tienen clases desbalanceadas, ya que en estos casos es común obtener valores altos en la sensibilidad pero valores bajos en la precisión, o viceversa, por lo que es necesario considerar ambas métricas para evaluar la calidad real del modelo.

Importancia de los Atributos

Al construir modelos predictivos una cuestión importante es identificar cuál o cuáles de los atributos analizados resultan más significativos para su construcción, de esta manera se puede saber cuál atributo es el más representativo en el estudio en cuestión. Para conocer estos atributos se calcula el efecto que tiene cada uno de ellos en la construcción del modelo, es decir, se calcula cuánto cambian las métricas de evaluación si los valores de un atributo se modifican. El procedimiento consiste en cambiar aleatoriamente todos los valores de un atributo (cambiar aleatoriamente todos los valores de la columna), crear un nuevo modelo, calcular las métricas

de evaluación y comparar estos resultados con el modelo original. Si las métricas de evaluación disminuyen significa que ese atributo en particular es importante para su construcción. El valor obtenido en la importancia representa la diferencia entre las métricas de evaluación del modelo original y del modificado, por lo tanto, entre mayor sea el valor obtenido en la importancia, el atributo es más representativo. Cabe destacar que la importancia no es un valor normalizado, es decir, si se obtiene un valor de 0.258 en la importancia de cierto atributo en dos estudios diferentes, no significa que tenga la misma importancia en ambos, el valor obtenido se debe comparar con el valor de las métricas de evaluación originales para determinar su importancia real, ya que no es lo mismo un cambio de 0.258 en un modelo cuya exactitud es de 0.654 a uno cuya exactitud es de 0.937, diríamos que es más importante en el modelo con menor exactitud. La importancia de los atributos puede ser calculada en el conjunto de entrenamiento o en el conjunto de prueba. Los atributos que son importantes en el conjunto de entrenamiento pero no en el conjunto de prueba pueden generar modelos sobreajustados.

2.2.4. Herramientas de Desarrollo

Para el desarrollo de este trabajo se usó Google Colaboratory[1], una herramienta de Google que permite crear modelos de aprendizaje automático en la nube. Google Colaboratory (Google Colab) es un entorno interactivo que permite escribir y ejecutar código de Python en el navegador fácilmente y es de libre acceso. En Google Colab se crean *notebooks* que permiten combinar código, texto e imágenes en un solo documento, estos *notebooks* se almacenan en Google Drive. Los *notebooks* de Colab están basados en los *notebooks* de Jupyter [2], pueden compartirse fácilmente y se puede acceder a ellos desde cualquier navegador. JupyterLab es un entorno interactivo basado en web para el desarrollo de código, los *notebooks* permiten la creación y ejecución de fragmentos de código para el procesamiento de datos de una forma interactiva.

Google Colab permite el uso de las bibliotecas más populares de Python para analizar y visualizar datos, tales como NumPy y matplotlib [16, 17]. Para el aprendizaje automático, Colab permite cargar y manipular *datasets* con ayuda de la API (Interfaz de programación de aplicaciones) Pandas [18, 19]; también es posible entrenar clasificadores y evaluar modelos con la ayuda

[1] https://colab.research.google.com/
[2] https://jupyter.org/

de algunas bibliotecas, tales como Scikit-Learn y Keras [20, 21, 22]. A continuación se presenta una breve descripción de las bibliotecas usadas en este trabajo.

NumPy: Es una biblioteca especializada de Python, "Numerical Python", que permite realizar una gran variedad de cálculos numéricos. NumPy representa los datos como arreglos multidimensionales, logrando con esto, que el almacenamiento y el procesamiento de los datos sea mucho más eficiente.

Matplotlib: Es una biblioteca especializada para la creación de gráficos en Python, permite crear y personalizar fácilmente los gráficos más comunes, tales como: gráficos de barras, histogramas, diagramas de dispersión, mapas de color, etc.

Pandas: Es una biblioteca especializada de Python para la manipulación y el análisis de datos, su nombre se deriva de "Panel Data". Pandas permite la manipulación de los datos como tablas numéricas y series temporales, ofrece estructuras poderosas y flexibles.

Scikit-Learn: Es una herramienta sencilla y eficiente para el análisis predictivo de datos, su funcionamiento se basa en las biblioteca de NumPy, SciPy y matplotlib. Scikit-Learn cuenta con algoritmos de clasificación, regresión y agrupamiento.

Keras: Es una biblioteca de Python especializada en construir y entrenar modelos de aprendizaje profundo. La construcción de los modelos en Keras es a través de bloques, estos bloques se conectan entre sí y pueden personalizarse fácilmente.

2.3. Minería de Datos Educativa

Dentro de las áreas de aplicación del DM se encuentra la educación. Una institución educativa implica tres “actores” principales: estudiantes, maestros y entorno. La interacción de los estudiantes y maestros con el entorno escolar genera grandes cantidades de información. Estos datos se almacenan en las instituciones y pueden ser usados para conocer más sobre el desarrollo y/o desempeño de las actividades académicas [23].

La información obtenida de las bases de datos académicas puede servir para buscar solución a problemas como: estudiantes con bajo desempeño escolar, rezago estudiantil, deserción, etc.

Sin embargo, los coordinadores y responsables académicos de las instituciones educativas se ven rebasados por la cantidad de información almacenada y requieren la implementación de herramientas más sofisticadas para su análisis. Esta necesidad dio como resultado la creación del campo de la minería de datos educativa (Educational Data Mining, EDM, por sus siglas en inglés).

El EDM es el campo de estudio del DM en el entorno educativo, convierte los datos provenientes de los sistemas educativos en información útil. El EDM es un puente entre dos disciplinas: la educación por un lado y las ciencias computacionales por otro. Basado en un estudio realizado por Romero y Ventura en 2009 [24] las técnicas más comunes en el EDM son la regresión, el agrupamiento y la clasificación.

El campo de estudio del EDM es muy extenso, existe una gran variedad de aplicaciones que pueden estar orientadas a diferentes actores o, incluso, a diferentes sistemas educativos. La Figura 2.12 muestra un esquema de las posibles aplicaciones del DM en el entorno educativo, se puede observar que las aplicaciones pueden ir desde recomendar actividades a los estudiantes hasta proveer información a los administradores o académicos para mejorar los sistemas educativos.

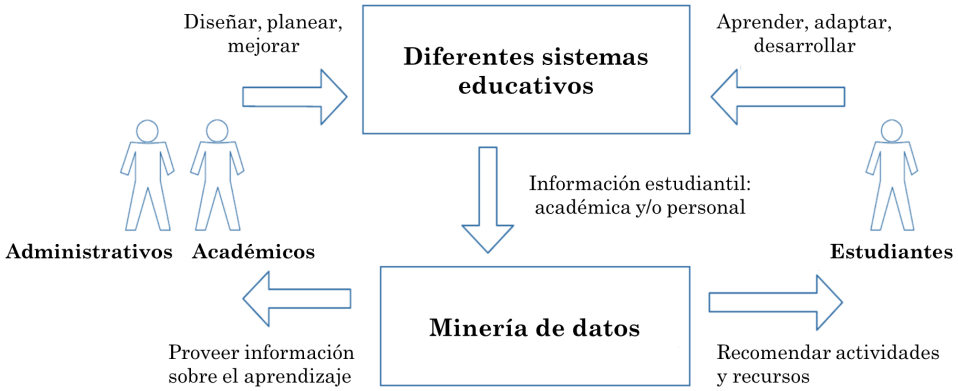

Figura 2.12. Aplicación del DM en el entono educativo, basada en [25].

Con referencia a los distintos sistemas educativos, el EDM puede ser aplicado a sistemas tradicionales o a educación a distancia. En los sistemas de educación tradicionales, los maestros tienen la capacidad de obtener información sobre el aprendizaje de los estudiantes mediante experiencias cara a cara, lo que permite evaluar continuamente sus métodos de enseñanza y obtener información sobre su desempeño académico. Sin embargo, los programas que trabajan en entornos electrónicos no tienen la misma facilidad y deben buscar otros medios para obtener

esta información. Afortunadamente, estas instituciones son las que almacenan mayor cantidad de datos acerca de los estudiantes y el entorno, por lo que el EDM se ha convertido en una herramienta fundamental para su estudio.

Por otro lado, las aplicaciones del EDM también pueden estar dirigidas a los diferentes actores del entorno educativo, como estudiantes, maestros, administradores e incluso investigadores. Los académicos y administradores son los responsables de diseñar, planear y mantener los sistemas educativos; sin embargo, los estudiantes son los que utilizan los sistemas e interaccionan con ellos. De acuerdo con la Figura 2.12, las aplicaciones del EDM pueden estar orientadas a diferentes actores:

- *Orientada a los estudiantes.* El objetivo es recomendar actividades y recursos que permitan mejorar su desempeño académico.
- *Orientada a los maestros.* El objetivo es ayudar a entender el proceso de aprendizaje de los estudiantes, y a mejorar sus métodos de enseñanza.
- *Orientada a los administradores.* El objetivo es proveer información significativa a las instituciones educativas para promover la mejora de los programas educativos.

Todos los casos de estudio mencionados previamente tienen diferentes objetivos, ningún estudio es igual, por eso es importante analizar el problema y elegir el método del EDM más adecuado.

2.3.1. Clasificación de las Aplicaciones de la Minería de Datos Educativa

Como se mencionó en la sección anterior, las aplicaciones del DM dentro del entorno educativo son muy variadas. Con el paso del tiempo los problemas de la educación que han sido abordados por el DM se han diversificado y con ello, han surgido diversos estudios que analizan y clasifican los trabajos desarrollados en esta área. A continuación, se presentan algunos estudios que demuestran como ha evolucionado la clasificación de estas aplicaciones con el tiempo, desde el año 2009 hasta el año 2018.

En 2009, en un trabajo realizado por Baker y Yacef, se propusieron cuatro áreas clave de

estudio en el EDM: modelos de los estudiantes, modelos de la estructura del conocimiento, modelos de apoyo pedagógico y perfeccionamiento de las teorías educativas [26].

- **Modelos de los estudiantes:** Representan información sobre las características de los estudiantes, conocimientos adquiridos, meta-cognición, motivación y actitudes. Estos modelos se han aplicado para predecir el fracaso estudiantil o la deserción.
- **Modelos de la estructura del conocimiento:** Mediante la combinación de modelos psicométricos y algoritmos de aprendizaje supervisado se han desarrollado modelos que permiten inferir el conocimiento de los estudiantes a partir de la evaluación de habilidades.
- **Modelos de apoyo pedagógico:** Descubrir cuáles tipos de apoyo pedagógico son más efectivos de acuerdo a la situación académica o personal de los estudiantes, o según las características del grupo de objetivo, ayuda a mejorar el aprendizaje.
- **Perfeccionamiento de las teorías educativas:** La búsqueda de información empírica para entender mejor el fenómeno educativo permite lograr un entendimiento más profundo sobre los factores clave que impactan el aprendizaje y diseñar mejores sistemas educativos.

En 2010, Romero y Ventura establecieron su propia categorización para las aplicaciones del EDM, ellos dividieron en 11 grupos los 300 artículos citados en su revisión, basándose en sus objetivos y técnicas empleadas [24]. La clasificación fue la siguiente: análisis y visualización de datos, retroalimentación al docente, recomendaciones para los estudiantes, predicción del desempeño estudiantil, modelos de los estudiantes, detección de conductas indeseables en los estudiantes, categorización de estudiantes, análisis de las redes sociales, desarrollo de mapas conceptuales, planeación y planificación, y construcción de *courseware*. La Figura 2.13 muestra las cinco categorías principales identificadas en el estudio y el número de artículos publicados de cada categoría hasta ese momento. A continuación, se describe brevemente cada una de estas categorías.

- **Análisis y visualización de datos:** El objetivo es resaltar la información importante y apoyar la toma de decisiones. La estadística y la visualización son las dos técnicas principales en esta área.

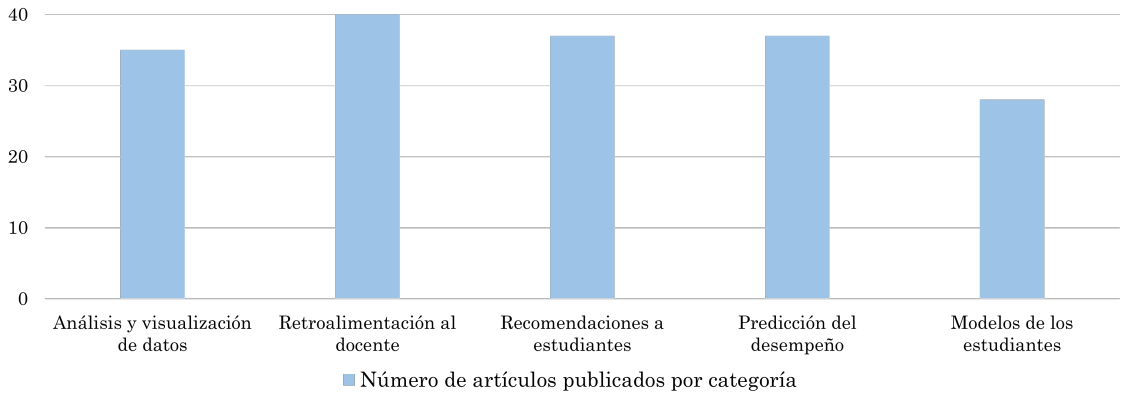

Figura 2.13. Principales categorías del EDM y número de artículos publicados hasta 2009, modificada de [24].

- **Retroalimentación al docente:** El objetivo es proveer información a los docentes y/o administradores de los cursos para que tomen mejores decisiones. Diversas técnicas del DM han sido empleadas, siendo las reglas de asociación las más comunes.
- **Recomendaciones para los estudiantes:** El objetivo es hacer recomendaciones a los estudiantes con respecto a sus actividades y también adaptar los contenidos de aprendizaje. Las técnicas más usadas en esta área son las reglas de asociación, el agrupamiento y los patrones secuenciales.
- **Predicción del desempeño estudiantil:** El objetivo es estimar el valor de las variables que describen al estudiante, que normalmente son el desempeño, el conocimiento y las calificaciones. Diversas técnicas has sido aplicadas, tales como: redes neuronales, métodos bayesianos, sistemas basados en reglas, regresión y análisis de correlación.
- **Modelos de los estudiantes:** El objetivo es desarrollar modelos cognitivos de los estudiantes, incluyendo modelos de sus habilidades y conocimiento declarativo. En esta área han sido aplicadas técnicas de aprendizaje supervisadas y no supervisadas.

En otra revisión sobre las aplicaciones del EDM realizada en 2018, Bakhshinategh et al. propusieron una nueva forma de clasificación [27]; identificaron 13 categorías, algunas de la cuales coinciden con las identificadas en estudios previos (Figura 2.14). Crearon una nueva taxonomía del EDM agrupando los trabajos de acuerdo a sus objetivos y métodos utilizados.

Esta taxonomía incluye 5 ramas de primer nivel: modelado del estudiante, apoyo a toma de

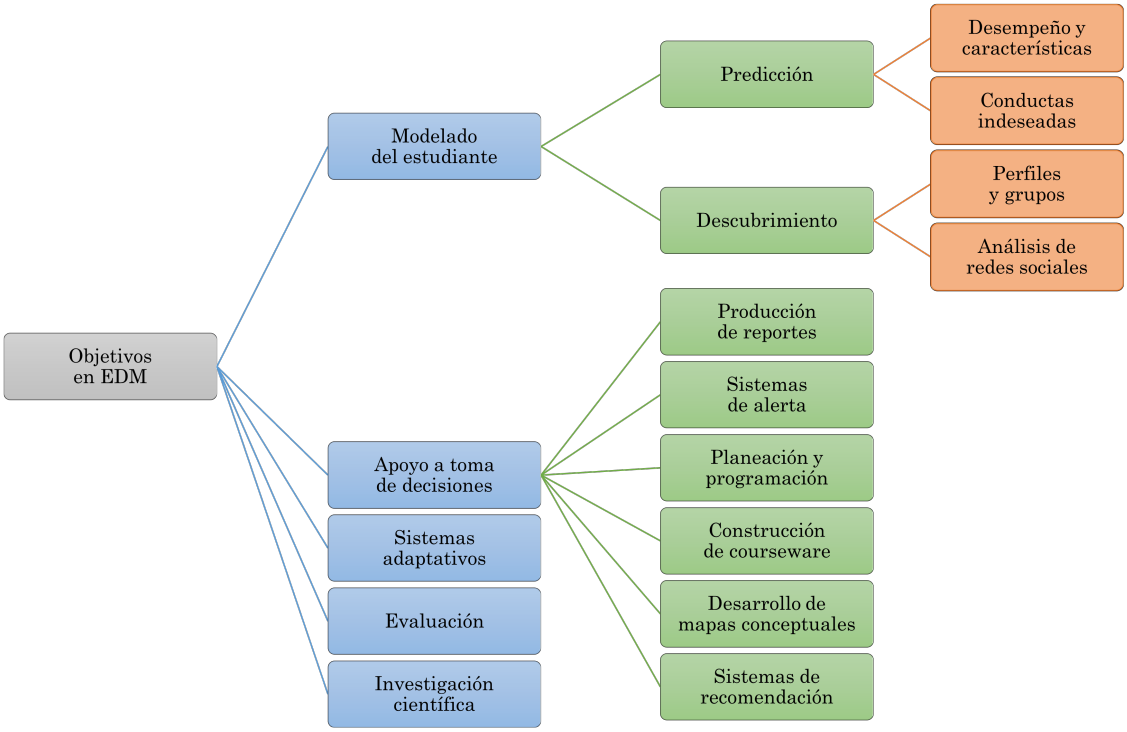

Figura 2.14. Taxonomía de las aplicaciones del EDM, traducida de [27].

decisiones, sistemas adaptativos, evaluación e investigación científica. El modelado del estudiante tiene dos subclasificaciones: predicción (desempeño y características, conductas indeseadas) y descubrimiento (perfiles y grupos, análisis de redes sociales). La rama de apoyo a toma de decisiones tiene seis subclasificaciones: producción de reportes, sistemas de alerta, planeación y programación, construcción de courseware, desarrollo de mapas conceptuales y sistemas de recomendación. Y finalmente, sin subclasificaciones, encontramos las ramas de sistemas adaptativos, evaluación e investigación científica.

Para entender mejor esta taxonomía, podemos observar el usuario objetivo en cada una de las clasificaciones. Los estudiantes se identifican como usuario objetivo en varias de ellas, sin embargo, la mayoría de las aplicaciones van dirigidas hacia los docentes y/o administrativos, ya que ellos son los responsables de crear, modificar y adaptar los sistemas y planes educativos, así como también de guiar a los estudiantes. La Tabla 2.1 presenta los posibles usuarios objetivo de cada aplicación.

Como se puede observar en los tres casos de clasificación presentados, el modelado de los estudiantes se identifica como área principal de estudio, dentro de la cual la predicción del desempeño

Tabla 2.1. Usuarios objetivo de las aplicaciones del EDM, traducida de [27].

	Estudiantes	**Docentes**	**Administradores**	**Investigadores**
Desempeño y características		x	x	
Detección de conductas indeseadas		x	x	
Perfiles y grupos de estudiantes	x	x		
Análisis de redes sociales		x	x	x
Producción de reportes	x	x	x	
Sistemas de alerta		x	x	
Planeación y programación	x	x	x	
Construcción de courseware		x		
Desarrollo de mapas conceptuales		x		x
Sistemas de recomendación	x	x		
Evaluación		x		
Sistemas adaptativos	x			
Investigación científica				x

y la deserción estudiantil resaltan por su importancia. Esto resulta relevante para nuestro trabajo, ya que, precisamente, en esta área de estudio se enfoca nuestro problema, predicción de la deserción escolar.

2.4. Trabajos Relacionados

La deserción escolar en el nivel superior se ha convertido en un tema de especial interés en los últimos años, no solo en nuestro país sino también en diversas partes del mundo. Estudios recientes se han basado en el DM para conocer los factores más importantes que influyen en la

Tabla 2.2. Información general sobre algunos trabajos relacionados.

Autores	**Año**	**Universidad**	**Tipo de datos**	**Número de registros**	**Exactitud máxima**	**Métodos**
Segura *et al.* [32]	2022	Madrid	Académicos Socioeconómicos	3428	85.59 % 88.72 %	SVM, DT, NN, KNN, LR
Opazo *et al.* [31]	2021	Chile	Académicos Socioeconómicos	5451	69 %	SVM, DT, NN, KNN, LR, RF NB, GBM
Yaacob *et al.* [29]	2020	Malasia	Académicos	64	90.8 %	DT, RF, NN, KNN, LR
Lázaro Álvarez *et al.* [30]	2020	Cuba	Académicos Socioeconómicos	546	60.53 % 85.08 % 96.49 %	DT, NN

deserción de los estudiantes. Una observación importante en este tipo de estudios, es que todos ellos tienen características diferentes, como el tipo de datos analizados, el número de registros, los métodos utilizados, etc. A continuación se presenta el resumen de algunos trabajos recientes realizados en esta área (ver Tabla 2.2).

En 2022, Segura y colaboradores [32] presentaron un estudio donde crearon modelos predictivos para el estudio de la deserción escolar en carreras de diversas áreas de la Universidad Complutense de Madrid. Los datos analizados corresponden a los alumnos de primer año de cinco áreas diferentes: Ciencias Sociales y Jurídicas, Ciencias, Ciencias de la Salud, Ingenierías, y Artes y Humanidades. Se realizaron dos experimentos, el primero de ellos se llevó a cabo con los datos de ingreso y el segundo, se realizó con la información correspondiente al término del primer semestre de la carrera. Se aplicaron cinco métodos de predicción: máquina de soporte vectorial, árboles de decisión, redes neuronales artificiales, k-vecinos más cercanos y regresión logística. Los atributos fueron agrupados en tres categorías: socio-económicos, de ingreso y académicos. Las categoría socio-económica incluye atributos como: género, edad, grado de estudios de los padres, municipio de residencia, etc. La segunda categoría considera: tipo de escuela (privada o pública), promedio de ingreso, número de preferencia para el área, etc. Y entre las variables del primer semestre se encuentran: porcentaje de materias aprobadas, el promedio del primer semestre, si el estudiante tiene un beca, etc. El conjunto de datos incluye la información de 3428 estudiantes de 10 carreras diferentes, de los cuales el 15.7 % deserta. Los resultados sugieren que los datos de ingreso no son suficientes para la predicción de la deserción, sin embargo, los resultados mejoran considerablemente al utilizar los datos del primer semestre de la carrera. La exactitud global media (incluyendo ambas clases) para el experimento realizado con los datos de ingreso es de aproximadamente 83.8 %, y para el experimento realizado con los datos del primer semestre es de 86.7 %; sin embargo, la exactitud de la clase que deserta tiene valores extremadamente bajos para ambos experimentos, casi todos por debajo del 30 %, lo que indica que los modelos clasifican incorrectamente a la mayoría de los estudiantes de esta clase. Por otro lado, los métodos de k-vecinos más cercanos y la regresión logística fueron los que generaron los mejores resultados. Para finalizar, en dicho estudio se identifican al promedio del primer semestre y al porcentaje de materias aprobadas, como las variables más importantes para la predicción de la deserción escolar, sin embargo, existen otros factores importantes, como la preferencia de los estudiantes

por el área en el que fueron admitidos.

En 2021, Opazo y colaboradores [31] presentaron un estudio sobre la predicción de la deserción escolar de estudiantes de ingeniería de dos universidades de Chile utilizando los datos de ingreso de los estudiantes. Los datos analizados corresponden a estudiantes del primer año de ingeniería de la Universidad Adolfo Ibáñez (UAI) y de la Universidad de Talca (U Talca) de Chile. La UAI tiene una deserción escolar del 12 % en el primer año de la carrera, mientras que la U Talca, tiene un 15 %. Para la construcción de los modelos predictivos se aplicaron ocho métodos diferentes: k-vecinos más cercanos, máquina de soporte vectorial, árboles de decisión, bosque aleatorio, potenciación del gradiente, Naïve Bayes, regresión logística y redes neuronales artificiales. El conjunto de datos analizado incluyó 5951 registros y 14 variables, de estos registros, 3750 pertenecen a estudiantes de ingeniería de la UAI, y el resto de la U Talca. Algunas de los atributos empleados en la predicción fueron: año de ingreso, género, tipo de escuela (privada, pública, subsidiada), promedio del bachillerato, promedio de matemáticas, promedio de lenguaje, puntuación de las pruebas nacionales, lugar de origen y lugar de residencia. Los resultados indicaron que el método que alcanzó la mejor predicción fue la potenciación de gradiente, obteniendo una exactitud de 69 %, mientras que los métodos de regresión logística y k-vecinos más cercanos, obtuvieron la menor exactitud, con un valor de 62 %. Adicionalmente, los modelos obtenidos permitieron identificar que el promedio obtenido en matemáticas en el examen nacional de ingreso a la universidad fue el más representativo en ambas universidades.

Incluso, existen estudios específicos sobre la deserción escolar en carreras del área de Ciencias de la Computación. Por ejemplo, un estudio presentado en 2020 por Yaacob y colaboradores [29], utiliza técnicas de DM para predecir la deserción escolar de estudiantes de la carrera de Ciencias de la Computación de la *Universiti Teknologi MARA* de Malasia. El estudio analiza los datos de los estudiantes al término del tercer año de la carrera y se utilizan cinco métodos de clasificación para la predicción. Los métodos aplicados fueron: árboles de decisión, regresión logística, bosque aleatorio, k-vecinos más cercanos y redes neuronales. El conjunto de datos analizado contiene la información de 64 estudiantes inscritos en el año 2016. Los atributos considerados fueron: promedio general, género y la calificación de 24 materias, la mayoría de ellas pertenecientes al área de Matemáticas, Tecnologías de la Información y Computación. Los modelos obtenidos alcanzaron una exactitud de entre 81.2 % y 90.8 % al realizar las predicciones, la regresión logística fue el

algoritmo con mayor exactitud, mientras que el modelo construido con árboles de decisión fue el que obtuvo la menor exactitud. Además, identificaron cuatro materias principales relacionadas con la deserción escolar, dos del área de Matemáticas (Matemáticas Discretas y Cálculo I) y dos del área de Computación (Programación Orientada a Objetos y Fundamentos de Estructuras de Datos).

Otro estudio realizado en 2020 sobre la deserción escolar en el área de Ciencias de la Computación fue el de Lázaro Álvarez y colaboradores [30], en el cual realizaron la predicción de la deserción de estudiantes de la carrera de Ingeniería en Informática en Cuba. La predicción se realizó en tres momentos diferentes de la carrera: al inicio (con los datos de ingreso), al término del primer semestre y al término del primer año. Los datos analizado incluyen la información de 456 estudiantes y se aplicaron dos métodos para la predicción, árboles de decisión y redes neuronales. De los 456 estudiantes, 47.4 % se graduaron al término del quinto año, 36.2 % desertaron en algún momento de la carrera y el resto de ellos no se había graduado al término del quinto año. La clasificación dividió al conjunto de los estudiantes en tres clases: aprobados (*promotion*), reprobados (*repetition*) y los que desertan (*dropout*). Los estudiantes que reprueban dos o más materias en el mismo año deben repetir el año académico, pueden repetir hasta dos años durante la carrera, pero solo pueden repetir el mismo año una vez. Los atributos que integran el conjunto de datos de ingreso incluyen datos personales como género y provincia; y académicos, como tipo de bachillerato, la calificación obtenida en matemáticas en el examen de admisión, el promedio del bachillerato y si la carrera de Ingeniería en Informática era su primera opción o no. Los atributos considerados al término del primer semestre y del primer año fueron las calificaciones de las materias de Matemáticas y Programación, y el porcentaje de materias aprobadas (general y por área). Los resultados obtenidos indicaron una diferencia considerable entre los modelos construidos con los datos de ingreso y con los datos del primer año de la carrera. Los modelos construidos con los datos de ingreso alcanzaron una exactitud aproximada de 60 %, mientras que los modelos correspondientes al primer semestre tuvieron una exactitud media de 84.9 % y los modelos construidos con los datos acumulados hasta el primer año lograron una exactitud media de 93.2 %. Respecto a los distintos métodos aplicados, no hubo una diferencia notable entre ellos, únicamente en los modelos construidos con los datos del primer año se pudo observar una diferencia cercana al 7 % entre la exactitud de ambos modelos, siendo superior la

de redes neuronales. Los árboles de decisión permitieron identificar que el porcentaje de materias aprobadas es el atributo académico más importante, mientras que para el conjunto de datos de ingreso, la provincia resulta el atributo principal.

2.5. Resumen

En este capítulo se presentó la toda la información necesaria para entender el entorno completo del problema de estudio, desde sus antecedentes y trabajos relacionados, hasta la parte computacional y herramientas de trabajo. El EDM es el campo de estudio del DM en el entorno educativo, dentro de sus principales aplicaciones se encuentra la predicción de la deserción escolar. Para el estudio de este problema se han utilizado diversos métodos de aprendizaje supervisados, tales como los árboles de decisión, bosque aleatorio, Naïve Bayes, máquina de soporte vectorial, redes neuronales artificiales, entre otros. A partir del trabajo relacionado es posible reconocer algunas características importantes en este tipo trabajos, características comunes, tales como los métodos y las métricas de evaluación utilizados, y características específicas, como el tipo de datos y el número de registros.

Capítulo 3

Preprocesamiento de Datos Académicos

De acuerdo con el proceso del descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD, por sus siglas en inglés) presentado en la Sección 2.1, las etapas previas a la minería de datos (Data Mining, DM) son: la selección, el preprocesamiento y la transformación del conjunto de datos. Estas tres etapas constituyen la primera parte del trabajo realizado en este estudio, el desarrollo de cada una de ellas se presenta a continuación.

Una parte fundamental previa al DM es conocer el conjunto de datos en cuestión, para ello es necesario examinar los datos recolectados antes de aplicar cualquier algoritmo de clasificación, así podremos reconocer las características de los elementos, identificaremos valores nulos o incompletos, sabremos el tamaño y dimensiones del conjunto, etc., lo cual es determinante para poder manipular los datos adecuadamente y elegir los métodos de predicción adecuados. Lo primero que debemos hacer es elegir un conjunto de datos (recolección de datos), después manipulamos este conjunto de datos para darle la estructura necesaria de acuerdo a nuestro estudio (preprocesamiento de datos) y como último paso, hacemos un análisis rápido de los datos para conocer qué información es capaz de proporcionarnos antes de aplicar cualquier algoritmo (análisis exploratorio).

3.1. Selección del Conjunto de Datos

El Área de Ciencias de la Computación de la Facultad de Ingeniería de la Universidad Autónoma de San Luis Potosí tiene 3 carreras: Ingeniería en Computación, Ingeniería en Informática e Ingeniería en Sistemas Inteligentes, y así como muchas instituciones educativas de nivel superior, enfrentan el problema de la deserción escolar. Este trabajo se enfocará únicamente en analizar la deserción escolar de las carreras de Ingeniería en Computación e Ingeniería en Informática, ya que la carrera de Ingeniería en Sistemas Inteligentes aún no cuenta con suficientes generaciones egresadas ni con la información necesaria para generar un conjunto de datos comparable con las otras dos carreras.

Se inició con la elección de los datos. Se seleccionaron dos conjuntos de datos compuestos por los datos académicos de los estudiantes de las carreras de Ingeniería en Computación e Ingeniería en Informática de las generaciones 2008 a 2013. Estas generaciones fueron seleccionadas ya que se consideró que la información de seis generaciones era suficiente para realizar este estudio, no se emplearon generaciones posteriores puesto que al momento de la recolección de los datos (año 2020) las generaciones posteriores a la 2013 no contaban con suficientes egresados. El primer conjunto de datos corresponde a los datos obtenidos en el proceso de admisión de los estudiantes y el segundo conjunto se conforma por el historial académico de los estudiantes durante la carrera (también conocido como kardex en la Facultad de Ingeniería).

Cabe señalar que los dos conjuntos de datos recolectados contienen únicamente datos académicos. Esta observación es importante ya que en muchas ocasiones la deserción escolar está relacionada con factores personales y la falta de esta información puede disminuir la exactitud de las predicciones.

3.1.1. Descripción del Conjunto de Datos del Proceso de Admisión

El conjunto de datos del proceso de admisión cuenta con 982 registros, de los cuales 537 son de la carrera de Ingeniería en Computación y 445 de la carrera de Ingeniería en Informática. Los datos están distribuidos como muestra la Figura 3.1.

Cada uno de estos registros cuenta con 9 atributos, representados por las columnas del con-

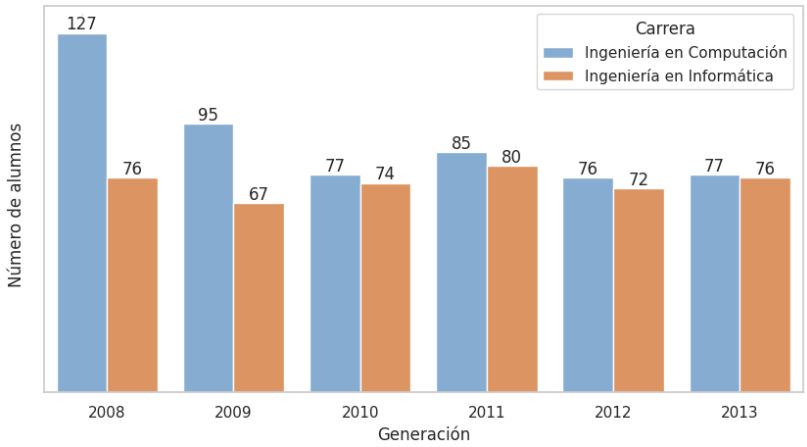

Figura 3.1. Distribución de los datos del proceso de admisión de acuerdo a la generación y carrera.

junto de datos. Los atributos son: clave de la UASLP, clave de la Faculta de Ingeniería (FI), generación, nombre, carrera, resultado general del examen de admisión, resultado del examen psicométrico, resultado del examen de conocimientos, y resultado del examen EXANI-II. La Tabla 3.1 muestra una breve descripción de estos atributos.

Tabla 3.1. Descripción de los atributos del examen de admisión.

Atributo	Descripción
Clave UASLP	Clave única asignada por la UASLP
Clave FI	Clave asignada por la Facultad de Ingeniería
Generación	Año de ingreso
Nombre	Apellido del estudiante
Carrera	Carrera a la que pertenece
Admisión total	Promedio general obtenido en el proceso de admisión
Psicométrico	Nivel obtenido en el examen psicométrico
Conocimientos	Puntuación obtenida en el examen de admisión de la Facultad
EXANI-II	Puntuación obtenida en el examen EXANI-II

Clave UASLP: Es una clave única de 6 dígitos asignada por la UASLP a cada estudiante.

Clave FI: Es una clave única de 12 dígitos asignada por la Facultad de Ingeniería a cada estudiante. Esta clave se integra por varios elementos: los primeros cuatro dígitos representan la generación, el siguiente dígito indica el género, después la clave de la carrera, a continuación la referencia de ingreso del estudiante (por ejemplo, si es trámite directo, reacomodo, cambio de carrera, etc.), y los últimos dígitos representan el lugar obtenido por el estudiante durante el proceso de admisión (con respecto a todos los aspirantes de la Facultad).

Generación: Es año de ingreso del estudiante a la Facultad.

Nombre: Es el nombre de los estudiantes. Por cuestiones de confidencialidad el conjunto de datos cuenta únicamente con el apellido paterno de los estudiantes.

Carrera: Es la carrera a la que pertenece el estudiante, Ingeniería en Computación o Ingeniería en Informática.

Admisión total: Es el promedio general obtenido por los estudiantes en el proceso de admisión, este valor se obtiene de una combinación de los 3 exámenes aplicados. Se calcula de acuerdo a las siguientes ponderaciones: 15 % del examen psicométrico, 40 % del EXANI-II y 45 % del examen de conocimientos.

Psicométrico: Es el nivel obtenido en el examen psicométrico aplicado por la UASLP en el proceso de admisión. Los estudiantes son clasificados en 6 niveles de acuerdo con la puntuación obtenida (I, II, III+, III-, IV y V).

Conocimientos: Es la puntuación obtenida en el examen de conocimientos de la Facultad aplicado en el proceso de admisión. Las materias evaluadas en este examen son Matemáticas, Física y Química. Los valores de este atributo pueden tener un mínimo de 0 y un máximo de 100.

EXANI-II: Es la puntuación obtenida en el Examen Nacional de Ingreso a la Educación Superior (EXANI-II) del Centro Nacional de Evaluación para la Educación Superior A.C. (Ceneval[1]) aplicado por la UASLP durante el proceso de admisión. El EXANI-II incluye las áreas de razonamiento lógico-matemático, razonamiento verbal, matemáticas, español, y tecnologías de información y comunicación. Los valores de este atributo también están entre 0 y 100.

Los datos obtenidos en el proceso de admisión podrían ser representativos para la identificación de alumnos con posible riesgo de deserción, por esta razón es importante tener un panorama general de los datos contenidos en este conjunto. El promedio general obtenido, *Admisión general*, es importante, ya que engloba la información de los otros tres atributos. La Figura 3.2a muestra el histograma de los valores obtenidos para este atributo. El valor mínimo encontrado

[1]https://ceneval.edu.mx/

es de 31.94 puntos y el máximo de 80.46 puntos, con una media de 49.31 puntos. El histograma de los resultados de cada uno de los tres exámenes también se presentan en la Figura 3.2. En la Figura 3.2b podemos observar que los valores del atributo *Psicométrico* son categóricos, la categoría con mayor frecuencia es el nivel III+, seguida del nivel III-. En la Figura 3.2c se puede ver el histograma del atributo *Conocimientos*, este atributo tiene su valor máximo en 71 y el mínimo en 18, siendo su media igual a 33 puntos. Por último, la Figura 3.2d presenta el histograma del atributo *EXANI-II*, con un valor máximo de 93 puntos, un valor mínimo de 33 puntos y una media de 61 puntos.

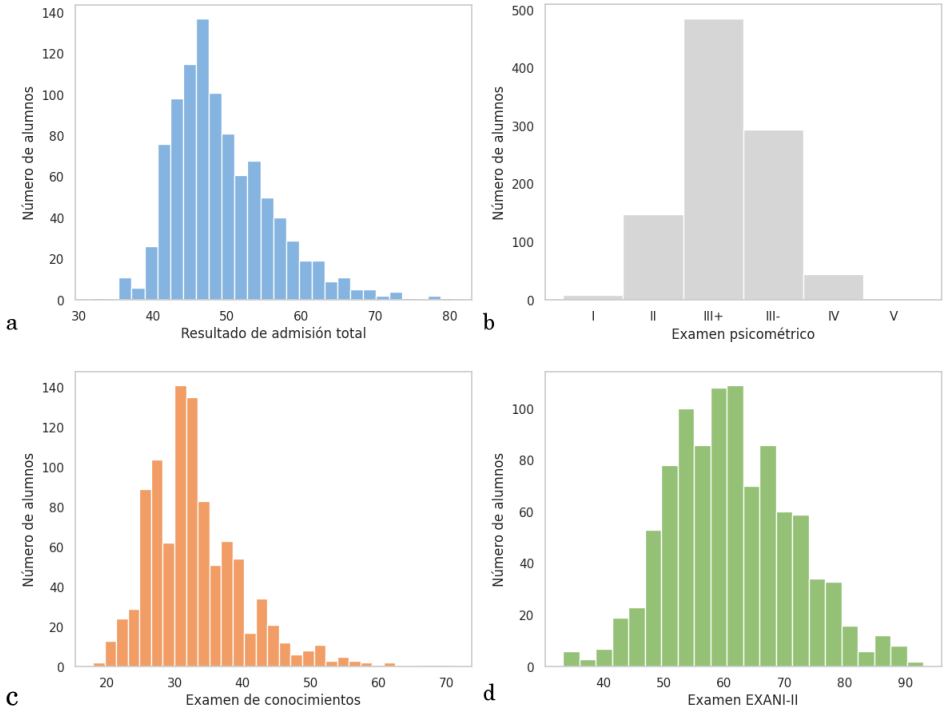

Figura 3.2. Histogramas de los datos obtenidos en el proceso de admisión. a) Resultado general del examen de admisión. b) Examen psicométrico. c) Examen de conocimientos. d) Examen EXANI-II.

Una observación importante respecto a este conjunto de datos, es que es fácil de describir y representar debido a que es bastante homogéneo, es decir, toda la información tiene la misma estructura, nomenclatura, etc. No requiere manipulación adicional para conocer la información contenida.

3.1.2. Descripción del Conjunto de Datos del Kardex

El segundo conjunto de datos recolectado corresponde al kardex de los estudiantes, aquí se almacenan los datos académicos generados desde el momento de su ingreso a la Facultad. El conjunto cuenta con la información de 939 estudiantes (509 de la carrera de Ingeniería en Computación y 430 de Ingeniería en Informática) y tiene 13 atributos, algunos de ellos coinciden con los atributos encontrados en el conjunto de datos del proceso de admisión y descritos en la Sección 3.1.1. El número de registros en este conjunto de datos es menor que en el descrito en la Sección 3.1.1 ya que son conjuntos de datos independientes y existen movimientos académicos que podrían verse reflejados únicamente en uno de los conjuntos de datos. Los 13 atributos que describen a los estudiantes son: clave de la UASLP, clave FI, generación, nombre, clave carrera, carrera, clave materia, materia, calificación, fecha de calificación, tipo de examen, semestre cursado y situación actual. La Tabla 3.2 presenta una breve descripción de cada atributo y a continuación se detallan aquellos que no han sido descritos previamente.

Tabla 3.2. Descripción de los atributos del kardex.

Atributo	Descripción
Clave UASLP	Clave única asignada por la UASLP
Clave FI	Clave asignada por la Facultad de Ingeniería
Generación	Año de ingreso
Nombre	Apellido del estudiante
Clave carrera	Clave de la carrera a la que se inscribe
Carrera	Carrera a la que pertenece
Clave materia	Clave de cada materia inscrita
Materia	Nombre de cada materia inscrita
Calificación	Calificación de cada materia inscrita
Fecha de calificación	Fecha de registro de la calificación de cada materia inscrita
Tipo de examen	Tipo de examen en que se presenta cada materia inscrita
Semestre cursado	Número de semestre en que se presenta cada materia inscrita
Situación actual	Ultima situación académica registrada del estudiante

Clave carrera: Es la clave de la carrera a la que pertenece, es una clave de dos dígitos asignada por la Facultad.

Clave materia: Es la clave de cada materia inscrita. Es una clave asignada por la Facultad y cada materia tiene una clave diferente.

Materia: Es el nombre de cada materia inscrita por el estudiante. Este atributo puede aparecer

repetido en el registro de un estudiante cuando el estudiante haya presentado la misma materia en examen ordinario, luego examen extraordinario o título, etc; la misma materia aparecerá varias veces en el kardex del estudiante.

Calificación: Es la calificación obtenida por el estudiante en cada materia inscrita. Este atributo no incluye únicamente valores numéricos, existen registros con valores categóricos, tales como: SA (sin asistencia), LR (laboratorio reprobado), AC (acreditado), ET (examen a título), NP (no presentó), ER (examen a regularización), etc. Los registros que sí tienen valores numéricos se encuentran entre 0 y 100.

Fecha de calificación: Es la fecha en que se registra en el sistema la calificación de cada materia inscrita.

Tipo de examen: Es el tipo de examen en que se presenta cada materia inscrita. Este atributo nos dice mediante qué tipo de examen fue obtenida la calificación registrada, si fue en examen ordinario (EO), examen extraordinario (EE), examen de regularización (ER), revalidación (RV), intersemestral ordinario (IO), etc.

Semestre cursado: Es el número de semestre en que se presenta cada materia inscrita. Cada calificación que se registra lleva asociado el número de semestre que cursa el estudiante en ese momento, este atributo es importante, ya que además de decirnos en qué semestre presentó cada materia, nos permite identificar cuál fue el último semestre cursado/inscrito por el estudiante. El plan de estudios de las carreras de Ingeniería en Computación e Ingeniería en Informática comprende 10 semestres, sin embargo, existe la posibilidad de que los estudiantes dispongan de más semestres para terminar. Los valores registrados para este atributo van desde el semestre 1 hasta el semestre 23.

Situación actual: Es la ultima situación académica registrada del estudiante. Este atributo nos permite conocer cuál fue la situación académica en la que el estudiante dejó la Facultad o si aún continúa. Las posibles categorías de este atributo son: inscrito, no inscrito, baja definitiva, baja temporal, baja académica, titulado, pasante y cambio de facultad. Cabe mencionar que la baja definitiva se presenta cuando el estudiante solicita su baja de la Facultad, mientras que la baja académica surge cuando el estudiante es dado de baja por la Facultad debido al incumplimiento de algún requisito académico. Este atributo es

particularmente importante para nuestro estudio, ya que almacena la información necesaria para clasificar a los estudiantes en dos clases: los estudiantes que terminan y los estudiantes que desertan. Por lo tanto, se convertirá en nuestro atributo objetivo.

Este conjunto de datos resulta particularmente difícil de describir e interpretar en su formato original debido a la heterogeneidad en algunos de sus atributos y también debido a que los registros están dados de manera continua y sin agrupar, por esta razón, no se muestran descripciones adicionales. Para poder describir sus características es necesario preprocesarlo, limpiarlo, homogeneizarlo y darle una estructura adecuada para nuestro estudio; este proceso completo se presenta a continuación.

3.2. Preprocesamiento del Conjunto de Datos

La etapa del preprocesamiento de los datos es fundamental ya que al final de esta etapa tendremos los datos listos para su análisis, si esta etapa no se realiza adecuadamente se podrían tener conflictos en etapas posteriores o podría haber pérdida de información. El objetivo del preprocesamiento de los datos es obtener un conjunto de datos de calidad y con características útiles para la extracción del conocimiento.

El preprocesamiento de los datos conlleva varias etapas, en este trabajo se realizaron cuatro pasos para la preparación de los datos: limpieza, transformación, reducción e integración del conjunto de datos (Figura 3.3).

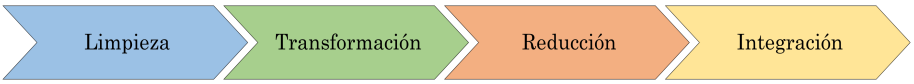

Figura 3.3. Etapas del preprocesamiento de los datos.

3.2.1. Limpieza

Es muy frecuente que los conjuntos de datos recolectados contengan datos con formatos no permitidos, valores nulos o faltantes, o haya ruido que interfiera con el análisis correcto de los datos. La limpieza de los datos ayuda a eliminar estos problemas, implica la rectificación de datos incorrectos, identificación de datos incompletos, etc.

Así, el primer paso que se realizó después de haber recolectado el conjunto de datos, fue la búsqueda de datos incompletos. En el conjunto de datos del proceso de admisión se encontraron 4 registros con los atributos vacíos, por lo que fueron eliminados. En el conjunto del kardex, se identificaron 531 renglones sin calificación de la materia, estos renglones fueron eliminados ya que no es posible conocer o calcular el valor faltante.

Estos fueron los únicos datos incompletos identificados, además de ellos, no se detectó ningún otro dato con formato incorrecto o incompleto, así que se procedió a la siguiente etapa.

3.2.2. Transformación

La segunda etapa del preprocesamiento consistió en la transformación del conjunto de datos. La transformación es la parte más larga de la preparación de los datos, implica la modificación de los atributos, creación de nuevos atributos, agrupación de datos, condensación de datos, normalización, generalización, etc.

Transformación del Conjunto de Datos del Proceso de Admisión

El conjunto de datos del proceso de admisión es muy homogéneo y no hay información adicional que se desee obtener a partir de los atributos existentes, por lo que la transformación de este conjunto fue rápida.

El único atributo que se modificó en este conjunto de datos fue el examen psicométrico. Existen algunos algoritmos de DM que no aceptan datos categóricos como entrada, y como se mencionó en la Sección 3.1.1, este atributo tiene valores categóricos (I, II, III+, III-, IV y V), por lo tanto, es necesario modificar estos valores y convertirlos en numéricos. El promedio general del proceso de admisión se calcula considerando los siguientes porcentajes: 15 % del examen psicométrico, 40 % del EXANI-II y 45 % del examen de conocimientos; por lo tanto, es posible conocer el valor numérico de este atributo realizando los cálculos pertinentes. Al llevar a cabo esta transformación se encontró que el valor máximo de este atributo es de 136 puntos y el mínimo de 66 puntos, la Figura 3.4 muestra el histograma correspondiente.

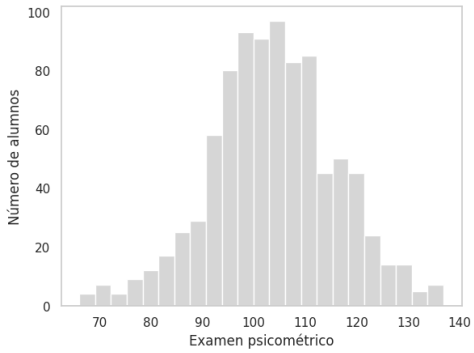

Figura 3.4. Histograma del atributo *Psicométrico* transformado.

Transformación del Conjunto de Datos del Kardex

El conjunto de datos del kardex contiene más atributos, más datos y es más heterogéneo, por lo tanto, su preparación fue mucho más complicada e implicó mucho más tiempo. Se transformaron varios atributos originales y se crearon nuevos atributos a partir de los existentes. Además, se agruparon los datos de cada estudiante y se clasificaron por semestre, esto con la finalidad de dar una estructura útil al conjunto de datos. A continuación, se describe la transformación realizada a cada uno de los atributos y la creación de nuevos atributos relacionados con ellos.

Clave UASLP

La clave única asignada por la UASLP sirvió para agrupar la información de los estudiantes. Los datos de cada estudiante fueron agrupados usando esta clave como referencia y se convirtió en el índice para los registros del conjunto de datos.

Materia

Con relación al atributo *Materia* se aplicaron varias transformaciones al conjunto de datos. Primero, se creó un diccionario donde se asoció cada materia con el número de créditos que aporta al ser acreditada. Existen materias que no aportan créditos al kardex, los estudiantes deben acreditarlas, sin embargo, no llevan asociado ningún valor numérico de créditos (como las materias de Inglés y los Seminarios); estos registros fueron eliminados ya que al no tener un valor numérico generarían un conjunto heterogéneo y podrían causar problemas posteriores. Con este diccionario se creó un nuevo atributo llamado *Créditos*, dicho atributo representa el número de créditos obtenidos por cada materia acreditada, y al ser sumados, se puede conocer el número de créditos acumulados por semestre o por año. Adicionalmente,

se creó un atributo derivado del atributo *Créditos*, este nuevo atributo establece si los estudiantes acumulan 45 créditos o no al término del primer año de la carrera. Dicho atributo se denominó *¿Cumple?* y sus valores son 0 y 1, el 0 indica que el estudiante no cumple con 45 créditos al término del primer año, mientras que el 1, representa que sí cumple con 45 créditos. En la sección 3.3 se explica la importancia de este atributo.

Por otro lado, este atributo *Materia* también se utilizó para agrupar algunas materias por área. Para analizar la posible influencia de algunas áreas académicas especificas en la deserción escolar se crearon dos diccionarios, uno donde se incluyeron las materias del Departamento de Físico-Matemáticas y otro para las materias del área de Programación. El Departamento de Físico-Matemáticas (DFM) de la UASLP apoya a la Facultad de Ingeniería al impartir las materias de las áreas de Cálculo, Álgebra, Física y Química que constituyen el tronco común de las carreras. Las materias del área de Programación son aquellas materias del plan de estudios relacionadas con programación, estas materias fueron elegidas ya que se ha observado que con frecuencia los estudiantes presentan problemas para aprobar estas materias. La Tabla 3.3 muestra las materias agrupadas en cada diccionario y el semestre al que pertenecen.

Tabla 3.3. Materias del Departamento de Físico-Matemáticas y del área de Programación.

Semestre	Departamento de Físico-Matemáticas	Materias de Programación
1	Álgebra A	Introducción a la programación
	Física A	
	Química A	
2	Álgebra B	Estructura de datos y algoritmos A
	Cálculo A	
3	Cálculo C	Estructura de datos y algoritmos B
	Cálculo D	
4		Programación orientada a objetos
5		Programación visual
6		Grafos
7		Compiladores e intérpretes A
8		Compiladores e intérpretes B

Calificación

Este atributo tuvo dos papeles fundamentales en el estudio: diferenciar las materias aproba-

das de las reprobadas y obtener calificaciones numéricas de las materias. Como se mencionó en la Sección 3.1.2, este atributo contiene valores numéricos y categóricos, debido a ello fue un poco complicada su manipulación y hubo la necesidad de crear nuevos atributos para su reemplazo. La primera transformación consistió en crear un nuevo atributo, *Situación materia*, con dos posibles valores: Aprobada o Reprobada. Este nuevo atributo permitió asociar el número de créditos a las materias aprobadas y también ayudó a determinar fácilmente el número de materias aprobadas.

Con relación a las calificaciones numéricas, es importante destacar que todas las materias aprobadas tienen calificación numérica, son algunas materias reprobadas las que presentan valores categóricos (EE, ET, ER, etc.). Con esta información, se creó un nuevo atributo llamado *Calificación final*, este atributo contiene únicamente valores numéricos, los valores categóricos fueron reemplazados por ceros para poder ser diferenciados y excluidos mediante condicionales al momento de extraer las calificaciones numéricas.

Habiendo obtenido las calificaciones numéricas, fue posible calcular promedios. Se crearon tres nuevos atributos calculando diversos promedios: *Promedio general*, *Promedio programación*, *Promedio DFM*.

Tipo de examen

El tipo de examen se utilizó para identificar el número de materias inscritas. Como se mencionó en la Sección 3.1.2, el atributo *Materia* puede tener valores repetidos si la materia se reprobó, entonces, para poder conocer el número de materias inscritas por el estudiante cada semestre se consideraron solamente aquellas materias cuyo tipo de examen fuera EO (examen ordinario) o IO (intersemestral ordinario).

Semestre cursado

Este atributo tuvo dos funciones en nuestro trabajo, ayudar a clasificar a las materias por semestre y conocer el último semestre cursado por cada estudiante. Para los estudiantes en situación de deserción, el último semestre registrado fue considerado como el último semestre cursado, a partir de ese semestre se consideró como baja. En relación con la clasificación, las materias de cada estudiante fueron agrupadas por semestre para poder analizar la información por periodos específicos, esto facilitó su manipulación y permitió hacer cálculos particulares.

Tras haber calculado los promedios previamente mencionados (*Promedio general*, *Promedio programación* y *Promedio DFM*), conociendo el número de créditos (atributo *Créditos*) y habiendo agrupado las materias por semestre, fue posible crear cuatro atributos más: *Rendimiento general*, *Rendimiento programación*, *Rendimiento DFM* y *Efectividad*.

El atributo *Rendimiento general* indica cual es el desempeño general del estudiante en función del promedio general obtenido y la razón de los créditos acumulados. Este valor se calculó usando los datos acumulados hasta el semestre en cuestión, considerando 45 créditos promedio por semestre. Este atributo tiene valores entre 0 y 10. Los cálculos se realizaron utilizando la Ecuación 3.1.

$$\textit{Rendimiento general} = \text{Promedio general} * \frac{\text{Número de créditos}}{45 * \text{Número de semestre}} \tag{3.1}$$

Los atributos *Rendimiento programación* y *Rendimiento DFM* representan el desempeño de los estudiantes en cada una de estas áreas. El valor de estos atributos se obtiene al multiplicar el promedio de las materias del área correspondiente por la razón del número de materias inscritas (también del área correspondiente). Los valores de estos atributos se obtienen con la Ecuación 3.2.

$$\textit{Rendimiento por área} = \text{ Promedio del área } * \frac{\text{Materias inscritas del área}}{\text{Materias totales del área}} \tag{3.2}$$

El atributo *Efectividad* es una relación entre las materias aprobadas y las materias inscritas por el estudiante. En los primeros semestres puede no haber mucha diferencia entre este atributo y el rendimiento general, ya que el primer semestre todos los estudiantes inscriben obligatoriamente el mismo número de materias, pero en semestres posteriores cada estudiante decide cuáles materias inscribir. Este atributo podría reflejar condiciones que otros atributos no, como por ejemplo, podría medir de cierto modo el compromiso del estudiante, tal vez por condiciones personales no puede inscribir la carga académica completa, sin embargo, cumple con las que puede/decide inscribir. El valor de este atributo se calcula con la Ecuación 3.3.

$$\textit{Efectividad} = \frac{\text{Número materias aprobadas}}{\text{Número de materias inscritas}} \tag{3.3}$$

Situación actual

Como se mencionó en la Sección 3.1.2, este atributo se transformará para convertirse en el atributo objetivo. El objetivo de este trabajo es el análisis de la deserción escolar, por lo tanto, debemos clasificar a los estudiantes en dos clases: los estudiantes que terminan la carrera y los estudiantes que desertan, este atributo nos permitirá crear dicha clasificación. En primer lugar, se identificaron los casos con situación académica indefinida para nuestro objetivo, estos fueron los estudiantes con situación de inscrito, baja temporal o cambio de facultad. Los registros con estas condiciones fueron eliminados, pues no sabemos con certeza si terminarán la carrera o no, se encontraron 38 registros: 20 inscritos, 11 con baja temporal y 7 de cambio de facultad. El resto de los estudiantes fueron clasificados con la creación de un nuevo atributo denominado *¿Termina?*. Los estudiantes con la situación de titulado o pasante fueron clasificados como positivos (es decir, sí terminan) y se les asignó el valor de 1. Los estudiantes con situación de no inscrito, baja definitiva o baja académica, fueron clasificados como negativos (no terminan) y se les asignó el valor de 0. De esta manera, el atributo *¿Termina?* se convierte en nuestro atributo objetivo, teniendo dos posible valores: 0 (deserta) y 1 (termina).

Con esto llegamos al final de la transformación de nuestros datos, en este punto el conjunto de datos tiene una estructura más apropiada para nuestro estudio y se han creado nuevos atributos, sin embargo, aún hace falta reducir el conjunto para eliminar los datos que ya no son relevantes o que fueron reemplazados por otros.

3.2.3. Reducción

El objetivo de la reducción del conjunto de datos consiste en obtener una representación reducida de los datos originales, pero manteniendo su estructura fundamental e integridad. Este paso es opcional, sin embargo, es recomendable realizarlo, pues reducir el tamaño del conjunto puede agilizar el proceso de los algoritmos de DM. Dos prácticas comunes dentro de la reducción es la selección de atributos y la selección de instancias (Figura 3.5). La selección de atributos permite eliminar atributos irrelevantes o redundantes, mientras que la selección de instancias consiste en tomar un subconjunto del conjunto original que permita representar la información del conjunto original.

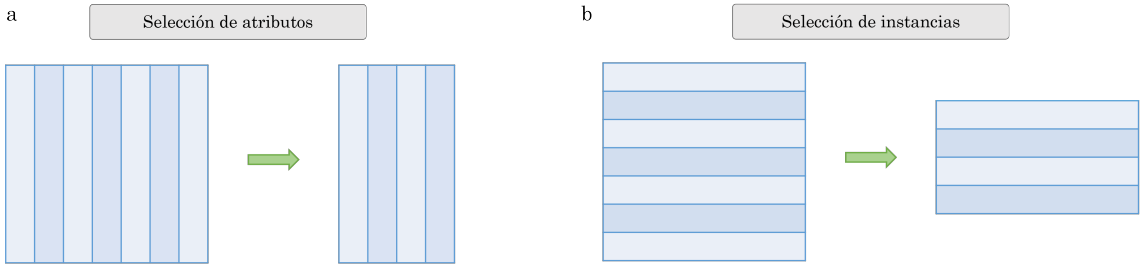

Figura 3.5. Reducción del conjunto de datos. a) Selección de atributos. b) Selección de instancias. Adaptada de [28].

Reducción del Conjunto de Datos del Proceso de Admisión

La reducción del conjunto del proceso de admisión se llevó a cabo bajo la selección de atributos únicamente, algunos de ellos fueron eliminados de acuerdo con las siguientes consideraciones. La *Clave UASLP* no aporta información relevante para el estudio, pero servirá como índice de los registros y se usará como referencia a lo largo de todo el trabajo. La *Clave FI* incluye información que podría ser útil (como el género, la generación, etc.; descripción completa presentada en la Sección 3.1.1), por lo tanto, ese atributo se conserva. La *Generación* su puede eliminar, ya que esta información está incluida en la clave de la Facultad; lo mismo ocurre con la *Carrera*. El *Nombre* se descartará, puesto que tampoco aporta información relevante. El atributo *Psicométrico* original, también se elimina, ya que fue reemplazado por sus valores numéricos. Los últimos cuatro atributos son importantes para nuestro estudio, por lo cual se conservan también. Así, nuestro conjunto se reduce considerablemente quedando solo 5 atributos: *Clave FI*, *Admisión total*, *Psicométrico* (numérico), *Conocimientos* y *EXANI-II* (Tabla 3.4).

Tabla 3.4. Atributos del proceso de admisión seleccionados.

Índice	Atributos
Clave UASLP	Clave FI
	Admisión total
	Psicométrico (numérico)
	Conocimientos
	EXANI-II

Reducción del Conjunto de Datos del Kardex

El conjunto de datos del kardex se reducirá bajo la selección de atributos y de instancias. De los 13 atributos originales, cinco de ellos coinciden con los del conjunto de admisión: *Clave UASLP*, *Clave FI*, *Generación*, *Nombre* y *Carrera*; para estos atributos se hicieron las mismas consideraciones mencionadas en la sección anterior, sin embargo, solo se conservará la *Clave UASLP* (como índice), la *Clave FI* será descartada para evitar atributos duplicados al hacer la integración con el conjunto de datos de admisión. Los atributos *Clave carrera*, *Clave materia* y *Fecha de calificación* fueron descartados, para el objetivo de nuestro estudio estos atributos no aportan información relevante. La información de los atributos *Materia*, *Calificación*, *Tipo de examen* y *Situación actual* fue condensada en los nuevos atributos creados, por lo tanto, estos cuatro atributos fueron eliminados. El atributo *Semestre cursado* sigue siendo importante e irreemplazable, por lo tanto, se quedará en nuestro conjunto de datos. Después de esta reducción, de los 13 atributos originales nos quedaremos únicamente con 2 de ellos: *Clave UASLP* y el *Semestre cursado*.

Ahora, en relación con los atributos creados a partir de los originales, nos quedaremos únicamente con los siguientes: *Créditos*, *Rendimiento general*, *Rendimiento programación*, *Rendimiento DFM*, *Efectividad*, *¿Cumple?* y *¿Termina?*. El resto de los atributos creados solo se utilizaron para obtener estos atributos más representativos, ya no necesitaremos la información que nos aportan. Al finalizar la selección de atributos, el conjunto de datos del kardex quedó con 8 atributos finales y la clave de la UASLP como índice, Tabla 3.5.

Tabla 3.5. Atributos del kardex seleccionados.

Índice	Atributos
Clave UASLP	Semestre cursado
	Créditos
	Rendimiento general
	Rendimiento programación
	Rendimiento DFM
	Efectividad
	¿Cumple?
	¿Termina?

Habiendo seleccionado los atributos representativos del conjunto, se procedió a la selección de instancias. Para llevar a cabo la reducción de los registros, los datos de los estudiantes fueron agrupados por clave y por semestre, quedando la información condensada de cada estudiante en un solo renglón.

3.2.4. Integración

La integración implica la fusión de dos o más conjuntos de datos provenientes de diferentes fuentes. Como se ha venido mencionando a lo largo de todo el capítulo, se cuentan con dos conjuntos de datos diferentes, uno procedente del proceso de admisión y otro que representa el historial académico de los estudiantes; estos conjuntos inicialmente tenían estructuras diferentes y no era posible unificarlos, sin embargo, después de haber realizado las transformaciones y reducciones correspondientes, fue posible realizar la integración de estos conjuntos. La *Clave UASLP* se estableció como índice de los registros en todos los conjuntos, esta clave es única e identifica a cada uno de los estudiantes, por lo tanto, este elemento fue usado como referencia para realizar la unión de los conjuntos de datos. La Tabla 3.6 muestra los atributos presentes en el conjunto de datos final.

Tabla 3.6. Atributos finales.

Índice	Atributos
Clave UASLP	Clave FI
	Admisión total
	Psicométrico (numérico)
	Conocimientos
	EXANI-II
	Semestre cursado
	Créditos
	Rendimiento general
	Rendimiento programación
	Rendimiento DFM
	Efectividad
	¿Cumple?
	¿Termina?

Después de haber realizado la unión de los conjuntos se procedió a realizar un análisis explo-

ratorio de ellos, esto con la finalidad de obtener información adicional que pudiera ser útil para plantear los experimentos correspondientes.

3.3. Análisis Exploratorio del Conjunto de Datos

El principal objetivo de nuestro trabajo es la predicción de la deserción escolar, por lo tanto, una práctica importante sería identificar puntos críticos temporales en la deserción escolar a lo largo de la carrera, para poder enfocarnos en los puntos previos a ellos. La exploración del conjunto de datos se inició con un análisis de la deserción escolar en cada uno de los primeros 8 semestres de la carrera. Se encontró que al terminar el primer semestre un poco más del 13 % de los estudiantes desertan, mientras que al terminar el segundo semestre, el número aumenta a más del 35 %, lo que revela que el mayor índice de deserción ocurre durante el primer año de la carrera (Figura 3.6). En semestres posteriores, el número de estudiantes que abandona los estudios se reduce notablemente, las cifras van del 1 % al 4.5 %, aproximadamente.

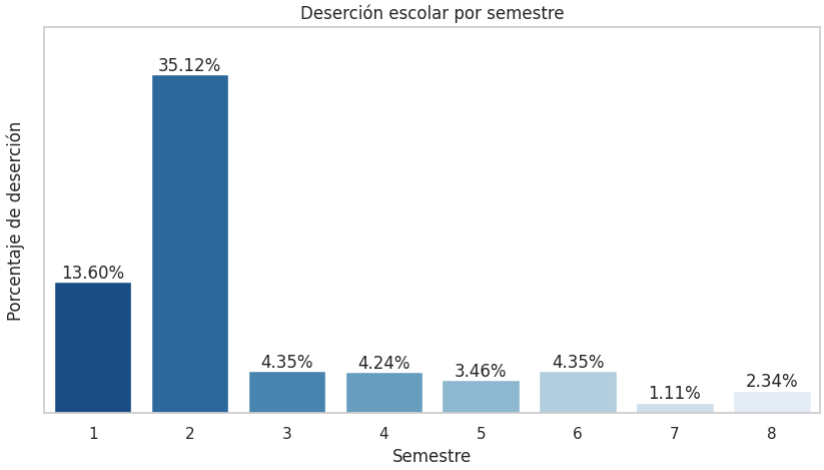

Figura 3.6. Porcentaje de deserción escolar en cada semestre.

Un requisito indispensable para que los estudiantes puedan inscribirse al tercer semestre, es haber aprobado como mínimo con 45 créditos, en caso de no cumplir con este requisito, son dados de baja de la Facultad. El 98 % de los estudiantes que abandona la carrera en el primer año lo hace por este motivo, únicamente el 2 % de los estudiantes abandona la carrera en el primer año habiendo cumplido los 45 créditos.

Con este análisis se pudo identificar el punto crítico de la deserción estudiantil en nuestro estudio. Usando esta información como referencia se decidió dividir el estudio en dos problemas fundamentales:

- Identificar a aquellos estudiantes que no cumplirán con el requisito de los 45 créditos.
- De los estudiantes que sí cumplen los 45 créditos, detectar a aquellos que abandonarán los estudios antes de terminar la carrera.

3.3.1. Antes del Requisito de los 45 Créditos

Poder identificar con antelación a los estudiantes que no cumplirán el requisito de los 45 créditos es un problema complicado, ya que para esto se cuenta únicamente con el historial académico del primer semestre y los datos procedentes del proceso de admisión. Realizar un análisis exploratorio del conjunto de datos puede aportar información importante acerca de los atributos y su relación con la variable objetivo. A continuación, se presenta el análisis realizado con los datos correspondientes.

La primera parte del análisis consistió en examinar la correlación entre las variables del problema y el atributo objetivo, para ello se calculó el coeficiente de correlación de Pearson y se construyó el mapa de correlación con dichos valores. El coeficiente de correlación de Pearson es una medida de la relación que existe entre dos variables, dicho coeficiente toma valores entre -1 y 1, los valores positivos representan correlación negativa (a medida que aumenta el valor de una variable, el valor de la otra disminuye), mientras que los valores positivos indican correlación positiva (si el valor de una variable aumenta, también lo hace el valor de la otra). Una correlación cercana a 0 sugiere que no existe una correlación entre las variables, y conforme este valor se acerca a |1| indica que hay una correlación entre ambas variables, entre mayor sea el valor absoluto del coeficiente, mayor es la correlación. Al realizar este análisis también se comprobó que todos los valores obtenidos fueran estadísticamente significativos (todos los coeficiente obtenidos tuvieron un valor p menor que 0.05).

De esta manera, se analizó la correlación entre los resultados de los exámenes del proceso de admisión y el cumplimiento de los 45 créditos, la Figura 3.7 muestra el mapa de correlación

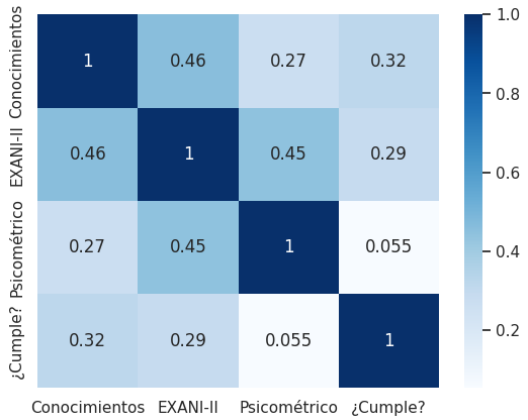

Figura 3.7. Mapa de correlación entre las variables del examen de admisión y el cumplimiento de los 45 créditos.

obtenido. El atributo *Admisión total* no fue considerado ya que en su lugar se tomó el resultado de cada uno de los exámenes de forma independiente. Los resultados obtenidos demostraron que la correlación entre estas variables y el cumplimiento de los 45 créditos es muy baja, principalmente con el examen psicométrico, su correlación es de 0.055; mientras que la correlación con el EXANI-II y con el examen de conocimientos es de 0.29 y 0.32, respectivamente. Para tener una representación visual de esta relación se creó la Figura 3.8, esta gráfica muestra la relación entre dichos elementos.

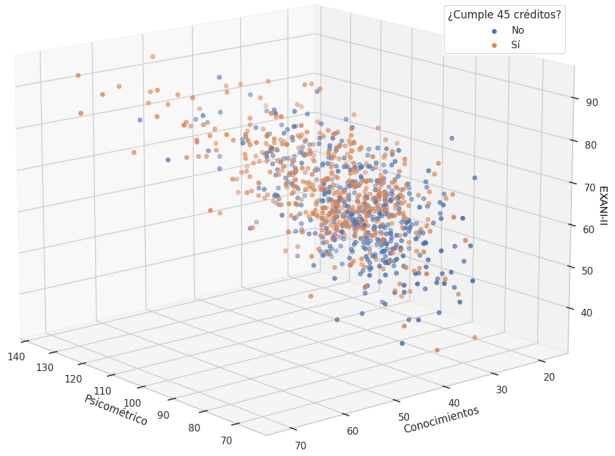

Figura 3.8. Relación entre los resultados del examen de admisión y el cumplimiento de los 45 créditos.

Cada uno de los ejes de la gráfica representa los resultados de los diferentes exámenes del proceso de admisión y el color de los puntos indica si el estudiante cumplió o no con los 45

créditos. Como puede observarse, no es posible identificar una región del espacio que identifique el cumplimiento de los 45 créditos en función de los resultados del proceso de admisión.

Los resultados de este análisis indican que, muy probablemente, los resultados del proceso de admisión no aporten información relevante para poder identificar a los estudiantes en riesgo de deserción al término del primer año de la carrera, sin embargo, no se descartará ningún atributo.

Después de este primer análisis, se procedió al análisis exploratorio de los datos académicos acumulados durante el primer semestre de los estudiantes. La primera parte consistió en analizar la correlación entre los atributos creados y el requisito de los 45 créditos (Figura 3.9). Los resultados demostraron una correlación mayor que la obtenida con los datos del proceso de admisión. Los valores de las correlaciones obtenidas están entre 0.58 y 0.74, lo que indica una alta correlación entre estas variables. El atributo que presenta la mayor correlación es la efectividad (0.74), seguida del rendimiento general (0.71); mientras que la correlación con el rendimiento por áreas es muy similar para ambas, 0.59 para Programación y 0.58 para el Departamento de Físico-Matemáticas.

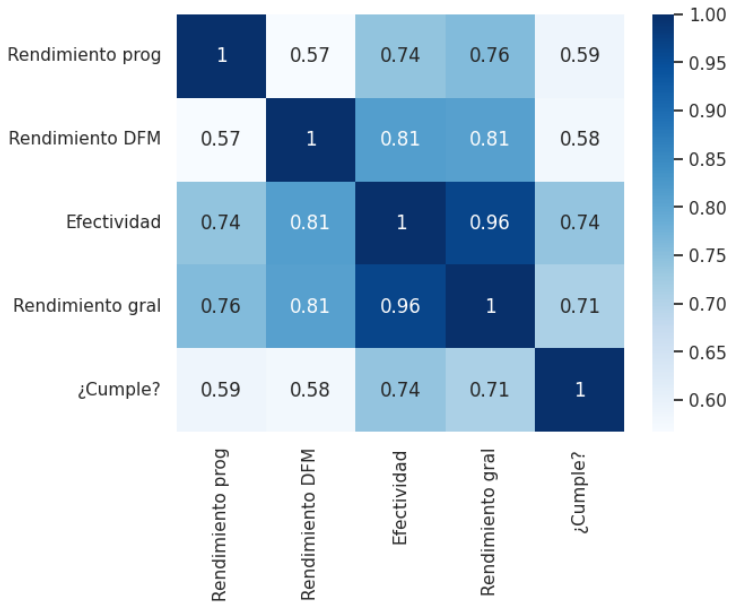

Figura 3.9. Mapa de correlación entre las variables académicas y el cumplimiento de los 45 créditos.

La representación gráfica de la relación entre el requisito de los 45 créditos y las variables académicas se muestra en la Figura 3.10. Los ejes representan los valores de las variables académicas, mientras que el color de los puntos indica la situación del cumplimiento de los 45 créditos.

Al observar las gráficas se puede notar una sensible separación entre los estudiantes que sí cumplen los 45 créditos y los que no, en la región superior se puede ver una acumulación mayor de puntos naranjas, lo que indica que los estudiantes que sí cumplen con los 45 créditos comparten ciertas características. Por otro lado, en la región inferior podemos ver que la concentración de puntos azules es mayor, esto indica que la mayor parte de los estudiantes que no cumplieron con el requisito de los 45 créditos tienen algunas similitudes entre ellos.

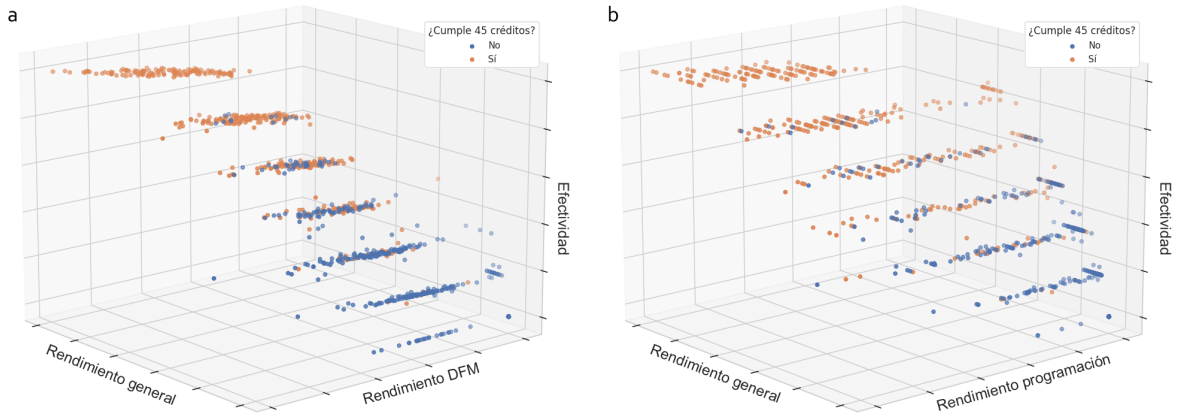

Figura 3.10. Relación entre las variables académicas y el cumplimiento de los 45 créditos. a) Rendimiento general, efectividad y rendimiento DFM. b) Rendimiento general, efectividad y rendimiento programación.

Después de este análisis sabemos que las variables académicas aportan mayor información relativa a la deserción escolar que los datos del proceso de admisión, a pesar de ello, todos los datos serán considerados para crear los modelos predictivos.

3.3.2. Después del Requisito de los 45 Créditos

De acuerdo con lo mencionado al inicio de la Sección 3, el segundo problema planteado fue el estudio de la deserción escolar posterior al primer año. El objetivo de esta parte del estudio es analizar la deserción de los estudiantes que sí cumplieron los 45 créditos, estudiantes que en semestres posteriores al primer año abandonaron los estudios sin concluir la carrera. En la Figura 3.6 se pudo observar que, posterior al primer año, no existe algún punto crítico en relación con el porcentaje de la deserción escolar, este valor tiene variaciones muy pequeñas entre un semestre y otro, por lo que todos los periodos serán analizados bajo las mismas condiciones. Dicho esto, podemos establecer nuestra nueva variable objetivo: el atributo *¿Termina?*, clasificaremos a los

estudiantes en dos clases, los que terminan la carrera y lo que no.

Empezaremos por hacer un análisis rápido de los datos que tenemos y su relación con nuestra variable objetivo. El primer análisis que se realizó fue el cálculo de la correlación entre las variables académicas y el atributo objetivo (Figura 3.11). Empezaremos describiendo la correlación del atributo *Rendimiento programación* con nuestra variable objetivo. El valor de la correlación entre ambas variables es mínimo en el Año 1 (0.29) y máximo en el Año 4 (0.46), esto podría deberse a que el número de materias de Programación cursadas durante el primer año son solo dos, y se acumulan conforme pasan los semestres hasta completar ocho materias al término del octavo semestre (Tabla 3.3). Su correlación no es muy alta, sin embargo, podría resultar relevante, principalmente en el Año 3 y Año 4.

La variable *Rendimiento DFM* es la que presenta la menor correlación con nuestra variable objetivo. A diferencia del atributo anterior, este atributo tiene su correlación máxima en el Año 2 (0.36) y disminuye con el tiempo, teniendo su valor mínimo en el Año 4 (0.11). Esto se debe a que, contrario al atributo anterior, las materias del Departamento de Físico-Matemáticas se cursan en los primeros semestres (Tabla 3.3), y por eso resultan más representativas en los dos primeros años.

Después viene la correlación con la variable *Efectividad*. El valor de la correlación con esta variable aumenta conforme pasan los semestres, su valor máximo es 0.61 y el mínimo 0.48, en el Año 1. La diferencia en estos valores se debe a que al inicio de la carrera todos los estudiantes inscriben el mismo número de materias, pero conforme pasan los semestres cada estudiante decide el número de materias que inscribe, por esta razón, es que en el Año 3 y 4 la correlación es mayor. A partir de Año 2 su correlación es mayor que 0.5, lo que indica una buena correlación con nuestra variable objetivo.

Finalmente, la variable *Rendimiento general*, la correlación entre esta variable y la variable objetivo permanece casi constante a partir del Año 2. En el Año 1 su valor es de 0.49, a partir de ese año, se mantiene casi constante en un valor de 0.58-0.59. Esta variable evalúa homogéneamente todos los periodos, promedio y créditos totales, debido a esto presenta un valor casi constante en el tiempo. Con estos valores sabemos que su correlación es alta y nos servirá para construir los modelos predictivos en todos los periodos.

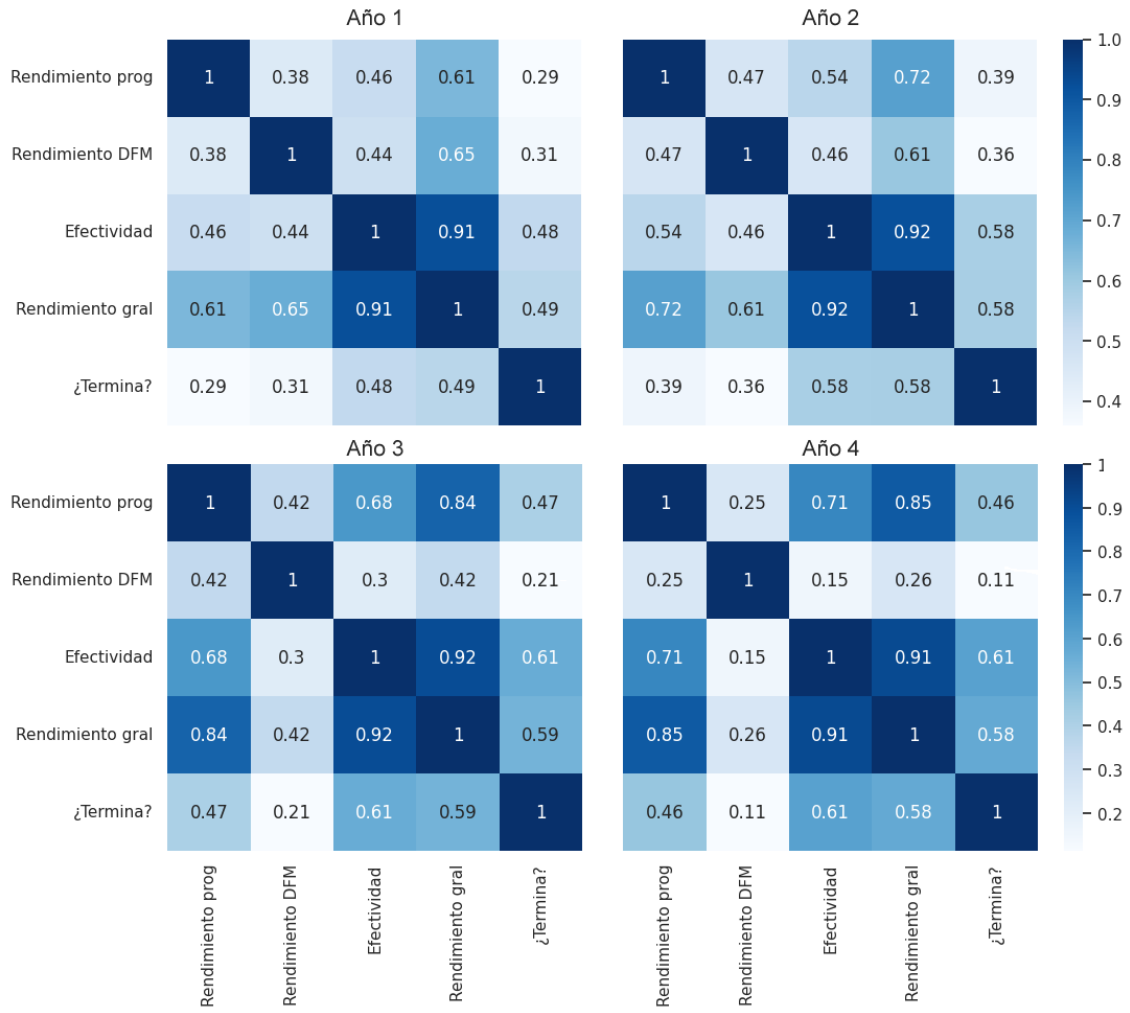

Figura 3.11. Mapa de correlación entre las variables académicas y el atributo *¿Termina?*.

Y para terminar este análisis, tenemos un ejemplo gráfico de la relación entre los atributos (*Rendimiento general*, *Rendimiento DFM*, *Rendimiento programación*, *Eficiencia*) y la variable objetivo de nuestro estudio. Se eligió como ejemplo representativo el Año 4, ya que debería observarse una tendencia más clara puesto que los estudiantes se encuentran más cerca del término de la carrera. La Figura 3.12 muestra la relación entre las variables académicas del Año 4 y si el estudiante termina o no la carrera. En ambas gráficas podemos ver que la separación entre las clases no es muy clara, sin embargo, en los extremos se observa la acumulación de puntos del mismo color, lo que indica que los elementos pertenecientes a una misma clase podrían presentar algunas similitudes en sus características.

Tras haber analizado la relación entre la variable *¿Termina?* y las variables académicas

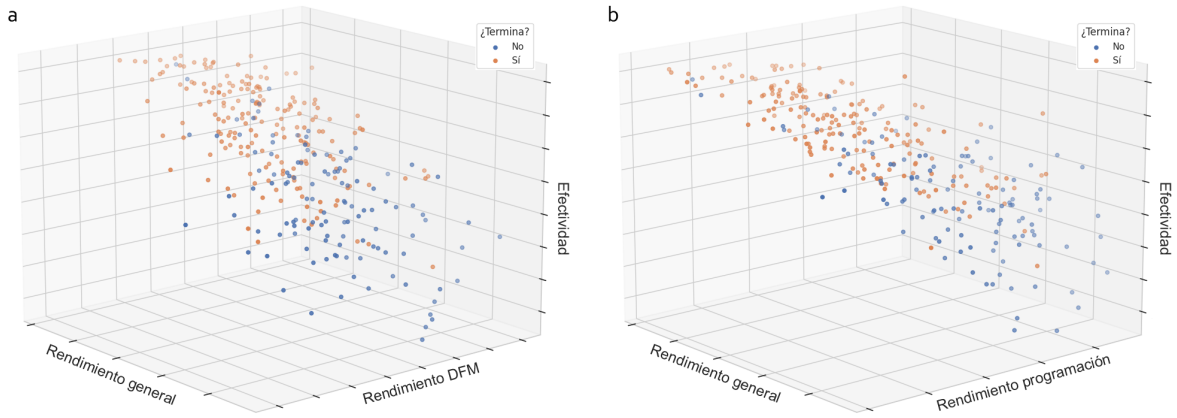

Figura 3.12. Relación entre las variables académicas del Año 4 y el atributo *¿Termina?* a) Rendimiento general, efectividad y rendimiento DFM. b) Rendimiento general, efectividad y rendimiento programación.

podemos decir que la correlación entre las variables se modifica en los distintos periodos, algunas sobresalen al inicio de la carrera y otras al final, debido a esto, a pesar de no tener correlaciones tan altas, podrían complementarse unas con otras al momento de construir los modelos predictivos.

Capítulo 4

Predicción de la Deserción Escolar

De acuerdo con la metodología presentada en la Sección 1.3, la etapa posterior al preprocesamiento de los datos es la aplicación de los métodos de clasificación, seguida del análisis de los resultados. En el capítulo anterior se describió la etapa del preprocesamiento, así, en este capítulo se presentarán las dos etapas posteriores: la aplicación de los métodos y la interpretación de los resultados.

La Facultad de Ingeniería de la UASLP, como parte de su reglamento establece que un estudiante debe aprobar 45 créditos, como mínimo, al término del primer año de la carrera, de lo contrario, es dado de baja de la Facultad. De acuerdo con los resultados del análisis exploratorio realizado al conjunto de datos, casi el 49 % de los estudiantes desertan durante el primer año de la carrera, de los cuales el 98 % de ellos lo hace debido al incumplimiento del requisito de los 45 créditos (Sección 3.3). Tomando esta información como referencia, se consideró importante plantear un problema que permitiera identificar a aquellos estudiantes en riesgo de abandonar la carrera debido al requisito del cumplimiento de los 45 créditos. El preprocesamiento de los datos y el análisis exploratorio realizado sobre el conjunto de datos final (Sección 3.3) permitió plantear dos problemas fundamentales:

- **Problema 1:** Predecir si el estudiante cumplirá los 45 créditos al terminar el primer año.
- **Problema 2:** De los estudiantes que sí cumplen los 45 créditos, predecir si terminarán la carrera.

Cada uno de estos problemas tiene sus propias características y son diferentes entre sí, por

ejemplo, el *Problema 1* debe abordarse al inicio de la carrera, mientras que el *Problema 2* puede ser abordado en semestres o años posteriores; por esta razón, es necesario diseñar experimentos específicos para cada uno de ellos. También hay que considerar que, tal como se mencionó en el capítulo anterior, el conjunto de datos preprocesado contiene datos procedentes del proceso de admisión y datos del historial académico de los estudiantes. Con todo esto, el primer paso consistió en diseñar los experimentos correspondientes. Para el *Problema 1* se diseñaron dos experimentos, el primero fue planteado para abordar el problema desde el ingreso de los estudiantes a la Facultad (*Experimento 1.1*), para ello, se usaron los datos del proceso de admisión; el segundo experimento se diseñó para ser aplicado al término del primer semestre usando los datos académicos generados durante ese periodo (*Experimento 1.2*). Para el *Problema 2*, también se plantearon dos experimentos, el primero de ellos se realizó con los datos del historial académico de los estudiantes acumulado hasta el término del cuarto semestre de la carrera (*Experimento 2.1*), mientras que el segundo experimento se realizó con el historial académico acumulado hasta el octavo semestre (*Experimento 2.2*). El plan de estudios de las carreras de Ingeniería en Computación e Ingeniería en Informática comprende 10 semestres, sin embargo, estos puntos de predicción fueron elegidos ya que al término del cuarto semestre los estudiantes deberían haber aprobado todas las materias del Departamento de Físico-Matemáticas y la mitad de las del área de Programación, y al término del octavo semestre deberían haber aprobado todas las materias del área de Programación, tal como se mostró en la Tabla 3.3 de la Sección 3.2.2. La Figura 4.1 muestra los puntos de predicción y los experimentos planteados para cada problema.

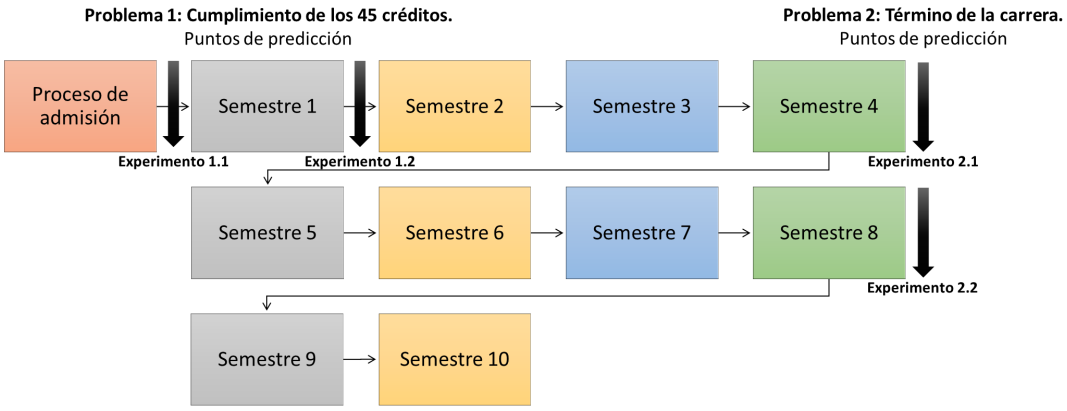

Figura 4.1. Puntos de referencia para la predicción de la deserción escolar en cada uno de los problemas planteados.

Los experimentos fueron desarrollados en el entorno de Google Colab, tal como se describió en la Sección 2.2.4, y se aplicaron los cinco métodos de clasificación mencionados en la Sección 2.2.2. Los métodos de clasificación aplicados en cada uno de los experimentos fueron: Árboles de Decisión (*Decision Tree*, DT, por sus siglas en inglés), Bosque Aleatorio (*Random Forest*, RF), Naïve Bayes (NB), Máquina de Soporte Vectorial (*Support Vector Machine*, SVM) y Redes Neuronales (*Neural Networks*, NN).

4.1. Problema 1: Cumplimiento de los 45 Créditos

De acuerdo con la información previa, el objetivo del *Problema 1* es intentar predecir si un estudiante cumplirá o no los 45 créditos al terminar el primer año. La predicción debe realizarse a más tardar al término del primer semestre de la carrera, por lo tanto, para lograr este objetivo se cuenta con la información procedente del examen de admisión y con el historial académico del primer semestre de los estudiantes. Poder identificar desde el inicio de la carrera a los estudiantes con posible riesgo de deserción sería de gran utilidad, para ello, se planteó el *Experimento 1.1*, diseñado para realizar la predicción utilizando los datos del proceso de admisión. Por otro lado, el *Experimento 1.2* corresponde a la predicción basada en el historial académico del primer semestre de los estudiantes, así podremos identificar variables importantes del proceso de admisión y del historial académico.

4.1.1. Experimento 1.1

Como se mencionó previamente, en este experimento se intentará predecir si un estudiante cumplirá los 45 créditos al término del primer año usando los datos del proceso de admisión. En la Sección 3.1.1 se presentó una descripción detallada de cada uno de los atributos de este conjunto de datos y en la Sección 3.2.2 se presentó el conjunto de datos con los atributos finales: *Admisión total*, *Psicométrico*, *Conocimientos* y *EXANI-II*. Para este experimento se tomaron como atributos las puntaciones obtenidas en los tres exámenes aplicados durante el proceso de admisión (*Conocimientos*, *Psicométrico*, *EXANI-II*). El atributo *Admisión total* no se consideró en este experimento ya que, como se mencionó en la Sección 3.1.1, este atributo se obtiene de una combinación de los 3 exámenes: 15 % del examen psicométrico, 40 % del EXANI-II y 45 % del examen

de conocimientos. La variable objetivo se definió como *¿Cumple?* que establece si el estudiante cumple o no con los 45 créditos al término del primer año; los valores de esta variable son *Sí* y *No*, por lo tanto, el experimento corresponde a una clasificación binaria. La Tabla 4.1 muestra algunos de los valores estadísticos de dichos atributos, únicamente para tener como referencia.

Tabla 4.1. Descripción estadística de los atributos del *Experimento 1.1*.

Atributo	Mínimo	Máximo	Media
Conocimientos	18.00	71.00	33.15
Psicométrico	66.11	136.84	103.71
EXANI-II	33.33	93.00	61.34

Por otro lado, el conjunto de datos incluye la información de 897 estudiantes, de los cuales el 51 % pertenece a la clase *No cumple* y el 49 % a la clase *Sí cumple*. De esta manera, el conjunto de datos empleado en este experimento incluye 897 registros con 3 atributos. Algunas características importantes del *Experimento 1.1* son:

- Los datos del conjunto se dividieron aleatoriamente en dos partes: el conjunto de entrenamiento y el conjunto de prueba. De acuerdo con la bibliografía revisada, los datos suelen dividirse en conjuntos de 70-30 % u 80-20 % para entrenamiento y prueba [32, 33]. Siguiendo esto, el conjunto de entrenamiento se creó con el 70 % de los registros y el 30 % restante conformó el conjunto de prueba. La validación cruzada de 10 iteraciones se realizó con el conjunto de entrenamiento para cada uno de los modelos.
- El porcentaje de estudiantes correspondiente a cada clase se mantuvo igual en los dos subconjuntos creados, el 51 % de los registros de ambos conjuntos pertenece a la clase *No cumple* y el 49 % a la clase *Sí cumple*, tal como en el conjunto original.
- Para la evaluación de los modelos se construyó la matriz de confusión correspondiente y se calcularon los valores de las métricas de evaluación utilizando el conjunto de prueba: exactitud, precisión, sensibilidad y valor F1 (para más detalles ver Sección 2.2.3).
- Los hiperparámetros de los modelos fueron ajustados y se utilizaron los más adecuados para nuestros datos. Para ello, se realizaron diversas pruebas repitiendo el mismo experimento modificando los hiperparámetros y evaluando los resultados obtenidos. Los hiperparámetros modificados se presentan en la Tabla 4.2, el resto se dejó con los valores predeterminados.

Tabla 4.2. Hiperparámetros ajustados en los distintos métodos del *Experimento 1.1*.

Método	Hiperparámetros
Árbol de Decisión	Profundidad máxima del árbol = 3
Bosque Aleatorio	Profundidad máxima de los árboles = 5, número de árboles = 5
Naïve Bayes	Sin modificaciones en los hiperparámetros
Máquina de Soporte Vectorial	C = 100, gamma = 0.001, tipo de kernel = *rbf*
Redes Neuronales	Una capa de entrada, tres capas ocultas y una de salida (Tabla 4.3)

Tabla 4.3. Estructura de la red neuronal empleada en el *Experimento 1.1*.

Número de capa	Tipo de capa	Número de neuronas	Función de activación
0	Entrada	3	-
1	Oculta	32	ReLU
2	Oculta	16	ReLU
3	Oculta	32	ReLU
4	Salida	1	Sigmoid

Resultados

En la primera parte de los resultados se analizan las matrices de confusión obtenidas para los modelos construidos con los distintos métodos, estas matrices se muestran en la Figura 4.2. Al comparar las matrices se puede observar que presentan ciertas similitudes entre ellas, gran parte de los estudiantes son clasificados en la clase incorrecta, principalmente los de la clase *Sí cumple*. La matriz de confusión del método de Bosque Aleatorio muestra resultados ligeramente superiores que el resto (Figura 4.2b), ya que logra clasificar correctamente a más número de estudiantes. De los 130 estudiantes de la clase *No cumple*, clasifica correctamente a 90 de ellos; mientras que para la clase *Sí cumple*, clasifica correctamente a 93 e incorrectamente a 47.

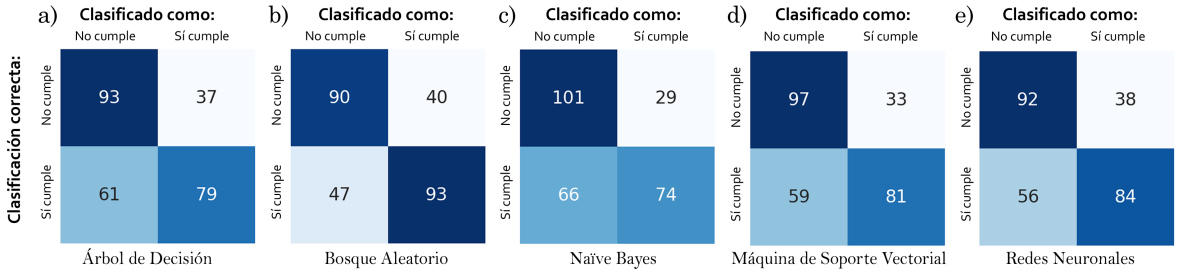

Figura 4.2. Matrices de confusión obtenidas para los diferentes modelos en el *Experimento 1.1*. a) Árbol de Decisión. b) Bosque Aleatorio. c) Naïve Bayes. d) Máquina de Soporte Vectorial. e) Redes Neuronales.

Los valores obtenidos en las métricas de evaluación de los distintos modelos se presentan en la Tabla 4.4, los valores resaltados en negrita representan los valores más altos obtenidos en cada una de las métricas. De acuerdo con esto, la exactitud máxima obtenida fue de 0.68 y se alcanzó con el método de Bosque Aleatorio, mientras que la exactitud mínima fue la presentada por el Árbol de Decisión, con un valor de 0.64. Al analizar el resto de las métricas, se observa que presentan valores muy similares entre sí, no existe una diferencia notable entre los resultados obtenidos con los distintos métodos, aunque es posible notar que el Bosque Aleatorio generó los mejores resultados.

Tabla 4.4. Métricas de evaluación obtenidas en el *Experimento 1.1*. Los valores en negrita representan los valores más altos de cada métrica.

Método	Exactitud	Precisión	Sensibilidad	Valor F1
Árbol de Decisión	0.64	0.68	0.56	0.62
Bosque Aleatorio	**0.68**	0.70	**0.66**	**0.68**
Naïve Bayes	0.65	**0.72**	0.53	0.61
Máquina de Soporte Vectorial	0.66	0.71	0.58	0.64
Redes Neuronales	0.65	0.69	0.60	0.64

Otra parte importante del análisis consiste en conocer la importancia que tiene cada uno de los atributos en la construcción de los modelos. Para ello, se extrajo el valor de la importancia de cada atributo en los distintos métodos[1], la Tabla 4.5 muestra los resultados obtenidos. La tabla muestra que el atributo más importante en este experimento es el examen de conocimientos, ya que presenta el valor más alto en todos los métodos; los valores más bajos se obtienen con el examen psicométrico, lo que sugiere que es el menos representativo en el experimento. Este resultado indica que dentro del proceso de admisión, el resultado del examen de conocimientos es el más significativo para el cumplimiento de los 45 créditos al término del primer año.

Como parte complementaria, se extrajeron las principales reglas de asociación derivadas del Árbol de Decisión correspondiente, estas reglas se muestran en la Tabla 4.6. Las reglas de asociación obtenidas indican que el punto crítico en el resultado del examen de conocimientos es la puntuación de 35.5, ya que los estudiantes que obtienen un resultado superior a éste tienen una probabilidad más alta de cumplir con el requisito de los 45 créditos en comparación con los que

[1]Excepto en redes neuronales, ya que en este método no es posible extraer directamente la importancia de los atributos como en los otros métodos.

Tabla 4.5. Importancia de cada uno de los atributos analizados en el *Experimento 1.1*. Los valores en negrita indican el valor máximo obtenido en cada método.

Método	Conocimientos	EXANI-II	Psicométrico
Árbol de Decisión	**0.133**	0.065	0.028
Bosque Aleatorio	**0.140**	0.092	0.032
Naïve Bayes	**0.064**	0.027	-0.015
Máquina de Soporte Vectorial	**0.132**	0.098	0.047

obtienen un resultado menor. Los estudiantes que obtienen un resultado mayor que 35.5 en el examen de conocimientos y mayor que 53.6 en el EXANI-II tienen un 77.4 % de probabilidad de cumplir el requisito de los 45 créditos; mientras que los estudiantes que obtienen una calificación menor que 35.5 en el examen de conocimientos y menor que 55.9 en el EXANI-II tienen un 77.7 % de probabilidad de no cumplir con los 45 créditos al término del primer año.

Tabla 4.6. Principales reglas de asociación obtenidas del Árbol de Decisión en el *Experimento 1.1*.

Núm.	Regla	Clase	Probabilidad	Número de muestras
1	Conocimientos > 35.5, Exani > 53.665	Sí	77.41 %	155
2	Conocimientos <= 35.5, Exani <= 55.915	No	77.77 %	171
3	Conocimientos > 35.5, Exani > 53.665, Psicométrico <= 108.49	Sí	86.96 %	69
4	Conocimientos <= 35.5, Exani <= 55.915, Psicométrico > 96.89	No	70.41 %	98

Tras haber analizado todos los resultados obtenidos en el *Experimento 1.1* se puede concluir que los datos procedentes del proceso de admisión no son suficientes para poder predecir el cumplimiento de los 45 créditos. A pesar de ello, en este experimento se logró identificar que el atributo *Conocimientos* es el más importante para el cumplimiento de los 45 créditos, ya que es el que presenta mayor importancia en todos los métodos y además se encuentra en el nodo raíz del árbol de decisión. Cabe recordar que el 45 % del resultado total del proceso de admisión a la Facultad corresponde al resultado del examen de conocimientos, lo que demuestra que el porcentaje asignado por la Facultad a este examen es acertado, ya que es el más relevante de los tres exámenes aplicados (Conocimientos, Psicométrico, EXANI-II). Como conclusión, se puede decir que los modelos obtenidos no permiten lograr el objetivo planteado para el *Experimento 1.1*, que es predecir si un estudiante cumplirá o no los 45 créditos al terminar el primer año usando los datos del proceso de admisión, ya que como se pudo observar en la Tabla 4.4 los valores alcanzados en la exactitud indican que de todas las predicciones realizadas solo el 68 % de ellas son acertadas.

Para finalizar con el *Experimento 1.1*, se presenta la predicción obtenida tras aplicar los cinco modelos a los datos de un estudiante que haya obtenido los valores medios en los tres exámenes del proceso de admisión (conocimientos: 33.15, psicométrico: 103.71, EXANI-II: 61.34; estos valores fueron presentados en la Tabla 4.1). Los resultados se muestran en la Tabla 4.7.

Tabla 4.7. Predicción realizada para un estudiante que haya obtenido el valor medio en los tres exámenes del proceso de admisión.

Método	Predicción
Árbol de Decisión	No cumple
Bosque Aleatorio	Sí cumple
Naïve Bayes	No cumple
Máquina de Soporte Vectorial	Sí cumple
Redes Neuronales	No cumple

El resultado de la predicción está dividido, 3 modelos predicen que el estudiante no cumplirá con los 45 créditos, mientras que otros 2, predicen lo contrario. Con este experimento de prueba reforzamos la conclusión presentada en el párrafo anterior, donde se dice que los datos del proceso de admisión no son suficientes para poder predecir si los estudiantes cumplirán o no con los 45 créditos al terminar el primer año de la carrera.

4.1.2. Experimento 1.2

De acuerdo con la definición del *Problema 1* presentada al inicio de la Sección 4.1, el objetivo de este problema consiste en intentar predecir si el estudiante cumplirá o no los 45 créditos al término del primer año. Los resultados del *Experimento 1.1* indicaron que los datos procedentes del proceso de admisión no son suficientes para lograr esta predicción, por lo tanto, el *Experimento 1.2* se diseñó para trabajar el mismo problema pero utilizando el historial académico del primer semestre de los estudiantes. De acuerdo con la información presentada en las secciones 3.2.2 y 3.2.3, los atributos más relevantes correspondientes al historial académico del primer semestre se resumen en: *Rendimiento general*, *Rendimiento programación*, *Rendimiento DFM* (Departamento de Físico-Matemáticas) y *Efectividad*. Estos cuatro atributos fueron utilizados para realizar la predicción en este experimento. La variable objetivo se definió como *¿Cumple?*, esta variable clasifica a los estudiantes que sí cumplen con los 45 créditos en la clase *Sí* y a los que no cumplen, en la clase *No*. La Tabla 4.8 muestra algunos valores estadísticos de los atributos empleados en

este experimento.

Tabla 4.8. Descripción estadística de los atributos del *Experimento 1.2*.

Atributo	Mínimo	Máximo	Media
Rendimiento general	0.00	9.92	3.06
Rendimiento programación	0.00	10.00	3.35
Rendimiento DFM	0.00	9.83	4.60
Efectividad	0.00	1.00	0.50

El conjunto de datos analizado incluye los datos académicos de 897 estudiantes, de los cuales el 51 % (458 estudiantes) no cumple con los 45 créditos y el 49 % (439 estudiantes) sí lo hace. El *Experimento 1.2* se llevó a cabo con las mismas características que el *Experimento 1.1*, excepto los hiperparámetros de los modelos. La Tabla 4.9 muestra los hiperparámetros modificados en este experimento.

Tabla 4.9. Hiperparámetros ajustados en los distintos métodos del *Experimento 1.2*.

Método	Hiperparámetros
Árbol de Decisión	Profundidad máxima del árbol = 3
Bosque Aleatorio	Profundidad máxima de los árboles = 3, número de árboles = 10
Naïve Bayes	Sin modificaciones en los hiperparámetros
Máquina de Soporte Vectorial	C = 0.1, gamma = 10, tipo de kernel = *lineal*
Redes Neuronales	Una capa de entrada, dos capas ocultas y una de salida (Tabla 4.10)

Tabla 4.10. Estructura de la red neuronal empleada en el *Experimento 1.2*.

Número de capa	Tipo de capa	Número de neuronas	Función de activación
0	Entrada	4	-
1	Oculta	32	ReLU
2	Oculta	32	ReLU
3	Salida	1	Sigmoid

Resultados

Para iniciar con el análisis de los resultados se presentan las matrices de confusión obtenidas en cada uno de los métodos (Figura 4.3). En general, se puede observar que las matrices presentan valores similares, el número de muestras clasificadas correctamente en ambas clases presentan variaciones muy pequeñas entre los métodos, lo que indica que también en las métricas de eva-

luación habrá poca diferencia. Al observar con más detalle todas la matrices, se puede ver que la matriz correspondiente al método de Bosque Aleatorio (Figura 4.3b) presenta la menor diferencia entre el número de muestras clasificadas correctamente para ambas clases, 115 muestras de la clase *No cumple* clasificadas correctamente y 110 de la clase *Sí cumple*, lo que sugeriría que este método logra clasificar mejor a los estudiantes de ambas clases.

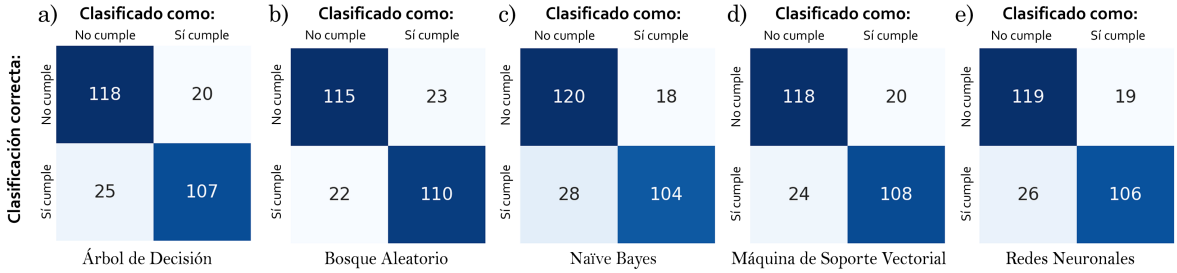

Figura 4.3. Matrices de confusión obtenidas para los diferentes modelos en el *Experimento 1.2*. a) Árbol de Decisión. b) Bosque Aleatorio. c) Naïve Bayes. d) Máquina de Soporte Vectorial. e) Redes Neuronales.

A continuación, en la Tabla 4.11, se presentan las métricas de evaluación obtenidas. La exactitud máxima alcanzada en el experimento es de 0.84 y corresponde al método de Máquina de Soporte Vectorial. El resto de los modelos presenta una exactitud de 0.83, sin embargo, existen pequeñas variaciones en los valores de las demás métricas, donde observamos que, tal como se mencionó en el párrafo anterior, el método de Bosque Aleatorio presenta las métricas más equilibradas, todas con un valor de 0.83. A partir de esto, se puede decir que los mejores métodos para la clasificación en este experimento son el de Máquina de Soporte Vectorial y el Bosque Aleatorio, ya que sus métricas de evaluación indican que son capaces de acertar el 83-84 % de las predicciones realizadas.

Tabla 4.11. Métricas de evaluación obtenidas en el *Experimento 1.2*. Los valores en negrita representan los valores más altos de cada métrica.

Método	Exactitud	Precisión	Sensibilidad	Valor F1
Árbol de Decisión	0.83	0.84	0.81	**0.83**
Bosque Aleatorio	0.83	0.83	**0.83**	**0.83**
Naïve Bayes	0.83	**0.85**	0.79	0.82
Máquina de Soporte Vectorial	**0.84**	0.84	0.82	**0.83**
Redes Neuronales	0.83	**0.85**	0.80	0.82

Después de las métricas de evaluación, se procede a analizar la importancia de cada uno de los atributos considerados en el experimento. La Tabla 4.12 muestra la importancia obtenida

para cada atributo en los distintos modelos. Como se puede observar en la tabla, el atributo *Rendimiento general* presenta la mayor importancia en todos los métodos, por lo tanto, se puede decir que es el atributo más significativo para la predicción del cumplimiento de los 45 créditos. El valor mínimo obtenido en cada uno de los modelos corresponde a un atributo diferente en todos ellos, debido a esto, no es posible establecer qué atributo es el menos importante. El atributo *Rendimiento DFM* aparece dos veces con el valor mínimo, aún así no se podría decir que es el atributo menos importante.

Tabla 4.12. Importancia de cada uno de los atributos analizados en el *Experimento 1.2*. Los valores en negrita indican el valor máximo en cada método.

Método	Rend_gral	Rend_prog	Rend_DFM	Efectividad
Árbol de Decisión	**0.277**	0.020	0.000	0.023
Bosque Aleatorio	**0.271**	0.003	0.007	0.036
Naïve Bayes	**0.075**	0.018	0.007	0.073
Máquina de Soporte Vectorial	**0.270**	0.115	0.057	0.000

Para complementar el análisis de los atributos se extrajeron las reglas de asociación del Árbol de Decisión generado, la Tabla 4.13 muestra algunas de las reglas más representativas. De acuerdo con las reglas obtenidas, el atributo principal es *Rendimiento general* con un valor crítico igual a 2.09, este valor divide a los estudiantes en dos grupos principales. Los estudiantes que alcanzan un rendimiento mayor que 2.09 en el primer semestre tienen un 80 % de probabilidad de cumplir con el requisito de los 45 créditos, mientras que aquellos que obtienen un rendimiento menor, tienen un 90 % de probabilidad de no cumplir los 45 créditos, y por lo tanto, ser dados de baja de la facultad.

Tabla 4.13. Principales reglas de asociación obtenidas del Árbol de Decisión en el *Experimento 1.2*.

Núm.	Regla	Clase	Probabilidad	Número de muestras
1	Rend_gral $>$ 2.09	Sí	80.28 %	350
2	Rend_gral $<$ 2.09	No	90.61 %	277
3	Rend_gral $>$ 4.57, Rend_DFM $>$ 7.07	Sí	100 %	62
4	Rend_gral $<$ 1.01	No	97.56 %	164

Como parte final del experimento se usaron los cinco modelos construidos para predecir el caso de un estudiante que tiene los valores medios de todos los atributos (Rendimiento general: 3.06, Rendimiento programación: 3.35, Rendimiento DFM: 4.60, Efectividad: 0.50; valores presentados

en la Tabla 4.8), la Tabla 4.14 muestra los resultados obtenidos. Las predicciones arrojadas sugieren que la probabilidad de que un estudiante con atributos promedio cumpla con los 45 créditos es muy alta, ya que 4 de los 5 modelos predicen que sí cumplirá.

Tabla 4.14. Predicción para un estudiante que tiene los valores medios de todos los atributos.

Método	Predicción
Árbol de Decisión	Sí cumple
Bosque Aleatorio	Sí cumple
Naïve Bayes	No cumple
Máquina de Soporte Vectorial	Sí cumple
Redes Neuronales	Sí cumple

Como conclusión del *Experimento 1.2* se puede decir que todos los modelos construidos generan resultados muy similares, todos logran clasificar a los estudiantes con una exactitud aproximada de 83 %, sin embargo, al considerar el resto de las métricas de evaluación se podría señalar al método de Bosque Aleatorio como el más adecuado para nuestros datos, aunque prácticamente se podría aplicar cualquiera de los métodos y los resultados serían muy similares. Adicionalmente, se pudo identificar al atributo *Rendimiento general* como el más importante en la predicción del cumplimiento de los 45 créditos, con un valor crítico de 2.09, un punto por debajo de su valor medio. Con todos estos resultados se puede concluir que los datos académicos procedentes del primer semestre de la carrera permiten identificar a los estudiantes en riesgo de no cumplir los 45 créditos con una confianza del 83 %.

El *Problema 1* se planteó con el objetivo de intentar predecir si un estudiante cumplirá o no los 45 créditos al terminar el primer año. Los resultados obtenidos en el *Experimento 1.1* demostraron que los datos del proceso de admisión no son suficientes para lograr este objetivo, mientras que los resultados del *Experimento 1.2* fueron notablemente superiores. Las predicciones realizadas con los datos del historial académico de los estudiantes alcanzaron una exactitud aproximada de 83 %, a diferencia del 65 % obtenida con los datos del proceso de admisión. Estos resultados indican que para el cumplimiento de los 45 créditos, los conocimientos y habilidades que tienen los estudiantes al ingresar a la carrera no son tan relevantes como los que adquieren en la Facultad y lo que viven a lo largo del primer semestre.

Para un estudio más general del *Problema 1* se diseñó también un experimento que incluyera los datos del proceso de admisión y los datos del historial académico, todos en un mismo expe-

rimento, sin embargo, al realizar dicha predicción se encontró que el resultado era ligeramente inferior al obtenido en el *Experimento 1.2*, la exactitud media fue de 82 %. Los resultados de este experimento no se mostraron ya que son prácticamente los mismos que los del *Experimento 1.2*, la importancia de los atributos del proceso de admisión es menor que la de los cuatro atributos del historial académico, incluso, las reglas de asociación derivadas del Árbol de Decisión son las mismas que las extraídas en el *Experimento 1.2*. Este resultado corrobora la conclusión del párrafo anterior.

4.2. Problema 2: Término de la Carrera

Después del cumplimiento de los 45 créditos, el siguiente problema que enfrenta el Área de Ciencias de la Computación de la Facultad es la deserción en los semestres posteriores al primer año. De los 439 estudiantes que sí cumplen con los 45 créditos y llegan al tercer semestre, únicamente el 42 % (185 estudiantes) sí termina la carrera, esta cifra representa otro problema de deserción importante. Con base en esta información, se planteó el *Problema 2* cuyo objetivo es intentar predecir si los estudiantes terminarán o no la carrera. Para abordar este problema se consideraron los datos del historial académico de los estudiantes que sí cumplieron con los 45 créditos y se plantearon dos experimentos distintos. Tal como se mencionó al inicio de este Capítulo 4 y se presentó en la Figura 4.1, los dos experimentos fueron planteados para realizar la predicción en diferentes puntos de la carrera, uno al término del segundo año y el otro, al término del cuarto año. Estos puntos fueron elegidos debido a que las materias del Departamento de Físico-Matemáticas se encuentran en los primeros cuatro semestres del plan de estudios y las materias del área de Programación (no optativas) abarcan hasta el octavo semestre (datos presentados en la Tabla 3.3). De esta manera, el *Experimento 2.1* corresponde a la predicción del término de la carrera empleando los datos del historial académico acumulado hasta el cuarto semestre, y el *Experimento 2.2* utiliza los datos acumulados hasta el octavo semestre.

4.2.1. Experimento 2.1

Como se mencionó previamente, el objetivo del *Experimento 2.1* es intentar predecir si los estudiantes terminarán o no la carrera, usando como atributos los datos del historial académico

acumulados hasta el cuarto semestre de la carrera. En este punto de la carrera los estudiantes deberían haber aprobado ya el 40 % de los créditos totales del plan de estudios, sin embargo, al analizar los datos se encontró que gran parte de los estudiantes había aprobado únicamente el 20 % de ellos, tal como se muestra en la Figura 4.4. Esta información es importante, ya que al hacer la predicción en este punto de la carrera, en realidad se está haciendo una predicción con datos académicos correspondientes al segundo y tercer semestre para gran parte de los estudiantes.

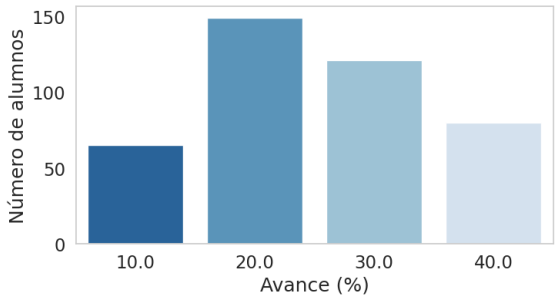

Figura 4.4. Avance académico de los estudiantes al terminar el cuarto semestre.

El conjunto de datos analizado se compone de los datos académicos de 414 estudiantes, de los cuales 185 de ellos sí terminan la carrera y los 229 restantes abandonan la carrera en semestres posteriores. Los cuatro atributos empleados fueron: *Rendimiento general*, *Rendimiento programación*, *Rendimiento DFM* y *Efectividad*; los principales valores estadísticos de estos atributos se muestran en la Tabla 4.15. La variable objetivo de este experimento es *¿Termina?*, de tal forma que, el 44.7 % de las muestras pertenece a la clase *Sí termina* y el 55.3 % pertenece a la clase *No termina*.

Tabla 4.15. Descripción estadística de los atributos del *Experimento 2.1*.

Atributo	Mínimo	Máximo	Media
Rendimiento general	0.87	10.22	4.20
Rendimiento programación	0.00	9.75	5.07
Rendimiento DFM	3.59	9.50	5.69
Efectividad	0.22	1.00	0.69

Para la realización del *Experimento 2.1*, se siguió el mismo procedimiento que en el *Experimento 1.1*, las características generales de ambos experimentos son las mismas (características expuestas en la Sección 4.1.1). Las características específicas de este experimento se presentan a

continuación:

- El porcentaje de estudiantes que pertenece a cada clase fue equivalente en los conjuntos de entrenamiento y de prueba, el 44.7 % de los registros pertenece a la clase *Sí termina* y el 55.3 % a la clase *No termina*, tal como en el conjunto original.
- La Tabla 4.16 muestra los hiperparámetros modificados en este experimento.

Tabla 4.16. Hiperparámetros ajustados en los distintos métodos del *Experimento 2.1.*

Método	Hiperparámetros
Árbol de Decisión	Profundidad máxima del árbol = 3
Bosque Aleatorio	Profundidad máxima de los árboles = 3, número de árboles = 7
Naïve Bayes	Sin modificaciones en los hiperparámetros
Máquina de Soporte Vectorial	C = 10000, gamma = 0.01, tipo de kernel = *rbf*
Redes Neuronales	Una capa de entrada, dos capas ocultas y una de salida (Tabla 4.17)

Tabla 4.17. Estructura de la red neuronal empleada en el *Experimento 2.1.*

Número de capa	Tipo de capa	Número de neuronas	Función de activación
0	Entrada	4	-
1	Oculta	32	ReLU
2	Oculta	32	ReLU
3	Salida	1	Sigmoid

Resultados

Al comparar las matrices de confusión obtenidas (Figura 4.5) se puede observar que, en general, todos los modelos clasificaron incorrectamente a más muestras de la clase *No termina* que de la clase *Sí termina*, lo que sugiere que existe alguna información dentro de los datos que hace que los modelos clasifiquen a estudiantes que no terminan, en la clase *Sí termina*. Por otro lado, a pesar de que todas las matrices de confusión presentan valores muy similares, se puede notar que la matriz de confusión correspondiente al método de Naïve Bayes (Figura 4.5c) muestra los valores más altos, mientras que los más bajos son alcanzados por el método de Redes Neuronales (Figura 4.5e).

A partir de las matrices de confusión se obtuvieron las métricas de evaluación correspondientes (Tabla 4.18). Al examinar la columna referente a la exactitud de los modelos, se puede observar

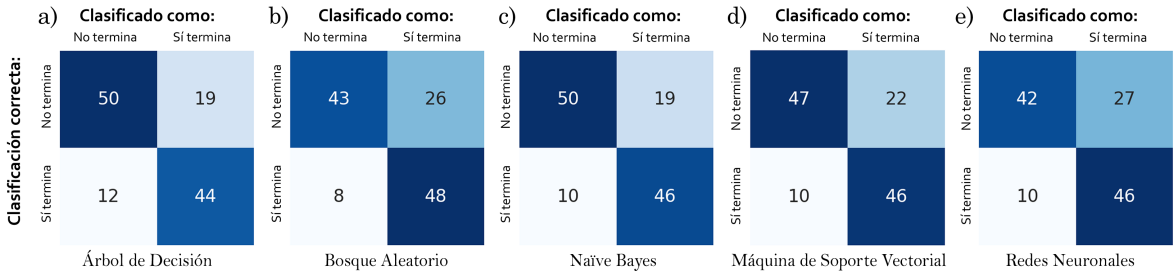

Figura 4.5. Matrices de confusión obtenidas para los diferentes modelos en el *Experimento 2.1*. a) Árbol de Decisión. b) Bosque Aleatorio. c) Naïve Bayes. d) Máquina de Soporte Vectorial. e) Redes Neuronales.

que la exactitud máxima alcanzada es de 0.77 y pertenece al método de Naïve Bayes, mientras que la mínima es de 0.70 y corresponde a Redes Neuronales. La exactitud media obtenida en este experimento es de 0.73, lo que indica que las predicciones realizadas con estos modelos son poco confiables, ya que solo el 73 % de las predicciones realizadas son acertadas. Además, los valores de la precisión son ligeramente menores que los de la exactitud, la precisión media es de 0.67, esto apunta a que todos los modelos tienden a clasificar estudiantes que no terminan la carrera en la clase *Sí termina*.

Tabla 4.18. Métricas de evaluación obtenidas en el *Experimento 2.1*. Los valores en negrita representan los valores más altos de cada métrica.

Método	Exactitud	Precisión	Sensibilidad	Valor F1
Árbol de Decisión	0.75	0.70	0.79	0.74
Bosque Aleatorio	0.73	0.65	**0.86**	0.74
Naïve Bayes	**0.77**	**0.71**	0.82	**0.76**
Máquina de Soporte Vectorial	0.74	0.68	0.82	0.74
Redes Neuronales	0.70	0.63	0.82	0.71

Por otra parte, para conocer la relevancia de cada uno de los atributos se calcularon los valores de su importancia en los distintos métodos, los resultados obtenidos se muestran en la Tabla 4.19. A partir de esta tabla se puede identificar al atributo *Rendimiento general* como el más importante para la predicción, ya que en 3 de los 4 métodos tiene la mayor importancia. El atributo *Efectividad* resultó ser el más importante en el método de Naïve Bayes, y de acuerdo con los resultados previos, este método presenta la mayor exactitud, lo que sugiere que también podría ser un atributo importante. Los atributos de *Rendimiento programación* y *Rendimiento DFM* tienen valores muy similares, son poco relevantes en todos los modelos.

Tabla 4.19. Importancia de cada uno de los atributos analizados en el *Experimento 2.1.* Los valores en negrita indican el valor máximo obtenido en cada método.

Método	Rend_gral	Rend_prog	Rend_DFM	Efectividad
Árbol de Decisión	**0.231**	0.000	0.016	0.000
Bosque Aleatorio	**0.193**	0.002	-0.013	0.011
Naïve Bayes	0.049	0.002	0.005	**0.058**
Máquina de Soporte Vectorial	**0.274**	0.087	0.084	0.035

Como parte complementaria, se analizaron las reglas de asociación extraídas del Árbol de Decisión obtenido, las reglas más destacadas se muestran en la Tabla 4.20. De acuerdo con las reglas presentadas, el principal atributo es *Rendimiento general* y presenta un punto crítico en el valor de 3.65. La Regla 1 señala que los estudiantes que obtienen un valor mayor que 3.65 tienen un 69.9 % de probabilidad de terminar la carrera, mientras que aquellos que obtienen un valor menor o igual, tienen una probabilidad de 80.2 % de no terminar la carrera (Regla 4). Además, la Regla 2 indica que la probabilidad de terminar la carrera aumenta considerablemente si los estudiantes alcanzan un rendimiento superior a 5.71, la probabilidad aumenta a 83.8 % .

Tabla 4.20. Principales reglas de asociación obtenidas del Árbol de Decisión en el *Experimento 2.1.*

Núm.	Regla	Clase	Probabilidad	Número de muestras
1	Rend_gral > 3.65	Sí	69.93 %	143
2	Rend_gral > 5.71	Sí	83.33 %	84
3	Rend_gral > 5.71, Efectividad > 0.96	Sí	95.24 %	21
4	Rend_gral <= 3.65	No	80.27 %	147
5	Rend_gral <= 2.72	No	90.36 %	83

Al final del experimento, se realizó una predicción para saber si un estudiante que tiene los valores medios de cada uno de los atributos termina o no la carrera. Los valores medios de los atributos fueron presentados en la Tabla 4.15 (*Rendimiento general*: 4.20, *Rendimiento programación*: 5.07, *Rendimiento DFM*: 5.69, *Efectividad*: 0.69). La Tabla 4.21 muestra los resultados obtenidos, donde se puede observar que 3 de los 5 modelos predicen que el estudiante no terminará la carrera, mientras que los últimos dos apuntan a que sí terminará. La predicción está dividida, sin embargo, dado que todos los modelos suelen clasificar erróneamente a estudiantes en la clase *Sí termina*, sería más acertado decir que la predicción final es que el estudiante *promedio* no termina la carrera.

Tabla 4.21. Predicción para un estudiante que tiene los valores medios de todos los atributos.

Método	Predicción
Árbol de Decisión	No termina
Bosque Aleatorio	No termina
Naïve Bayes	No termina
Máquina de Soporte Vectorial	Sí termina
Redes Neuronales	Sí termina

A partir de los resultados obtenidos y del análisis realizado, podemos concluir que los datos del historial académico de los estudiantes acumulados hasta el cuarto semestre, no son suficientes para predecir si un estudiante terminará o no la carrera. Utilizando esta información se pueden realizar predicciones con una exactitud máxima de 77 %, lo cual representa un valor poco confiable.

4.2.2. Experimento 2.2

Los resultados obtenidos en el *Experimento 2.1* indicaron que los datos académicos de los estudiantes acumulados hasta el cuarto semestre no son suficientes para cumplir apropiadamente con el objetivo del *Problema 2*. Así, para continuar con el estudio de este problema se planteó el *Experimento 2.2*, este experimento utiliza los datos del historial académico de los estudiantes acumulados hasta el octavo semestre de la carrera. En este punto, los estudiantes deberían haber aprobado el 80 % de los créditos, sin embargo, el análisis de los datos indica que pocos estudiantes alcanzan esta cifra, la Figura 4.6 muestra el avance de los estudiantes que terminan el octavo semestre. El 33 % de los estudiantes presenta un avance correspondiente al séptimo y octavo semestre, el 36 % tiene un avance correspondiente al quinto y sexto semestre, y el resto de los estudiantes muestra un avance correspondiente a semestres anteriores. Esta observación es importante, ya que implica que la predicción realizada en este experimento incluye datos académicos muy variados.

El conjunto de datos aplicado a este experimento, incluye la información académica de 302 estudiantes, integrada por tres atributos (*Rendimiento general*, *Rendimiento programación*, *Efectividad*) y la variable objetivo, *¿Termina?* El atributo *Rendimiento DFM* se eliminó de este conjunto de datos, ya que de acuerdo con la Figura 3.11 la correlación con la variable objetivo

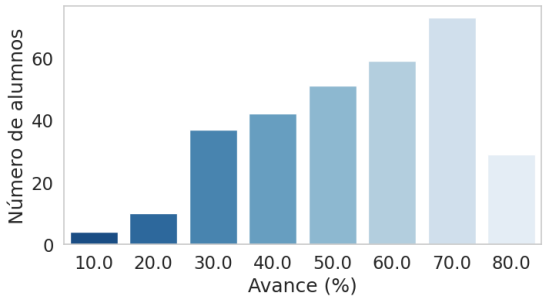

Figura 4.6. Avance académico de los estudiantes al terminar el octavo semestre.

es muy baja, su valor es de 0.11. De los 302 estudiantes que terminan el octavo semestre, 185 de ellos (61.3 %) sí terminan la carrera, mientras que los 117 restantes (37.7 %) abandonan la carrera en semestres posteriores. Así, el 61.3 % de las muestras pertenece a la clase *Sí termina* y el 37.7 % pertenece a la clase *No termina*. Los principales valores estadísticos de los atributos se muestran en la Tabla 4.22.

Tabla 4.22. Descripción estadística de los atributos del *Experimento 2.2*.

Atributo	Mínimo	Máximo	Media
Rendimiento general	0.71	9.73	4.71
Rendimiento programación	0.87	9.75	5.38
Efectividad	0.28	1.00	0.76

El *Experimento 2.2* se realizó con las mismas características generales que el *Experimento 2.1*. Las características específicas, así como los hiperparámetros modificados, se presentan a continuación:

- El porcentaje de estudiantes correspondiente a cada clase fue el mismo en los conjuntos de entrenamiento y de prueba, el 61.3 % de los registros pertenece a la clase *Sí termina* y el 38.7 % a la clase *No termina*, igual que en el conjunto original.
- Los hiperparámetros ajustados en este experimento se muestran en la Tabla 4.23.

Resultados

Para iniciar con el análisis de los resultados del *Experimento 2.2*, en la Figura 4.7 se presentan las matrices de confusión obtenidas para cada uno de los modelos. A grandes razgos, se observa

Tabla 4.23. Hiperparámetros ajustados en los distintos métodos del *Experimento 2.2*.

Método	Hiperparámetros
Árbol de Decisión	Profundidad máxima del árbol = 3
Bosque Aleatorio	Profundidad máxima de los árboles = 3, número de árboles = 17
Naïve Bayes	Sin modificaciones en los hiperparámetros
Máquina de Soporte Vectorial	C = 0.1, gamma = 10, tipo de kernel = *linear*
Redes Neuronales	Una capa de entrada, dos capas ocultas y una de salida (Tabla 4.17)

que las matrices de confusión son muy similares entre sí, existe una variación muy pequeña entre el número de estudiantes clasificados correctamente en cada uno de los modelos. Analizando con más detalle, se puede notar que los métodos de Bosque Aleatorio y Máquina de Soporte Vectorial (Figura 4.7b y Figura 4.7d) consiguen clasificar correctamente a un mayor número de estudiantes. Por su parte, los métodos de Naïve Bayes y Redes Neuronales (Figura 4.7c y Figura 4.7e) son aquellos que presentan más errores en la clasificación.

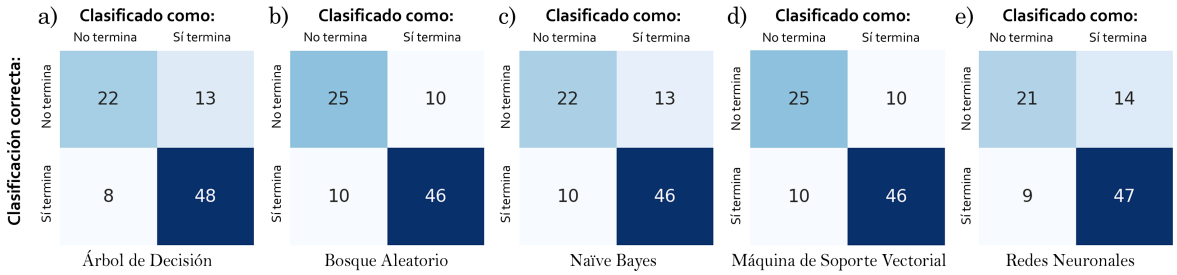

Figura 4.7. Matrices de confusión obtenidas para los diferentes modelos en el *Experimento 2.2*. a) Árbol de Decisión. b) Bosque Aleatorio. c) Naïve Bayes. d) Máquina de Soporte Vectorial. e) Redes Neuronales.

A continuación, se calcularon las métricas de evaluación correspondientes, los valores obtenidos se muestran en la Tabla 4.24. Los resultados muestran que el valor medio de la exactitud es de 0.76, sin embargo, en este experimento es conveniente considerar el valor F1 como principal métrica de evaluación, ya que el conjunto de datos se encuentra ligeramente desbalanceado (61.3 % pertenece a la clase *Sí termina* y 38.7 % a la clase *No termina*). El valor F1 tiene un valor medio de 0.81, donde los métodos de Árbol de Decisión, Bosque Aleatorio y Máquina de Soporte Vectorial alcanzan los mejores resultados, todos ellos con un valor de 0.82.

A continuación, se calculó la importancia de los atributos en cada uno de los métodos, la Tabla 4.25 muestra los resultados obtenidos. De acuerdo con estos valores, se puede identificar al atributo *Rendimiento general* como el más relevante en la predicción del término de la carrera, ya

Tabla 4.24. Métricas de evaluación obtenidas en el *Experimento 2.2*. Los valores en negrita representan los valores más altos de cada métrica.

Método	Exactitud	Precisión	Sensibilidad	Valor F1
Árbol de Decisión	0.77	0.79	**0.86**	**0.82**
Bosque Aleatorio	**0.78**	**0.82**	0.82	**0.82**
Naïve Bayes	0.75	0.78	0.82	0.80
Máquina de Soporte Vectorial	**0.78**	**0.82**	0.82	**0.82**
Redes Neuronales	0.75	0.77	0.84	0.80

que se identifica como atributo principal en la mayoría de los métodos. Por su parte, el atributo *Rendimiento programación* presenta los valores más bajos en todos los métodos, lo que sugiere que, de los tres atributos analizados, es el atributo menos relevante para la predicción del término de la carrera.

Tabla 4.25. Importancia de cada uno de los atributos analizados en el *Experimento 2.2*. Los valores en negrita indican el valor máximo obtenido en cada método.

Método	Rend_gral	Rend_prog	Efectividad
Árbol de Decisión	**0.313**	0.000	0.054
Bosque Aleatorio	**0.324**	0.019	0.088
Naïve Bayes	0.061	0.042	**0.069**
Máquina de Soporte Vectorial	**0.295**	0.001	0.000

Después de conocer la importancia de los atributos, se procedió a extraer las reglas de asociación del Árbol de Decisión. La Tabla 4.26 presenta las principales reglas de asociación obtenidas. De acuerdo con la Regla 1 y la Regla 3 de la tabla, el atributo *Rendimiento general* es el atributo más importante y tiene su punto crítico en el valor de 4.18, a partir de este valor se genera la principal división para la clasificación de los estudiantes. Los estudiantes que obtienen un valor mayor que 4.18 tienen una probabilidad de 88.9 % de terminar la carrera, mientras que aquellos que obtienen un valor menor o igual que 4.18, tienen un 74.1 % de probabilidad de no terminar la carrera. El atributo *Efectividad* tiene una importancia menor, sin embargo, en las reglas 2 y 4, se puede observar que este atributo permite hacer clasificaciones más específicas aumentando la probabilidad de pertenencia a la clase correspondiente.

Para finalizar el *Experimento 2.2*, se realizó la predicción sobre un estudiante *promedio* que termina el octavo semestre. Para ello, se emplearon los valores medios de los atributos utilizados en este experimento, dichos valores fueron presentados en la Tabla 4.22 (*Rendimiento general*:

Tabla 4.26. Principales reglas de asociación obtenidas del Árbol de Decisión en el *Experimento 2.2*.

Núm.	Regla	Clase	Probabilidad	Número de muestras
1	Rendimiento general > 4.18	Sí	88.98 %	118
2	Rendimiento general > 4.18, Efectividad > 0.83	Sí	93.10 %	87
3	Rendimiento general <= 4.18	No	74.19 %	93
4	Rendimiento general <= 4.18, Efectividad <= 0.64	No	89.13 %	46

4.71, *Rendimiento programación*: 5.38, *Efectividad*: 0.76). Las predicciones obtenidas se muestran en la Tabla 4.27. De acuerdo con estos resultados, se puede predecir que un estudiante que tiene valores medios en todos sus atributos al terminar el octavo semestre, sí termina la carrera, ya que los cinco modelos coinciden en este resultado.

Tabla 4.27. Predicción para un estudiante que tiene los valores medios de todos los atributos.

Método	Predicción
Árbol de Decisión	Sí termina
Bosque Aleatorio	Sí termina
Naïve Bayes	Sí termina
Máquina de Soporte Vectorial	Sí termina
Redes Neuronales	Sí termina

El *Experimento 2.2* se concluye resumiendo que los métodos de Árbol de Decisión, Bosque Aleatorio y Máquina de Soporte Vectorial logran los mejores resultados, presentando un valor F1 de 0.82. Además, se identificó al atributo *Rendimiento general* como atributo principal y a la *Efectividad* como segundo atributo relevante. De acuerdo con las reglas obtenidas en el Árbol de Decisión, el atributo *Rendimiento general* tiene un punto crítico en el valor de 4.18, los estudiantes con un rendimiento menor a 4.18 tienen una probabilidad muy alta de no terminar la carrera, en contraste con aquellos que obtienen un valor mayor. Con esta información se puede concluir que los datos académicos de los estudiantes, acumulados hasta el octavo semestre, permiten predecir si los estudiantes terminarán o no la carrera con un 82 % de certeza.

Como conclusión del *Problema 2* se puede decir que los datos académicos acumulados hasta el cuarto semestre no son suficientes para predecir, con buena exactitud, si los estudiantes terminarán o no la carrera. Al considerar los datos académicos acumulados hasta el octavo semestre, los resultados mejoran y la calidad de las predicciones aumentan aproximadamente 10 %, alcanzando valores F1 de 0.82. También es importante recordar que los datos académicos analizados son muy variados, como se pudo observar en la Figura 4.4 y Figura 4.6 gran parte de los estudiantes

presenta rezago en el avance académico, existen estudiantes que no han aprobado ni la mitad de los créditos correspondientes, lo que repercute en las predicciones realizadas.

Capítulo 5

Conclusiones

La deserción escolar es un problema que enfrentan muchas instituciones educativas de nivel superior, incluida la Facultad de Ingeniería de la UASLP. Poder identificar a los estudiantes con alto riesgo de deserción sería una herramienta muy útil que ayudaría en los procesos de tutoría. En este trabajo se analizaron los datos académicos de los estudiantes de las carreras de Ingeniería en Computación e Ingeniería en Informática, de las generaciones 2008 a 2013, con la finalidad de crear modelos predictivos que permitan la identificación de los estudiantes en riesgo de deserción. De acuerdo con el análisis de los datos realizado en este trabajo, se encontró que casi el 49 % de los estudiantes que ingresan a estas dos carreras, desertan durante el primer año. Este alto índice de deserción se debe al requisito del cumplimiento de los 45 créditos establecido por la Facultad. Para abordar este problema, se planteó el primer objetivo del trabajo, tratar de predecir si los estudiantes cumplirán o no los 45 créditos al término del primer año. El problema se analizó desde dos perspectivas diferentes: realizar la predicción con los datos procedentes del proceso de admisión y usando los datos académicos del primer semestre de la carrera de los estudiantes. Los resultados demostraron que los datos del proceso de admisión no son suficientes para predecir si los estudiantes cumplirán o no los 45 créditos. Sin embargo, al utilizar los datos académicos del primer semestre, los resultados mejoran considerablemente y se logra realizar la predicción con un 84 % de exactitud. Dentro de este experimento, se logró reconocer que el *Rendimiento general* de los estudiantes es la mejor métrica para identificar a los estudiantes con alta probabilidad de deserción. Estos resultados indican que la deserción escolar durante el primer año de la carrera depende principalmente de lo que los estudiantes viven durante el primer semestre de la carrera,

y no de las aptitudes con las que ingresa a la Facultad.

Por otra parte, a través del análisis de los datos, también se observó que la deserción escolar es un problema continuo, ya que no solo ocurre durante el primer año de la carrera, sino que continúa a lo largo de los semestres posteriores. De los estudiantes que logran llegar al tercer semestre, el 42 % de ellos deserta en semestres posteriores, la tasa de deserción en semestres avanzados disminuye, manteniendo valores de entre 3 % y 4 %, aproximadamente. Con base en esta información, se planteó el segundo objetivo de este trabajo, tratar de predecir si los estudiantes terminan o no la carrera. Para ello, se consideró la información académica de los estudiantes que sí cumplieron los 45 créditos y se inscribieron al tercer semestre. La primera predicción se realizó con los datos académicos acumulados hasta el cuarto semestre de la carrera, y posteriormente, se llevó a cabo la predicción con los datos académicos acumulados hasta el octavo semestre de la carrera. Una observación importante, es que gran parte de los estudiantes presenta un rezago académico notable, lo que genera que la información de los conjuntos de datos analizados sea un poco heterogénea. De los resultados obtenidos, se puede concluir que la información acumulada hasta el cuarto semestre no permite predecir con certeza si los estudiantes terminarán o no la carrera, estos datos permiten clasificar a los estudiantes con un 77 % de exactitud. Al realizar la predicción con los datos académicos acumulados hasta el octavo semestre los resultados mejoran, las predicciones realizadas con estos datos alcanzan un valor F1 de 0.82, lo cual representa un buen valor, ya que indica que el 82 % de las predicciones realizadas son acertadas. Por otra parte, también se observó que el *Rendimiento general* de los estudiantes es el mejor atributo para predecir si los estudiantes terminan o no la carrera. El rendimiento de los estudiantes en el área de Programación y en el DFM, no resultó tan relevante para la predicción como el *Rendimiento general*, estos dos atributos tuvieron poca importancia en todos los modelos generados. Los resultados obtenidos en estos experimentos demostraron que la calidad de las predicciones mejora conforme los semestres avanzan, es posible identificar a gran parte de los estudiantes que presentan alto riesgo de deserción, sin embargo, se podría considerar la información de semestres posteriores para aumentar el número de predicciones correctas.

A partir de estas conclusiones es posible encontrar algunas similitudes con los trabajos relacionados presentados en la Sección 2.4. Por ejemplo, en los trabajos de Opazo [31] y Lázaro Álvarez [30] se realizó la predicción de la deserción escolar utilizando los datos de ingreso, obteniendo una

exactitud máxima de 69 % y 60 %, respectivamente; mientras que en nuestro estudio al realizar la predicción del cumplimiento de los 45 créditos utilizando los datos del proceso de admisión se obtuvo una exactitud máxima de 68 %. A pesar de que todos ellos fueron realizados bajo condiciones muy diferentes, los resultados presentan cierta similitud en la exactitud obtenida. Por otro lado, en el estudio realizado por Yaacob y colaboradores [29] se presentó la predicción de la deserción escolar utilizando datos académicos del tercer año de los estudiantes, la exactitud obtenida en dicho estudio tuvo valores entre 81.2 % y 90.8 %, mientras que en nuestro estudio realizado con los datos académicos acumulados hasta el cuarto año, se alcanzó un valor F1 de 82 %, lo que sugiere, que a pesar de ser experimentos muy diferentes, existe cierta similitud en las métricas de evaluación obtenidas. A partir de esto, podemos decir que nuestro estudio resulta similar, en cuanto a métodos y resultados, a otros trabajos relacionados.

Desde el punto computacional, se puede concluir que existen factores que fueron fundamentales al realizar la predicción de la deserción escolar, tales como la elección de un conjunto de datos apropiado, la selección de las herramientas y métodos convenientes, la manipulación adecuada del conjunto de datos, la extracción de los atributos representativos, la aplicación de los métodos de predicción apropiados, el ajuste correcto de los hiperparámetros de cada uno de ellos y la elección de los instrumentos pertinentes para el análisis de los resultados.

Para finalizar, podemos concluir que es difícil predecir si un estudiante terminará o no la carrera utilizando únicamente información académica, en muchas ocasiones, la deserción escolar se debe a factores externos al entorno académico que no es posible considerar y dificulta la predicción. Con los datos académicos empleados en este trabajo se logró clasificar correctamente a más del 80 % de los estudiantes, lo cual representa un buen resultado. Los resultados podrían mejorar si se añadiera información personal o socioeconómica, sin embargo, también es posible hacer nuevas manipulaciones al conjunto de datos para crear nuevos atributos y evaluar su importancia en la deserción escolar.

5.1. Trabajo Futuro

El resultado obtenido a partir de cualquier modelo predictivo depende esencialmente de la manipulación previa del conjunto de datos utilizado y de los método aplicados, por ello, a partir de un

mismo conjunto de datos es posible obtener distintos modelos, y por ende, diferentes resultados. Con base en esto, y de acuerdo con los resultados obtenidos en este trabajo, existe la posibilidad de crear modelos que permitan obtener predicciones más acertadas para nuestro problema de investigación. La combinación de posibles modificaciones es muy variada, existen diversas características que podrían ayudar a obtener mejores predicciones, algunas de ellas se mencionan a continuación.

- **Incorporación de datos socioeconómicos.** Como se mencionó previamente, existen factores personales que favorecen la deserción escolar, si se pudiera añadir algún tipo de información socioeconómica de los estudiantes al conjunto de datos, se podría mejorar la calidad de las predicciones realizadas por los modelos.

- **Actualización del conjunto de datos.** Los datos utilizados en este trabajo corresponden a los estudiantes de las generaciones 2008 a 2013 de las carreras de Ingeniería en Computación e Ingeniería en Informática, sin embargo, este estudio puede ser ampliado utilizando la información de generaciones posteriores. Por otro lado, también es posible realizar un estudio similar para los estudiantes de la carrera de Ingeniería en Sistemas Inteligentes, el único inconveniente, es que actualmente la información de los estudiantes de esta carrera podría ser insuficiente, ya que la primera generación ingresó en el año 2017, habría que esperar un par de años para contar con un conjunto de datos más extenso, principalmente para la predicción del término de la carrera.

- **Manipulación del conjunto de datos.** La creación de nuevos atributos y la modificación de los atributos ya existentes son un buen punto para continuar con este trabajo de investigación. El conjunto de datos original puede ser manipulado nuevamente para la creación de nuevos atributos o los atributos ya creados pueden ser modificados. Algunas modificaciones que podrían realizarse a los atributos originales, sería, por ejemplo, la discretización, existen atributos continuos que podrían discretizarse y analizar su funcionalidad para la creación de los modelos. Otra opción, sería analizar el rendimiento de los estudiantes por semestre, no de forma general como se hizo en este trabajo, eso podría ayudar a identificar semestres críticos y, tal vez, podría mejorar la calidad de las predicciones. Incluso, se podría modificar la variable objetivo, en lugar de trabajar una clasificación binaria (*Sí termina*/*No*

termina), podría ser una clasificación multiclase añadiendo niveles de riesgo, *Alto*, *Medio*, *Bajo*, etc. Otra perspectiva interesante podría ser la predicción utilizando como referencia el número de créditos que han acumulado los estudiantes, por ejemplo, realizar la predicción cuando los estudiantes han cumplido con el 20 % o 40 % de los créditos, en lugar de emplear el semestre cursado como referencia. Estos son solo algunos ejemplos de lo que podría modificarse, sin embargo, existen muchas posibilidades.

- **Selección de los métodos de predicción.** Los métodos de predicción, así como el ajuste de los hiperparámetros también son fundamentales para el desempeño de los modelos. En este trabajo se utilizaron cinco modelos de predicción distintos, sin embargo, solo el Bosque Aleatorio es un método basado en el ensamble de modelos. Los métodos de ensamble ayudan a mejorar el rendimiento de los modelos, de hecho, en los resultados se pudo observar que el Bosque Aleatorio fue uno de los métodos que obtuvo mejores resultados en todos los experimentos. Por esta razón, aplicar otros métodos de ensamble podría ayudar a mejorar los resultados del problema de investigación.

Referencias

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[2] Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. Future generation computer systems, 13(2-3), 99-115.

[3] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. Ieee Access, 5, 15991-16005.

[4] Alturki, S., Hulpu, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. Technology, Knowledge and Learning, 1-33.

[5] Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. Mining educational data to improve students' performance: a case study, 2(2).

[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

[7] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. & Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá: Ediciones Universidad Cooperativa de Colombia, 63-86.

[8] 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.

[9] Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees: theory and applications. World scientific.

[10] Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. ACM Computing Surveys (CSUR), 53(3), 1-34.

[11] Géron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: Concepts. Tools, and Techniques to build intelligent systems.

[12] Bramer, M. (2016). Principles of data mining. London: Springer.

[13] Bagnato, J. I. (2020). Aprende Machine Learning en español: teoría+ práctica Python. Amazon Italia Logistica.

[14] Aggarwal, C. C. (2015). Data mining: the textbook. New York: springer.

[15] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. New York: springer, 2, 1-758.

[16] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

[17] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

[18] McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, 445, 51-56.

[19] Reback, J., McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Klein, A., ... & Seabold, S. (2020). pandas-dev/pandas: Pandas 1.0. 5. Zenodo.

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. The Journal of machine Learning research, 12(1), 2825-2830.

[21] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project, 108-122.

[22] Chollet, F. (2015). keras.

[23] Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004, June). A case study of knowledge discovery on academic achievement, student desertion and student retention. In ITRE 2004. 2nd International Conference Information Technology: Research and Education (pp. 150-154). IEEE.

[24] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

[25] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

[26] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM | Journal of Educational Data Mining, 1(1), 3-17.

[27] Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies, 23(1), 537-553.

[28] García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining.

[29] Yaacob, W. W., Sobri, N. M., Nasir, S. M., Norshahidi, N. D., & Husin, W. W. (2020, March). Predicting student drop-out in higher institution using data mining techniques. In Journal of Physics: Conference Series (Vol. 1496, No. 1, p. 012005). IOP Publishing.

[30] Lázaro Alvarez, N., Callejas, Z., & Griol, D. (2020). Predicting Computer Engineering students' dropout in Cuban Higher Education with pre-enrollment and early performance data. JOTSE: Journal of Technology and Science Education, 10(2), 241-258.

[31] Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. Mathematics, 9(20), 2599.

[32] Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?. Mathematics, 10(18), 3359.

[33] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education: Artificial Intelligence, 3, 100066.