BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

Facultad De Ciencias De La Computación

TESIS

"Desarrollo de un Sistema de Clasificación y Predicción del Aprovechamiento Académico de Estudiantes de Nivel Medio Superior"

Presenta: Marlem Martínez Castillo

Para obtener el grado de: Licenciatura en Ciencias de la Computación

Director: Yolanda Moyao Martinez
Codirector: Carmen Cerón Garnica

Puebla, Pue, diciembre 2025

Dedicado a

mis papas y hermanos

Agradecimientos

Quiero agradecerle a dios y a la vida por permitirme la oportunidad de lograr una de mis metas a largo plazo. A mis padres María del Rayo y José Cesar, a quienes agradezco por todo el amor, apoyo, paciencia y sacrificio que hicieron para apoyarme incondicionalmente en este largo camino.

A mis hermanos y sobrino, agradezco todos sus consejos y enseñanzas.

A Hugo JG. agradezco por estar siempre conmigo, ser mi soporte, y apoyo.

A mis compañeros y amigos por todo el apoyo incondicional, consejos y alentó brindado durante toda esta etapa.

A Keysi y Katy, por su alegría y compañía en todas las noches de trabajo y desvelo.

Agradezco a mis asesoras Yolanda Moyao Martinez y Carmen Cerón Garnica por la paciencia, dedicación, enseñanza y colaboración que me permitió el desarrollo y elaboración de este trabajo.

También agradezco a la Vicerrectoría de investigación y Estudios de Posgrado (VIEP) por el apoyo, a través del financiamiento otorgado para realizar este trabajo de investigación.

"A todos muchas gracias"

Resumen

La deserción escolar es un problema primordial, que afecta el desarrollo académico y social del estudiante en especial la educación Media Superior en México, en promedio estudiantes que culminan los estudios son del 68% mientras que un 32% opta por abandonar. Por lo general, el desempeño es evaluado mediante pruebas que ayudan a medir el rendimiento y aprendizaje académico.

En este contexto, las evaluaciones proporcionan información parcial sobre si el desempeño a alcanzado los resultados esperados, esto demuestra la importancia de medir el nivel de conocimiento adquirido y validar si el aprendizaje ha sido exitoso. Por lo anterior, es fundamental desarrollar soluciones que permitan clasificar y predecir el aprovechamiento académico de los estudiantes de manera eficiente y precisa.

El presente trabajo tiene como objetivo el desarrollo de un sistema basado en el análisis de datos y algoritmos de aprendizaje automático que permitan identificar patrones que influyen en la posibilidad de un abandono estudiantil. Esta herramienta permite generar informes detallados acerca del desempeño de los estudiantes, teniendo así una visión integral que permita identificar problemas de manera oportuna y puntual, y así poder aplicar intervenciones que ayuden en la mejora del aprovechamiento académico.

ÍNDICE

ÍNDICE FIGURAS

Introducción

En la actualidad entender y comprender el panorama estudiantil juega un papel importante, desde la educación básica hasta la educación superior, la implementación de sistemas de clasificación y predicción de aprovechamiento académico es un indicador clave para comprender el nivel estudiantil y nivel de compresión de cada uno. La predicción del rendimiento académico es un campo de investigación fundamental y relevante, ya que contribuye a una mejora al desarrollo académico, permite la valoración del aprendizaje, y facilita la toma de decisiones basada en datos reales.

Al basar el análisis en nivel medio superior, se resalta que el estudiante, se aproxima al nivel superior (universidad) donde se define su futuro académico y la incorporación al mundo laboral. Gracias al avance de tecnologías y algoritmos, hoy en día es posible analizar, identificar, clasificar y predecir el rendimiento académico basado en información estudiantil, permitiendo la implementación de estrategias e intervenciones tempranas.

Este estudio tiene como objetivos el desarrollo de un Sistema de Clasificación y Predicción del Aprovechamiento Académico de Estudiantes de Nivel Medio Superior, basado en técnicas de minería de datos y aprendizaje automático, con el objetivo principal de la detección de estudiantes en peligro de deserción estudiantil, aplicando técnicas de minería de datos, aprendizaje automático y algoritmos de ----, con el fin de evaluar en entornos reales la precisión de resultados. Este estudio no solo busca demostrar la eficacia del desarrollo del sistema, sino también la importancia de mejorar la calidad de aprendizaje, el acompañamiento y la orientación educativa de los estudiantes.

CAPITULO 1: ANALISIS Y PLANTEAMIENTO

1.1 Definición del problema

En México, en pleno 2024, la deserción estudiantil sigue siendo uno de los problemas que atañe a la sociedad en general. Este problema se refiere a la interrupción de los estudios en cualquier nivel educativo, lo cual trae graves consecuencias en el desarrollo personal y profesional de los estudiantes. Muchas instituciones educativas carecen de herramientas para anticipar los posibles cambios en el rendimiento académico de los estudiantes, y como consecuencia, no se cuenta con estrategias de apoyo y orientación para evitar la deserción estudiantil.

La educación media superior comprende a niveles de bachillerato, bachiller técnico, preparatoria y similares, en un grupo de edad de 15-17 años. Según el Instituto Nacional de Estadística y Geografía (INEGI), la deserción estudiantil es muy cambiante en los últimos ciclos escolares. En el ciclo de 2021-2022 el índice de abandono fue de 11.3%; en 2022-2023 el índice disminuye a 9.7 % y en 2023-2024 desciende a 8.7%.

La educación tradicional evalúa mediante exámenes y calificaciones acumulativas, sin tomar en cuenta otros factores importantes que intervienen en el desempeño académico. Existen estudios que identifican diferentes factores, como problemas económicos, falta de apoyo familiar, bajo rendimiento y desmotivación, y que en gran medida contribuyen a este fenómeno social de decesión escolar y que impactara e incrementara la posibilidad de vulnerabilidad y marginación. Debido a la falta de identificación y seguimiento temprano de riesgos, no es posible brindar acciones de atención y orientación oportuna que contribuyan a reducir o eliminar el problema de deserción escolar.

Ante esta situación, es fundamental desarrollar soluciones que permitan clasificar y predecir el aprovechamiento académico de los estudiantes de manera temprana, identificando patrones en la información académica correspondiente a los alumnos en estudio. Con esta herramienta, las instituciones pueden clasificar el rendimiento académico y aplicar medidas de apoyo para reducir la deserción escolar.

1.2 Factores influyentes

En base a la Encuesta Nacional de Deserción de la Educación Media Superior (EDEMS 2012) realizada por la secretaria de Educación Pública, factores influyentes que pueden estar relacionados con deserción estudiantil son:

Factores personales: Los factores personales son aquellos que dependen directamente solo del alumno y de su capacidad, tales como:

- Falta de interés o motivación.
- Indisciplina, reprobación o inasistencia.
- Reprobación, suspensión o expulsión.
- Embarazo, unión o matrimonio.

Factores sociales y familiares: Las condiciones de estudio en el hogar de igual manera influyen en el entorno físico del alumno. En situaciones como:

- Falta de apoyo familiar: Bajo interés y valor en la educación del estudiante y mayor importancia a tareas del hogar, provoca que el estudiante no perciba como necesario el hecho de estudiar.
- Violencia intrafamiliar: Ambientes conflictivo e intranquilos desmotivan al estudiante.
- Falta de dinero: El estudiante tiende a abandonar los estudios para ayudar al sustento del hogar.

Factores académicos

Las condiciones y ubicación geográfica de la institución educativa influyen en la permanencia del alumnado.

- Ubicación escolar: La institución se encuentra demasiado lejos o no existe.
- Falta de atención: La escuela no cuenta con estrategias para resolver las dificultades presentadas durante el proceso de aprendizaje.

Factores económicos: El factor socioeconómico, es uno de los principales factores externos asociado a la interrupción de estudios.

- Falta de dinero: El estudiante tiende a abandonar los estudios para aportar dinero al hogar.
- Falta de acceso a recursos: El alumno no puede sustentar materiales escolares necesarios.

Se observa, que los factores personales, familiares y económicas integran los principales detonadores de la deserción estudiantil. Así que, identificarlos y analizarlos evitara la mayoría de los casos de deserción estudiantil.

1.3 Exploración y recolección de datos

La recolección de datos utilizados para este estudio fue recolectada principalmente, a través del formulario "*Análisis Comparativo de Modelos de Clasificación para la Predicción del Aprovechamiento Académico de Estudiantes de dos Preparatoria*", con el objetivo de analizar factores que influyen en el rendimiento académicos de estudiantes de nivel medio superior.

El formulario está compuesto con un total de 31 preguntas de opción múltiple o respuesta abierta, obteniendo tanto datos cuantitativos como cualitativos.

Se estructuró en varias categorías para recolección de información relevante sobre el estudiante:

- Datos personales: Edad, sexo, zona y preparatoria.
- Entorno familiar: Miembros en la familia, estado de convivencia, calidad de relación familiar, nivel educativo y profesiones de padres.
- Hábitos de estudio: Tiempo dedicado a estudiar, materias reprobadas, apoyo escolar o familiar, clases particulares, actividades extraescolares y ausencias escolares.
- Situación social: Relaciones románticas, tiempo libre semanal y frecuencia de salida de amigos.
- Salud: Estado de salud y consumo de alcohol.
- Aspiraciones académicas: Intenciones de estudiar una licenciatura.

El formulario fue distribuido de manera digital y directamente a estudiantes de la preparatoria Benito Juárez García y al Complejo Regional Tehuacán, con el fin de alcanzar al público objetivo "estudiantes de nivel medio superior".

El formulario estuvo disponible en un periodo de 31 días, con fecha de inicio 7 de mayo del 2025 y cierre 7 de junio de 2025, finalmente al cierre de recolección de datos se obtuvieron un total de 606 respuestas.

1.4Objetivo general

Desarrollar un sistema para clasificar y predecir el rendimiento académico de los estudiantes utilizando técnicas de minería de datos y aprendizaje automático, con el fin de identificar patrones y proponer estrategias de mejora y prevención.

1.4.1 Objetivos particulares

- Analizar factores cruciales en el aprovechamiento académico de estudiantes de nivel medio superior.
- Investigar, elegir y aplicar algoritmos y técnicas idóneas de minería de datos para limpieza, procesamiento y tratamiento de datos.

- Diseño y entrenamiento de modelos de clasificación y predicción, para estimar el rendimiento académico de los estudiantes.
- Desarrollo de una interacción amigable para el acceso del sistema de predicción a través de una interfaz intuitiva para el usuario.
- Implementar y validar el sistema en simulación o en un contexto real.

1.5 Justificación

El aprovechamiento académico es clave para la formación educativa de todo estudiante, sin embargo, como se menciona antes se presentan diversos factores y retos que llegan a afectar significativamente el desempeño estudiantil.

Este estudio de investigación proporcionara una herramienta efectiva al personal académico que, a través del uso de técnicas de minería de datos y aprendizaje automático, les ayude a identificar a los estudiantes que se encuentran bajo riesgo de alguna deserción escolar debido a su bajo rendimiento académico e incluso debido a otros factores.

Esta investigación contribuirá a:

- Anticipar posibles dificultades y riesgos académicos y la adecuada toma de decisiones en el nivel medio superior.
- Implementar intervenciones personalizadas y tempranas que ayuden a prevenir la deserción estudiantil.
- Mejorar el rendimiento académico de los estudiantes y, finalmente, a elevar el nivel educativo, las oportunidades sociales y económicas de los estudiantes.

CAPITULO 2: ESTADO DEL ARTE

2.1 Investigación de trabajos relacionados

Para entender la problemática, se han revisado diferentes trabajamos relacionados que han abordado el tema desde distintas perspectivas y soluciones:

El estudio "Predicción del Rendimiento Académico usando Inteligencia Artificial", aborda la predicción en el rendimiento académico de los estudiantes de nivel superior a través de un análisis de factores definidos como (educacional, familiar, socioeconómicos, entre otros). Utilizan técnicas de inteligencia artificial (clasificadores bayesianos) diseñan una metodología de aprendizaje para entrenar un sistema capaz de clasificar el rendimiento del estudiante, y predecir casos de estudiantes con problemas graves de rendimiento académico. Este sistema fue realizado y probado en estudiantes de la Universidad pública en Colombia [1].

En el enfoque de estudio "Árboles de Decisión como Metodología para Determinar el Rendimiento Académico en Educación Superior", utilizan métodos y técnicas de modelado estadístico como árboles de decisión y regresión lineal múltiple, con el objetivo de analizar variables independientes asociadas al rendimiento académico y detectar alumnos que requieren algún tipo de apoyo académico o atención externa.

Este estudio fue realizado en una institución de nivel superior ubicada en la zona urbana de Pánuco, Veracruz, se llegó a la conclusión de que es necesario reducir la sobrecarga de trabajo de los alumnos; sin embargo, este fenómeno de sobrecarga de trabajo está más marcado cuando se realiza en fecha de exámenes, donde los alumnos suelen tener niveles más altos de estrés. Por lo tanto, el estudio

se realizó sobre sujetos con estas condiciones, con el objetivo de realizar un análisis más preciso y obtener resultados más certeros [2].

Existen diferentes métodos de análisis y enfoques para analizar dicha problemática. Por ejemplo, el artículo "Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos", utiliza datos socio demográficos y resultados de exámenes de admisión de 415 alumnos de las carreras referidas al área de tecnología de la Universidad Autónoma de Yucatán (México), inscritos en un periodo del año de 2016 al año 2019.

Estos datos fueron considerados para el estudio y el análisis para crear modelos predictivos de riesgo académico con ayuda de métodos de minería de datos. El algoritmo de aprendizaje automático obtuvo una precisión del 75.42% en sus predicciones, además de que su curva ROC de 0.805 garantiza la calidad del modelo de clasificación para la detección de estudiantes en riesgo alto en comparación de estudiantes sin riesgos.

Estos resultados sugieren que es posible, a través de la minería de datos y de algoritmos de aprendizaje automático, el desarrollo de un sistema para identificar estudiantes con características que ponen en riesgo su deserción en una etapa temprana, y además diseñar estrategias de intervención que sirvan como apoyo a los estudiantes en riesgo [3].

Se observa que en el trabajo "Un modelo analítico para la predicción del rendimiento académico de "estudiantes de ingeniería", se utiliza principalmente la minería de datos para comprender de forma eficiente no solo los procesos de aprendizaje, sino que también el contexto donde se llevan a cabo.

Este trabajo se enfoca en la estrategia "Learning analytics", que consiste en medir, recolectar, analizar los datos recolectados en la comunidad educativa, es un método centrado en la mejora de la educación, mejora en los factores de enseñanza y mejora en los lineamientos educativos, con el objetivo de mejorar la

experiencia educativa a través del estudio de todos los datos recolectados y analizados.

Posteriormente, se utilizaron herramientas de learning analytics para crear modelos que identifiquen la deserción académica de alumnos de Ingeniería de la Universidad de Chile. Estos resultados arrojaron una clasificación del 86% de los casos y una clasificación de niveles bajos con una precisión de 38%. Así que con ello fue posible identificar los casos, donde fue necesario aplicar la intervención vocacional en estudiantes con riesgo de deserción [4].

Por otro lado, se tiene la aplicación de un enfoque de evaluación basado en Aprendizaje Automático (Machine Learning), "Predicción de rendimiento académico de alumnos, usando machine learning", el cual tiene como objetivo principal la predicción del rendimiento académico en estudiantes de nivel superior.

Se enfoca en el empleo de técnicas de aprendizaje automático consideradas tales como árboles de decisión, bosques aleatorios, redes neuronales y máquinas de soporte vectorial. Estas herramientas combinadas con técnicas de ciencia de datos permiten realizar un análisis de sentimientos, cuyo objetivo principal es evaluar las emociones y opiniones de los estudiantes. Además, de implementar un sistema de recomendaciones con el objetivo de detectar preferencias y con ello proporcionar orientación personalizada para cada estudiante [5].

A manera de conclusión, la Tabla 1. Muestra un comparativo entre los diferentes modelos, presentados previamente.

Tabla 1.

Comparación de sistemas de predicción.

Sistema	*Características útiles*	*Limitaciones*

Predicción del Rendimiento Académico usando Inteligencia Artificial	○ Uso de técnicas de IA para la predicción de rendimiento. ○ Considera variedad de factores influyentes. ○ Gran porcentaje de efectividad. ○ Permite detección temprana a estudiantes en riesgo. ○ Modelo preciso con alto porcentaje de eficacia.	× No existen intervenciones en caso de estudiantes en riesgo. × Sistema probado solo en una pequeña parte de la población estudiantil.
Árboles de Decisión como Metodología para Determinar el Rendimiento Académico en Educación Superior	○ Uso de árboles de decisión y regresión lineal para identificar variables asociadas al rendimiento académico. ○ Detección de variables clave. ○ Propone estrategias de mejora, además de encuestas en los momentos de menor estrés.	× Enfoque solo en las asignaturas relacionadas a programación. × Resultados basados en una pequeña área de estudiantes.
Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos	○ Uso de la minería de datos para generar modelos predictivos en estudiantes de bajo rendimiento. ○ Uso de técnicas de clasificación para identificar variables predictivas.	× Resultados basados en un área pequeña. × Enfoque realizado con solo un modelo de predicción. × No existen intervenciones específicas para la detección temprana.
Predicción de rendimiento académico de alumnos usando machine learning	○ Uso de diversas técnicas de aprendizaje automático (arboles de decisión, bosques aleatorios, redes neuronales y máquinas de soporte vectorial) para la predicción de rendimiento académico.	• El estudio se enfoca solo en resultados previos, pero no aborda cambios a mayor plazo.

Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería	o Uso innovador de learning analytics para predecir la deserción y rendimiento. o Modelo predictivo del 86% de precisión. o Busca predecir la deserción temprana para una adecuada prevención.	× Enfoque limitado solo en alumnos de primer año de nivel superior. × Enfoque realizado solo en plan común de ingeniería. × Uso de datos limitados, es decir información hasta segundo semestre.

Fuente: Elaboración propia.

2.2 Métodos y tecnologías

Para el trabajo de investigación a desarrollar, es importante revisar método, algoritmos y tecnologías utilizables para el sistema de predicción.

La minería de datos junto con el aprendizaje automático son herramientas para la mejora de resultados. Ya que la minería de datos descubre patrones y el aprendizaje automático tiene por objetivo aprender de ellos para hacer predicciones y automatizar decisiones.

Tecnologías para el aprendizaje automático y la minería de datos:

- **Herramienta R**: Entorno de software y lenguaje de programación para análisis estadístico y gráficos. Se utiliza para el análisis de datos, así como para el desarrollo de minería de datos y aprendizaje automático.

  Paquetes para el análisis de datos y modelado estadísticos:

  - Caret: Facilita construcción de modelos predictivos, con herramientas para el proceso de datos, para el ajuste de modelos y

para la evaluación del rendimiento. Además de comparar diferentes algoritmos de aprendizaje automático.

- randomForest: Implementación del algoritmo "Random Forest", algoritmo basado en arboles de decisión donde se crea una serie de árboles de decisión aleatorios para desarrollar predicciones.
- Nnet: Utilizado para redes neuronales (perceptrón multicapa), por la proporción de modelado no lineal y clasificación.

- **Weka:** Plataforma de código abierto muy utilizada para la minería de datos y el aprendizaje de datos:
  - Algoritmos de aprendizaje automático: Arboles de decisión, Maquinas de soporte vectorial, redes neuronales, k-means, regresión logística y random forest.
  - Procesamiento de datos: Weka ofrece herramientas para la limpieza de datos y transformación de datos. Realiza tareas como discretización, normalización y variables categóricas.
  - Visualización: La creación de graficas ayuda a una visualización clara de los datos y resultados de los modelos.
  - Formato de datos: El formato utilizado por weka es .arff, sin embargo, admite otros formatos para archivos, como .csv y para base de datos admite sql.

- **RapidMiner:** Herramienta parecida a weka en el aspecto de análisis de datos, minería de datos y aprendizaje automático. A diferencia que RapidMiner se centra más en el proceso de modelado de datos.
  - Algoritmos de clasificación (clasificación, regresión, agrupación, análisis de texto, redes neuonales, etc.)
  - Procesamiento de datos: Ayuda al preprocesamiento de limpieza, transformación y preparación de datos para el modelado, como

normalizar, manejo de valores nulos, fusión y división de conjuntos de datos.

- ❖ Automatización: Optimiza procesos de minería de datos, útil para situaciones de grandes volúmenes de datos.
- ❖ Modelado predictivo: Enfoque de utilidad centrado en modelos predictivos a través de datos para la toma de decisiones.

- **Phyton**: Lenguaje popular y fácil para la minería de datos y el aprendizaje automático. Sus bibliotecas son clave para la manipulación, transformación y visualización de datos.
  - ❖ Panda: Biblioteca de phyton para la manipulación y análisis de datos, permite agrupaciones, fusiones y pivotado de datos.
  - ❖ NumPy: Biblioteca para la manipulación de matrices y cálculos numéricos.
  - ❖ Matplolib y Seaborn: Permite la visualización de datos a través de gráficos.
  - ❖ Scrapy y BeautifulSoup: Ayuda a la extracción de datos de páginas web y xml.
  - ❖ Sciki-learn: Ofrece variedad de algoritmos para machine learning.
  - ❖ TensorFlow: Bibiblioteca desarrollada por Google, con objetivo de la creación y entrenamiento de modelos de aprendizaje automático, enfocado principalmente en las redes neuronales profundas.

2.3 Algoritmos de clasificación

Técnicas para asignar categorías a conjuntos de datos. Se utiliza para predecir el rendimiento, para clasificar a los estudiantes en distintas categorías de acuerdo con su nivel de desempeño.

- Arboles de decisión: División de datos según características, como asistencia, calificaciones y factores económicos.

- Redes neuronales: Detección de patrones en base en las relaciones entre los datos.

Algoritmos de regresión:

Técnicas que modela relaciones entre variables dependientes o variables independientes. Se utiliza para predecir valores, como las calificaciones. Emplea instancias como:

- Regresión lineal: Es una técnica que se utiliza para relacionar variables como calificaciones, asistencias, etc, con el objetivo de predecir una nota final.

Algoritmos de Clustering:

Técnicas que agrupas datos en grupos basado en similitudes, llamados "clusters". Se utiliza para la detección de patrones de rendimiento. Emplea instancias como:

- k-means: Agrupación en clústeres según el rendimiento.

2.4 Análisis de tecnologías

Existen diferentes propuestas de herramientas y tecnologías para dar solución a dicha problemática. Presentada en diferentes categorías y esquemas de tablas para describir ventajas y desventajas.

Minería de datos

Se presenta en la Tabla 2. Un comparativo entre las diferentes tecnologías, para la minería de datos.

Tabla 2.

Tecnologías para manejar Minería de datos.

Fuente: Elaboración propia.

Tecnología	*Ventajas*	*Desventajas*
Weka	✓ Plataforma completa con variedad de algoritmos. ✓ Permite distintos formatos de datos. ✓ Permite realizar procesamiento de datos ✓ No se necesita programación avanzada. ✓ Permite visualizar gráficos.	× Interfaz anticuada. × No está preparado para grandes volúmenes de datos.
RapidMiner	✓ Incluye variedad de algoritmos. ✓ Ofrece herramientas avanzadas para el aprendizaje automático. ✓ Permite diversos formatos de datos. ✓ Adecuada para grandes volúmenes de datos.	× Requiere licencia para opciones avanzadas.
Phyton (Pandas, Numpy, Sciki-learn)	✓ Manipulación de datos. ✓ Integración con más bibliotecas de phyton. ✓ Algoritmos de machine Learning. ✓ Proporciona validación y evaluación de modelos. ✓ Mayor rendimiento.	× Requiere mayor conocimiento en programación.

Aprendizaje automático

Se presenta en la Tabla 3. Un comparativo entre las diferentes tecnologías, para aprendizaje automático.

Tabla 3.

Tecnologías para aprendizaje automático.

Tecnología	*Ventajas*	*Desventajas*
TensorFlow	✓ Diversidad de modelos y algoritmos. ✓ Ideal para redes neuronales. ✓ Arquitectura flexible.	× Complejo de usar. × No recomendable para modelos de alto rendimiento.
Scikit-learn	✓ Variedad de algoritmos preimplementados. ✓ Fácil de utilizar. ✓ Documentación detallada. ✓ Evaluación y validación de modelos. ✓ Integrable con otras bibliotecas.	× No adecuado con redes neuronales. × No es recomendable para alto volumen de datos

Fuente: Elaboración propia.

Algoritmos de predicción

Se presenta en la Tabla 4. Un comparativo entre los diferentes algoritmos de predicción.

Tabla 4.

Algoritmos de predicción.

Algoritmo	*Ventajas*	*Desventajas*
Redes neuronales	✓ Eficiente para problemas complejos. ✓ Adaptable a ajuste de datos. ✓ Alta capacidad para detección de errores.	× Eficiente a mayor cantidad de datos. × Mayor consumo de recursos computacionales.
Arboles de decisión	✓ Interpretación intuitiva y fáciles.	× Inestable, pequeño ajuste en datos cambia todo.

	✓ Maneja bien valores faltantes. ✓ Maneja datos categóricos y numéricos.	
Regresión logística	✓ Fácil de interpretar. ✓ Poco tiempo de entrenamiento. ✓ Útil para clasificación binaria	× No existen relaciones lineales

Fuente: Elaboración propia.

2.4 Limites

Es de importancia señalar las diversas limitaciones que pueden afectar el funcionamiento y precisión de las herramientas y tecnologías, tales como:

- **Calidad de datos:** La calidad de datos nulos, respuestas inconsistentes o datos con sesgo afectan la precisión de respuesta de los modelos predictivos.
- **Desbalance de datos:** Si se obtienen más datos de ciertos grupos en particular o de un tipo específico de estudiantes, es probable que el modelo no detecte estudiantes con características que lo orillen a la deserción estudiantil.
- **Tipo de almacenamiento de datos:** Los datos pueden ser almacenados de distintas formas o formatos, esto resulta problemático a la hora de recolección para el análisis de información.
- **Datos no contables:** Hay aspectos académicos que no son fácilmente cuantificados para un análisis completo, como la motivación, responsabilidad, esfuerzo, etc.

CAPITULO 3: MARCO TEÓRICO

3.1 Minería de datos

La minería de datos toma relevancia ante el avance de la sociedad donde los negocios se vuelven más populares y los datos incrementan de forma voraz. Es aquí donde los datos adquieren un papel importante, un nuevo recurso estratégico para la toma de decisiones.

En este sentido, la información de las organizaciones ha buscado dar mayor valor a la información almacenada en sus bases de datos, impulsando la automatización de procesos con el fin de descubrir conocimientos útiles que, de otro modo, permanecerían sin aprovechar o se perderían.

La Minería de datos descubre patrones, tendencias y relaciones con el único propósito de dar las mejores tomas de decisiones con mayor conocimiento.

Es fundamental establecer una base teórica y sólida para sustentar el análisis y la implementación del desarrollo de un sistema de predicción de aprovechamiento académico.

El proceso KDD ((Knowledge Discovery in Databases) se refiere al proceso iterativo de la búsqueda de conocimiento en base de datos, utilizado para extraer conocimiento útil en grandes conjuntos de datos.

Proceso KDD:

El proceso interactivo e iterativo del KDD son los siguientes:

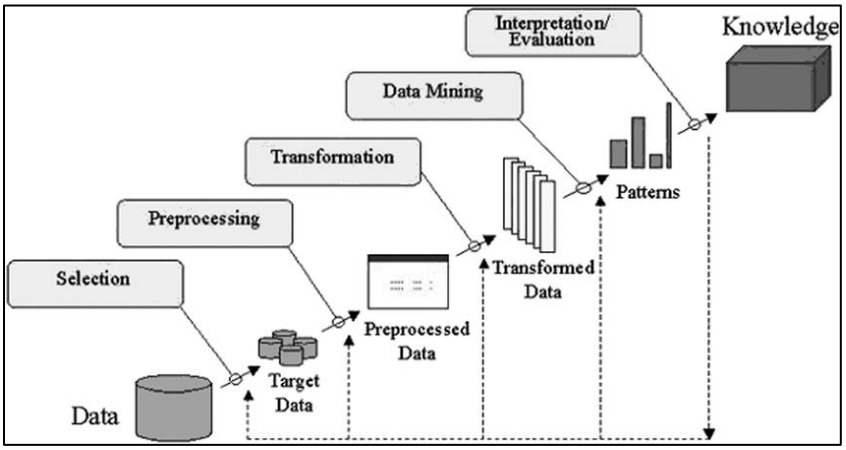
- Fase de selección: Fase donde se selecciona e identifica los datos relevantes. para trabajar a futuro, tanto fuentes internas como externas. El objetivo de esta es crear una fuente de datos que contenga la suficiente información para ser utilizada.

- Fase de procesamiento: El objetivo de esta fase es procesar y mejorar la calidad de datos. Es decir, extraer la información seleccionar, descartar, limpiar y transformar en un conjunto de datos de calidad para trabajar en la siguiente fase.
- Fase de Minería de datos: En la fase de minería datos se aplican técnicas y modelos (arboles de decisión, regresión, etc.) para extraer patrones, relaciones y modelos basados en la fase de procesamiento.
- Fase de evaluación e interpretación: Una vez encontrados patrones de la fase de minería de datos es importante evaluar y validar el conocimiento obtenido.
- Fase de difusión y uso: Finalmente se presentan resultados y hallazgos del conocimiento adquirido.

  Se presenta en la Figura 2. Las etapas del proceso KDD

Figura 1. Etapas del proceso KDD

Tomado de Fayyad, Piatetsky-Shapiro y Smyth (1996).



En el proceso KDD es posible distinguir al menos seis etapas importantes, estas son:

- Recolección de datos: Extracción de información de diversas fuentes, base de datos o archivos.

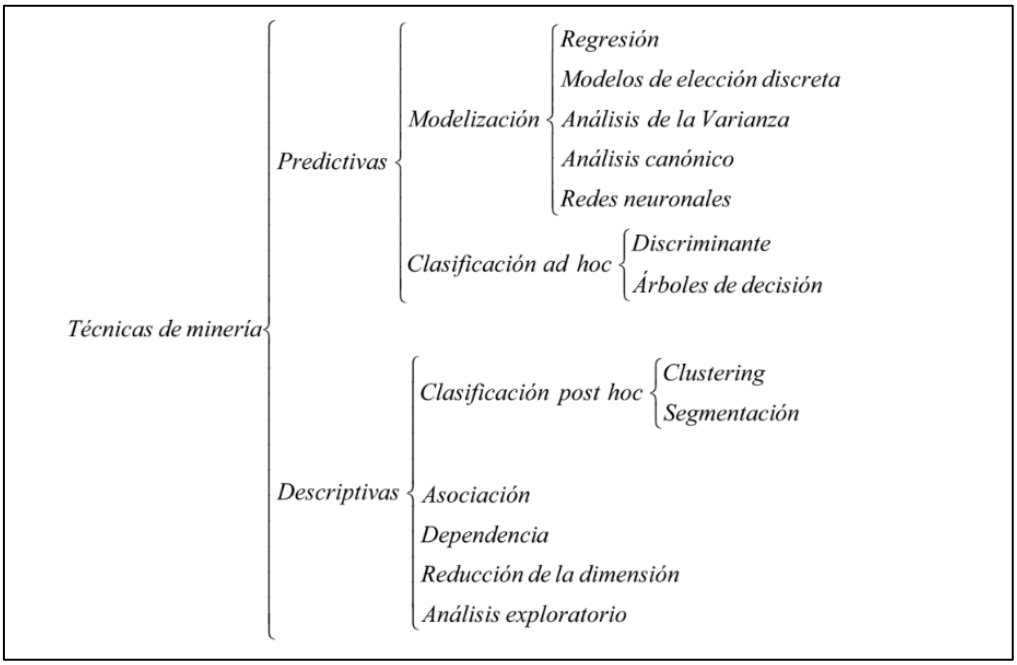
- Selección: Selección de datos más relevantes y útiles.
- Limpieza: Eliminación de datos erróneos, Inconsistentes e irrelevantes
- Transformación de datos: Eliminación de
- Minería de datos: Aplicación de algoritmos y técnicas para descubrir relaciones significativas.
- Evaluación y validación: Análisis de resultados obtenidos, se valida la precisión del modelo además de validar

3.2 Técnicas de minería de datos

Las técnicas de la minería de datos se dividen en predictivas y descriptivas, figura 2. Cada una de ellas permite abordar distintos tipos de problemas dependiendo del objetivo de análisis.

Figura 2. Clasificación de las técnicas de minería de datos

Tomado de *Minería de datos: Técnicas y herramientas*, por C. Pérez López y D. Santín González, s.f., p. 498



Técnicas predictivas: Usadas cuando el objetivo es predecir un valor o categoría desconocida a partir de datos existentes. Basados en técnicas estadísticas y matemáticas, para estimar resultados futuros.

Estas técnicas se subdividen en dos grupos:

I. **Modelización** es el proceso de construir modelos que predigan el comportamiento de una variable dependiente a partir de otras variables independientes.
   - **Regresión:** Usada para predecir valores continuos o numéricos basados en variables independientes.
   - **Modelos de elección discreta:** Usada para predecir una decisión o resultado categórico, es decir "Se utilizan para explicar o predecir una elección entre un conjunto de dos o más alternativas discretas (es decir, distintas y separables; mutuamente excluyentes)" {12].
   - **Análisis de la varianza:** Usada para comparar las varianzas entre las medias de diferentes grupos.
   - **Análisis canónico:** Usada para explorar relaciones entre conjuntos de variables, determina si existe alguna combinación entre las variables.
   - **Redes neuronales:** Son modelos de aprendizaje automático que simula el funcionamiento del cerebro humano. Imitando una serie de neuronas conectada que procesan información.

II. **Clasificación ad hoc** se refiere a soluciones que se centran en asignar observaciones o soluciones de un conjunto predefinido de categorías.
   - Discriminante: Usada para explorar combinaciones lineales de variables, es decir reconocimiento de patrones.
   - Arboles de decisión: Modelo usado en estructura de árbol para representar de manera esquemática alternativas para facilitar la toma de decisiones.

Técnicas descriptivas: Estas técnicas se utilizan para caracterizar datos, encontrando relaciones, patrones o tendencias.

III. **Clasificación post hoc** se refiere a pruebas estadísticas, es decir se descubren directamente de los datos.
   - **Clustering** organiza y clasificar un grupo de objetos para crear subconjuntos de datos.
   - **Segmentación** es asignar o agrupar individuos en segmentos homogéneos

IV. Descriptivas

- Asociación usada en la identificación de patrones y relación entre las variables, además de buscar predecir comportamientos y tomar decisiones futuras.
- Dependencia es usada para entender y modelar la relación de dependencia entre distintas variables.
- Reducción de la dimensión usada para disminuir el n número de variables o dimensiones en un conjunto de datos para mejor rendimiento.
- Análisis exploratorio utilizado para analizar e investigar anomalías y formular hipótesis.

3.3 Algoritmos de clasificación

Los algoritmos de clasificación y predicción son punto clave en la minería de datos porque permiten predecir, asignar y etiquetar datos, basados en los patrones aprendidos. Existen diferentes algoritmos para extraer patrones e información útil de grandes volúmenes de datos.

Entre los algoritmos más comunes se encuentra los árboles de decisión (algoritmo de aprendizaje supervisado), estructura jerárquica de un árbol donde divide el conjunto de datos en función de las características y cada rama es la respuesta a una pregunta, mientras que el nodo representa una decisión. Este algoritmo emplea estrategias como el divide y vencerás para identificar puntos de división viables en el árbol.

La regresión logística, es utilizada para clasificar y analizar de forma predictiva, utiliza las matemáticas para la búsqueda de instancias que pertenezcan a una clase (si/no, verdadero/falso).

Otro algoritmo de predicción y clasificación es el de redes neuronales, el cual se refiere a redes que están inspiradas en el cerebro humado porque estas están compuestas de nodos, es decir neuronas organizadas en capas. Estas aprenden, reconocen patrones y relaciones entre los datos para finalmente tomar

decisiones. Mientras que en el algoritmo de reglas de asociación se busca identificar relaciones, patrones o asociaciones entre las variables de un volumen de datos.

Todos estos algoritmos tienen por objetivo buscar, predecir, aprender y relacionar datos, es decir, identifican a estudiantes con similitudes o patrones que estén afectando su desarrollo académico, desde asistencias, calificaciones, o comportamientos que indiquen algún índice de deserción. Contribuyendo así, a que las instituciones puedan brindar el apoyo y orientación adecuados para mejorar el bienestar de cada estudiante.

3.4 Flask

Flask es un microframework de enfoque minimalista y ligero escrito en Python, un microframework es un conjunto de herramientas y librerías que proporciona las funcionalidades mínimas e imprescindibles para construir una aplicación web. Con flask podemos crear aplicaciones web funcionales, incluyendo componentes como el enrutamiento (routing) y el manejo de solicitudes HTTP.

Ventajas:

- ✓ Proporciona estructura de proyecto.
- ✓ Fácil de adaptar bibliotecas.
- ✓ Incluye un servidor web de desarrollo.
- ✓ Cuenta con depurador y soporte integrado para pruebas unitarias.
- ✓ Ágil y rápida.

Usos comunes:

- ✓ Prototipos rápidos que permite desarrollar aplicaciones web pequeñas o de forma rápida y con mínimo código.
- ✓ Microservicios para construir la lógica de backend que devuelve datos en formato JSON para aplicaciones frontend.

CAPITULO 4: DISEÑO E IMPLEMENTACIÓN

En este capítulo se presenta la implementación del **Sistema de Clasificación y Predicción del Aprovechamiento Académico**, así como los resultados obtenidos durante su desarrollo y validación. Se detallan todas las fases de desarrollo, con el objetivo de demostrar su funcionalidad y efectividad.

4.1 Fase 1: Recolección y preparación de datos

Fase 1: Recolección y preparación de datos.

***Objetivo*:** Preparar y organizar los para su análisis y modelado predictivo.

Actividades:

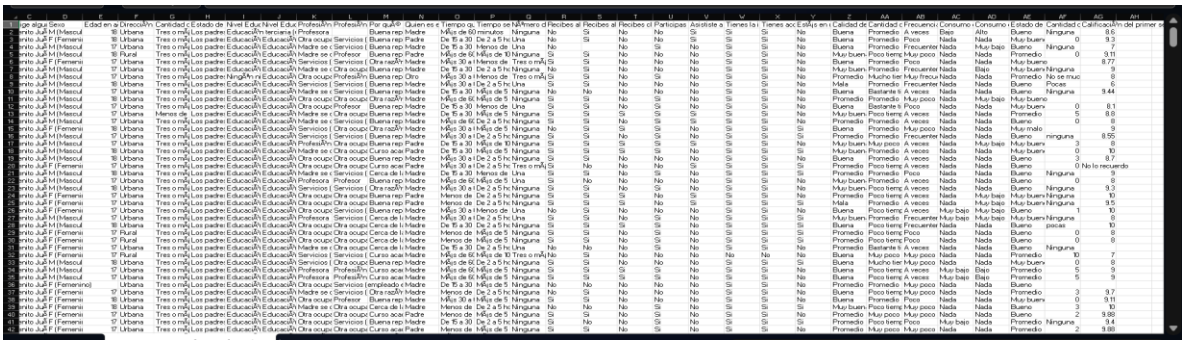
- Recolección de la información en archivos CSV.
- Carga de dataset archivo CSV.
- Limpieza de datos: Eliminación de columnas irrelevantes, eliminación de datos nulos o inconsistentes.
- Codificación de variables categóricas.
- Normalización de datos, ponderación de datos, y selección de variables relevantes.

Resultado: Dataset limpio y estructurado (faseUno.csv), listo para la fase dos de minería de datos.

Nota: En la figura 3 se presenta la información obtenida a partir del formulario titulado "*Análisis Comparativo de Modelos de Clasificación para la Predicción del Aprovechamiento Académico de Estudiantes de dos Preparatoria*"

Figura 3. Información antes de limpieza y normalización.

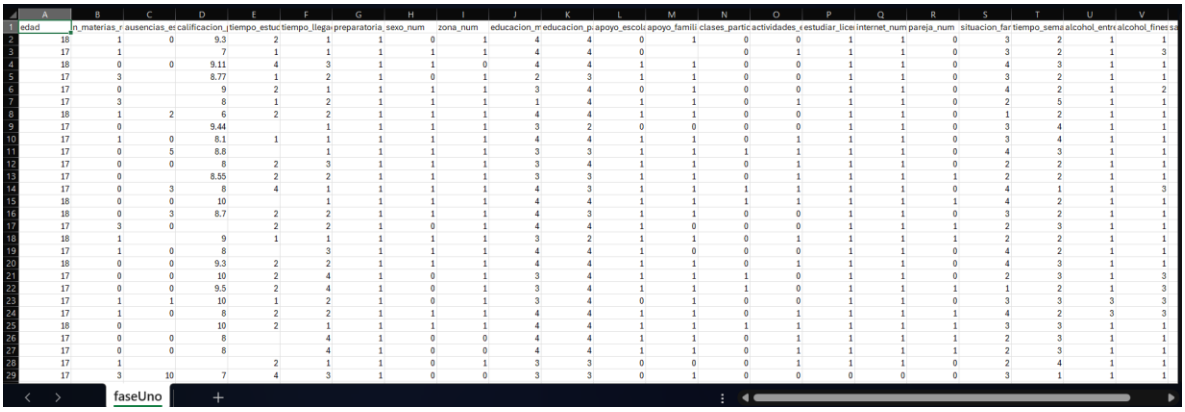
Fuente: Elaboración propia



La información obtenida del formulario se comparó con el nuevo dataset limpio generado en la Fase 1, con el objetivo de verificar y denotar la limpieza, normalización y organización de la información, asegurando un uso adecuado y consistente de los datos para posteriormente trabajar en la demás. Figura 4.

Figura 4. Información después de limpieza y normalización**.**

Fuente: Elaboración propia



El código 1 presentado en el permite cargar y limpiar el dataset de estudiantes. Dentro del proceso, se eliminan valores faltantes, se normalizan las columnas numéricas y se codifican las variables categóricas, asegurando que la información esté organizada y lista para las siguientes fases.

Código 1. Carga y limpieza de datos

```
import pandas as pd


def procesar_csv(ruta_csv):
    # Cargar el archivo CSV
    df = pd.read_csv(ruta_csv)
    print("Columnas originales:")
    print(df.columns.tolist())

    # Eliminar columnas innecesarias
    df.drop(columns=[
        'Marca temporal', 'Acepta participar', 'Asististe a la guardería',
        'Estado de convivencia de los padres', 'Quien es el tutor',
        'Por qué elegiste esta preparatoria', 'Profesión de la madre',
        'Profesión del padre', 'Cantidad de miembros en tú familia', 'Frecuencia
con la que  sales con amigos'
    ], inplace=True)

    # Renombrar columnas
    renombrar_columnas = {
        'Sexo' : 'sexo',
        'Elige alguna preparatoria': 'preparatoria',
        'Edad en años': 'edad',
        'Dirección ': 'zona',
        'Nivel Educativo de la madre': 'educacion_madre',
        'Nivel Educativo del padre': 'educacion_padre',
        'Tiempo que te lleva llegar a la escuela': 'tiempo_llegada',
        'Tiempo semanal dedicado al estudio': 'tiempo_estudio',
        'Número de materia reprobadas': 'n_materias_reprobadas',
        'Recibes algún apoyo escolar': 'apoyo_escolar',
        'Recibes algún apoyo familiar': 'apoyo_familiar',
        'Recibes clases particulares': 'clases_particulares',
```

```
        'Participas en actividades extraescolares':
'actividades_extraescolares',

        'Tienes la intención de estudiar la licenciatura':
'estudiar_licenciatura',

        'Tienes acceso a Internet en casa': 'internet',

        'Estás en una relación romántica': 'pareja',

        'Calidad de las relaciones familiares': 'situacion_familiar',

        'Cantidad de tiempo libre semanal': 'tiempo_semanal',

        'Consumo de alcohol durante la semana': 'alcohol_entre_semana',

        'Consumo de alcohol durante el fin de semana': 'alcohol_fines_semana',

        'Estado de salud ': 'salud',

        'Cantidad de ausencias escolares': 'ausencias_escolares',

        'Calificación del primer semestre': 'calificacion_primer_semestre'

    }

    df.rename(columns=renombrar_columnas, inplace=True)


    # Normalizar respuestas

    df['sexo'] = df['sexo'].astype(str).str.strip().str.upper().replace({'M
(MASCULINO)': 'M', 'F (FEMENINO)': 'F'})


    # Educación padres

    df['educacion_madre'] =
df['educacion_madre'].astype(str).str.strip().str.lower()

    df['educacion_padre'] =
df['educacion_padre'].astype(str).str.strip().str.lower()

    edu_map = {

        'educación terciaria (universitaria)': 'universidad',

        'educación primaria': 'primaria',

        'educación secundaria': 'secundaria',

        'ningún nivel educativo': 'ninguno'

    }

    df['educacion_madre'] = df['educacion_madre'].map(edu_map)

    df['educacion_padre'] = df['educacion_padre'].map(edu_map)


    # Materias reprobadas y prepa
```

```
    df['n_materias_reprobadas'] =
df['n_materias_reprobadas'].replace({'Ninguna': 0, 'Una': 1, 'Tres o más': 3})

    df['preparatoria'] = df['preparatoria'].replace({'Benito Juárez García':
'Benito_Juarez', 'Complejo Regional Tehuacán': 'Complejo_Tehuacan'})



    # Ausencias escolares

    df['ausencias_escolares'] = df['ausencias_escolares'].replace({

        'Ninguna': 0,'cero': 0,'Pocas': 2,'No faltó': 0,'Nada': 0,'Nunca ':
0,'Ninguna, Cero ': 0

    })



    # Mapas ordinales

    mapa_tiempo_estudio = {'Menos de 2 horas': 1,'De 2 a 5 horas': 2,'De 5 a 10
horas': 3,'Más de 10 horas': 4}

    mapa_tiempo_llegada = {'De 15 a 30 minutos': 1,'Más 30 a 60 minutos': 2,'Más
de 60 minutos': 3,'Menos de 15 minutos':4}

    df['edad'] = pd.to_numeric(df['edad'].astype(str).str.extract(r'(\d+)')[0],
errors='coerce')

    df['tiempo_estudio_num'] = df['tiempo_estudio'].map(mapa_tiempo_estudio)

    df['tiempo_llegada_num'] = df['tiempo_llegada'].map(mapa_tiempo_llegada)



    # Mapas numéricos

    df['preparatoria_num'] = df['preparatoria'].map({'Benito_Juarez': 1,
'Complejo_Tehuacan': 0})

    df['sexo_num'] = df['sexo'].map({'M': 1, 'F': 0})

    df['zona_num'] = df['zona'].map({'Urbana': 1, 'Rural': 0})

    edu_num = {'ninguno': 1, 'primaria': 2, 'secundaria': 3, 'universidad': 4}

    df['educacion_madre_num'] = df['educacion_madre'].map(edu_num)

    df['educacion_padre_num'] = df['educacion_padre'].map(edu_num)



    # Variables binarias

    binarios =
['apoyo_escolar','apoyo_familiar','clases_particulares','actividades_extraescola
res','estudiar_licenciatura','internet','pareja']

    for col in binarios:

        df[col + '_num'] =
df[col].astype(str).str.strip().str.lower().map({'si': 1, 'no': 0})
```

```
    # Otras ordinales

    df['situacion_familiar_num'] = df['situacion_familiar'].map({'Mala': 1,
'Promedio': 2, 'Buena': 3, 'Muy buena': 4})

    df['tiempo_semanal_num'] = df['tiempo_semanal'].map({'Muy poco tiempo
libre': 1,'Promedio': 2,'Poco tiempo libre': 3,'Bastante tiempo libre': 4,'Mucho
tiempo libre': 5})

    df['alcohol_entre_semana_num'] = df['alcohol_entre_semana'].map({'Nada':
1,'Bajo': 2,'Muy bajo': 3,'Promedio': 4})

    df['alcohol_fines_semana_num'] = df['alcohol_fines_semana'].map({'Nada':
1,'Bajo': 2,'Muy bajo': 3,'Promedio': 4})

    df['salud_num'] = df['salud'].map({'Malo': 1,'Promedio': 2,'Bueno': 3,'Muy
bueno': 4})



    # Columnas numéricas

    num_cols =
['calificacion_primer_semestre','ausencias_escolares','n_materias_reprobadas','t
iempo_estudio_num','tiempo_llegada_num']

    for col in num_cols:

        df[col] = pd.to_numeric(df[col], errors='coerce')



     # Columnas numéricas

    num_cols =
['calificacion_primer_semestre','ausencias_escolares','n_materias_reprobadas','t
iempo_estudio_num','tiempo_llegada_num']

    for col in num_cols:

        df[col] = pd.to_numeric(df[col], errors='coerce')



    # Categorización calificación primer semestre en Bajo, Medio, Alto

    bins = [0, 7, 8, 10]  # límites para categorías

    labels = ['Bajo', 'Medio', 'Alto']

    df['calificacion_primer_semestre_cat'] =
pd.cut(df['calificacion_primer_semestre'], bins=bins, labels=labels,
include_lowest=True)



    # Mapeo de categorías a números para evitar errores al usar modelos

    mapa_calif = {'Bajo': 0, 'Medio': 1, 'Alto': 2}
```

```
    df['calificacion_primer_semestre_cat_num'] =
df['calificacion_primer_semestre_cat'].map(mapa_calif)


        # Eliminar columnas de texto originales
    texto_cols = [
        'sexo','zona','educacion_madre','educacion_padre','apoyo_escolar','apoyo
_familiar',
        'clases_particulares','actividades_extraescolares','estudiar_licenciatur
a','internet',
        'pareja','situacion_familiar','tiempo_semanal','alcohol_entre_semana','a
lcohol_fines_semana','salud',
        'tiempo_estudio','tiempo_llegada','preparatoria'
    ]
    df.drop(columns=texto_cols, inplace=True)


    # Reemplazar espacios vacíos por
    df.replace(r'^\s*$', pd.NA, regex=True, inplace=True)


    # Mostrar información de registros antes del filtro
    print("Registros totales:", len(df))
    print("Valores nulos por columna:")
    print(df.isnull().sum())


    # Filtrar filas con demasiados NaN (opcional)
    df = df[df.isnull().sum(axis=1) <= 5]  # más flexible


    # Guardar CSV limpio
    df.to_csv("faseUno.csv", index=False, encoding="utf-8-sig")
    print("Archivo faseUno.csv generado con éxito. Registros finales:", len(df))


    return df
```

4.2 Fase 2: Análisis exploratorio con minería de datos

Fase 2. Análisis exploratorio de datos con Minería de datos

Objetivo: Identificar patrones y relaciones entre las variables académicas para seleccionar las más relevantes.

Actividades:

- Carga de dataset limpio de la fase 1.
- Análisis descriptivo: cálculo de promedios, desviaciones estándar y percentiles.
- Visualización de datos: gráficos de dispersión, histogramas y boxplots.
- Análisis de patrones y correlaciones: determinación de qué variables influyen más en el desempeño académico.
- Segmentación de estudiantes según su nivel de aprovechamiento.

***Resultado*:** Conjunto de variables relevantes y visualizaciones que sirven como base para entrenar y evaluar los modelos de predicción.

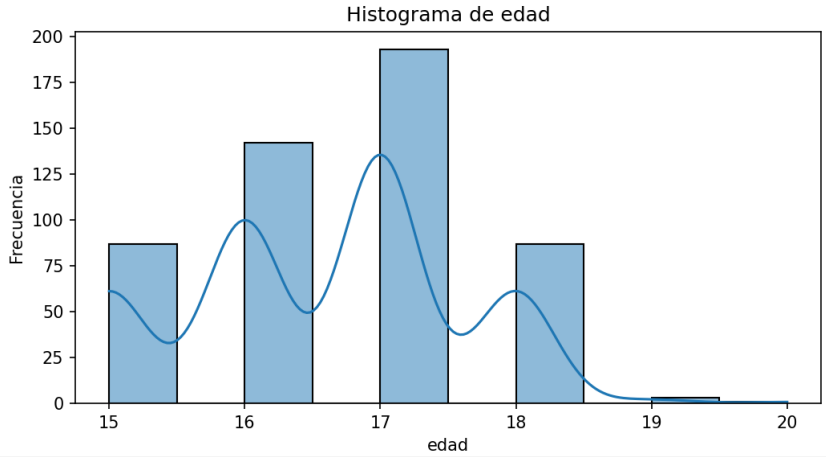
Análisis estadísticos de los datos:

Edad de los estudiantes

- La edad promedio es de 16.57 años, con un rango de 15 a 20 años.
- La mediana es de 17 años, lo que indica que la mayoría de los estudiantes está en primeros semestres.
- La baja desviación estándar (0.99) sugiere que no hay mucha variabilidad en la edad.

Figura 5. Histograma de edad.

Fuente: Elaboración propia



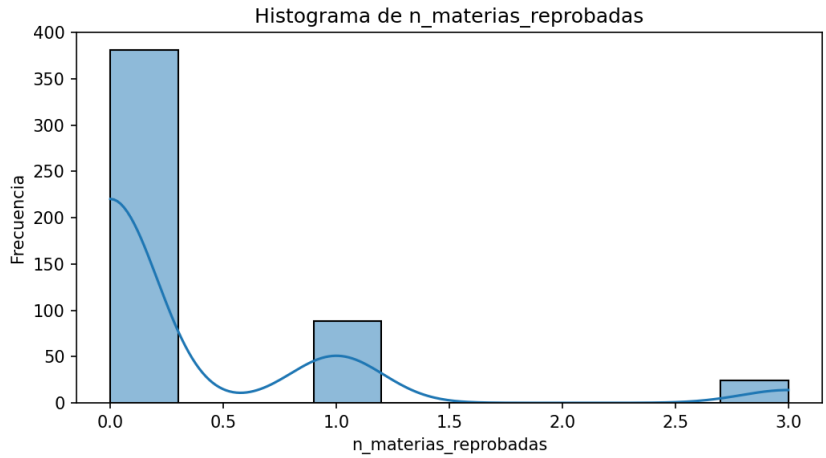
Número de materias reprobadas

- La media es 0.32 y la mediana es 0, con un máximo de 3. Lo que nos indica que la mayoría de los estudiantes no ha reprobado materias.

Nota: Es importante tomar en cuenta este análisis, ya que el número de materias reprobadas puede afectar directamente el desempeño académico futuro. Figura 6.

Figura 6. Histograma de materias reprobadas.

Fuente: Elaboración propia

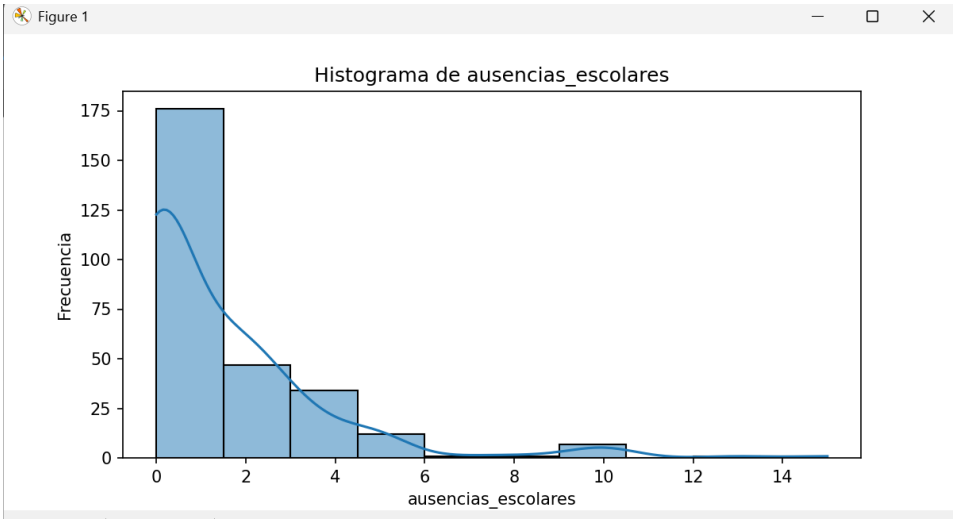


Ausencias escolares

- Se muestra que la media es de 1.5 con un máximo de 15 ausencias, pero la mediana es 1, lo que indica que la mayoría de los estudiantes tiene pocas ausencias. Mientras que la desviación estándar relativamente alta (2.29) sugiere que hay un grupo reducido de estudiantes con ausencias relevantes.

Figura 7. Histograma de ausencias escolares**.**

Fuente: Elaboración propia

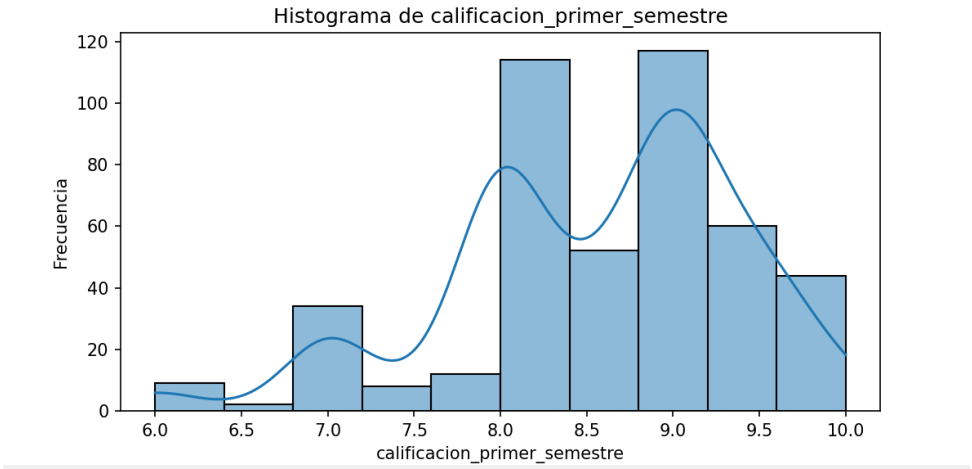


Calificaciones del primer semestre

- Se presenta que el promedio mínimo es 8.54, la mediana de 8.76 y máximo de 10. Esto demuestra que la mayoría de los estudiantes tienen un buen desempeño académico.
- Aunque la desviación estándar baja (0.85) refleja que la mayoría de los alumnos se concentra en un rango estrecho de calificaciones.

Figura 8. Histograma de calificaciones del primer semestre.

Fuente: Elaboración propia.



Se muestra la tabla 5 para entender la clasificación de las categorías del tiempo de estudio.

Tabla 5.

Categorías del tiempo de estudio

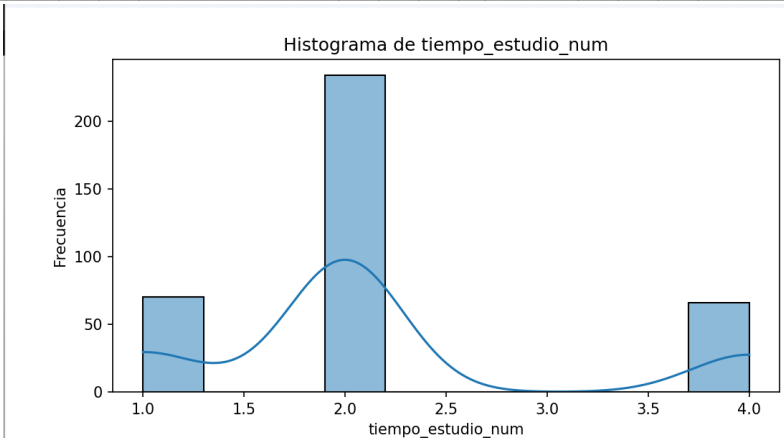
Valor	Significado
1	Menos de 2 horas
2	De 2 a 5 horas
3	De 5 a 10 horas
4	Más de 10 horas

Fuente: Elaboración propia.

- La figura 8 muestra que la variable tiempo_estudio_num la mayoría de los estudiantes dedican poco tiempo al estudio. Es decir que, la media (2.17) y la mediana (2) son casi iguales, lo que indica que la mitad de los estudiantes solo estudian de 2 a 5 horas. En cambio, muestra que muy pocos estudian más de 10 horas.

Figura 9. Histograma de tiempo de estudio.

Fuente: Elaboración propia.



En una escala representada en la tabla 6.

Tabla 6.

Categorías de tiempo de llegada

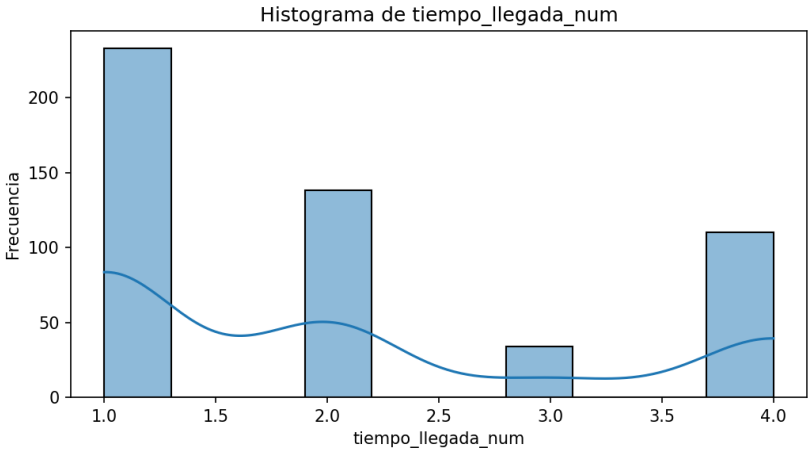
Valor	Significado
1	De 15 a 30 minutos
2	Más 30 a 60 minutos
3	Más de 60 minutos
4	Menos de 15 minutos

Fuente: Elaboración propia.

- En la figura 9, se observa que la mayoría de los estudiantes tarda entre 30 y 60 minutos en llegar a la escuela, mientras que un mínimo de estudiantes llega más rápido (En menos de 15 min) o más tarde (Más de 60 minutos) como se muestra en la tabla 6. Esto sugiere que el tiempo de llegada podría ser un factor relevante para considerar en el análisis del aprovechamiento académico, ya que los estudiantes que llegan tarde podrían afectar su rendimiento o incluso las asistencias.

Figura 10. Histograma de tiempo de llegada

Fuente: Elaboración propia.



Se analizaron información de 2 preparatorias donde

Tabla 7.

Categorías de preparatoria

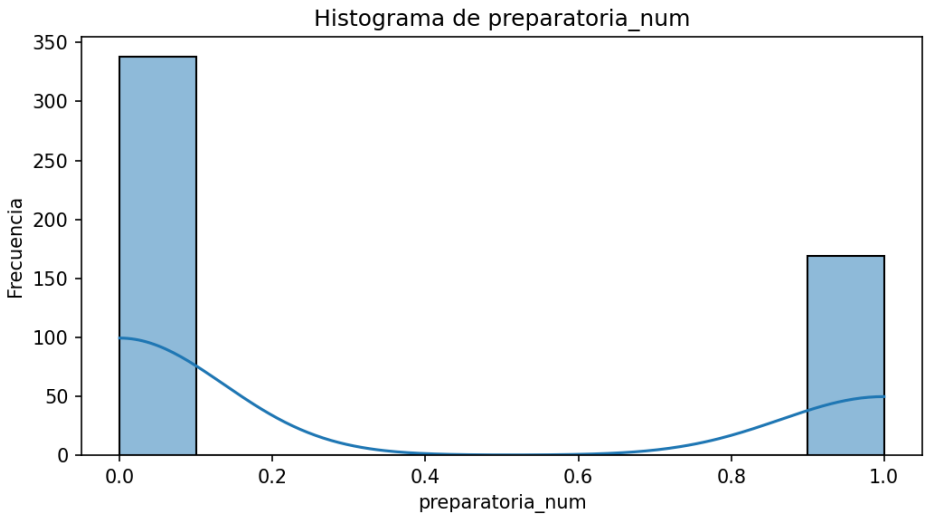
Valor	Significado
0	Complejo de Tehuacán
1	Preparatoria Benito Juárez

Fuente: Elaboración propia.

Se muestra que la mayoría de los estudiantes pertenece a la preparatoria 0 (Complejo de Tehuacán), mientras que un tercio pertenece a la preparatoria 1 (Preparatoria Benito Juárez).

Figura 11. Histograma de preparatoria

Fuente: Elaboración propia.



En la tabla 8 se muestra la categoría de cada zona:

Tabla 8.

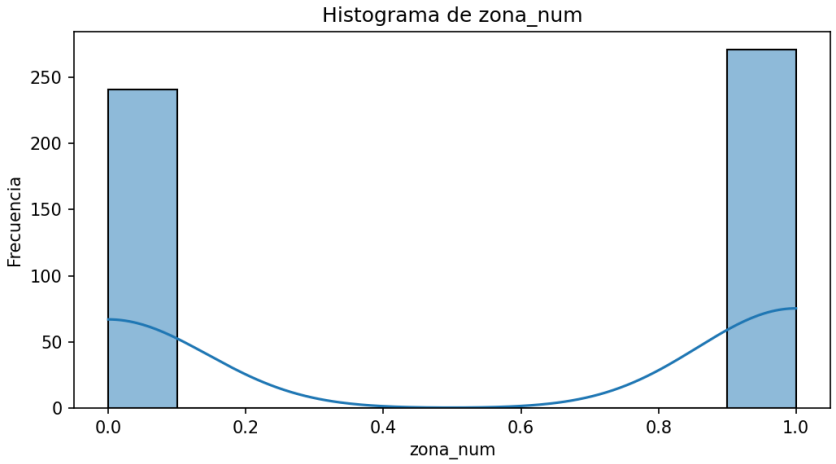
Categorías de zona.

Valor	Significado
0	Rural
1	Urbana

La mayoría de los estudiantes se encuentran en la categoría 1 (Urbana) con un 53%, con un 47% de estudiantes que pertenecen a la categoría 2 (Rural). Esta variable puede ser útil para analizar si la zona geográfica tiene alguna relación con el rendimiento académico, asistencia o hábitos de estudio.

Figura 12. Histograma de zona de estudiantes

Fuente: Elaboración propia.



En la tabla 9 se muestra la categoría de educación

Tabla 9.

Categorías de educación

Valor	Significado
1	Ninguno
2	Primaria
3	Secundaria
4	Universidad

- En la figura 13 y 14 se observa el nivel de estudios de ambos padres donde la mayoría tanto de los padres y madres de los estudiantes tienen educación secundaria, mientras que un grupo más mayoritario cuenta con nivel de estudios de Universidad. Es de recalcar que esta variable es relevante para el análisis de aprovechamiento académico, ya que a veces muestran que el nivel educativo de los padres suele correlacionar con el desempeño de los hijos.

Figura 13. Histograma de educación de la madre.

Fuente: Elaboración propia.

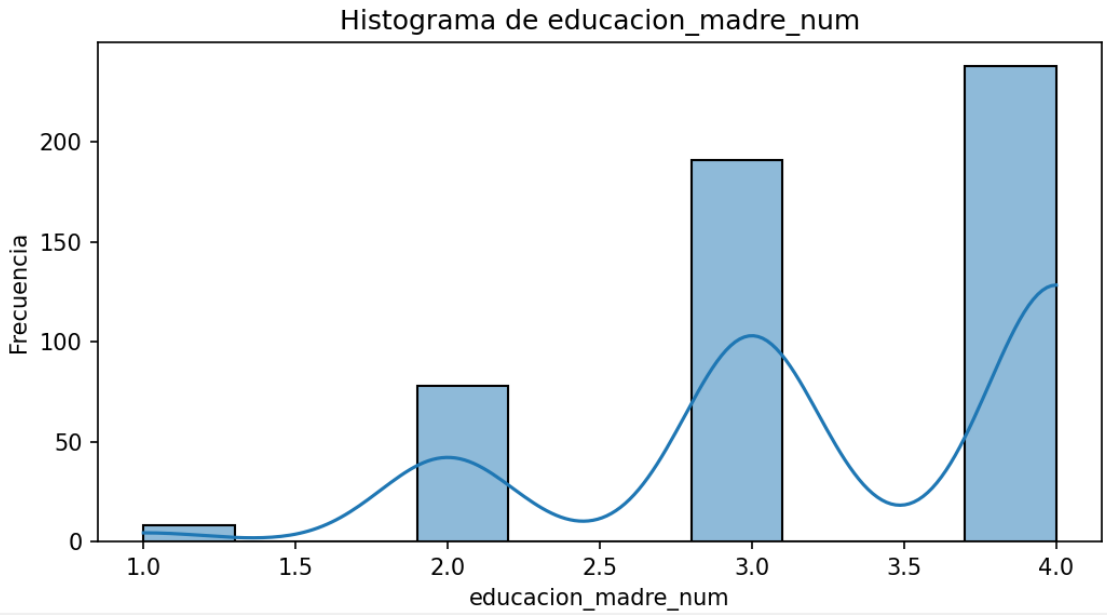
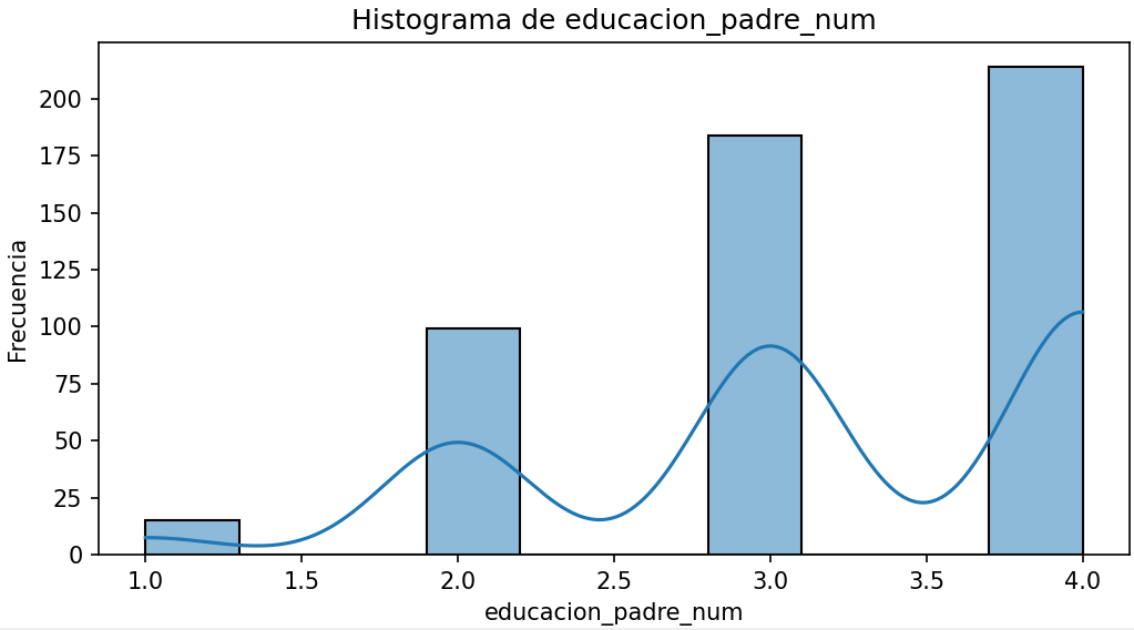


Figura 14. Histograma de educación del padre.

Fuente: Elaboración propia.



En la tabla 10 se muestra la categoría de apoyo escolar

Tabla 10.

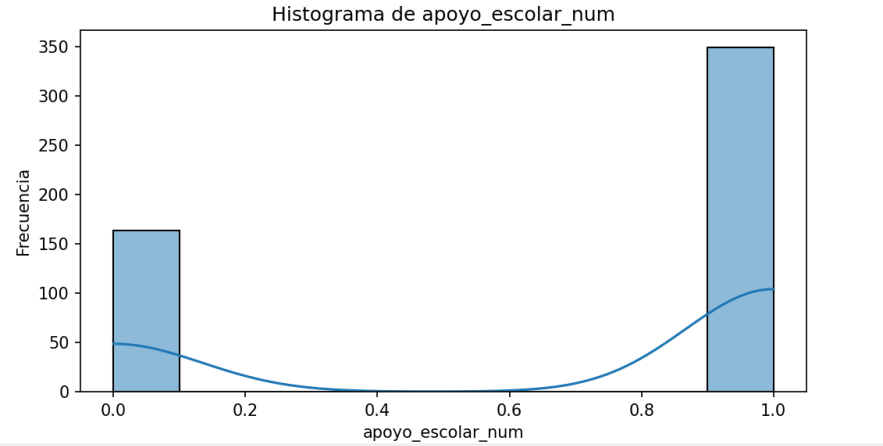
Categorías de apoyo

Valor	Significado
0	No
1	Sí

- En la figura 15, se presenta que la mayoría de los estudiantes son apoyados en sus estudios, lo que podría ser positivo para su rendimiento académico. Aunque existe un grupo menor que no cuenta con apoyo, lo que puede representar un riesgo de menor desempeño o necesidad de intervención.

Figura 15. Histograma de apoyo escolar

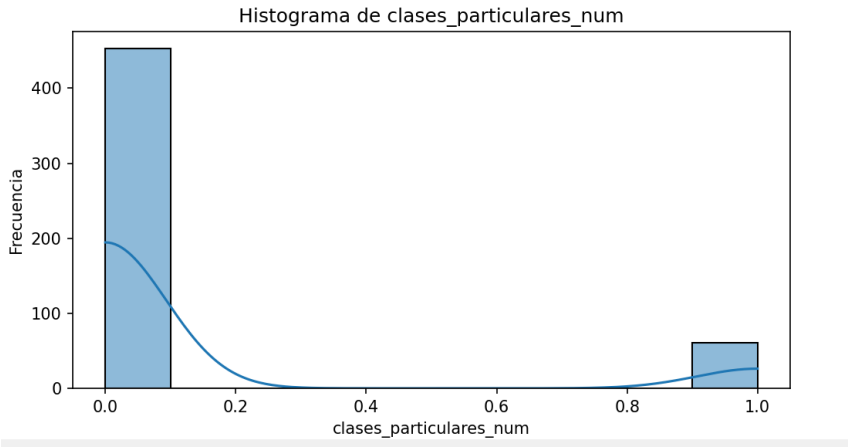
Fuente: Elaboración propia.



Se observa en la figura 16, donde 0 significa "no" y 1 significa "sí" que la mayoría de los estudiantes no cuentan con clases particulares, estos podrían indicar que sea por factores económicos o de tiempo.

Figura 16. Histograma de estudiantes que toman clases particulares

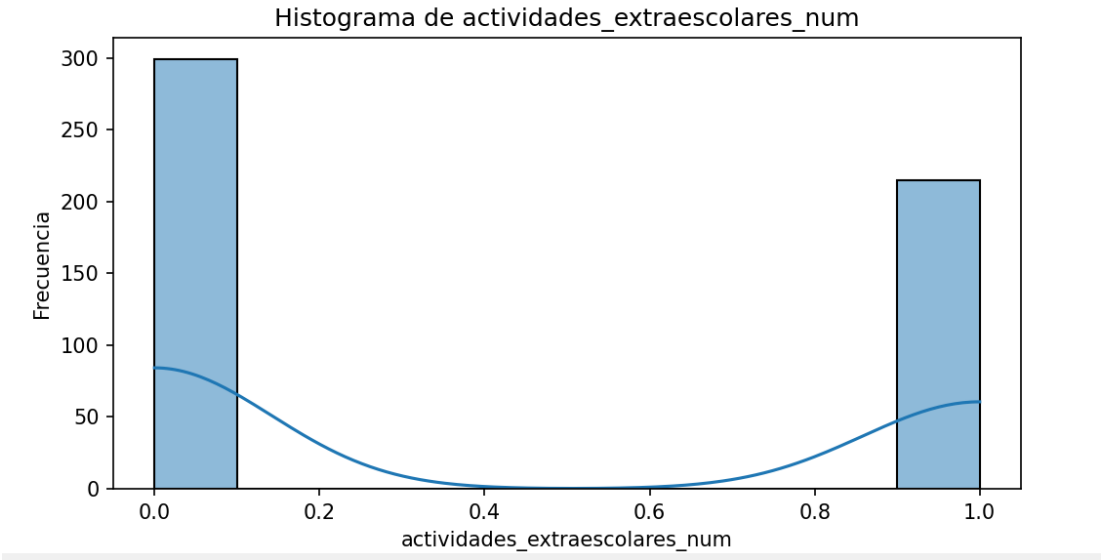
Fuente: Elaboración propia.



Se observa en la figura 17, que menos de la mitad de los estudiantes realizan actividades extraescolares, es decir un 75% si participa mientras que un mínimo de 25% no participa en actividades.

Figura 17. Histograma de actividades extraescolares

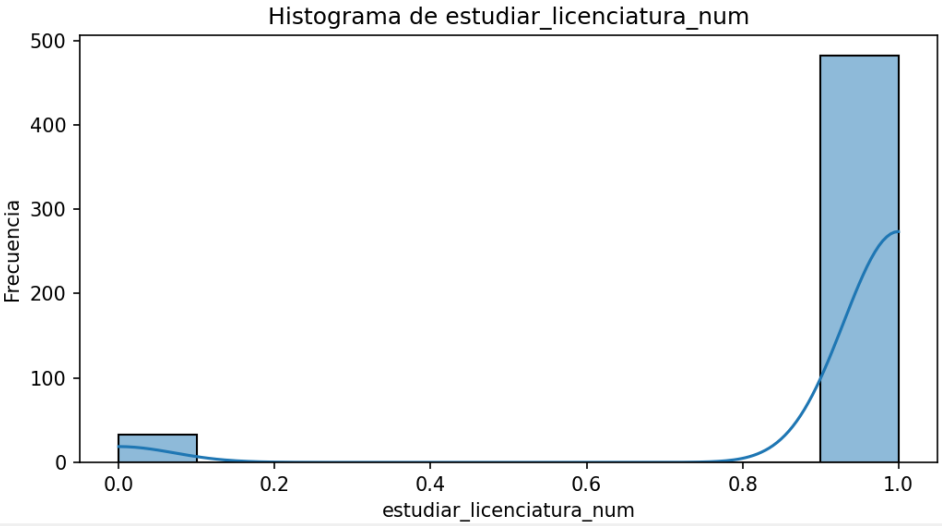
Fuente: Elaboración propia.



El saber las estadísticas acerca si los estudiantes tienen planeado continuar con sus estudios es de gran ayuda. En la figura 18 muestra que la gran mayoría planea continuar con sus estudios. Mientras que un mínimo grupo no tiene intención de cursar una licenciatura.

Figura 18. Histograma de si se desea estudiar una licenciatura

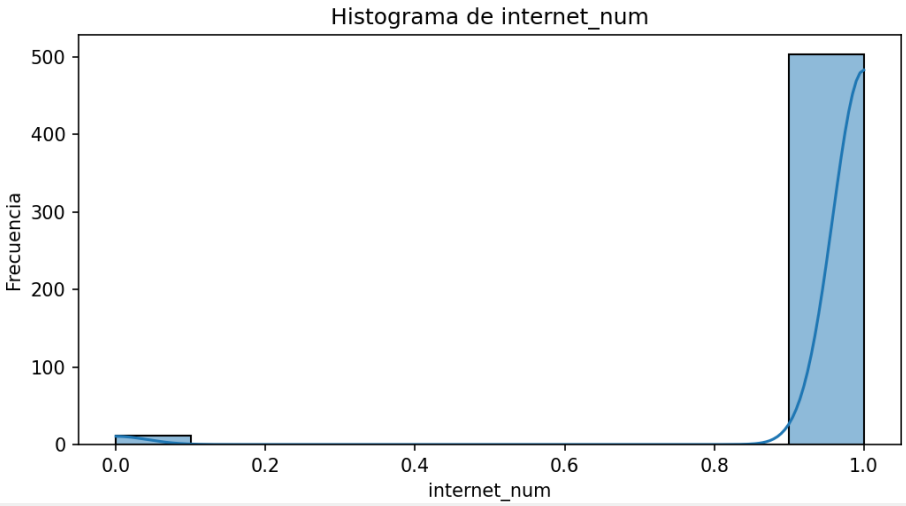
Fuente: Elaboración propia.



La gran mayoría de estudiantes cuenta con acceso a internet, esta condición es favorable para un aprendizaje autónomo que favorece positivamente el aprovechamiento académico. Sin embargo, existe un mínimo grupo sin acceso a internet como se muestra en la figura 19.

Figura 19. Histograma de estudiantes con internet

Fuente: Elaboración propia.



Tener pareja mientras se estudia puede influir de diferentes formas. Tanto positivamente como negativamente, en la figura 19 donde no es 0 y si es 1 la mayoría de estudiante no mantienen una relación sentimental, mientras que un mínimo de estudiantes si se encuentran en una relación.

Figura 20. Histograma de estudiantes con pareja

Fuente: Elaboración propia.

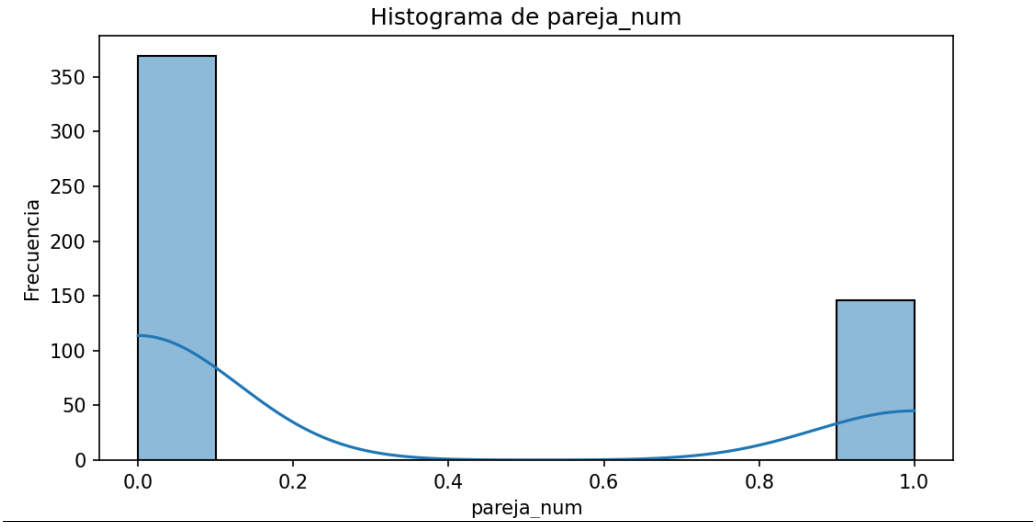


Tabla 11.

Categorías de situación familiar.

Valor	Significado
1	Muy buena
2	Buena
3	Promedio
4	Mala

La situación familiar es un factor clave en el rendimiento académico, ya que la familia es un pilar de motivación y apoyo. Sin embargo, en la figura 19, el 50%

de los estudiantes tienen una relación "promedio", mientras que un 25% una relación "mala" y el restante tiene una valoración de" muy buena".

Figura 21. Histograma situación familiar

Fuente: Elaboración propia.

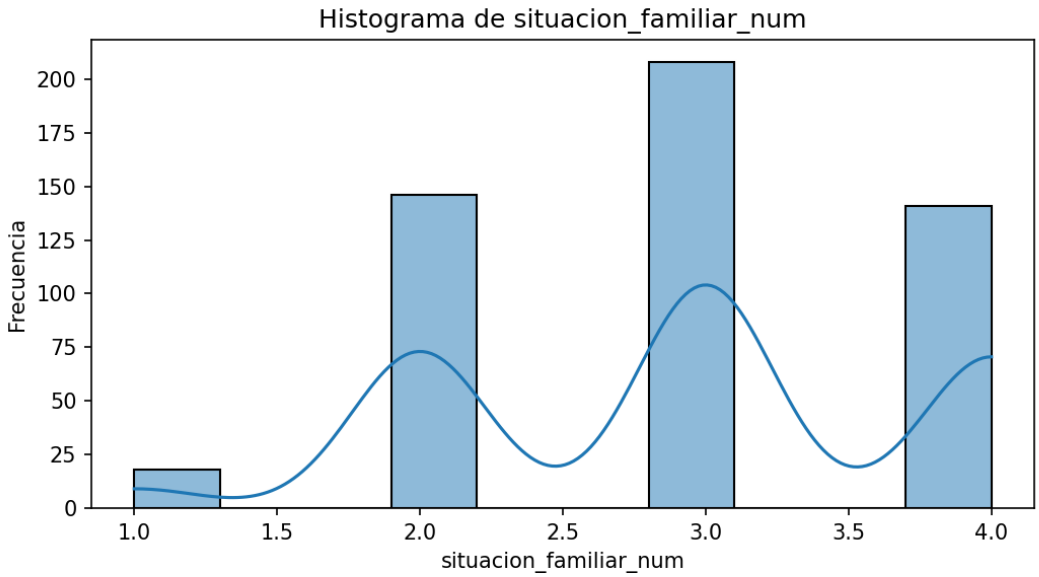


Tabla 12.

Categorías de tiempo semanal.

Valor	Significado
1	Mucho tiempo libre
2	Bastante tiempo libre
3	Poco tiempo libre
4	Promedio
5	Muy poco tiempo libre

De acuerdo con las categorías de la tabla 12, observamos en la figura 19 que la mayoría de los estudiantes tienen bastante tiempo libre durante la semana. Muy pocos alumnos cuentas con muy poco tiempo libre, casos muy

minoritarios. En general, los estudiantes tienden a mantener una amplia disponibilidad de tiempo libre, lo cual podría favorecer la organización de sus estudios y actividades personales.

Figura 22. Histograma tiempo semanal

Fuente: Elaboración propia.

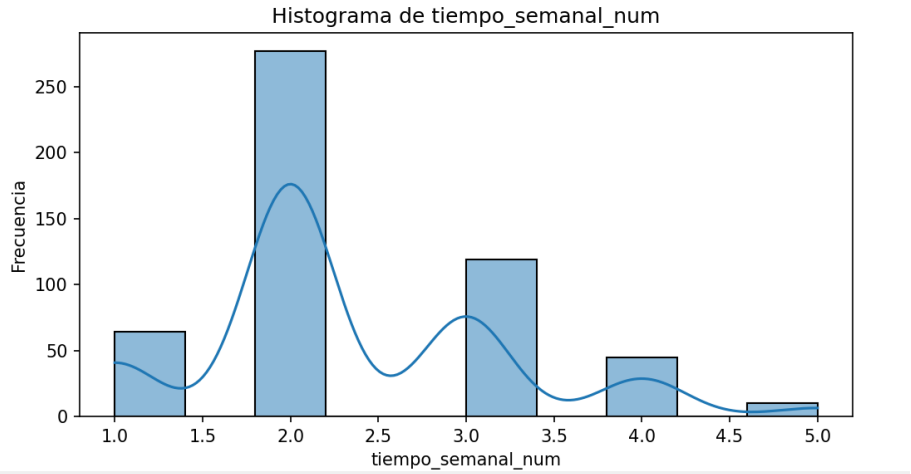


Tabla 13.

Categorías de alcohol consumido entre semana.

Valor	Significado
1	Promedio
2	Muy bajo
3	Bajo
4	Nada

Basado en los valores de la tabla 13, se observó que la mayoría de los estudiantes consumen alcohol durante la semana de forma promedio o moderadamente. La desviación estándar de la figura 20 indica que existe una ligera variabilidad, es decir, aunque la mayoría mantiene un consumo bajo, hay cierta dispersión entre quienes consumen más y quienes no consumen nada.

Si comparamos con el alcohol consumida fines de semana figura 21, el valor promedio coincide con un consumo promedio no se observan valores elevados que indiquen una práctica de consumo alto.

Figura 23. Histograma de alcohol consumido entre semana

Fuente: Elaboración propia.

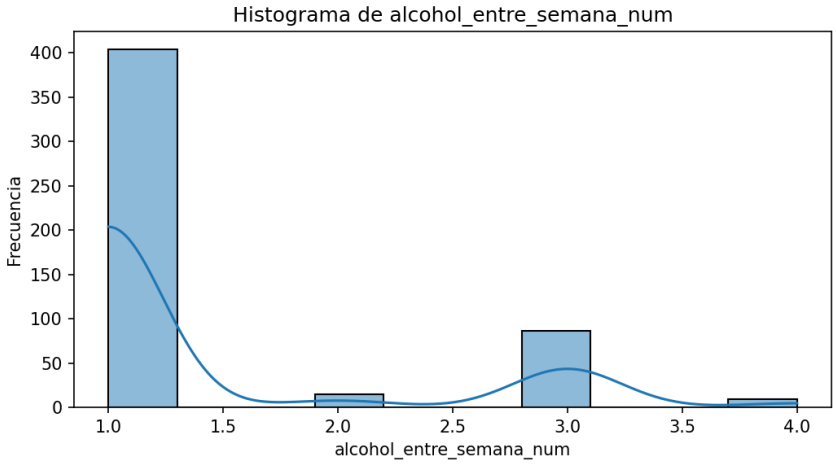


Figura 24. Histograma de alcohol consumido fines de semana

Fuente: Elaboración propia.

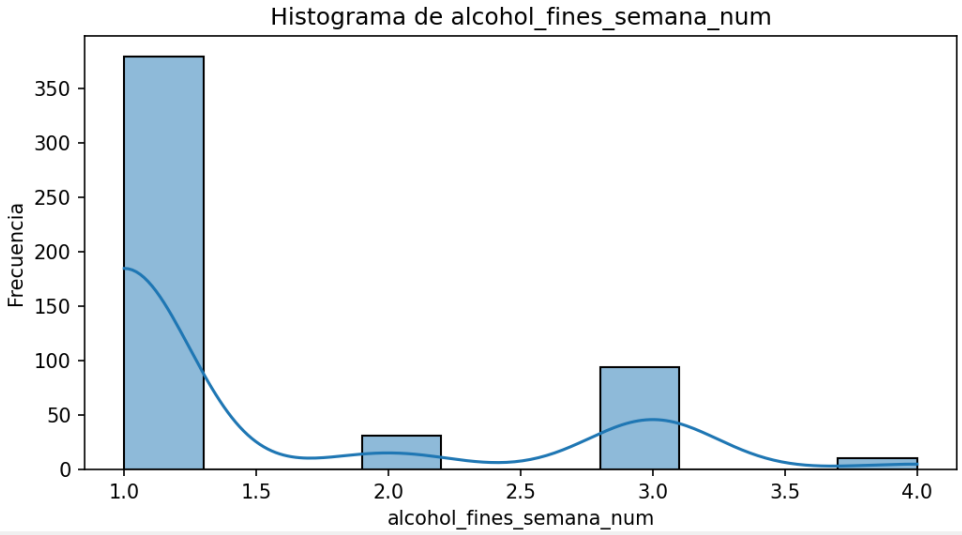


Tabla 14.

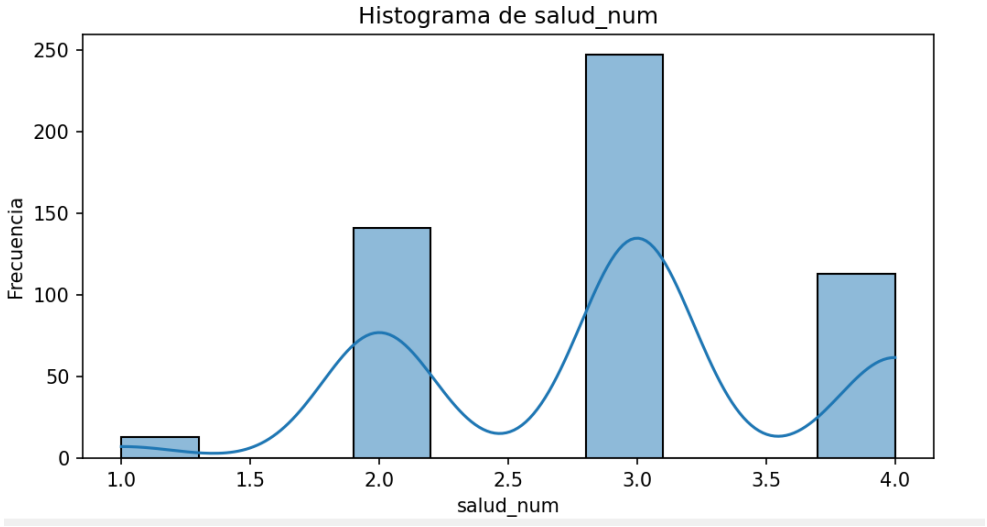
Categorías de estado de salud

Valor	Significado
1	Muy buena
2	Buena
3	Promedio
4	Mala

En promedio se muestra en la figura 22 que la salud de los estudiantes es buena, aunque la mediana muestra que la mayoría percibe su estado de salud como promedio. En general los estudiantes mantienen un aceptable estado de salud ni muy buena ni muy mala.

Figura 25. Histograma estado de salud de los estudiantes.

Fuente: Elaboración propia.



El mapa de correlación figura 23, permite identificar el grado de relación lineal existente entre las variables numéricas consideradas en el estudio. Se observa

una correlación positiva entre el nivel educativo del padre y de la madre, lo cual sugiere una similitud en los niveles de formación dentro del entorno familiar.

De igual forma, se identificó una ligera relación entre el apoyo escolar y el apoyo familiar, lo que podría reflejar que los estudiantes que reciben acompañamiento institucional también cuentan con respaldo y apoyo en casa. En cuanto a las variables académicas, la calificación del primer semestre presenta correlaciones bajas con el resto de los factores, lo que sugiere que el aprovechamiento académico depende de una combinación de variables y no de un único elemento determinante.

Figura 26. Mapa de correlación entre variables numéricas

Fuente: Elaboración propia.

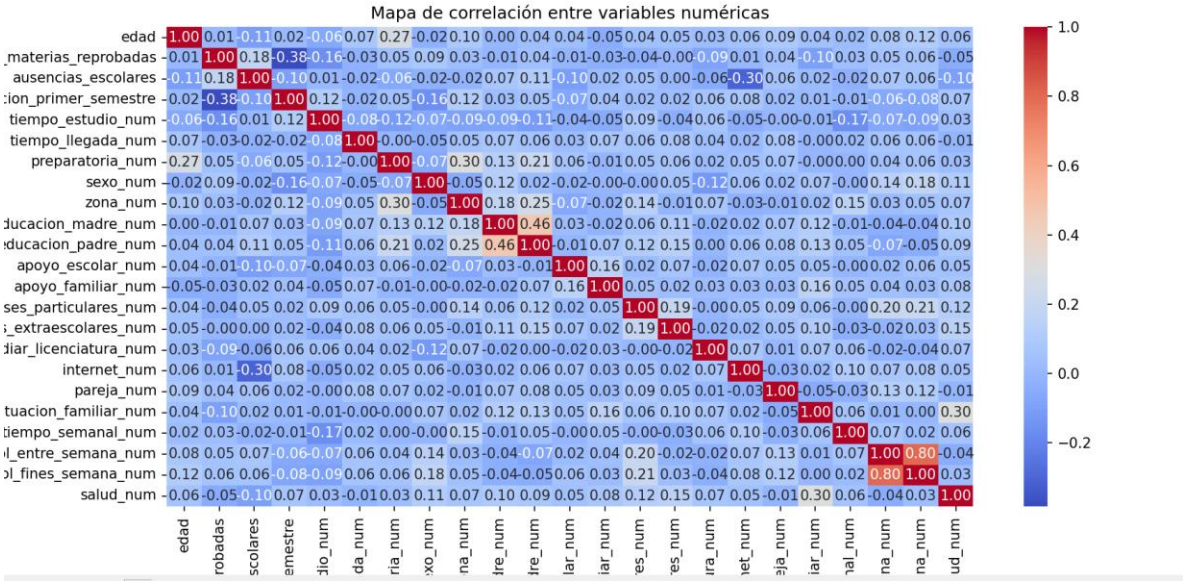
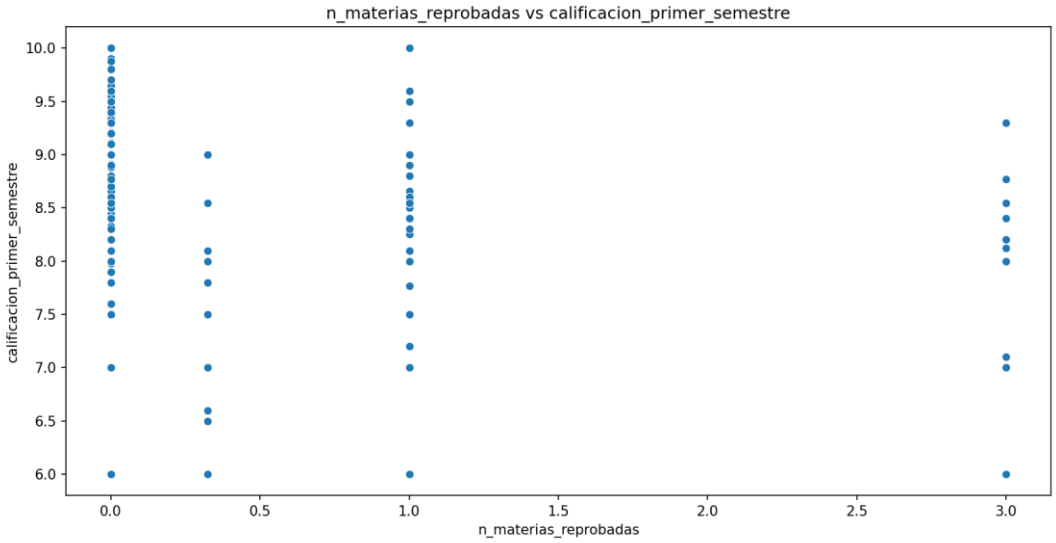


Figura 27. Materias reprobadas vs calificación primer semestre

Fuente: Elaboración propia.



Código 2. Mineria de datos

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

def mineria(ruta_csv="faseUno.csv"):
    # Crear carpeta para guardar gráficas si no existe
    carpeta_graficas = "graficas"
    os.makedirs(carpeta_graficas, exist_ok=True)

    # Cargar CSV limpio
    df = pd.read_csv(ruta_csv)

    # Selección de columnas numéricas
    num_cols = df.select_dtypes(include='number').columns.tolist()
    print("Columnas numéricas:", num_cols)

    # Estadísticas descriptivas
    desc_stats = df[num_cols].describe(percentiles=[0.25, 0.5, 0.75])
    print("\nEstadísticas descriptivas completas:")
    print(desc_stats)

    # Histogramas
    for col in num_cols:
        plt.figure(figsize=(8,4))
        sns.histplot(df[col], kde=True, bins=10)
        plt.title(f'Histograma de {col}')
        plt.xlabel(col)
```

```
        plt.ylabel("Frecuencia")
        plt.tight_layout()
        plt.savefig(f"{carpeta_graficas}/hist_{col}.png")
        plt.close()

    # Matriz de correlación
    for col in num_cols:
        df[col].fillna(df[col].mean(), inplace=True)
    corr_matrix = df[num_cols].corr()
    plt.figure(figsize=(10,6))
    sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f")
    plt.title("Mapa de correlación entre variables numéricas")
    plt.tight_layout()
    plt.savefig(f"{carpeta_graficas}/correlacion.png")
    plt.close()

    # Variable objetivo
    target = 'calificacion_primer_semestre_cat_num'

    # Correlación con la variable objetivo
    correlation_with_target = corr_matrix[target]

                                        selected_features                    =
correlation_with_target[correlation_with_target.abs()                        >
0.2].index.tolist()

    if target in selected_features:
        selected_features.remove(target)

    print("Variables seleccionadas para Fase 3:", selected_features)

    # Scatterplots vs variable objetivo
    for col in selected_features:
        plt.figure(figsize=(6,4))
        sns.scatterplot(data=df, x=col, y=target)
        plt.title(f'{col} vs {target}')
        plt.tight_layout()
        plt.savefig(f"{carpeta_graficas}/scatter_{col}_vs_{target}.png")
        plt.close()

    # PCA si hay más de una variable seleccionada
    if len(selected_features) > 1:
        X_scaled = StandardScaler().fit_transform(df[selected_features])
        pca = PCA(n_components=len(selected_features))
        X_pca = pca.fit_transform(X_scaled)
                print("Varianza   explicada   por   cada   componente:",
pca.explained_variance_ratio_)
    else:
        print("PCA no se aplica: solo hay una variable seleccionada.")

    return df, selected_features
```

4.3 Fase 3: Aprendizaje automático

Fase 3: Aprendizaje automático.

***Objetivo*:** Preparar y organizar los para su análisis y modelado predictivo.

Actividades:

- Discretización de calificaciones del primer semestre en categorías (Bajo, Medio y Alto)
- División del conjunto de datos para entrenar el modelo.
- Entrenamiento del modelo Random Forest Classifier con 100 árboles.
- Evaluación de modelo.

Resultado: Modelo entrenado con tabla final de predicción por cada estudiante.

En esta fase 3, se desarrolló un modelo de aprendizaje automático supervisado con el objetivo de predecir el aprovechamiento académico de los estudiantes a partir de las variables seleccionadas de la fase 2. Para esto se trabajo en un flujo de fases y procesos para implementar el entrenamiento del modelo Random Forest Classifier o bosque aleatorio.

El dataset se dividio en conjuntos de entrenamiento 70% y prueba 30%, el modelo Random Forest Classifier se entrenó con 100 árboles, cada árbol selecciono de forma aleatoria una muestra del dataset, posteriormente determina por votación la clase más votada entre todos los árboles es decir (Si de 100 árboles, 70 predicen “Medio”, 20 “Alto” y 10 “Bajo”) el modelo seleccionara “Medio” como predicción final, por ser el más votado.

Código 3. Aprendizaje Automático

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.impute import SimpleImputer
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score


def fase3_aprendizaje_por_alumno(ruta_csv="faseUno.csv",
selected_features=None, carpeta_graficas="graficas"):
    # Crear carpeta para gráficas si no existe
    os.makedirs(carpeta_graficas, exist_ok=True)

    # Cargar CSV
    df = pd.read_csv(ruta_csv)

    # Variable objetivo que usamos en fase 2
    target = 'calificacion_primer_semestre_cat_num'

    if df[target].isna().sum() > 0:
        df[target].fillna(df[target].mode()[0], inplace=True)  # Usamos la
moda porque es categórica

    etiquetas_map = {0: 'Bajo', 1: 'Medio', 2: 'Alto'}
    df['aprovechamiento_label'] = df[target].map(etiquetas_map)

    if selected_features is None:
        num_cols = df.select_dtypes(include='number').columns.tolist()
        selected_features = [col for col in num_cols if col != target]
```

```
X = df[selected_features]
y = df[target]

# Imputar valores faltantes en features con la media
imputer = SimpleImputer(strategy='mean')
X = pd.DataFrame(imputer.fit_transform(X), columns=selected_features)

# División train-test
X_train, X_test, y_train, y_test, df_train, df_test = train_test_split(
    X, y, df, test_size=0.3, random_state=42, stratify=y
)

# GridSearchCV con RandomForest
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'class_weight': ['balanced']
}

rf = RandomForestClassifier(random_state=42)

grid_search = GridSearchCV(
    estimator=rf,
    param_grid=param_grid,
    cv=5,
    scoring='accuracy',
    n_jobs=-1
)

grid_search.fit(X_train, y_train)

best_rf = grid_search.best_estimator_
```

```
    y_pred = best_rf.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    print(f"Accuracy en conjunto de prueba: {accuracy:.4f}")

    df_test = df_test.copy()
    df_test['prediccion'] = y_pred
    df_test['prediccion_label'] = df_test['prediccion'].map(etiquetas_map)

    report = classification_report(y_test, y_pred, target_names=['Bajo',
'Medio', 'Alto'])
    cm = confusion_matrix(y_test, y_pred)

    # Guardar matriz de confusión
    plt.figure(figsize=(6, 4))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
                xticklabels=['Bajo', 'Medio', 'Alto'],
                yticklabels=['Bajo', 'Medio', 'Alto'])
    plt.xlabel("Predicción")
    plt.ylabel("Real")
    plt.title("Matriz de Confusión Random Forest (GridSearchCV)")
    plt.tight_layout()
    plt.savefig(f"{carpeta_graficas}/matriz_confusion.png")
    plt.close()

    # Importancia de variables
    importances = pd.Series(best_rf.feature_importances_,
index=selected_features).sort_values(ascending=False)
    plt.figure(figsize=(8, 5))
    sns.barplot(x=importances.values, y=importances.index)
    plt.title("Importancia de variables - Random Forest (GridSearchCV)")
    plt.tight_layout()
    plt.savefig(f"{carpeta_graficas}/importancia_variables.png")
    plt.close()
```

```
    df_test['calificacion_primer_semestre_cat'] =
df_test['calificacion_primer_semestre_cat'].fillna('Desconocido')

    df_test['ausencias_escolares'] = df_test['ausencias_escolares'].fillna(-
1)

    print("Valores nulos en df_test para columnas finales:")

    print(df_test[['ausencias_escolares',
'calificacion_primer_semestre_cat',
'aprovechamiento_label','prediccion_label']].isnull().sum())



    columnas_resultado = selected_features + ['ausencias_escolares',
'calificacion_primer_semestre_cat',
'aprovechamiento_label','prediccion_label']

    df_resultados = df_test[columnas_resultado].reset_index(drop=True)



    print("Mejores parámetros encontrados por GridSearchCV:")

    print(grid_search.best_params_)

    print(f"Mejor accuracy validación cruzada:
{grid_search.best_score_:.4f}")



    print("Distribución de clases predichas en prueba:")

    print(pd.Series(y_pred).value_counts(normalize=True))



    return best_rf, df_resultados, report, cm, importances
```

4.4 Fase 4: Sistema final

Fase 4: Sistema final

***Objetivo*:** Implementar una interfaz para ingresar datos de estudiantes y mostrar las predicciones del rendimiento académico en tiempo real.

Actividades:

- Desarrollo de una interfaz web utilizando el framework de Flask.
- Visualización de resultados y generación de reportes automáticos.
- Representación gráfica del desempeño académico de los estudiantes.

Resultado: Sistema funcional con interfaz, capaz de predecir el nivel de aprovechamiento académico de estudiantes y presentar los resultados de forma visual y útil para toma de decisiones.

Para el desarrollo de la interfaz web y la integración del modelo predictivo, se utilizó Flask, framework de Python que permite construir aplicaciones web de manera rápida y flexible. Flask fue elegido por su simplicidad, su compatibilidad con bibliotecas de ciencia de datos.

Flujo de trabajo:

1. Vinculación de cada fase (Fase 1 -> Fase 2 -> Fase 3 -> Fase 4)
2. Creación de vistas HTML.
3. Crear rutas de vista que usara el usuario
4. Mostrar login donde el usuario accede solamente si tiene una cuenta.

5. Mostrar input de carga de archivo, donde el usuario puede ingresar el archivo csv.
6. Recibir datos procesarlos por las fases hasta llegar a la tres con el modelo entrenado y hacer predicciones.
7. Mostrar resultados y estadísticas al usuario.

Código 4. Sistema web con flask

```
from flask import Flask, render_template, request, redirect, url_for,
send_from_directory
import os
import shutil
from faseUno import procesar_csv
from faseDos import mineria
from faseTres import fase3_aprendizaje_por_alumno
from db import get_conexion

app = Flask(__name__)
app.config['UPLOAD_FOLDER'] = 'uploads'
app.config['GRAFICAS_FOLDER'] = 'graficas'

# Crear carpetas si no existen
os.makedirs(app.config['UPLOAD_FOLDER'], exist_ok=True)
os.makedirs(app.config['GRAFICAS_FOLDER'], exist_ok=True)

@app.route('/')
def login():
    return render_template('login.html')

#valida credenciales
@app.route('/validar_login', methods=['POST'])
def validar_login():
    email = request.form['email']
    password = request.form['password']

    try:
        conn = get_conexion()
        cur = conn.cursor()
        cur.execute("SELECT * FROM usuarios WHERE email=%s AND password=%s",
(email, password))
        usuario = cur.fetchone()

        if usuario:
            return redirect(url_for('index'))
        else:
            return render_template('login.html', error='Credenciales
inválidas')

    except:
```

```
        return render_template('login.html', error='Error de conexión')
    finally:
        cur.close()
        conn.close()

@app.route('/graficas/<filename>')
def graficas(filename):
    return send_from_directory(app.config['GRAFICAS_FOLDER'], filename)

@app.route('/descargar_excel')
def descargar_excel():
    from flask import send_file
    try:
        # Verificar si existe el archivo de resultados
        archivo_excel = os.path.join(app.config['UPLOAD_FOLDER'],
'resultados_alumnos.xlsx')
        if os.path.exists(archivo_excel):
            return send_file(archivo_excel, as_attachment=True,
download_name='resultados_alumnos.xlsx')
        else:
            return "No hay resultados disponibles para descargar", 404
    except Exception as e:
        return f"Error al descargar el archivo: {str(e)}", 500

@app.route('/home', methods=['GET', 'POST'])
def index():
    print("Método:", request.method)
    if request.method == 'POST':
        print("Se recibió un archivo CSV.")
        # Validar archivo
        if 'csv_file' not in request.files:
            return redirect(request.url)
        file = request.files['csv_file']
        if file.filename == '':
            return redirect(request.url)

        # Guardar archivo subido
        ruta_csv = os.path.join(app.config['UPLOAD_FOLDER'], file.filename)
        file.save(ruta_csv)

        # Limpiar carpeta de gráficas antes de generar nuevas
        if os.path.exists(app.config['GRAFICAS_FOLDER']):
            shutil.rmtree(app.config['GRAFICAS_FOLDER'])
        os.makedirs(app.config['GRAFICAS_FOLDER'], exist_ok=True)

        # ===== FASE 1: Limpieza del CSV =====
        df_fase1 = procesar_csv(ruta_csv)

        # ===== FASE 2: Minería de datos y selección de variables =====
        df_fase2, selected_features =
mineria(ruta_csv=os.path.join(os.getcwd(), "faseUno.csv"))

        # ===== FASE 3: Random Forest y resultados por alumno =====
        modelo, df_resultados, report, cm, importances =
fase3_aprendizaje_por_alumno(
```

```
            ruta_csv=os.path.join(os.getcwd(), "faseUno.csv"),
            selected_features=selected_features,
            carpeta_graficas=app.config['GRAFICAS_FOLDER']
        )

        # Guardar resultados finales en CSV y Excel para descarga
        df_resultados.to_csv(os.path.join(app.config['UPLOAD_FOLDER'],
'resultados_alumnos.csv'), index=False)
        df_resultados.to_excel(os.path.join(app.config['UPLOAD_FOLDER'],
'resultados_alumnos.xlsx'), index=False, engine='openpyxl')

        # Renderizar resultados en HTML
        return render_template(
            'resultados.html',
            tablas=[df_resultados.to_html(classes='table table-striped',
index=False, justify='center')],
            graficas=os.listdir(app.config['GRAFICAS_FOLDER']),
            folder_graficas=app.config['GRAFICAS_FOLDER']
        )

    return render_template('index.html')

if __name__ == '__main__':
    app.run(debug=True)
```

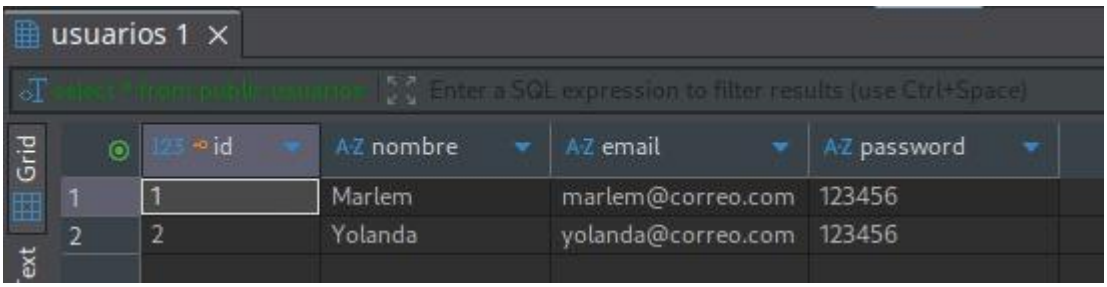
Se utilizo posgrest DB junto con una prueba gratuita de render para la conexión del sistema a la base de datos.

En la base de datos se creo una tabla de **usuarios** para el login del sistema, con atributos:

1. Serial Id
2. Varchar nombre
3. Varchar email
4. Varchar password

Figura 28. Tabla usuarios

Fuente: Elaboración propia.



Así mismo fue necesario crear una conexión en el sistema para vincularlas como se muestra en el código 5.

Código 5. Conexión a BD

```
import psycopg2

def get_conexion():
    conn = psycopg2.connect(
        host="dpg-d3m8q0t6ubrc73ejgnu0-a.oregon-postgres.render.com",
        database="sistema_wzzl",
        user="marlee",
        password="jpKLfHWYalFl6vV8fIHIX09Kzb2ajgpr",
        port="5432"
    )
    return conn
```

CAPITULO 5: PRUEBAS

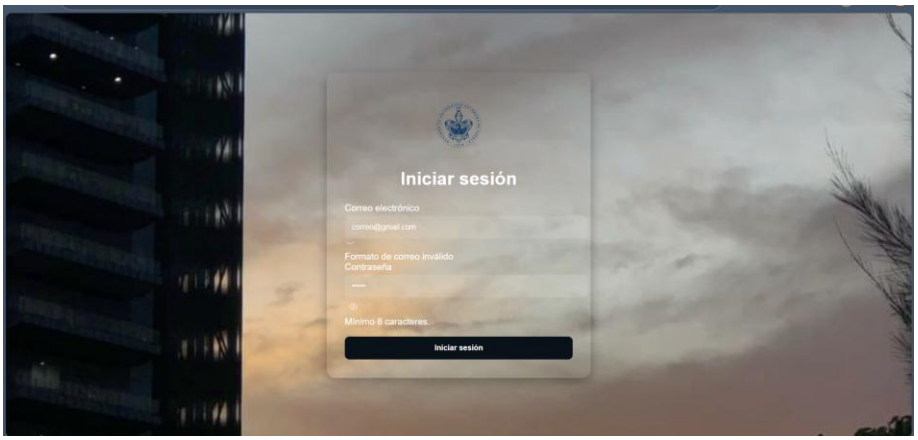
Como pruebas comenzamos iniciando sesión figura 28, con un usuario registrado en la base de datos.

Datos

Usuario: marlem@correo.com con password: 123456

Figura 29. Login de sistema

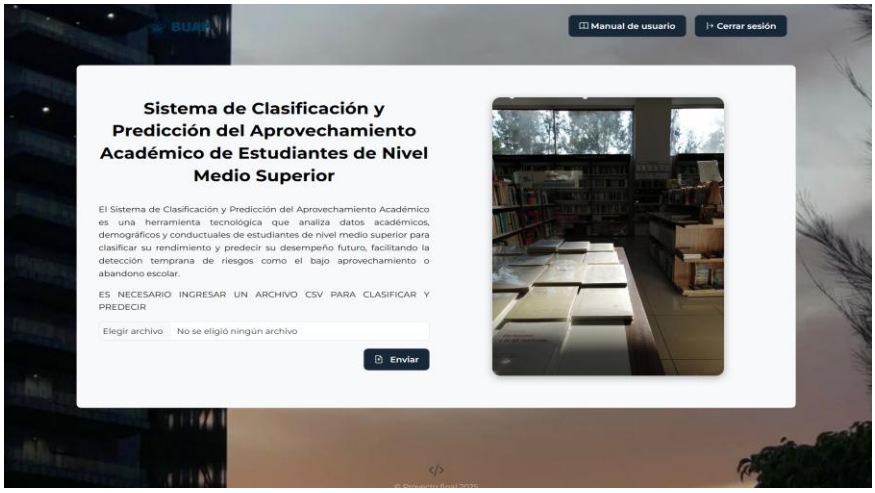
Fuente: Elaboración propia.



En la figura 29, se observa el inicio del sistema junto con una breve descripción asi mismo el input para subir el archivo CSV a analizar.

Figura 30. Interfaz inicio

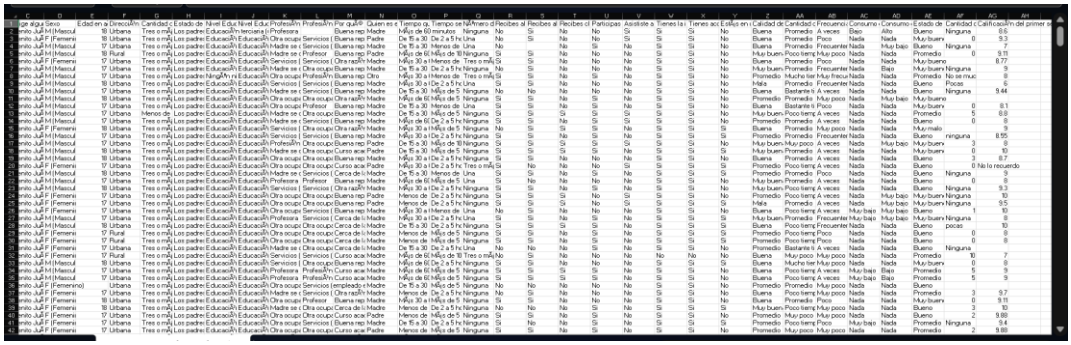
Fuente: Elaboración propia.



Una vez cargado un archivo CSV como se muestra en la figura 30, pasara por 4 etapas para poder ser interpretado y analizado por el modelo de predicción.

Figura 31. Archivo cargado en sistema

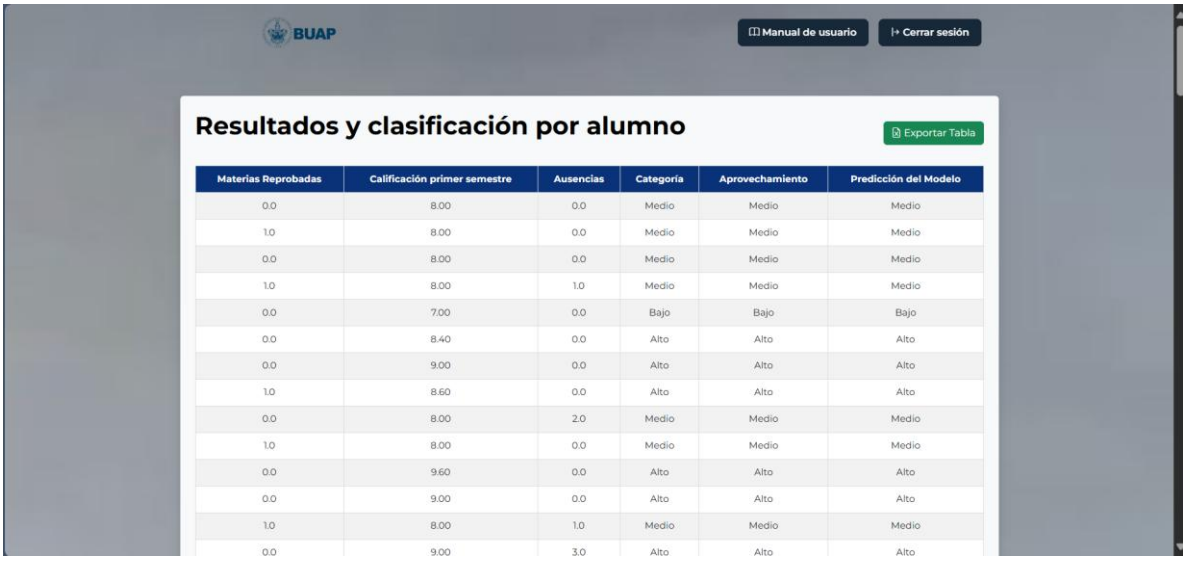
Fuente: Elaboración propia.



La predicción del sistema figura 31 y figura 32, son los resultados de todas las fases explicadas anteriormente. Compuesta por materias reprobadas, calificación de primer semestre, ausencias, categoría, resultados de aprovechamiento y predicción del modelo.

Figura 32. Resultados por alumno

Fuente: Elaboración propia.



Materias Reprobadas	Calificación primer semestre	Ausencias	Categoría	Aprovechamiento	Predicción del Modelo
0.0	8.00	0.0	Medio	Medio	Medio
1.0	8.00	0.0	Medio	Medio	Medio
0.0	8.00	0.0	Medio	Medio	Medio
1.0	8.00	1.0	Medio	Medio	Medio
0.0	7.00	0.0	Bajo	Bajo	Bajo
0.0	8.40	0.0	Alto	Alto	Alto
0.0	9.00	0.0	Alto	Alto	Alto
1.0	8.60	0.0	Alto	Alto	Alto
0.0	8.00	2.0	Medio	Medio	Medio
1.0	8.00	0.0	Medio	Medio	Medio
0.0	9.60	0.0	Alto	Alto	Alto
0.0	9.00	0.0	Alto	Alto	Alto
1.0	8.00	1.0	Medio	Medio	Medio
0.0	9.00	3.0	Alto	Alto	Alto

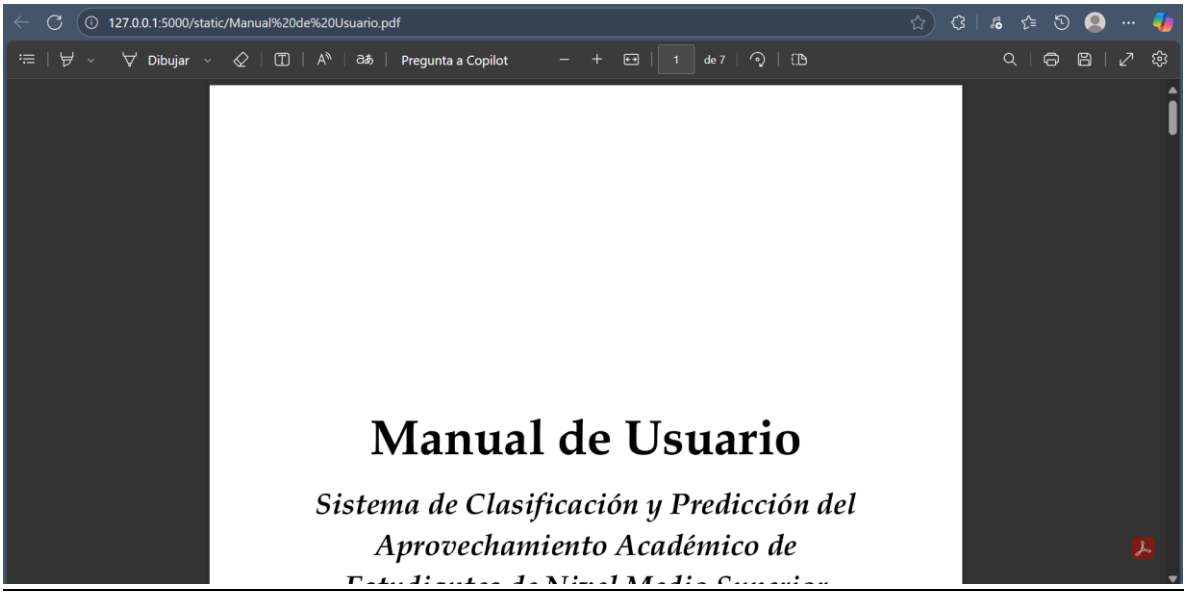
Figura 33. Estadísticas de resultados

Fuente: Elaboración propia.

Se incluye en el sistema un manual de usuario figura 33, demostrando la forma correcta de utilizar el sistema.

Figura 34. Manual de Usuario

Fuente: Elaboración propia.



CONCLUSIONES

El desarrollo del sistema de clasificación y predicción de aprovechamiento académico permitió demostrar que con el avance tecnológico de hoy en día contribuye en el análisis educativo. A través de cuatro fases (limpieza, minería de datos, aprendizaje automático y diseño web) se logró automatizar el procesamiento de archivos CSV con datos académicos, para generar resultados confiables de predicciones y estadísticas de un grupo de estudiantes de su desempeño académico. Como mejoras futuras se propone implementar nuevos algoritmos para predecir y clasificar, así como la admisión de diferente tipo de archivos.

BIBLIOGRAFIA

[1] Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. Formación Universitaria, 13(1), 57-66. https://doi.org/10.4067/S0718-50062020000100093

[2] Martínez, M. A. D., De los Angeles Ahumada-Cervantes, M., & Melo-Morín, J. P. (2021). Árboles de decisión como metodología para determinar el rendimiento académico en educación superior. *Dialnet*. https://dialnet.unirioja.es/servlet/articulo?codigo=8843574

[3] Franco, E. A., Martínez, R. E. L., & Domínguez, V. H. M. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. *Revista de Educación A Distancia (RED)*, *21*(66). https://doi.org/10.6018/red.463561

[4] Celis, Sergio & Moreno, Luis & Poblete, Patricio & Villanueva, Javier & Weber, Richard. (2015). Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería. Revista Ingeniería de Sistemas. 29. 5-24.

[5] Gutiérrez Villaverde, H., Linares Barbero, M., Agüero Correa, A. y Pérez Nuñez, J. (2022). Predicción de rendimiento académico de alumnos usando machine learning. Universidad de Lima, Facultad de Ciencias Empresariales y Económicas, Carrera de Negocios Internacionales.

[6] Witten, I. H., & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann.

[7] Chen, M. (2024, 25 noviembre). *What is machine learning?* https://www.oracle.com/mx/artificial-intelligence/machine-learning/what-is-machine-learning/

[8] TecnoDigital. (2024, marzo 7). *La minería de datos y el análisis de datos*. Informatec Digital. Recuperado de https://informatecdigital.com/la-mineria-de-datos-y-el-analisis-de-datos/

[9] Amazon Web Services. (n.d.). *¿Qué es la regresión logística?*. AWS. Recuperado el 13 de febrero de 2025, de https://aws.amazon.com/es/what-is/logistic-regression/

[10] De Estadística y Geografía, I. N. (s. f.). *Tabulados Interactivos-Genéricos*. https://www.inegi.org.mx/app/tabulados/interactivos/?px=Educacion_11&bd=Educacion

[11] Orozco-Rodríguez, C. (2022). *Factores que influyen en el abandono escolar de la licenciatura en Matemáticas de la Universidad de Guadalajara*. *Revista Mexicana de Investigación Educativa*, *27*(92), 259–287. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662022000100259

[12] Hernández Robles, A. K., & Vargas Valle, E. D. (2016). *Condiciones del trabajo estudiantil urbano y abandono escolar en el nivel medio superior en México*. *Estudios Demográficos y Urbanos*, *31*(3), 663–696. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0186-72102016000300663

[13] De Estadística y Geografía, I. N. (s. f.). *Tabulados Interactivos-Genéricos*. https://www.inegi.org.mx/app/tabulados/interactivos/?px=Educacion_11&bd=Educacion

[14] Galán, K. C. F. (s. f.). *Mineria de datos*. Scribd. https://es.scribd.com/document/346232447/Mineria-de-Datos

[15] NotasMD.pdf

[16] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases*. AI Magazine, 17(3), 37–54.

[17] Admin. (2020, 6 abril). *¿Qué es el KDD o Proceso de descubrimiento de conocimiento?* DiagramasUML.com. https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/#google_vignette

[18] *Minería de datos. Técnicas y herramientas*. (s. f.). Google Books https://books.google.com.mx/books?id=wzD_8uPFCEC&lpg=PR4&dq=tecnicas%20de%20mineria%20de%20datos&lr&hl=es&pg=PR4#v=onepage&q=tecnicas%20de%20mineria%20de%20dato&f=false

[19] Muguira, A. (2023, 15 mayo). *¿Qué es el modelo de elección discreta?* QuestionPro. https://www.questionpro.com/blog/es/modelo-de-eleccion-discreta/

[20] Muñoz, J. D. (2017, 17 noviembre). Qué es Flask. *OpenWebinars.net*. https://openwebinars.net/blog/que-es-flask/