EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Minería de Datos Educativa para la Detección de Patrones en una Institución Educativa Superior

presentada por

Lic. Brenda Erika Benítez Morales

como requisito para la obtención del grado de

Maestra en Ciencias de la Ingeniería

Director de tesis

Dr. Vitervo López Caballero

Codirector de tesis

Dr. Jesús Arce Landa

Cuernavaca, Morelos, México. Diciembre 2025.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo tecnológico
Coordinación de Ciencias de la Ingeniería

Cuernavaca, Mor., 10/ noviembre/ 2025
OFICIO No. MCI/135/2025
CENIDET-AC-004-M14-OFICIO
Asunto: **Aceptación de documento de tesis**

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de la estudiante **Brenda Erika Benítez Morales**, con número de control M23CE119, de la Maestría en Ciencias de la Ingeniería, le informamos que hemos revisado el trabajo de tesis de grado titulado **"MINERÍA DE DATOS EDUCATIVA PARA LA DETECCIÓN DE PATRONES EN UNA INSTITUCIÓN EDUCATIVA SUPERIOR"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DR. VITERVO LÓPEZ CABALLERO
Director de tesis

DR. JESÚS ARCE LANDA
Codirector de Tesis

DRA. ALICIA MARTÍNEZ REBOLLAR
Revisor 1

MTRO. VÍCTOR JOSUÉ RUIZ MARTÍNEZ
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante
JAL/ang

2025
Año de
La Mujer
Indígena

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Subdirección Académica

Cuernavaca Mor, 28/noviembre/2025
Oficio No. SAC/275/2025
Asunto: Autorización de impresión de tesis

BRENDA ERIKA BENÍTEZ MORALES
CANDIDATA AL GRADO DE MAESTRA EN CIENCIAS DE LA INGENIERÍA
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"MINERÍA DE DATOS EDUCATIVA PARA LA DETECCIÓN DE PATRONES EN UNA INSTITUCIÓN EDUCATIVA SUPERIOR"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

c.c.p. Coordinación de Ingeniería
Departamento de Servicios Escolares

CMAZ/lmz

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4104,
e-mail: acad_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet Centro Nacional de Investigación y Desarrollo Tecnológico
Norma Mexicana en Igualdad Laboral y No Discriminación
(NMX-SAA-14001-IMNC-2015) ISO 14001 ISO 14001:2015
ESPACIO 100% LIBRE DE PLÁSTICO DE UN SOLO USO

Dedicatoria

Con profunda gratitud, reconozco la culminación de mis estudios de maestría, un logro que atribuyo gracias a Dios. Agradezco la oportunidad de continuar mi formación, lo que me impulsa a perfeccionarme constantemente y a alcanzar una versión superior de mí misma.

A mis queridos y amados padres, Manuela Atocha Morales Carta y José de Jesús Benítez Zepeda, con profundo amor y gratitud, les dedico este logro. Su inquebrantable amor, infinita paciencia y constante apoyo han sido el cimiento de mi formación. Agradezco especialmente su respaldo incondicional que me permitió perseverar en mis estudios, superando cada desafío. Esta tesis es también el reflejo de los valores que me inculcaron y de su incansable esfuerzo por mi bienestar. Gracias por todo lo vivido juntos; este éxito es, en gran medida, suyo.

A mis queridos y amados hijos, Melany Alejandra Zúñiga Benítez y José Manuel Zúñiga Benítez, con todo mi amor, les dedico este logro. Su paciencia, comprensión y apoyo incondicional fueron el motor que me impulsó a seguir adelante durante este proceso de estudio y la elaboración de esta tesis. Cada sacrificio y cada hora de dedicación valieron la pena gracias a su invaluable presencia. Son mi mayor inspiración y esta tesis es también un testimonio de su amor.

Expreso mi más sincero agradecimiento a mis hermanos y al resto de mi familia por su constante apoyo y comprensión a lo largo de este proceso de estudio de la maestría. Su aliento y respaldo fueron fundamentales para culminar esta etapa académica.

Agradecimientos

Deseo expresar mi más sincero agradecimiento a la Secretaría de Ciencias, Humanidades, Tecnología e Innovación (SECIHTI), por la invaluable oportunidad de estudio y el generoso apoyo que me han brindado. Su contribución ha sido fundamental para mi desarrollo académico y para alcanzar mis metas profesionales.

Deseo expresar mi más sincero agradecimiento a mi director de tesis, el Dr. Vitervo López Caballero por su invaluable orientación en el campo de la minería de datos y por compartir generosamente sus amplios conocimientos. Su asesoramiento ha sido un pilar fundamental en mi formación académica y sus conocimientos en disciplinas afines, ha enriquecido significativamente mi trabajo.

Expreso mi más sincero y profundo agradecimiento al Dr. Jesús Arce Landa, mi codirector de tesis. Su constante apoyo. Valoro enormemente su orientación precisa en la elaboración y estructura de este trabajo, así como su invaluable paciencia para guiarme a lo largo de este complejo proceso de investigación. Su asesoramiento ha enriquecido significativamente la calidad de esta tesis.

Expreso mi más sincero agradecimiento a la Dra. Alicia Martínez Rebollar y al Mtro. Víctor Josué Ruíz Martínez por formar parte de mi comité revisor, por sus valiosas asesorías y conocimientos compartidos. Su experiencia y orientación han sido fundamentales para la culminación exitosa de mi trabajo, y aprecio enormemente la claridad y profundidad de sus aportaciones.

Con un especial reconocimiento, agradezco al Tecnológico Nacional de México Campus Iztapalapa III por facilitar el acceso a los datos cruciales para la realización de esta tesis. Su valiosa colaboración ha sido indispensable para la obtención de resultados significativos.

Deseo agradecer a la Dra. Hilda Lee del Carmen Faviel y a la Dra. Paola Eunice Gallegos Ortiz por su apoyo constante y su gran compañerismo, los cuales fueron fundamentales durante esta etapa.

Resumen

La Minería de Datos Educativos es una disciplina en crecimiento que aplica técnicas de la ciencia de datos para descubrir patrones y tendencias ocultos en grandes volúmenes de datos académicos. Su objetivo central es fundamentar la toma de decisiones basada en evidencia en entornos educativos, con énfasis en la mejora del rendimiento estudiantil y la reducción de la deserción.

Esta *tesis* presenta un caso práctico realizado en El Tecnológico Nacional de México Campus Iztapalapa III con el fin de identificar patrones académicos predictivos del éxito o fracaso estudiantil a nivel de ingeniería. La metodología implementada se dividió en cuatro etapas clave: (I) recopilación de datos, (II) preprocesamiento, (III) selección y validación del modelo, e (IV) interpretación de los resultados. Se utilizaron técnicas de aprendizaje no supervisado, específicamente Análisis de Componentes Principales y el agrupamiento K-means, que revelaron tres perfiles estudiantiles distintos. El análisis identificó un grupo con indicios de desvinculación emocional o académica, lo que subraya la necesidad de implementar estrategias de intervención temprana como asesoramiento psicológico, mentoría personalizada y orientación vocacional.

Finalmente, se entrenaron cuatro algoritmos de aprendizaje supervisado para clasificar a los nuevos estudiantes en función de sus características académicas. Los modelos más efectivos fueron el Perceptrón Multicapa, las Máquinas de Vectores de Soporte y la Regresión Logística, logrando métricas de rendimiento elevadas, incluyendo una precisión del 95%, una exactitud del 96%, una sensibilidad del 96% y un puntaje F1 del 96%. Estos resultados demuestran la eficacia de la Minería de datos Educativa para la detección de patrones y reducir la deserción estudiantil.

Palabras Clave: Minería de Datos Educativos, Deserción Estudiantil, Rendimiento Académico, Aprendizaje Supervisado, Agrupamiento K-means, Predicción de Éxito/Fracaso.

Abstract

Educational Data Mining is a growing discipline that applies data science techniques to uncover hidden patterns and trends in large volumes of academic data. Its central objective is to support evidence-based decision-making in educational settings, with an emphasis on improving student performance and reducing dropout rates.

This thesis presents a practical case study conducted at the National Technological Institute of Mexico, Iztapalapa III Campus, to identify academic patterns predictive of student success or failure at the engineering level. The implemented methodology was divided into four key stages: (I) data collection, (II) preprocessing, (III) model selection and validation, and (IV) interpretation of results. Unsupervised learning techniques were used, specifically Principal Component Analysis and K-means clustering, which revealed three distinct student profiles. The analysis identified a group with signs of emotional or academic disengagement, highlighting the need to implement early intervention strategies such as psychological counseling, personalized mentoring, and vocational guidance.

Finally, four supervised learning algorithms were trained to classify new students based on their academic characteristics. The most effective models were the Multilayer Perceptron, Support Vector Machines, and Logistic Regression, achieving high performance metrics, including 95% precision, 96% accuracy, 96% sensitivity, and an F1 score of 96%. These results demonstrate the effectiveness of Educational Data Mining for pattern detection and reducing student dropout rates.

Keywords: Educational Data Mining, Student Dropout, Academic Performance, Supervised Learning, K-means Clustering, Success/Failure Prediction.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

Capítulo 1
Introducción

La Minería de Datos Educativa es una disciplina clave que busca mejorar los procesos educativos de forma automatizada al desarrollar métodos para analizar los datos únicos generados en entornos de aprendizaje. La Minería de Datos Educativa integra técnicas de campos como la minería de datos, el aprendizaje automático, la psicometría, la estadística, la visualización de información y el modelado computacional. Su objetivo principal es descubrir información valiosa, patrones, relaciones y excepciones dentro de grandes volúmenes de datos educativos, lo que requiere de procesos específicos para depurar la información y generar conocimiento confiable y excepciones dentro de estos grandes volúmenes de datos. (F. De et al., 2020; Sulla Torres, 2021)

La educación superior se encuentra inmersa en una era de transformación digital, donde la gran cantidad de datos generados por las actividades académicas y administrativas representan un activo valioso y, a menudo, subutilizado. El desafío fundamental para las Instituciones de Educación Superior ya no radica solo en la recolección de estos datos, sino en la capacidad de transformarlos en conocimiento accionable. La Minería de Datos Educativa se enfoca en la detección de patrones, tendencias y correlaciones significativas, hasta ahora ocultas, sobre los datos históricos de una institución, como el TecNM Iztapalapa III. Su capacidad predictiva y prescriptiva facilita la toma de decisiones basada en evidencia. Al analizar los datos históricos, se puede identificar a estudiantes en riesgo de abandono antes de que deserten, optimizar la asignación de recursos, evaluar la eficacia de los programas de estudio o personalizar las trayectorias de aprendizaje. En este contexto, la presente tesis sienta las bases para explorar cómo la Minería de Datos Educativa, mediante técnicas como la clasificación o agrupamiento (clustering) y la asociación de reglas, no solo puede mejorar la eficiencia operativa de la Institución de Educación Superior, sino, fundamentalmente, potenciar el éxito estudiantil y la calidad académica, marcando un avance hacia una gestión universitaria verdaderamente inteligente y proactiva.

1.1 Descripción del problema de investigación

En la actualidad uno de los principales problemas que preocupan a la mayoría de las Instituciones de Educación Superior, es la deserción estudiantil o interrupción de los estudios antes de la finalización del programa académico, debido a su naturaleza multifactorial y compleja, lo que dificulta encontrar una solución única y sencilla.

En el TecNM Campus Iztapalapa III, se ha identificado un número preocupante en la tasa de deserción durante los últimos 5 años. La tasa oscila entre un 25% y el 30% anual promedio, superando el promedio institucional del 15%. La tasa de abandono para los estudiantes de nivel ingeniería de las carreras en Civil, Gestión Empresarial e Informática; se ha observado que la mayor parte de bajas ocurren durante el primer y segundo semestre, lo que sugiere posibles problemas de adaptación y/o manejo de las asignaturas fundamentales, causas académicas, causas sociales, causas institucionales o personales.

Dada la tendencia creciente y las graves consecuencias mencionadas, en el contexto de la deserción estudiantil en programas con un enfoque en la educación sostenible, los enfoques actuales de Minería de Datos Educativa se centran en la detección de patrones para implementar intervenciones de retención y al mismo tiempo evitar la pérdida de futuros profesionales esenciales para el desarrollo de infraestructuras, tecnologías y negocios sostenibles en la región metropolitana.

1.2 Objetivos

1.2.1 Objetivo General

Aplicar técnicas de Minería de Datos Educativa a una Institución Educativa Superior para identificar patrones asociados al riesgo de deserción estudiantil, mediante modelos de agrupamiento evaluados con un índice de Calinski–Harabasz, con el fin de diseñar estrategias de intervención focalizadas que contribuyan a reducir el abandono escolar.

1.2.3 Objetivos Específicos

- Diseñar, validar y aplicar un instrumento de recopilación de datos (formularios) que capture las variables académicos, socioeconómicos e históricos relevantes de los estudiantes de una Institución Educativa Superior, de algunos semestres y egresados asegurando la representatividad de la muestra para el análisis de la deserción.

- Codificar y transformar las variables categóricas nominales y ordinales recopiladas (incluyendo datos del formulario) mediante la técnica de OneHotEncoding en Python, garantizando la conversión total a formato numérico y la imputación de valores nulos y generar un conjunto de datos limpios y listos para el entrenamiento del modelo de Minería de Datos.

- Interpretar y analizar los resultados obtenidos del modelo de agrupamiento de deserción, identificando las principales tendencias y patrones de riesgo académico y conductual y facilitar la toma de decisiones.

1.3 Justificación

La presente ***tesis*** aborda una problemática en el ámbito de la educación superior: la predicción del éxito o fracaso académico de estudiantes universitarios. En un contexto donde las Instituciones de Educación Superior buscan optimizar el rendimiento estudiantil y reducir la deserción, la Minería de Datos Educativa emerge como una herramienta fundamental. Esta ***tesis*** se justifica por su capacidad de cerrar la brecha entre la teoría y la práctica, ofreciendo soluciones analíticas con aplicabilidad directa en entornos educativos reales.

Mediante la detección de los patrones para aplicar algoritmos de aprendizaje automático, tanto supervisado como no supervisado, se ha desarrollado un caso de estudio sistemático desde la recolección y preprocesamiento de los datos hasta la selección, validación e interpretación de modelos.

Los análisis revisados evidencian una preferencia metodológica consolidada hacia el empleo de modelos predictivos que equilibran la interpretabilidad con la robustez, destacando el uso frecuente de algoritmos como Redes Neuronales, Árboles de Decisión, Máquinas de Soporte Vectorial y Naive Bayes. Particularmente, la amplia adopción de las redes neuronales subraya la necesidad de modelar eficazmente relaciones complejas inherentes a los datos estudiados.

Adicionalmente, el entrenamiento de modelos de aprendizaje supervisado demostró una excepcional capacidad predictiva. Algoritmos como el Perceptrón Multicapa, las Máquinas de Soporte Vectorial y la Regresión Logística alcanzaron métricas de desempeño sobresalientes: una precisión global del 95 %, una precisión del 96 %, una sensibilidad del 96 % y un puntaje F1 del 96 %.

Estos resultados validan la alta fiabilidad de los modelos para clasificar con precisión a estudiantes de nuevo ingreso, basándose en la detección de patrones de datos históricos.

La implementación de estos modelos predictivos no solo permitirá a las Instituciones de Educación Superior tomar decisiones informadas para optimizar el apoyo académico y las intervenciones, también contribuirá significativamente a la retención estudiantil y al mejoramiento continuo de la calidad educativa.

Esta ***tesis***, por lo tanto, no solo es relevante desde una perspectiva metodológica, sino que ofrece un impacto tangible al proporcionar las herramientas robustas para la proactiva gestión del rendimiento académico, facilitando un entorno más propicio para el éxito de sus estudiantes.

1.4 Estructura del documento

La ***tesis*** está organizada en seis capítulos, en los que se detallan las principales actividades y metodologías implementadas. En esta sección se presenta una introducción fundamental al tema, abordando la importancia de la Minería de Datos Educativa, que emerge como una herramienta crucial para la detección temprana y la comprensión de la deserción estudiantil, constituyendo la base conceptual de este proyecto. En el Capítulo 2, se establece un marco conceptual sólido, definiendo con precisión los términos clave que sustentan el estudio de esta ***tesis***. En el Capítulo 3 se presenta el estado del arte sobre el tema de la minería de datos educativa, un concepto fundamental para esta ***tesis*** ofreciendo una definición formal y clara, ya que es la fuente de la información esencial para la solución sugerida. En el Capítulo 4, se detalla la metodología utilizada. En el Capítulo 5 se presenta un caso de estudio que demuestra cómo se aplica la solución. En el Capítulo 6 se muestran las conclusiones obtenidas y los trabajos futuros, destacando en la misma línea de investigación.

Capítulo 2

Marco conceptual

En el presente capítulo se establece el Marco Conceptual que sustenta los principios y el léxico utilizados en esta ***tesis***. Se definirán los conceptos clave para proporcionar una base sólida y un entendimiento claro del lenguaje empleado a lo largo del estudio. Adicionalmente, se analizaron los trabajos previos relevantes que se relacionan directamente con el contexto de esta investigación, situando el estudio dentro del panorama académico existente.

2.1 Fundamentos conceptuales de la Minería de Datos Educativa

El presente marco conceptual establece las bases teóricas y metodológicas que sustentan la aplicación de la Minería de Datos Educativa para comprender, predecir, detectar los patrones y mitigar la deserción estudiantil.

La minería de datos educativa es una disciplina que se enfoca en el desarrollo de métodos para explorar los datos únicos que provienen de entornos educativos. Su objetivo es comprender a los estudiantes y sus entornos de aprendizaje. Su aplicación más destaca en el contexto de la gestión universitaria se centra en la reducción del abandono escolar, los algoritmos más comunes en la Minería de Datos Educativa para estos fines son, Máquinas de Soporte Vectorial, Regresión Logística, Random Forest, Redes Neuronales (Perceptron Multicapa) y clasificadores Bayesianos (Naive Bayes).(David et al., 2018).

Técnicas de Minería de Datos Educativa aplicadas a la detección de Patrones

La Minería de Datos Educativa ofrece un conjunto robusto de técnicas orientadas a la extracción de conocimiento, siendo la base metodológica para la detección de patrones de riesgo de deserción. Estas técnicas se clasifican generalmente en dos categorías: algoritmos de aprendizaje supervisado (con variable objetivo) y aprendizaje no supervisado (sin variable objetivo). Se detallan los principales conceptos de los algoritmos y las métricas de evaluación de estos algoritmos que se aplicarán en esta ***tesis*** son las siguientes:

2.2 Algoritmos de aprendizaje automático supervisado

El aprendizaje automático es considerado una rama de la inteligencia artificial, ya que aplica diversas técnicas en aprendizaje del pasado. Se relaciona con las ciencias computacionales, las ingenierías, la estadística, la minería de datos y con reconocimiento de patrones. Los componentes clave para obtener resultados ideales de acuerdo con el problema en cuestión, son tres, a saber: la representación, la evaluación y la optimización. A su vez, se ramifica en aprendizaje supervisado y no supervisado.(Castillo Aráuz & Martínez, 2023).

Se conoce como aprendizaje supervisado aquel en que los datos son etiquetados con el objetivo de entrenar un modelo (Almasri et al., 2020); esto conduce a una categorización o clase. El algoritmo aprende de las características de los datos para realizar predicciones. Muchos algoritmos son agrupados en esta categoría, como por ejemplo los siguientes:

Regresión Logística

Logistic Regression (regresión logística) Es una técnica estadística de varias variables, que facilita el análisis de la relación que se encuentra entre un grupo de variables independientes y una variable dependiente, esta última es dicotómica, es decir que solo puede tomar dos valores, que generalmente son cero y uno, donde cero es ausencia y uno presencia, es decir la variable dependiente es discreta, a diferencia del grupo de variables independientes la cuales pueden ser tanto cuantitativas como cualitativas; además la función de partida es exponencial, lo que ayuda a representar mejor ciertos fenómenos (Amat Rodrigo, 2025).

Para convertir la puntuación z, se utiliza la siguiente fórmula (que puede variar de $(-\infty\ a + \infty)$ en una probabilidad P (que debe estar entre 0 y 1), se utiliza la función sigmoide (σ):

$$\sigma(z) = P(Y = 1|X) = \frac{1}{1 + e^{-z}} \quad (1)$$

Random Forest

Es un método de Machine Learning conjunto que permite la clasificación y la regresión, el cual está conformado por un conjunto de árboles de decisión, construidos mediante bagging, cuya variación se controla considerando a los datos de entrenamiento. En el proceso de clasificación, cada árbol emite su voto de forma individual sobre la clase predicha, en la que finalmente se considera aquella clase con votación mayoritaria usando la siguiente fórmula. (Faw Agregar, Gaber y Elyan, 2014).

$$\hat{y} = \text{mod}\, o\{h_i(x)\}_t^T = 1 \tag{2}$$

Fórmula clave en la construcción del Árbol

La determinación de la "mejor división" en un nodo en fórmulas matemáticas, como:

Impureza de Gini (Gini Index) es la siguiente:

$$G_{ini} = 1 - \sum_{k=1}^{k} (p_k)^2 \tag{3}$$

Donde Pk es la proporción de observaciones de la clase k (e.g., "Deserta" o "No Deserta") en el nodo, y K es el número de clases. El objetivo es minimizar Gini.

En la Predicción Individual, Una vez que los árboles están entrenados, se introduce la información de un nuevo estudiante al que se quiere predecir. Cada árbol de decisión en el bosque produce su propia predicción: "Deserta" o "No Deserta". Para obtener la predicción final del Random Forest, se aplica la votación por mayoría. La predicción de clase que haya sido elegida por la mayoría de los árboles de decisión individuales se convierte en la predicción final del modelo para el estudiante.

Máquina de soporte Vectorial

Support Vector Machine (máquinas de vectores de soporte): para este algoritmo los datos son trazados en un espacio n dimensional (basado en el número de características) y un límite de decisión (o hiperplano) que divide los datos en clases. Cuanto más lejos estén los puntos de datos graficados del límite de decisión, más seguro estará el algoritmo sobre la predicción. Los puntos de datos más cercanos al hiperplano se denominan vectores de soporte. (Mazurett González, 2025).

Para la predicción y detección de patrones utilizamos la siguiente formula:

$$f(xnuevo) = \text{sign}\left(\sum_{i=1}^{N} a_i yi^k\left(x_{i,} \times nuevo\right) + b\right) \quad (4)$$

Donde:

$f(x)$: $Funsión\ de\ decisión$, el resultado de la función es la etiqueta de clase predicha para el nuevo dato x. El sing ( ) indica si el resultado es positivo (+1, una clase) o negativo (-1, la otra clase).

sign (.): Funsión signo, devuelve el signo del valor dentro del paréntesis. Si el valor es > 0, la predicción es +1. Si es < 0, es -1. Si es 0, el punto que esta justo en el límite.

$\sum_{i=1}^{N}$ $Sumatoria$: suma los resultados para cada uno de los N vectores de soporte identificados durante el entrenamiento.

N: Número de vectores de soporte, la cantidad de puntos de datos del conjunto de entrenamiento que están más cerca del hiperplano y que definen el margen. Son los puntos críticos.

$\alpha i: Multiplicadores\ de\ Lagrange,$ son los coeficientes calculados durante el proceso de optimización de SVM. Solo los vectores de soporte tienen un $\alpha i > 0$; para el resto es cero.

$y_i: Etiquetas\ de\ clase\ del\ Vector\ de\ Soporte,$ la clase real ala que pertenece el vector de soporte $x_{i,}$ puede ser +1 o -1.

$k(x_i, x): Funsión\ Kernel,$ calcula la similitud entre un vector de soporte $x_{i,}$ y el nuevo punto de datos *x*. Transforma implícitamente los datos a un espacio de mayor dimensión (si se usa un Kernel no lineal).

$x_{i:Vector\ de\ Soporte,}$ es un punto de datos específico del conjunto de entrenamiento que es crucial para definir el hiperplano.

x: Nuevo punto de datos, es el punto de datos cuya clase se requiere predecir o clasificar.

b: Termino de sesgo (bias)/ Intercepto, es un valor escalar que desplaza el hiperplano del origen, ajustando su posición óptima en el espacio de características.

Una vez entrenado, el modelo utiliza la función de decisión para predecir la clase de un nuevo estudiante (si es probable desertor o no).

Redes Neuronales - Perceptrón Multicapa

Multilayer Perceptron (perceptrón multicapa): es la red neuronal artificial (RNA) más común. Tiene dos capas conectadas directamente con el entorno, conocidas como entrada y salida; entre las dos están las intermedias que se denominan ocultas. La señal que transmiten las neuronas sigue una única dirección (de la entrada hasta la salida) sin formar bucles. Esta estructura es llamada feed fordward. También utiliza un algoritmo llamado reprogramación para minimizar los errores entre los resultados del modelo y los resultados esperados (Marius-Constantin et al., 2009). Se utilizaron las siguientes fórmulas.

$$z_i = \phi\left(w(1)x_i + b^{(1)}\right) \tag{5}$$

$$y^i = v\left(w(2)z_i + b^{(2)}\right) \tag{6}$$

2.3 Aprendizaje automático no supervisado

Por otro lado, en el aprendizaje no supervisado, los datos no son etiquetados. Lo que el algoritmo realiza es un agrupamiento para descubrir patrones en ellos. Algunos métodos son:

Análisis de componentes Principales

*Principal Component Analysis (PCA):*es una técnica utilizada para reducir la dimensionalidad de los datos buscando que la pérdida de información sea la menor posible. El conjunto de datos se transforma en un sistema de coordenadas. Se elige un primer eje en la dirección de mayor variación en los datos. Se escoge un segundo eje ortogonal al primer eje y con la mayor varianza posible. Este proceso se repite hasta que todos los datos estén todos incluidos en los componentes principales generados; presenta una mejora en el rendimiento de los algoritmos y reduce el tiempo de entrenamiento (Hasterok et al., 2019).

K-Means

K-Means: es un algoritmo para formar clústeres con características similares. El centro de cada conglomerado se llama centroide y es la distancia media de los valores en cada conglomerado. El algoritmo encuentra k conglomerados únicos y cada muestra se asigna al conglomerado con el centro más cercano. Es un algoritmo iterativo, es decir, ejecuta ciclos o iteraciones cada vez que se actualizan los centros de los clústeres (Alférez et al., 2022).

2.4 Métricas para la evaluación de algoritmos de Machine Learning

Las métricas para la evaluación de algoritmos de Machine Learning, se encuentra la precisión, que es la relación entre el número de casos positivos correctamente clasificados y

el total de casos positivos predichos. La métrica de exactitud, que se obtiene a partir de la división entre la suma de clasificaciones correctas y el total de los casos clasificados. Borja et al. (2020) indica que, a pesar de su cálculo y compresión, su mayor desventaja es la de producir menos valores discriminatorios para el caso de multiclase desbalanceado. Otra métrica importante es la exhaustividad, también denominada como Recall o sensibilidad, que es la relación entre el número de casos positivos correctamente clasificados y el total de casos positivos reales.

2.5 Herramientas aplicadas

Python

Python es un lenguaje de programación con una gran facilidad de uso y de alto nivel, Upasani y Virendra (2020) detallan que cuenta con una gran variedad de librerías útiles para la carga, procesamiento y visualización de datos, además de librerías capaces de ejecutar algoritmos de machine learning para la clasificación, predicción, clustering, etc.

Google Collab

Google Collab es un entorno de Jupyter Notebook basado en la nube que se ejecuta en diversos navegadores, Colab nos permite escribir y ejecutar código de Python o de R para aplicaciones en Inteligencia Artificial, garantizando por ejemplo el desarrollo y entrenamiento de redes neuronales, entre otros algoritmos. También permite compartir el código con múltiples usuarios proporcionando edición simultánea del código.

Scikit Learn

Scikit-Learn es una biblioteca de tipo código abierto en constante desarrollo y mejora. En Upasani y Virendra (2020) mencionan que esta librería proporciona múltiples algoritmos de Machine Learning, para aprendizaje supervisado y no supervisado, y otras múltiples herramientas científicas que la han permitido llegar a ser un instrumento ideal para el desarrollo de aplicaciones de Machine Learning.

Capítulo 3

Estado del arte

El Estado del Arte se establece mediante el análisis riguroso de la literatura reciente, incluyendo contribuciones clave, modelos predominantes y hallazgos empíricos. Dicho análisis es fundamental para el marco teórico de esta ***tesis***, contextualizar el estudio y determinar la frontera actual del conocimiento en el campo.

3.1 Clasificación de los trabajos del estado del arte

Las categorías utilizadas para clasificar las técnicas de la Minería de Datos Educativa para la detección de patrones en una institución educativa superior son las siguientes:

1. Minería de Datos Educativa. Es una disciplina clave que surge de la necesidad de analizar grandes cantidades de datos para resolver problemas educativos

2. Analítica del Aprendizaje. Es un modelo de análisis de datos educativos enfocado en la mejora y optimización de los procesos de aprendizaje y la toma de decisiones.

3.2 Minería de Datos Educativa

En el contexto de la Minería de Datos Educativa para la Detección de Patrones para una Institución de Educación Superior, la aplicación de modelos supervisados para predecir el patrón de interés, se aplicará técnicas de selección de características para identificar las variables más predictivas de ese patrón e integrando fuentes de datos mixtas para lograr una predicción de alta precisión y ofrecer una visión completa del estudiante.

Minería de Datos para detección de patrones estresores en la educación superior (Urbina-Nájera et al., 2020)

En este artículo se aplicó la minería de datos educativa para identificar los factores que inciden en la deserción escolar universitaria en el estado de Puebla, México. Se utilizó un instrumento adaptado de 56 preguntas en 9 categorías (desde datos demográficos hasta servicios e infraestructura), cuya fiabilidad fue alta (Coeficiente de Cronbach de 0.8767). La población fue de 230,788 estudiantes, y la muestra calculada se administró en línea a 500 estudiantes (300 de Instituciones de Educación Superior públicas y 200 de privadas) usando muestreo aleatorio simple.

Posteriormente, se aplicó la selección de atributos con evaluadores GainRatioAttributeEval e InfoGainAttributeEval, identificando 27 atributos como los más relevantes de las 56 iniciales. El algoritmo de árboles de decisión y se aplicó en dos fases: con 56 atributos, obteniendo un 74.6% de exactitud y con los 27 atributos más relevantes, logrando un aumento significativo del desempeño con un 92.6% de exactitud. Los resultados de los árboles de decisión permitieron establecer patrones que definen la permanencia o deserción. El patrón de "Lealtad" se relaciona con un ambiente estudiantil satisfactorio, casi nula falta de asesoría y seguimiento académico, servicios en general adecuados y no haber vivido una situación de impacto. Las cinco principales causas de deserción encontradas fueron la falta de asesorías, inadecuado ambiente estudiantil, falta de seguimiento académico, deficiente calidad educativa y servicio en general. Estos hallazgos contrastan con la literatura previa, donde predominaban factores como el disgusto por la carrera y los recursos económicos. El estudio concluye que la deserción es causada por un conjunto de factores e insta a las Instituciones de Educación Superior a crear mecanismos para mejorar las asesorías, el seguimiento académico y el clima educativo positivo, sugiriendo ampliar la muestra a otros estados para enriquecer los patrones encontrados.

Minería de Datos identificando causas de deserción en las Instituciones Públicas de Educación superior en México (López Pedraza et al., 2019)

En este artículo el problema de la deserción escolar en las Instituciones de Educación Superior públicas de México es grave y multifactorial, y se está abordando con minería de datos para identificar los patrones y mejorar las decisiones institucionales. La metodología utilizada para este fin es CRISP-DM (43% de uso en 2014) y SEMMA , además del proceso KDD. .La técnica se divide en predictivas (supervisadas) y descriptivas (no supervisadas).En 2016, los algoritmos más populares fueron Regresión (67%), Clustering (57%) y Árboles de Decisión/Reglas (55%).El software líder en 2018 fue Python (65,6%) , seguido por RapidMiner (52,7%) y R (48,5%).

Los estudios realizados en México, que han empleado metodologías como el descubrimiento de conocimiento en bases de datos y algoritmos como J48, coinciden en que las principales causas de deserción son de tipo personal/individual, académico y socioeconómico. Factores específicos identificados incluyen: Patrones Personales y Académicos: Planes de matrimonio, exámenes extraordinarios, poco tiempo de estudio, no haber elegido la carrera como primera opción. Patrones Académicos: Calificaciones obtenidas en el primer año de estudios, siendo los algoritmos Naive Bayes Tree y J48 los más confiables para la predicción. Patrones Socioeconómicos/Individuales: Edad, ingresos familiares (para alumnos más jóvenes), y el nivel de inglés (para alumnos mayores). A pesar de que estos estudios identifican las causas y proponen acciones (principalmente tutorías), ninguno ha documentado los resultados obtenidos de la implementación de dichas acciones para confirmar si realmente lograron disminuir el abandono escolar. Además, las investigaciones se han centrado en variables del estudiante. Por ello, los autores del artículo están investigando la posible influencia de las características y el desempeño de los docentes en la deserción estudiantil.

Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. (Pérez-Gutiérrez, 2020)

En este artículo se abordó la predicción de deserción en 762 estudiantes del programa de Ingeniería de Sistemas en una universidad colombiana, analizando registros académicos de 7 años (2004-2010). La metodología se basó en el proceso CRISP-DM (Cross Industry Standard Process for Data Mining). Para la tarea de clasificación binaria (desertor/no desertor) se compararon cuatro modelos de minería de datos: Árboles de Decisión, Regresión Logística, Naive Bayes y Bosques Aleatorios (Random Forest), además de contrastar los hallazgos con la herramienta Watson Analytics de IBM. En la etapa de preparación, se generó un dataset de 31 columnas enfocado en variables dinámicas del rendimiento académico, como promedios de notas por campo académico (Ingeniería de Sistemas - SE, Matemáticas - MAT, etc.), el promedio académico acumulado (GPA) y, crucialmente, la desviación

estándar del promedio acumulado del semestre académico para medir la irregularidad del rendimiento. Los resultados se evaluaron mediante el Área Bajo la Curva (AUC) ROC , concluyendo que el modelo de Bosques Aleatorios fue el más efectivo, ya que alcanzó un AUC de 0,97 en el último semestre y de 0,91 en el tercer semestre. Se demostró que el uso de algoritmos simples es suficiente para alcanzar niveles ideales de precisión y que la capacidad predictiva mejora conforme el estudiante avanza en su carrera. Los factores de deserción más influyentes fueron la desviación estándar de las notas (Std GPA), que presentó el mayor impacto positivo según el modelo de Regresión Logística, junto con el fallo en cursos de Matemáticas (Falló MAT) y Administración (Falló MGMT). Los modelos de Árbol de Decisión y Watson Analytics identificaron que un GPA bajo (menor a 3,57) y un bajo promedio en cursos de Ingeniería de Sistemas (Prom SE) son indicadores críticos para el abandono.

Uso de la analítica del aprendizaje de los estudiantes para minimizar la perdida escolar en las diferentes modalidades.

(T. De & -Ecuador, 2022)

En este artículo, el principal objetivo de estas investigaciones es aportar con el diseño de modelos de gestión de análisis de aprendizaje para identificar y reducir el riesgo académico y la repetición estudiantil. Se aplicó una metodología empírico-analítica de enfoque cuantitativo cuasiexperimental, con un fuerte énfasis en el corte longitudinal para analizar el progreso de los estudiantes a lo largo de múltiples periodos académicos. La metodología se basa en la recolección de grandes volúmenes de datos de sistemas de información institucionales y Entornos Virtuales de Aprendizaje . En su fase inicial, la investigación se apoya en una revisión sistemática de literatura (utilizando el flujo PRISMA) para determinar los criterios y enfoques más relevantes de la analítica del aprendizaje y Minería de Datos Educativa, con publicaciones clave entre 2019 y 20224. Las técnicas de minería de datos comúnmente aplicadas para la predicción incluyen algoritmos de clasificación como el árbol de decisión CHAID. La Analítica del Aprendizaje facilita la abstracción de tendencias y el análisis comparativo del rendimiento académico. En los resultados y factores determinantes:

Se identificó en el resultado el desarrollo de un modelo aplicable que permite a las instituciones identificar áreas de riesgo y mejorar las estrategias de intervención. Los modelos confirman que la analítica del aprendizaje, es una herramienta esencial que permite la predicción y el monitoreo preciso del rendimiento académico, generando retroalimentación oportuna a docentes y estudiantes. Los factores que mejor explican y predicen la deserción universitaria, identificados, son principalmente las razones socioeconómicas del estudiante, el puntaje de ingreso a la universidad, y el riesgo de repetición o pérdida académica.

Gestión del Proceso Educativo utilizando Analíticas del Aprendizaje y Big Data en la Universidad. (Villarroel-Henríquez & Gallardo-Aguayo, 2025)

En este artículo la gestión del proceso educativo mediante Analíticas del Aprendizaje y Big Data es esencial para el desarrollo de una cultura de aprendizaje institucional en la Educación Superior. La investigación empleó un enfoque mixto de tipo anidación, siendo el componente cuantitativo el principal (diseño no experimental, transversal y correlacional), complementado por un componente cualitativo para profundizar en los hallazgos. La muestra consistió en 30 directivos de universidades chilenas (50% privadas y 50% estatales) , y se utilizó un cuestionario online validado de 16 preguntas (cerradas en escala Likert y abiertas). El análisis se realizó mediante estadística descriptiva (SPSS) y análisis de contenido cualitativo (Nvivo). Los resultados principales indican que gran parte de los directivos están familiarizados con la recolección (90.1%) y el uso (76.6%) de datos para la mejora educativa. La información que más se recopila es la del perfil de ingreso del estudiante (sociodemográficos, historial académico) y el rendimiento académico/progresión (tasas de aprobación, reprobación, repitencia) para implementar medidas correctivas. El principal hallazgo que denota un desafío es la menor frecuencia en el uso de datos para predecir el desempeño futuro del egresado en el mundo laboral (Media: 6.09; 56.7% de respuesta positiva) , lo que señala una oportunidad de mejora en el levantamiento de información desde el mundo del trabajo.

Las universidades están recolectando y usando datos para el aseguramiento de la calidad y el seguimiento de la trayectoria, se requiere de una mayor sistematización y recursos para avanzar hacia la toma de decisiones y la predicción del desempeño laboral post-egreso.

3.3 Patrones y Problemas Abordados en la Minería de Datos Educativa

Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa. (Urbina-Nájera et al., 2021)

En este artículo empleó un enfoque cuantitativo con un diseño no experimental, transeccional descriptivo , utilizando técnicas de Minería de Datos Educativa donde se analizaron 10,635 registros de estudiantes recolectados entre otoño de 2014 y primavera de 2019. El conjunto de datos estaba desequilibrado (3,233 desertores y 7,402 no desertores). Se analizaron 12 atributos académicos y demográficos, incluyendo Período, Programa, Porcentaje de asistencia y Porcentaje de materias reprobadas. Se aplicó validación cruzada (80% entrenamiento / 20% prueba). Y se probaron varios clasificadores, destacando los Árboles de Decisión, Bosques Aleatorios, Máquina de Soporte Vectorial y Bayes Ingenuo. La evaluación se centró en la Exactitud y, crucialmente, en la Tasa de Error Balanceado para compensar el desequilibrio de las clases. Se utiliza la Selección de Atributos (método Relief) para optimizar el modelo. El estudio identificó patrones de deserción con una alta precisión, focalizando el riesgo en las etapas iniciales de la carrera. Como mejor clasificador a los algoritmos de Árbol de Decisión (REPTree y C4.5) resultaron ser los más efectivos, el modelo optimizado con REPTree alcanzó una exactitud del 94.2145% y una baja TEB (12.85%), superando los mejores reportes en la literatura previa. Los atributos que identifican a un potencial hijo desertor fueron (Porcentaje de materias reprobadas, porcentaje de asistencia, último semestre cursado, créditos cursados, periodo y programa.) El análisis de las reglas de decisión de los árboles determinantes que el fenómeno de la deserción está concentrado en los primeros cinco semestres de estudio.

3.4 Técnicas y Metodologías de Detección de Patrones

Análisis de la predicción temprana de los factores determinantes en la deserción de la educación superior desde el enfoque de Machine Learning – Un análisis desde el contexto colombiano.(Herrera Flores, 2025)

Este artículo analizó el fenómeno de la deserción universitaria entre 2014 y 2025 en Medellín, con el objetivo de consolidar el uso de la ciencia de datos como estrategia continua para la predicción y prevención del abandono académico. Se basó en una revisión crítica de estudios y datos oficiales en el sistema de información especializado para el análisis de la permanencia en la educación superior colombiana a partir del seguimiento a la deserción estudiantil (SPADIES). Se identificó y aplicó un enfoque metodológico sólido, destacando: como Enfoque de Análisis: Minería de Datos Educativa, siguiendo la metodología del Proceso Estándar Intersectorial para Minería de Datos. CRISP-DM, Modelos Predictivos: Las técnicas más utilizadas incluyen la Regresión Logística, Árboles de Decisión y Redes Neuronales Artificiales (Machine Learning) para generar alertas tempranas, en las fuentes y herramientas se emplearon datos de SPADIES y sistemas internos (UNAD, UdeA), con lenguajes de programación como Python y R para el modelado y Power BI/Tableau para la visualización. Encontrando factores determinantes, los principales indicadores de riesgo son el promedio académico, el número de créditos aprobados y el estrato socioeconómico. El mayor riesgo de deserción se concentra en los primeros tres semestres del programa, Se utilizó un modelo de Regresión Logística implementado demostró ser altamente efectivo su nivel de acierto general superior al 70%. Un análisis específico con umbral ajustado alcanzó una exactitud del 94.4% y una sensibilidad (recall) del 97.1% en la detección de desertores, confirmando su alta capacidad para identificar a tiempo a los estudiantes en riesgo. En cuanto al desafíos persiste una brecha entre la teoría y la práctica, con retos en la interoperabilidad de plataformas y la necesidad de capacitación técnica.

3.5 Fuentes de Datos y Herramientas

Modelos de inteligencia artificial en minería de datos educativa para predecir la deserción en Educación Superior: una revisión integral.(Pérez Niño et al., 2024)

Este artículo utilizó una metodología de revisión integral de literatura científica, de tipo documental y aplicada, enfocada en analizar artículos publicados entre 2015 y 2024 que implementan modelos de Inteligencia Artificial y Minería de Datos Educativos para predecir la deserción estudiantil en la educación superior. La revisión se centró en comparar los modelos de IA basándose en su rendimiento (precisión y exactitud) y frecuencia de uso en los estudios consultados, además de identificar las variables predictivas más recurrentes. En cuanto a los resultados comparados con la tesis de minería de datos, se encontró que los modelos de Redes Neuronales Artificiales obtuvieron el mejor rendimiento general con una precisión promedio del 95.92% y una exactitud del 92.73%. Los modelos de Bosques Aleatorios también destacaron por su alto rendimiento y estabilidad, siendo los más equilibrados con una precisión promedio del 94.01% y una exactitud del 91.86%. A pesar de ser el algoritmo más frecuentemente citado en la literatura revisada (14 menciones), el Árbol de Decisión presentó un desempeño inferior, con una precisión promedio del 88.46%. Finalmente, las Redes Neuronales Profundas y las Redes de Memoria a Corto Plazo también demostraron ser prometedoras, con precisiones promedio superiores al 91%. Las variables identificadas como más significativas en los modelos predictivos de patrones de deserción fueron las de tipo académicas, demográficas y socioeconómicas.

3.6 Casos de Estudio Relevante

Prototipo de minería de datos en la detección oportuna de estudiantes en riesgo de deserción escolar orientación escolar gestión universitaria integral del abandono (David et al., 2018).

Este artículo buscó predecir de manera temprana a los estudiantes en riesgo para implementar apoyos e intervenciones que reduzcan o impidan el abandono. La metodología se centró en el desarrollo de un prototipo web que simula y aplica la encuesta del proyecto Alfa-GUÍA (Gestión Universitaria Integral del Abandono). Esta encuesta consta de 34 preguntas que miden 82 variables, clasificadas en cinco factores de deserción: Individual, Académico, Sociocultural, Económico e Institucional. Los datos recopilados se almacenan en una base de datos MySQL y luego se procesan mediante un análisis estadístico de Minería de Datos para identificar patrones y describir el perfil de abandono. El estudio experimental se realizó en el Tecnológico Nacional de México en Roque en la carrera de Ingeniería en Tecnologías de la Información y Comunicación (estudiantes de primer año, los más propensos al abandono). El éxito y los resultados obtenidos con las metodologías y algoritmos propuestos lograron describir el perfil de abandono en la institución. El análisis estadístico arrojó un porcentaje de probabilidad de abandono del 24% en la muestra estudiada. Al examinar los factores, se encontró que el factor individual es el más destacado con un 29.7% de probabilidad de abandono, seguido por el socioeconómico (26.72%), el académico (23.99%) y el institucional (20.69%). Se identificaron variables específicas en la muestra que describen el perfil de abandono: aumento de la probabilidad con la edad (intervalos de 19 a 25 años), vivir con cónyuge o pareja, no disponer de recursos económicos o beca, no elegir la carrera por vocación, no participar en grupos, y ser padre o madre.

Los principales factores que favorecen la deserción identificados fueron el Académico, el Socioeconómico y el Individual.

Esta tesis busca abordar esta brecha, planteando una solución innovadora que constituye la principal contribución de este trabajo. En el Capítulo 4, dentro de la Metodología de la solución, se detallará la propuesta desarrollada para llenar este vacío, la cual se fundamenta en los hallazgos de la revisión de literatura.

Capítulo 4

Metodología de solución

En el presente capítulo se detalla la metodología de esta ***tesis*** adoptada, la cual se fundamenta en un proceso sistemático y riguroso. Para asegurar la validez y fiabilidad de los resultados, se ha estructurado el enfoque en cuatro fases secuenciales: Recolección de Datos, Preprocesamiento de Datos, Modelado y Validación, e Interpretación de Resultados. Se describirá cada una de estas etapas para comprender el procedimiento seguido en la investigación.

El éxito y la fiabilidad de cualquier iniciativa de análisis o proyecto dependen directamente de la rigurosidad metodológica empleada especialmente cuando se basa en el uso de patrones. Por ello, la Figura 4-1 detalla exhaustivamente las Fases de la Metodología de Solución que se aplicarán en la presente **tesis** Esta estructura metodológica no es solo una secuencia de pasos, sino un marco de trabajo sistemático y modular que define los procedimientos específicos y los criterios de validación necesarios en cada etapa. La adhesión estricta a este enfoque basado en patrones es crucial, pues garantiza la coherencia del proceso, la trazabilidad de las decisiones y la máxima fiabilidad de los resultados obtenidos, facilitando así la toma de decisiones estratégicas informadas.

4.1 Recolección de datos

Para esta fase del estudio, se elaboró un instrumento de recolección de datos conocido como cuestionario, cuyo objetivo es capturar información detallada sobre diversos patrones que influyen en el fenómeno de estudio. El cuestionario se estructuró siguiendo patrones definidos, lo que permite un análisis sistemático y la identificación de patrones de respuesta claros en los participantes. Este enfoque estructurado se organizó en cuatro patrones de información principales: patrones académicos, patrones socioeconómicos, patrones psicológicos y motivacionales, y patrones institucionales, compuesto por 31 preguntas, fue sometido a un riguroso proceso de validación por expertos para asegurar su pertinencia y precisión. Para confirmar la calidad y consistencia del cuestionario, y garantizar que los patrones de información capturados sean fiables, se aplicaron dos métricas psicométricas clave:

Coeficiente V de Aiken: Se utilizó para validar el contenido del instrumento, asegurando que cada pregunta sea relevante y que mida adecuadamente el constructo deseado.

Alfa de Cronbach: Esta métrica fue empleada para evaluar la consistencia interna, garantizando que las preguntas del cuestionario estén interrelacionadas y midan de manera fiable un mismo concepto. Este proceso metodológico garantiza que los datos recolectados son válidos y fiables, proporcionando una base sólida para el análisis posterior.(Escurra M, 2020) (del Carmen del Amo Chicharro et al., 2025).

4.2 Preprocesamiento de datos

En esta fase, se aplica una técnica de preprocesamiento de datos denominada OneHotEncoder que se utiliza en aprendizaje automático para convertir variables categóricas a una representación numérica binaria, esto para que los algoritmos de aprendizaje automático lo puedan entender.(Hernández G, 2019)

Después de validar el cuestionario con expertos y de aplicarlo a los estudiantes, se recopiló la información de la muestra y se procedió a codificar las variables categóricas a un formato binario mediante la técnica OneHotEncoding. Este proceso resultó en un dataset final compuesto por 430 instancias y 148 atributos.

4.3 Selección de modelos y validación

Tras procesar la información y convertirla a una representación numérica apta para los algoritmos, se utilizaron técnicas de aprendizaje no supervisado para explorar los datos. Específicamente, se aplicaron algoritmos de reducción de dimensiones y buscar simplificar la complejidad de los datos, identificando los patrones de variabilidad más importantes y de agrupamiento de objetos (clustering) revelando patrones de comportamiento o segmentos naturales que serán cruciales para el modelado posterior. El objetivo de este paso fue descubrir los patrones y estructuras ocultas dentro del conjunto de datos.

Una vez recopilada y procesada la información, se procede a su división en dos conjuntos. El 80% de los datos se destina al entrenamiento de los modelos de aprendizaje supervisado. Estos modelos usarán este conjunto para aprender los patrones de relación entre las variables predictoras y la variable objetivo. mientras que el 20% restante se reserva para la evaluación y prueba de su rendimiento. Este conjunto de prueba asegura que el modelo no solo memorice los patrones del conjunto de entrenamiento, sino que sea capaz de generalizarlos a datos no vistos.

Para el análisis, se han seleccionado los siguientes modelos de aprendizaje supervisado, ampliamente reconocidos y utilizados en el ámbito del análisis de datos académicos según la literatura revisada: Perceptrón Multicapa, Máquinas de Soporte Vectorial, Regresión Logística y Random Forest. La elección de estos algoritmos se basa en su probada eficacia para abordar problemáticas similares, lo que asegura una metodología robusta y alineada con las prácticas estándar de esta ***tesis***.

Para validar la efectividad de los modelos de aprendizaje supervisado, y confirmar su capacidad para reconocer y clasificar correctamente los patrones de datos inherentes al fenómeno de estudio, se empleó una metodología de evaluación robusta.

Esta matriz de confusión es una herramienta visual clave que permite:

1. Cuantificar la Calidad del Patrón de Predicción: Permite comparar las predicciones del modelo (los patrones de salida que el modelo infiere) con los resultados reales (los patrones de verdad presentes en el conjunto de prueba).
2. Identificar Errores de Clasificación de Patrones: Muestra de forma clara la proporción de aciertos y errores, revelando los patrones de error donde el modelo se equivoca (por ejemplo, casos mal clasificados como Falsos Positivos o Falsos Negativos).

El análisis de esta matriz es fundamental para calcular métricas de rendimiento detalladas, asegurando que los patrones aprendidos por el modelo sean sólidos y generalizables.

A partir de la matriz de confusión, se calcularon las siguientes métricas para obtener una evaluación cuantitativa del rendimiento:

• Exactitud (*Accuracy)*: Mide la proporción de predicciones correctas sobre el total de predicciones.

• *Precisión*: Evalúa la calidad de las predicciones positivas, indicando la proporción de resultados positivos verdaderos de entre todas las instancias que el modelo clasificó como positivas.

• Sensibilidad (*Recall*): Mide la capacidad del modelo para identificar todas las instancias relevantes, es decir, la proporción de positivos verdaderos que el modelo identificó correctamente.

• *F1-score*: Proporciona una media armónica de la precisión y la sensibilidad, ofreciendo una métrica única que equilibra ambos valores y brinda una visión integral del rendimiento del modelo.

Este conjunto de métricas garantiza una evaluación completa y confiable del desempeño de cada modelo en la tarea de clasificación.

4.4 Interpretación de resultados

Una vez que los modelos de aprendizaje automático han procesado la información y arrojado sus resultados, es fundamental realizar una interpretación experta de los mismos.

Sin embargo, aunque los modelos identifiquen patrones de predicción (como ¿qué variables se asocian con un resultado específico?) o patrones de agrupamiento (¿cómo se segmenta la población?): No son capaces de generar recomendaciones por sí mismos.

Es en este punto donde el experto en datos asume un papel crucial. Su función es analizar los hallazgos de los modelos, contextualizar los patrones y traducirlos en recomendaciones prácticas y accionables para el cliente.

Este proceso de interpretación va más allá de los números; implica comprender el significado de los datos en el contexto del problema, identificar patrones relevantes y prever las implicaciones de los resultados.

De esta manera, la fase final de la metodología se convierte en un puente entre el análisis técnico y la toma de decisiones estratégicas. Al combinar la capacidad predictiva de los modelos con el conocimiento y la experiencia del analista, es posible tomar decisiones informadas que generen un impacto positivo y significativo.

Este proceso de interpretación es vital para que el cliente pueda tomar decisiones informadas y efectivas. Con esto, se concluye este capítulo y se dará inicio al Capítulo 5, donde profundizaremos en la aplicación de estos conceptos a través de un caso de estudio práctico, aplicado a una Institución de Educación Superior (IES).

Capítulo 5

Caso de estudio

Con el objetivo primordial de fortalecer la vinculación entre los fundamentos teóricos y la aplicación práctica de la Minería de Datos Educativos, se ha seleccionado un caso de estudio real y pertinente. En este sentido, esta ***tesis*** se centrará como caso de estudio al Instituto Tecnológico de Iztapalapa III, una Institución de Educación Superior (IES).

Crear un cuestionario como instrumento de medida requiere de rigor. Según Tapia (2010), un cuestionario es un conjunto de distintos ítems que pueden plantearse de diversas formas con varias alternativas, formato, orden de preguntas y contenido determinado teniendo en cuenta el tema que queremos investigar.

Un cuestionario es un instrumento tradicional que permite recoger gran número de datos y permite llegar a multitud de participantes de manera poco costosa. La comparabilidad de la información es su finalidad.(Escurra M, 2020)

5.1 Recolección de datos de información y selección de la muestra

En este apartado se describe la metodología empleada para la construcción del cuestionario y los criterios utilizados para determinar el tamaño de la muestra, elementos cruciales para la validez de la investigación.

5.1.1 Construcción del cuestionario

En esta actividad se construye un cuestionario que contiene 31 preguntas agrupadas en cuatro factores como son: patrones académicos, patrones socioeconómicos, patrones psicológicos y motivacionales y patrones institucionales. En la Figura 5-1, se muestran los pasos generales para llegar a la recolección de los datos.

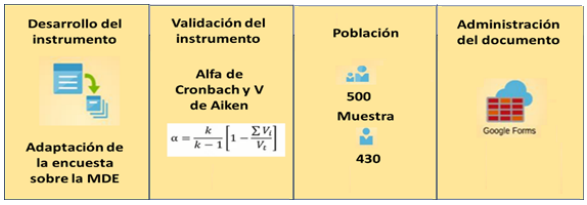


Figura 5-1. Pasos generales para llegar a la recolección de los datos.

El Tecnológico Nacional de México (TecNM) Campus Iztapalapa III se fundó el 14 de septiembre de 2009 como una medida urgente para satisfacer la demanda de Educación Superior Tecnológica en la Ciudad de México, particularmente en zonas marginadas.

La institución, que inició su actividad con solo dos programas, ha mostrado un patrón de crecimiento constante y estratégico y actualmente ofrece las ingenierías: Civil, Informática y Gestión Empresarial. Su filosofía institucional se centra en brindar atención de calidad a toda su matrícula, asegurando un patrón de servicio centrado en el estudiante. El compromiso esencial del Instituto es formar profesionales de nivel licenciatura que sean capaces de desarrollar ciencia y tecnología pertinente, contribuyendo al desarrollo sustentable del país y alineándose con las necesidades sociales.

Este compromiso define el patrón de egreso deseado: profesionales que no solo posean conocimientos técnicos, sino que también sean agentes de cambio con un claro patrón de impacto social y tecnológico. De esta manera, la oferta educativa se adapta y responde a los patrones de demanda del mercado laboral y del desarrollo nacional.

Uno de los valores es el ser humano, es considerado en la institución como el elemento en el que se enfoca todo nuestro objetivo de vida. Por lo tanto, se desea reducir la deserción estudiantil y para ello se utilizó una herramienta para recabar la información, se llama Teams de Microsoft. Por medio de esta herramienta se obtiene la recopilación de los comentarios de los alumnos de Iztapalapa III desde 2do (105 alumnos), 3ro (12 alumnos), 4to (177 alumnos), 5to(37alumnos), 8°(50 alumnos), 9º (16 alumnos), egresado / egresada (15 alumnos) y sin egresar (18 alumnos), de las 3 Ingenierías como son Informática (167 alumnos), Gestión Empresarial (93 alumnos) y Civil (170 alumnos), con la finalidad de obtener retroalimentaciones sobre proyectos, actividades, formas de evaluaciones hacia los alumnos, sondeos económicos, opiniones, eventos o decisiones de los alumnos que nos facilitan la toma de decisiones rápidas, previniendo con ello la deserción estudiantil y la mejora de enseñanza en los docentes.

Sin embargo, al llenar el formulario, se les informa a los alumnos sobre el tratamiento de sus datos personales, el cual se realizará en estricto apego a los principios establecidos en la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados (LGPDPPSO). Esta Ley menciona que se consideran datos personales cualquier información relativa a una persona física identificada o identificable. Una persona es identificable cuando su identidad puede determinarse, ya sea directamente (por su nombre, por ejemplo) o indirectamente (mediante otros datos que, en conjunto, permitan identificarla).

Es importante señalar que, en concordancia con la *Fracción I del artículo 113 de la Ley Federal de Transparencia y Acceso a la Información Pública*, la información que contenga datos personales de una persona física identificada o identificable se considera confidencial. Esto incluye, entre otros ejemplos, el domicilio particular, número de teléfono personal y la Clave Única de Registro de Población (CURP) de los alumnos.

5.1.2 Validación del cuestionario

El coeficiente V de Aiken es una medida utilizada para cuantificar la validez de contenido de un conjunto de ítems (como preguntas de un cuestionario o elementos de una prueba) basada en la evaluación de jueces expertos y su propósito es confirmar que el instrumento sigue el patrón de contenido teórico y operacional establecido. Específicamente, el coeficiente V de Aiken ayuda a:

a. Asegurar la Coherencia del Patrón: Garantiza que cada ítem del instrumento (cada pregunta) sea pertinente, claro y representativo del constructo o del patrón de información que se busca medir.
b. Cuantificar el Consenso del Patrón: Permite cuantificar el grado de acuerdo o consenso entre los jueces expertos sobre si el conjunto de ítems sigue el patrón de validez esperado para el dominio de estudio.

El objetivo fundamental de incorporar el coeficiente V de Aiken en el proceso de diseño de un cuestionario dirigido a la población estudiantil del Tecnológico de Iztapalapa III es garantizar que los reactivos sigan el patrón de validez de contenido requerido.

Esta metodología permitirá evaluar de manera rigurosa la pertinencia, Si el ítem se ajusta al patrón temático que debe medir, claridad si la formulación del ítem es comprensible y no genera patrones de ambigüedad en las respuestas. y representatividad si el conjunto de ítems en su totalidad representa fielmente el patrón de los constructos teóricos o variables de interés. De este modo, se busca optimizar la calidad del instrumento, asegurando la recolección de datos altamente confiables y válidos, esenciales para la consecución de los objetivos de la investigación y la toma de decisiones informadas por los patrones de comportamiento identificados en la población estudiantil.

En esencia, la V de Aiken busca determinar el grado en que un grupo de expertos está de acuerdo en que los ítems de un instrumento son relevantes y representativos del constructo o dominio que se pretende medir. Este se evalúa antes de aplicar el cuestionario a los estudiantes y su propósito es validar el contenido de las preguntas que conforman el cuestionario.

Al utilizar la V de Aiken, se consideraron los siguientes puntos:

- Medir la validez de contenido y el patrón teórico: A diferencia del coeficiente de Cronbach que mide la consistencia interna o el patrón de correlación entre ítems, el coeficiente de V de Aiken se enfoca en si el contenido de los ítems es adecuado para medir patrón de constructo teórico que se requiere, según el juicio de expertos. Es decir, valida si los ítems siguen el patrón de contenido correcto. Participación de 8 jueces expertos en la evaluación de patrones: Se solicita a un panel de jueces expertos que evalúen cada ítem en una escala ordinaria con respecto a su relevancia, claridad o coherencia. Los jueces analizan si el ítem corresponde al patrón de criterios establecido. La escala utilizada es discreta categórica, los jueces analizan si el ítem corresponde al patrón de criterios establecido. (como No cumple con el criterio, Poco relevante, Neutral, Preciso y Calidad relevante).

Esta retroalimentación se utiliza para revisar y mejorar la elaboración del cuestionario por medio de las observaciones escritas por los expertos.

De la evaluación que realizaron los expertos aplicando la fórmula de la Ecuación 1, el resultado es el siguiente:

$$V = \frac{\sum(x_i - l)}{N(k - l)} \quad (1)$$

Donde

x_i = *Puntuación otorgada por el experto i.*

l = Menor puntuación posible en la escala.

k= Mayor puntuación posible en la escala.

N=Número de expertos que calificaron el ítem.

5.1.3. Selección de la muestra

Se construyó y validó un cuestionario con el apoyo de un grupo de expertos en psicología, garantizando que el instrumento siguiera los patrones de contenido y fiabilidad requeridos.

Como se muestra en la Figura 5-2, se obtuvieron las respuestas de 430 estudiantes provenientes de diversas ingenierías y semestres. Estos estudiantes mostraron interés y disposición para responder a un listado de 31 preguntas, conformando el patrón de muestra final para el análisis.

Esta distribución es crucial, ya que permite identificar posibles patrones de comportamiento o patrones de factores que puedan variar significativamente a lo largo de la trayectoria académica del estudiante.

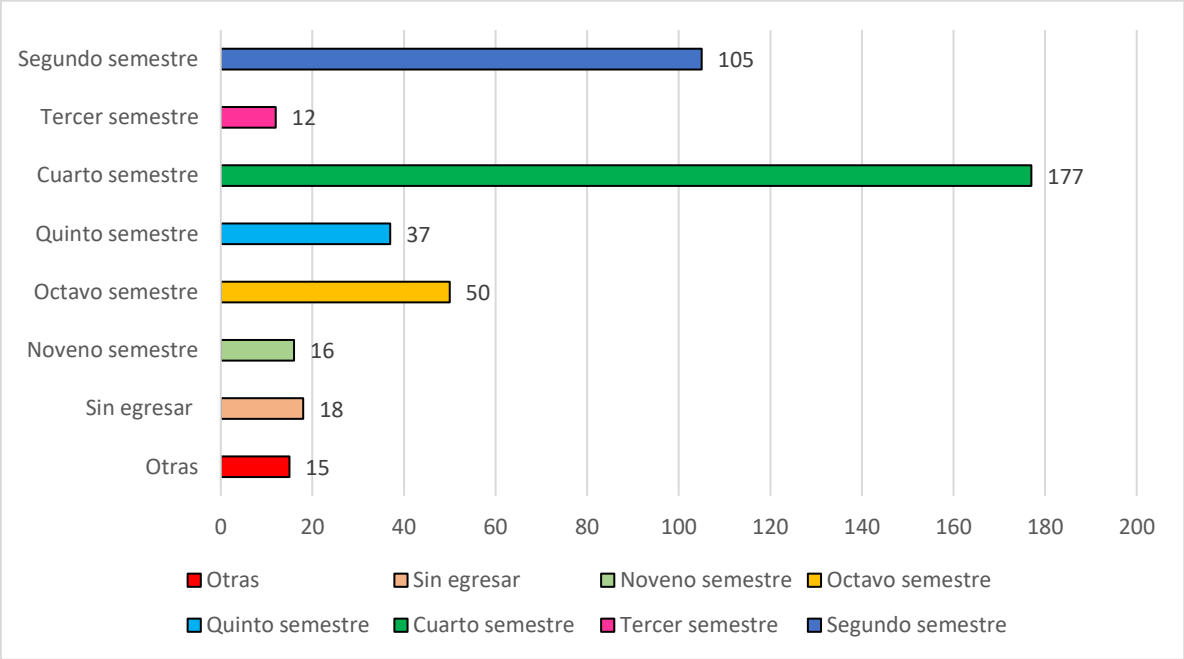


Figura 5-2. Respuestas de 430 estudiantes provenientes de diversas ingemierías y semestres.

La Figura 5-2 presenta la consolidación de los datos recopilados, conformando el conjunto de datos (dataset). Este integra las respuestas de 430 estudiantes provenientes de diversas ingenierías y semestres, quienes completaron la totalidad del instrumento de evaluación.

- **Segundo semestre**: 105 estudiantes
- **Tercer semestre**: 12 estudiantes
- **Cuarto semestre**: 177 estudiantes
- **Quinto semestre**: 37 estudiantes
- **Octavo semestre**: 50 estudiantes
- **Noveno semestre**: 16 estudiantes
- **No egresados (as):** 18 estudiantes
- **Egresados y titulados**: 15 estudiantes

Para validar las respuestas de los estudiantes, se aplicó el coeficiente de Cronbach (α) a una prueba piloto, ya que es una medida estadística utilizada para evaluar la consistencia interna

o confiabilidad de un instrumento de medición, como un cuestionario. Este coeficiente determina el grado en que un conjunto de ítems (preguntas) se correlacionan entre sí, midiendo de manera coherente un único constructo teórico o característica subyacente. Un valor de alfa de Cronbach cercano a 1 indica una alta confiabilidad, lo que sugiere que las preguntas del cuestionario miden de forma consistente el mismo concepto.

Su objetivo principal es validar la consistencia interna y fiabilidad del conjunto de preguntas del cuestionario. La métrica del alfa de Cronbach indica el mejor valor de alfa cuando $\propto$>0.80 lo que expresa una "muy buena consistencia", por otro lado, cuando $\propto$>0.70 indica que es "buena consistencia", a continuación, se muestra los siguientes valores:

$\alpha \geq 0.90$: Excelente consistencia interna.
$0.80 \leq \alpha < 0.90$: Muy buena consistencia interna.
$0.70 \leq \alpha < 0.80$: Buena consistencia interna.
$0.60 \leq \alpha < 0.70$: Cuestionable consistencia interna.
$0.50 \leq \alpha < 0.60$: Pobre consistencia interna.
$\alpha < 0.50$: Inaceptable consistencia interna.

El Alfa de Cronbach se calcula utilizando la Ecuación 2:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum v_i}{v_t}\right] \quad (2)$$

Donde:

α: Alfa de Cronbach
k: Número de ítems
v_i: Varianza de cada ítem
v_t: Varianza del total

A continuación, se muestran las preguntas donde se aplicó el Alfa de Cronbach en una prueba piloto de 6 preguntas. Esta prueba se aplicó a 13 estudiantes y consta de 6 ítems.

Tabla 5-1.Preguntas de la prueba piloto

ALUMNOS	Q1 ¿Cómo calificarías tu nivel de comprensión de los contenidos en tu carrera?	Q2 ¿Cuál es el grado de satisfacción con la enseñanza en la universidad?	Q3 ¿Cuál es tu nivel de satisfacción con los métodos de enseñanza?	Q4 ¿Qué tan presionado te sientes por la carga académica?	Q5 ¿En qué medida te preocupa financiar tus estudios en el futuro?	Q6 ¿Qué tan motivado te sientes para terminar tu carrera?	TOTAL
1	1	5	4	3	2	4	19
2	4	3	3	4	3	4	21
3	3	4	4	4	4	5	24
4	2	3	3	3	3	5	19
5	4	3	5	3	4	5	24
6	4	5	4	3	2	5	23
7	4	3	3	3	3	3	19
8	3	3	3	2	2	5	18
9	5	4	4	5	4	4	26
10	1	3	3	2	1	1	11
11	3	4	4	2	4	3	20
12	1	3	3	2	2	4	15
13	2	5	4	4	5	5	25
VARIANZA	1.67	0.67	0.39	0.84	1.23	1.3	16.5207

La varianza de los ítems individuales, denotada como v_i, es de 6.1065. La varianza total de la prueba, calculada a partir de la suma de los puntajes de los ítems, es de 16.5207.

Sustituyendo los valores en la Ecuación 2, se obtiene como resultado que el alfa de Cronbach es igual a= 0.756447

$$\alpha = \frac{6}{6-1}\left[1 - \frac{6.1065}{16.5207}\right] = 0.756447 \quad (2)$$

En la Tabla 5-1, preguntas de la prueba piloto, se observa que, para el conjunto de los 6 Ítems, la varianza muestra el 0.75, esto quiere decir que las preguntas están de acuerdo a la siguiente pauta: $0.70 \leq \alpha < 0.80$: Aceptable consistencia interna, las preguntas son aceptables para formar parte del cuestionario.

Tabla 5-2.Cálculo del coeficiente de Alfa de Cronbach en prueba piloto.

No.	Q1	Q2	Q3	Q4	Q5	Q6	TOTAL
1	1	5	4	3	2	4	19
2	4	3	3	4	3	4	21
3	3	4	4	4	4	5	24
4	2	3	3	3	3	5	19
5	4	3	5	3	4	5	24
6	4	5	4	3	2	5	23
7	4	3	3	3	3	3	19
8	3	3	3	2	2	5	18
9	5	4	4	5	4	4	26
10	1	3	3	2	1	1	11
11	3	4	4	2	4	3	20
12	1	3	3	2	2	4	15
13	2	5	4	4	5	5	25
Varianza	1.67	0.67	0.39	0.84	1.23	1.3	16.5207

En la Tabla 5-2, podemos observar que el ítem 4 al aplicar la varianza de forma individual, el Análisis de la consistencia de la fiabilidad, nos indica que la varianza muestra el 0.84, esto quiere decir que la pregunta está de acuerdo a la siguiente pauta: $0.80 \leq \alpha < 0.90$: Buena consistencia interna. Se ha desarrollado un diccionario de datos para estandarizar las respuestas en la aplicación de un cuestionario existente sobre Minería de Datos Educativa. Este recurso optimiza la legibilidad y reduce la redundancia, lo que garantiza una mayor coherencia en la información recopilada.

En la Figura A-1. Diseño del instrumento de medición o diccionario (Data set), muestra el uso de variables o atributos abreviados mismo que optimiza el manejo de preguntas extensas. Es crucial que estas abreviaturas sean claras, consistentes y estén correctamente documentadas para garantizar la integridad y la correcta interpretación de los datos.

En la creación del diccionario de datos, cada pregunta se define como una variable. Esta variable especifica tanto el tipo de escala como las opciones de respuesta asociadas, información que se localiza dentro de los Anexos como la Figura A-1. Diseño del instrumento

de medición. Una vez finalizado el diccionario, el archivo se guarda en un sistema de almacenamiento en la nube. Este documento, que contiene los atributos o variables que se utilizarán en la base de datos, queda listo para ser importado en Google Colab para su posterior programación en Python.

5.2 Preprocesamiento de la información

En esta ***tesis*** se utilizó la técnica de preprocesamiento de datos OneHotEncoder. Este método se emplea para convertir variables categóricas en una representación numérica binaria, lo cual facilita su procesamiento por parte de los algoritmos de aprendizaje automático. El preprocesamiento de datos es una fase fundamental para preparar los conjuntos de datos antes de su análisis. Este proceso es necesario porque los datos iniciales a menudo presentan imperfecciones y ruido. En esencia, el preprocesamiento de datos se enfoca en mejorar la calidad de la información, asegurando que sea confiable y apta para un análisis preciso.

En la Tabla 5-2 se muestra un Ejemplo de Codificación OneHotEncoder. La forma de cómo se codificaron las respuestas de los estudiantes para su análisis.

La pregunta de ejemplo es la siguiente: ¿Cuál fue tu promedio en el último semestre?, sus posibles respuestas son: 5 o menos-reprobado (x1), 6 suficiente (x2), 7-aprobado (x3), 8-bien (x4), 9-notable (x5) y 10-sobresaliente (x6).

El proceso consiste en crear una variable dummy binarizada para cada categoría de respuesta. Cada nueva columna se completa con un valor de 0 o 1, donde el 1 indica la presencia de esa categoría en un registro específico y el 0 su ausencia. Este método es crucial para evitar que el modelo asuma una relación numérica o jerárquica entre las categorías que no existe realmente, lo cual podría sesgar los resultados del análisis.

En la Tabla 5-3 se ilustra el método de codificación OneHotEncoder aplicado a la pregunta de ejemplo. En este formato específico:

- Cada fila (identificada como X1, X2,X3,X4,X5,X6 representa a un alumno.
- Cada columna (identificada como x1,x2,x3,x4,x5,x6 representa una categoría o un grupo.

La identificación del alumno se basa en el número de la fila (X1 hasta X6).

X1 se refiere al Alumno 1.

X2 se refiere al Alumno 2.

...y así continuar hasta X6, que se refiere al Alumno 6.

Tabla 5-3.Ejemplo de Codificación OneHotEncoder.

Promedio	X1	X2	X3	X4	X5	X6
x1	1	0	0	0	0	0
x2	0	1	0	0	0	0
x3	0	0	1	0	0	0
x4	0	0	0	1	0	0
x5	0	0	0	0	1	0
x6	0	0	0	0	0	1

En la Figura 5-3, se presenta el código Python que se utilizó para codificar la pregunta de ejemplo.

El proceso de codificación de variables categóricas con la técnica OneHotEncoder se lleva a cabo en varias etapas. Etapa 1, se instancia la clase OneHotEncoder de la biblioteca Scikit-learn. Etapa 2, se aplica el método fit_transform () a la columna "promedio" del DataFrame de Pandas. Este método ajusta el codificador a los datos y simultáneamente los transforma en un vector binario.

Etapa 3, el vector binario resultante se convierte de nuevo a un DataFrame de Pandas y se le asignan nombres de columna descriptivos. Etapa 4, este nuevo DataFrame codificado (encoded_df) se concatena con el DataFrame original (df), lo que permite agregar los nombres de los índices y completar el proceso de preparación de los datos para su posterior análisis. El proceso se muestra en la siguiente Figura 5-3.

```
#CODIFICAMOS LOS ATRIBUTOS CATEGÓRICOS DE LA BASE DE DATOS
      Q5_promedio  = df_datos_academicos[[ Q5_promedio ]]

oh_encoder = OneHotEncoder()

df_datos_academicos_type_oh = oh_encoder.fit_transform(  Q5_promedio   )

df_datos_academicos_encoded = pd.DataFrame(df_datos_academicos_type_oh.toarray(), columns=["10: Sobresaliente",
                                                                                           "5 o Menos: Reprobado",
                                                                                           "6: Suficiente",
                                                                                           "7: Aprobado",
                                                                                           "8: Bien",
                                                                                           "9: Notable"])

#RESETEAMOS LOS INDICES PARA REALIZAR LA CONCATENACIÓN
df_datos_academicos_encoded.reset_index(inplace=True,drop=True)
df_datos_academicos_prep.reset_index(inplace=True,drop=True)
df_datos_academicos_prep= pd.concat([df_datos_academicos_prep,df_datos_academicos_encoded],axis=1)
df_datos_academicos_prep
```

	Q5_promedio	Q6_promedio_semestre	Q7_materias_no_aprobadas	Q8_satis_ense	Q9_satis_met_ense	Q10_comunicar_capaz_prof_compa	Q11_horas_estudio_fuera
0	8: Bien	8: Bien	Ninguna materia	Neutral	Satisfecho	Neutral	1-2 horas
1	8: Bien	8: Bien	Ninguna materia	Neutral	Neutral	Neutral	Menos de una hora
2	8: Bien	8: Bien	Ninguna materia	Satisfecho	Satisfecho	Capaz	3-4 horas
3	9: Notable	8: Bien	Ninguna materia	Satisfecho	Satisfecho	Capaz	1-2 horas
4	8: Bien	7:Aprobado	Ninguna materia	Satisfecho	Satisfecho	Totalmente capaz	Menos de una hora
...	...	...	...	...	...	...	...

Figura 5-3.Código Phyton de codificación ONEHOTENCODER de la variable Promedio.

5.3 Selección de modelos y validación

Tras el preprocesamiento de los datos, que incluyó la codificación de variables categóricas mediante One-Hot Encoding, el conjunto de datos educativos fue transformado a una matriz de 430 instancias por 148 atributos. Este incremento en la dimensionalidad se debe a que la técnica de One-Hot Encoding convierte cada categoría de una variable en un nuevo atributo binario, lo que expande la representación del conjunto de datos. El siguiente paso en el análisis consiste en aplicar modelos de aprendizaje no supervisado para identificar patrones y estructuras latentes en estos datos académicos de alta dimensión.

Dado el elevado número de atributos, una visualización efectiva de los datos resulta indispensable para la detección de tendencias y agrupaciones. En este contexto, se implementa la técnica de Análisis de Componentes Principales (PCA), un método de reducción de dimensionalidad que transforma el conjunto de datos original en un nuevo sistema de coordenadas. Las nuevas variables, denominadas componentes principales, capturan la varianza máxima de los datos, permitiendo una representación visual más comprensible y facilitando la identificación de estructuras latentes.

Su objetivo es reducir las dimensiones del espacio de observación en el que se estudian los objetos dados. Esta reducción se obtiene mediante la creación de nuevas combinaciones lineales de variables que caracterizan los objetos estudiados. Estas combinaciones, denominadas componentes principales, deben cumplir ciertas condiciones matemáticas y estadísticas (Maćkiewicz & Ratajczak, 1993).

Para esta **tesis**, se estableció un umbral del 95% de la varianza total como el criterio para determinar el número adecuado de componentes. Como se ilustra en la Figura 5-4, al analizar la curva de la varianza explicada, se observa que aproximadamente de 72 a 73 componentes principales son suficientes para capturar el 95% de la varianza acumulada.

Este resultado permite una reducción de la dimensionalidad considerable, de 148 a 73 atributos, lo que facilita el análisis posterior sin comprometer la integridad de la información original.

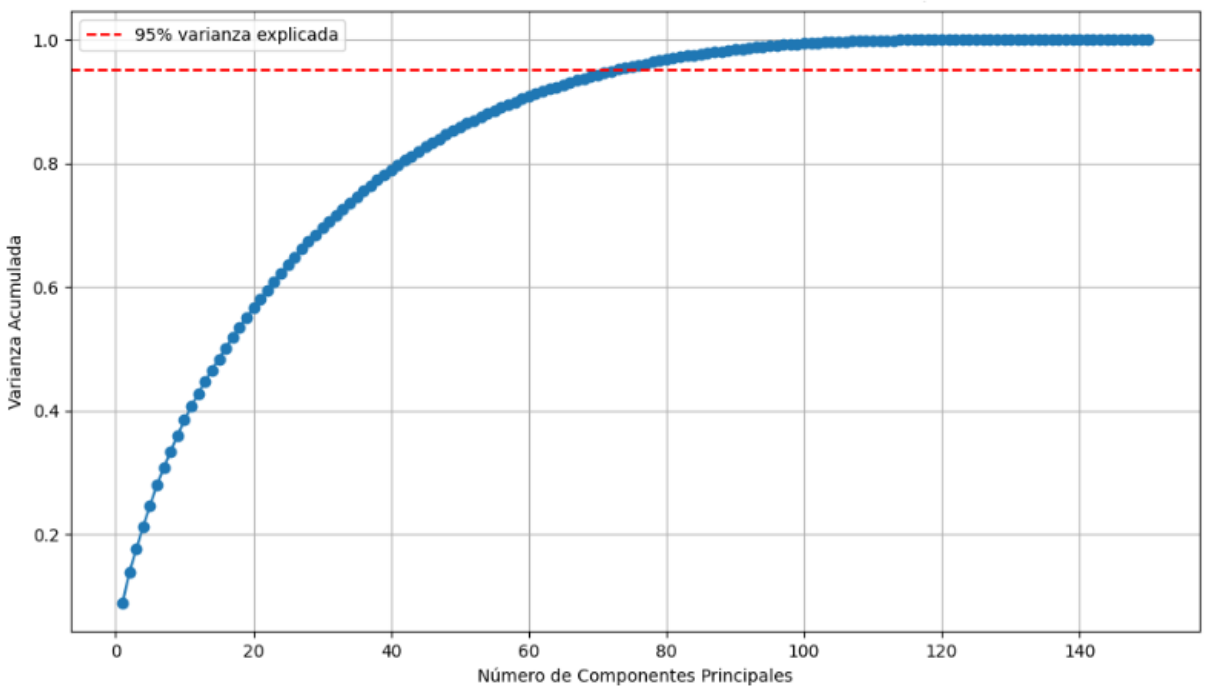


Figura 5-4. Curva de la varianza explicada acumulada.

El siguiente paso en el análisis consiste en aplicar el algoritmo de aprendizaje no supervisado K-Means para detectar patrones y estructuras latentes en los datos académicos. Este algoritmo es particularmente útil para agrupar observaciones similares en clústeres, basándose en la distancia entre sus características.

Un parámetro fundamental para la correcta ejecución del algoritmo K-Means es la especificación del número de grupos (K). La determinación del valor óptimo de K es crucial, ya que un número inadecuado de clústeres puede llevar a una interpretación errónea de los datos, para asegurar que el algoritmo K-means refleje con precisión la estructura subyacente de los datos.

Para ello, se aplicaron dos métodos complementarios: el método del codo (elbow method) y el coeficiente de silueta promedio.Como se muestra en la Figura 5-5, el análisis se apoyó en dos técnicas complementarias, para obtener la determinación del número óptimo de clusters (k).

1. **Método del Codo**: Este procedimiento consiste en graficar la inercia (la suma de las distancias internas en cada grupo) frente al número de clusters. A medida que aumenta k, la inercia disminuye, pero llega un punto donde la mejora se vuelve marginal, formando un "codo" en la curva. Ese punto indica el número óptimo de grupos. En este caso, la curva comienza a estabilizarse en k = 3 y k = 4, lo que sugiere que añadir más clusters no aporta beneficios significativos.
2. **Coeficiente de Silueta Promedio**: Esta métrica evalúa la calidad del agrupamiento midiendo qué tan bien se ajusta cada instancia a su cluster en comparación con otros. Valores cercanos a 1 indican una separación clara entre grupos. El análisis muestra que los valores más altos se encuentran también en el rango de k = 3 a k = 4, confirmando la estructura natural de los datos.

   Ambos criterios coinciden en que k = 3 y k = 4 son las opciones más adecuadas, por lo que se procederá a realizar el agrupamiento con estos valores para comparar resultados y seleccionar el más robusto.

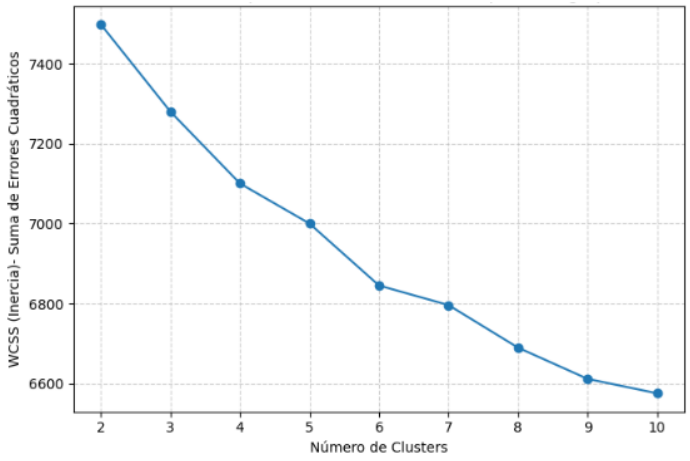


Figura 5-5. Método del Codo para determinar el número óptimo de grupos (k).

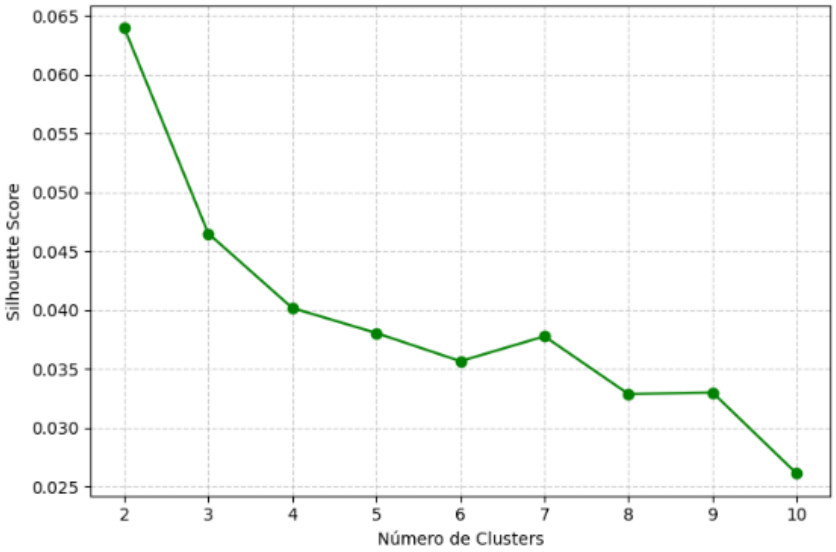


Figura 5-6. Coeficiente de Silueta Promedio para determinar el número óptimo de grupos (k).

Una vez determinado el valor óptimo de K, se procedió a la implementación del algoritmo K-means de Scikit-learn para agrupar los datos académicos. Para asegurar la reproducibilidad de los resultados, el algoritmo fue configurado con los siguientes parámetros: un número de clústeres de K=3, un estado aleatorio fijo (random_state=42) para garantizar la consistencia en cada ejecución, y 10 inicializaciones (n_init=10) para seleccionar la mejor partición posible y evitar mínimos locales.

Como se observa en la Figura 5-7. La visualización del agrupamiento se realizó utilizando solo los dos primeros componentes principales del Análisis de Componentes Principales (PCA), lo que permite una representación gráfica bidimensional de los resultados. Esta visualización reveló la existencia de tres agrupamientos principales bien definidos: el grupo 0 (azul), el grupo 1 (naranja) y el grupo 2 (verde).

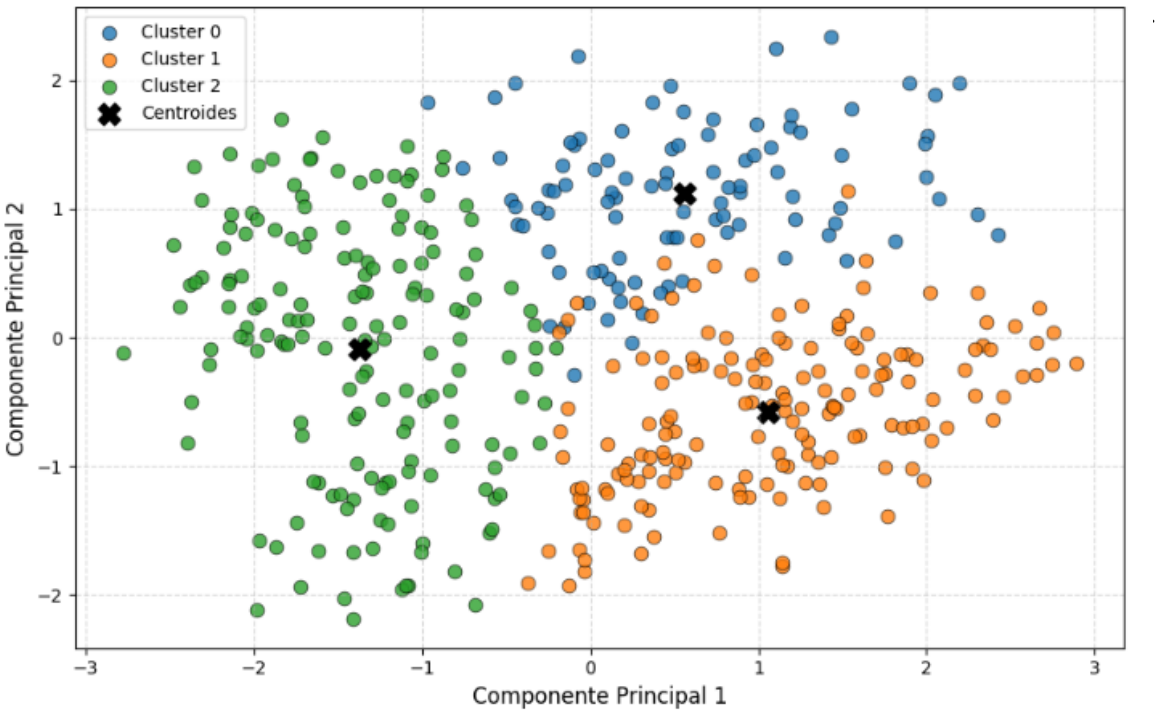


Figura 5-7. Visualización de los grupos con el algoritmo K-means después de aplicar PCA.

Para interpretar y dar significado a cada uno de los grupos identificados por el algoritmo K-means, se recurrió a dos enfoques complementarios: el cálculo de la moda y un análisis cualitativo de las variables dominantes dentro de cada clúster. La moda, al identificar las respuestas más frecuentes, permite caracterizar los perfiles típicos de los estudiantes en cada grupo.

La siguiente sección presenta una interpretación detallada de cada clúster, basada en el análisis de las variables que mostraron una mayor frecuencia (moda). Este enfoque nos permite asignar una etiqueta o un perfil descriptivo a cada grupo, lo que facilita la comprensión de los patrones de comportamiento y características subyacentes en la población estudiantil.

La observación del gráfico revela un clúster de color naranja que se encuentra visualmente inmerso dentro del área predominantemente azul (Clúster 0), a pesar de que su color lo identifica como parte del Clúster 1. Este fenómeno es una ocurrencia común y esperable en el agrupamiento K-Means, y su presencia puede explicarse de las siguientes maneras:

- Proximidad al Centroide: El algoritmo K-Means asigna cada punto de datos al clúster cuyo centroide está más cerca de ese punto en el espacio multidimensional (en este caso, el espacio de las componentes principales). Aunque visualmente el clúster naranja parezca estar rodeado por puntos azules, su distancia euclidiana (o métrica de distancia utilizada) al centroide del Cluster 1 (naranja) es menor que su distancia al centroide del Clúster 0 (azul).

- Naturaleza Iterativa de K-Means: ya que es un algoritmo iterativo que busca minimizar la suma de las distancias cuadradas entre los puntos y los centroides de sus clusters asignados. Durante el proceso de optimización, algunos puntos pueden quedar "atrapados" en un cluster al que son asignados en una iteración temprana y su reasignación a otro cluster podría no reducir significativamente la inercia total, o la configuración final de los centroides lo mantiene en su cluster actual.

- Lo que se observa como un "clúster anaranjado dentro del azul" no es literalmente un clúster incrustado en otro, sino una zona de solapamiento o intersección entre los clústeres. Específicamente, en la región central del gráfico, alrededor de los valores de PC1 cercanos a 0.5 y PC2 cercanos a 1.0.

- Forma y Densidad de los Clústeres: Los clústeres no son siempre perfectamente esféricos o totalmente separados. Pueden tener formas irregulares y sus límites pueden ser difusos, especialmente en áreas donde la densidad de puntos de datos es menor o donde las características que los definen son menos distintivas.

- K-Means es un algoritmo iterativo que parte de una inicialización aleatoria de los centroides. Puede converger a un óptimo local, lo que significa que la configuración final de los clusters no es necesariamente la mejor a nivel global. En este caso el punto naranja podría haber sido "arrastrado" hacia el centroide azul en una iteración temprana y el algoritmo convergió antes de que pudiera ser reasignado correctamente a su cluster naranja.

A continuación, se presenta una descripción detallada de cada patrón, basándose en la moda de las respuestas.

Esto permite una comprensión profunda de las características académicas, de comportamiento y sociodemográficas que definen a cada clúster.

La caracterización de cada grupo proporciona información valiosa que puede ser utilizada para el desarrollo de intervenciones o políticas específicas, adaptadas a las necesidades de cada segmento de la población estudiantil.

El patrón 0 (Tabla 5-4). Descripción del patrón 0. Es un grupo de estudiantes de alto rendimiento que se caracteriza por un patrón de factores positivos y de apoyo que culmina en un rendimiento académico robusto. Este patrón se repite consistentemente en las instancias asignadas a esta agrupación de ingeniería civil, mayores de 21 años, hombres, con buen rendimiento académico (promedio de 8), alta satisfacción con la enseñanza, fuerte apoyo familiar y un buen equilibrio entre la vida personal y académica, la deserción universitaria no parece ser un problema.

Sin embargo, siempre hay oportunidades para mejorar y asegurar que estos estudiantes no solo permanezcan, sino que prosperen.

Tabla 5-4. Descripción del patrón 0.

PATRÓN 0	
Estudiantes de alto rendimiento	
Edad:	Más de 21 años
Carrera:	Ingeniería civil
Sexo:	Hombres
Promedio:	Generalmente buenos (“8: bien)
Satisfacción con enseñanza y métodos:	Alta (Satisfechos)
Motivación y apoyo familiar:	Alta (“Muy motivados”, “Siempre”)
Equilibrio vida-escolar:	Bien logrado
Abandono de estudios:	No, nunca

Algunas recomendaciones para continuar impulsando su éxito y compromiso son la siguientes: Fomentar la especialización y proyectos aplicados, programas de mentoría avanzada, oportunidades de liderazgo, desarrollo de habilidades complementarias, monitoreo continuo y personalizado. Al implementar estas estrategias, las instituciones educativas pueden no solo evitar la deserción en este grupo de estudiantes ya exitosos, sino también potenciar su desarrollo integral, asegurando que se conviertan en profesionales altamente capacitados y comprometidos con su campo. Gracias a estas características, logran mantener un buen equilibrio en su vida. No tienen historial de abandono de estudios y se sienten satisfechos con su profesión.

El patrón 1 (Tabla 5-5). Descripción del patrón 1, representa a estudiantes con posibles casos de desvinculación emocional o académica. Se sugiere atención mediante programas de orientación psicológica, mentoría personalizada y asesoría vocacional.

Tabla 5-5. Descripción del grupo1.

PATRÓN 1	
Estudiantes en alto riesgo (Desvinculación Académica/Emocional)	
Edad:	Más de 21 años
Carrera:	Ingeniería civil
Sexo:	Hombres
Promedio:	8: bien, es aceptable con percepción neutral en muchas respuestas
Satisfacción con enseñanza y métodos:	Neutral
Motivación y apoyo familiar:	Neutral
Equilibrio vida-escolar:	Neutral
Abandono de estudios:	Posiblemente

La confluencia de diversos factores, incluyendo la neutralidad de los factores, una insuficiente conexión con el proceso de enseñanza, niveles de motivación y apoyo poco sólidos, y un desequilibrio en la vida personal y académica, posiciona a los estudiantes en una categoría de alto riesgo. Esta situación incrementa significativamente la probabilidad de que se produzca el abandono escolar.

El patrón 2 (Tabla 5-6). Descripción del patrón 2, representa a estudiantes con alto compromiso y desempeño, que sugiere una fase de madurez y preparación para retos académicos superiores.

Tabla 5-6. Descripción del grupo 2.

PATRÓN 2	
Estudiantes de Alto Potencial	
Edad:	20 años
Carrera:	Ingeniería en Informática
Sexo:	Mujer
Promedio:	Generalmente buenos (“8: bien”)
Satisfacción con enseñanza y métodos:	Altas (“Satisfechas”)
Motivación y apoyo familiar:	Alta (“Muy motivado”, “Siempre”)
Equilibrio vida-escolar:	Bien logrado
Abandono de estudios:	No, nunca

Este grupo está compuesto por mujeres con una edad promedio de 20 años, cursando la carrera de Ingeniería en Informática. La estructura subyacente en esta agrupación está definida por variables modales que indican alto desempeño y un fuerte sentido de pertenencia o responsabilidad. Es una tipología con un historial académico sólido . La consistencia de este patrón sugiere que son estudiantes en transición hacia cursos avanzados.

Al tener el compromiso internalizado, las recomendaciones deben enfocarse en programas que capitalicen este potencial de liderazgo y excelencia, como el reconocimiento académico y el desarrollo de líderes estudiantiles.

Para validar los resultados del agrupamiento, se utilizaron dos métricas ampliamente reconocidas en el campo del aprendizaje automático para cuantificar qué tan bien se han formado los grupos: el Índice de Calinski-Harabasz y el Coeficiente de Silueta.

El Coeficiente de Silueta, que una métrica muy útil para evaluar la coherencia de los objetos dentro de sus propios clusters y su separación de los clusters vecinos. Para cada punto de dato, el coeficiente de silueta mide cuán similar es ese punto a otros puntos en su propio cluster en comparación con puntos en otros clusters.

De manera análoga al Índice de Calinski-Harabasz, se emplean los datos X_pca y las asignaciones de clusters para su evaluación. Una vez que los datos han sido etiquetados con el algoritmo de aprendizaje no supervisado K-means, el siguiente paso crucial es validar la calidad y coherencia de estos clústeres utilizando modelos de aprendizaje supervisado. Este enfoque permite evaluar si las agrupaciones generadas de forma automática son lo suficientemente robustas y predictivas.

Para validar los resultados de agrupamiento, se seleccionaron cuatro modelos de clasificación supervisada. Una vez agrupada la información, esta es dividida en 2 conjuntos, el 80% para entrenar los modelos de aprendizaje supervisado y el 20 % para probar los modelos. Una vez etiquetados nuestros datos con el algoritmo de agrupamiento K-means, el siguiente paso es validar nuestros resultados con algoritmos de aprendizaje supervisados. Para este caso se seleccionan cuatro modelos: Perceptrón Multicapa (Multilayer Perceptron, MLP), las Máquinas de Soporte Vectorial (Support Vector Machines, SVM), la Regresión Logística (Logistic Regression, LR) y Random Forest (FR). El desempeño de cada algoritmo se puede observar en la Tabla 5-6, se presenta una comparativa de rendimiento de diferentes algoritmos de aprendizaje automático, evaluados mediante diversas métricas.

Cada fila corresponde a un algoritmo específico, mientras que las columnas detallan las métricas de evaluación.

Tabla 5-7. Se muestran los resultados del algoritmo de aprendizaje.

Algoritmo	Accuracy	Precision (macro)	Recall (marco)	F1-score (macro)	Precision (weighted)	Recall (weighted)	F1-Score (weighted)
LR	0.9535	0.9587	0.9543	0.9563	0.9543	0.9535	0.9536
FR	0.9302	0.9380	0.9310	0.9290	0.9369	0.9302	0.9280
SVM	0.9535	0.9658	0.9467	0.9542	0.9583	0.9535	0.9538
MLP	0.9535	0.9576	0.9599	0.9582	0.9550	0.9535	0.9536

Como se observa en la Tabla 5-7, se muestran los resultados del algoritmo de aprendizaje que obtuvo mejores resultados fue el Perceptrón Multicapa, seguido de la Máquina de Soporte Vectorial y la Regresión Logística.

Para calcular cada una de las métricas de evaluación se utilizó la matriz de confusión.

En el trabajo (Cotrina-Teatino et al., 2025) el autor menciona que la matriz de confusión es una herramienta que permite visualizar el desempeño de un modelo al predecir cada clase dentro del conjunto de datos de prueba. Una matriz de confusión es una tabla que resume el rendimiento de un algoritmo de clasificación. Compara las predicciones del modelo con los valores reales (observados o verdaderos) para un conjunto de datos. Para un problema de clasificación binaria (dos clases, por ejemplo, "positivo" y "negativo"), la matriz de confusión típicamente se organiza en una tabla de 2×2 como se observa en la Tabla 5-7 matriz de confusión.

Tabla 5-8. Matriz de confusión.

Valor:	Predicción: Positivo	Predicción: Negativo
Real: Positivo	Verdaderos Positivos (VP)	Falsos Negativos (FN)
Real: Negativo	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Donde: Verdaderos Positivos (VP): Casos en los que el modelo predijo correctamente la clase positiva. El modelo predijo "positivo" y el valor real era "positivo".

- Verdaderos Negativos (VN): Casos en los que el modelo predijo correctamente la clase negativa. El modelo predijo "negativo" y el valor real era "negativo".
- Falsos Positivos (FP) (Error de Tipo I): Casos en los que el modelo predijo la clase positiva, pero el valor real era negativo. Es una "alarma falsa". Falsos Negativos (FN) (Error de Tipo II): Casos en los que el modelo predijo la clase negativa, pero el valor real era positivo. Es una "falla en la detección".

En la Figura 5-8, se muestra la Matriz de Confusión del algoritmo de Regresión Logística, que ilustra la fase de validación de los resultados obtenidos mediante la aplicación de algoritmos de aprendizaje supervisado. Específicamente, se presenta la matriz de confusión correspondiente al algoritmo de Regresión Logística.

Esta representación visual detallada confirma que el modelo de Regresión Logística alcanzó una precisión del 0.9535, lo que subraya su robustez y eficacia en la tarea de clasificación abordada. El algoritmo de regresión logística acertó 82 veces, y se equivocó 4 veces. Etiquetó a dos estudiantes que pertenecían al patrón 0 cuando realmente pertenecía al patrón 1. Etiquetó a un estudiante que pertenecía al patrón 1, cuando realmente pertenecía al patrón 0 y finalmente etiquetó a un estudiante que pertenecía al patrón 1 cuando realmente pertenecía al patrón 2, dando un total de 4 desaciertos.

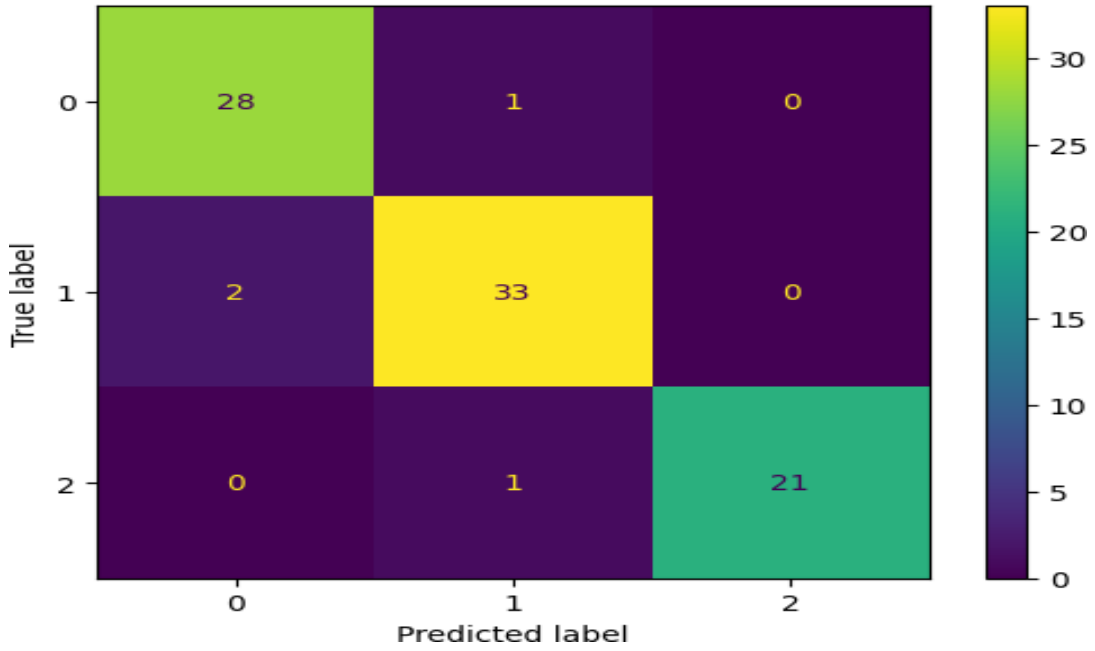


Figura 5-8. Matriz de Confusión del algoritmo de Regresión Logística.

El algoritmo Random Forest (Bosques Aleatorios), cuya matriz de confusión se presenta en la Figura 5-9, Matriz de confusión del algoritmo de Random Forest es una herramienta clave para comprender el desempeño de este modelo de clasificación. Si bien su exactitud global fue de 0.9302, ligeramente inferior a la de otros clasificadores como Regresión Logística, SVM y MLP, el análisis detallado de su matriz de confusión es fundamental para desglosar la distribución de aciertos y errores. En este sentido, revela que el algoritmo Random Forest realizó 80 clasificaciones correctas y 6 incorrectas. Específicamente, se identificó que cuatro estudiantes, que en realidad pertenecían al patrón 0, fueron incorrectamente etiquetados como pertenecientes al patrón 1. Adicionalmente, dos estudiantes del patrón 0, fueron erróneamente clasificados en el patrón 2. Este desglose pormenorizado de los Verdaderos Positivos, Falsos Negativos, Falsos Positivos y Verdaderos Negativos para cada clase proporciona una visión precisa de las áreas donde el modelo exhibe fortalezas y dónde persisten oportunidades de mejora en su capacidad predictiva.

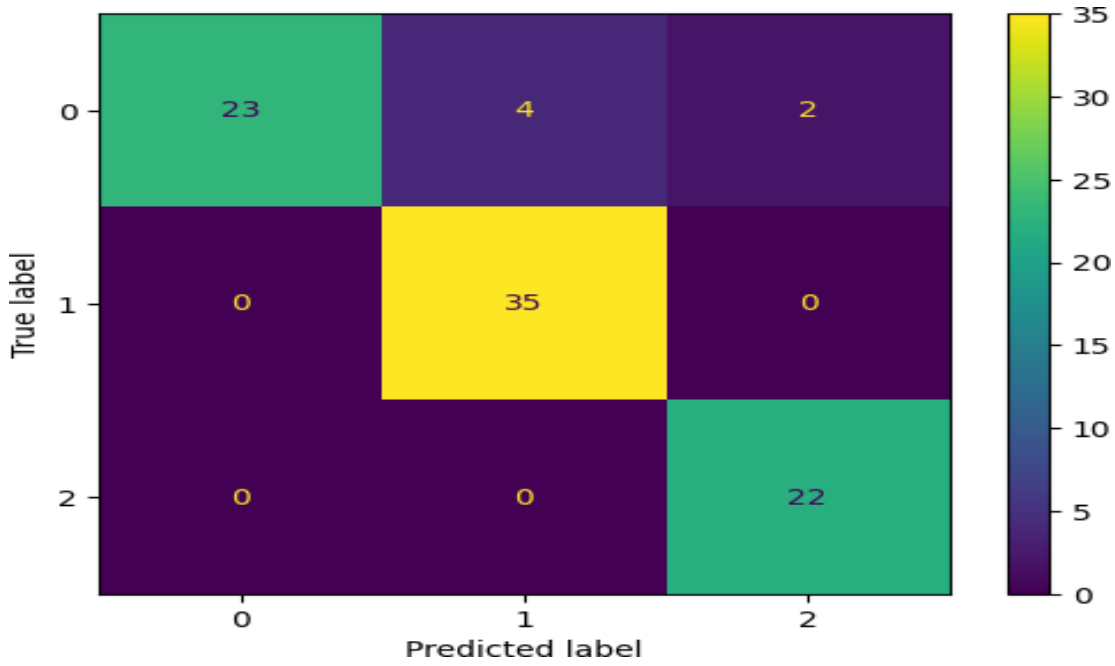


Figura 5-9. Matriz de confusión del algoritmo de Random Forest.

La Matriz de confusión del algoritmo de Máquinas de Soporte Vectorial (SVM), que se muestra en la Figura 5-10, es una herramienta indispensable para evaluar y comprender el rendimiento de este clasificador. En este caso, la matriz revela que el algoritmo SVM realizó 82 clasificaciones correctas y únicamente 4 incorrectas. Un análisis más granular de estos errores permite identificar que el modelo etiquetó erróneamente a dos estudiantes que, en realidad, pertenecían al patrón 2, asignándolos incorrectamente al patrón 1. Adicionalmente, dos estudiantes genuinamente pertenecientes al patrón 0 fueron incorrectamente clasificados dentro del patrón 1. La matriz de confusión es una herramienta diagnóstica esencial que trasciende la simple cuantificación del rendimiento, ofreciendo una perspectiva cualitativa invaluable sobre el comportamiento predictivo del modelo SVM en situaciones reales.

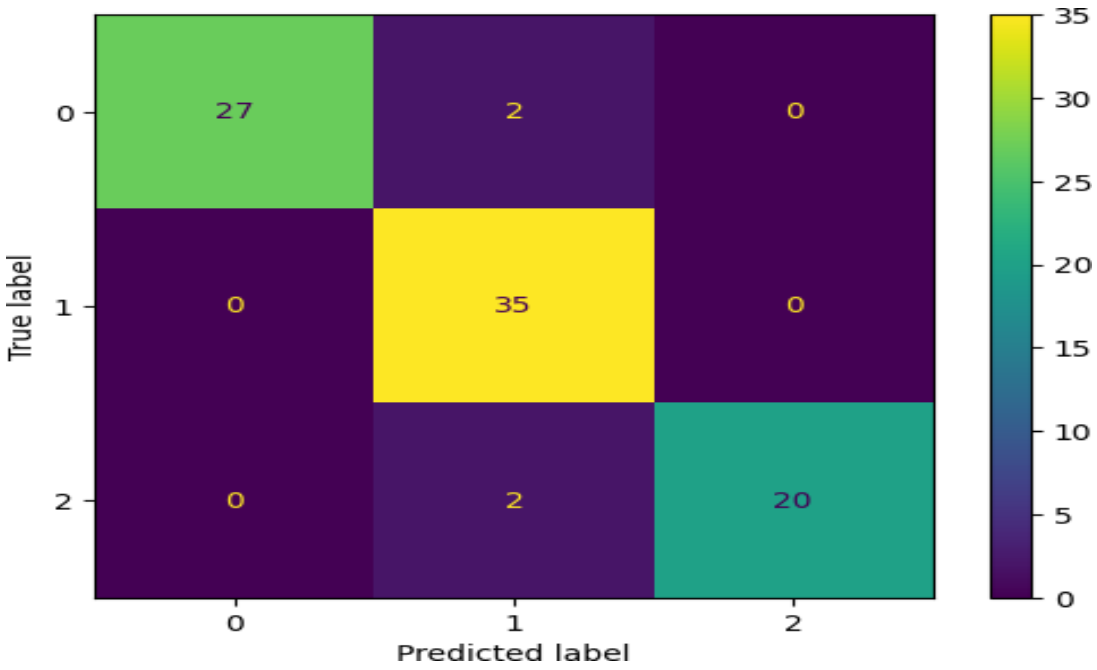


Figura 5-10. Matriz de confusión del algoritmo de Máquinas de Soporte Vectorial.

Finalmente, la Figura 5-11 presenta la Matriz de confusión del algoritmo Redes Neuronales (Perceptrón Multicapa), la cual revela un desempeño robusto. Este modelo logró 82 clasificaciones correctas y únicamente 4 predicciones erróneas, el MLP se posicionó como un algoritmo de alto desempeño, logrando una exactitud del 0.9535, comparable a la obtenida por la Regresión Logística y las Máquinas de Soporte Vectorial. Un análisis más detallado de los errores indica que el algoritmo clasificó incorrectamente a tres estudiantes del patrón 1, asignándolos erróneamente al patrón 0. Asimismo, se identificó un caso en el que un estudiante perteneciente al patrón 0 fue incorrectamente etiquetado como parte del patrón 1. Esta granularidad en la visualización de la matriz de confusión es fundamental para comprender las especificidades del comportamiento del modelo MLP y para futuras optimizaciones de su capacidad predictiva.

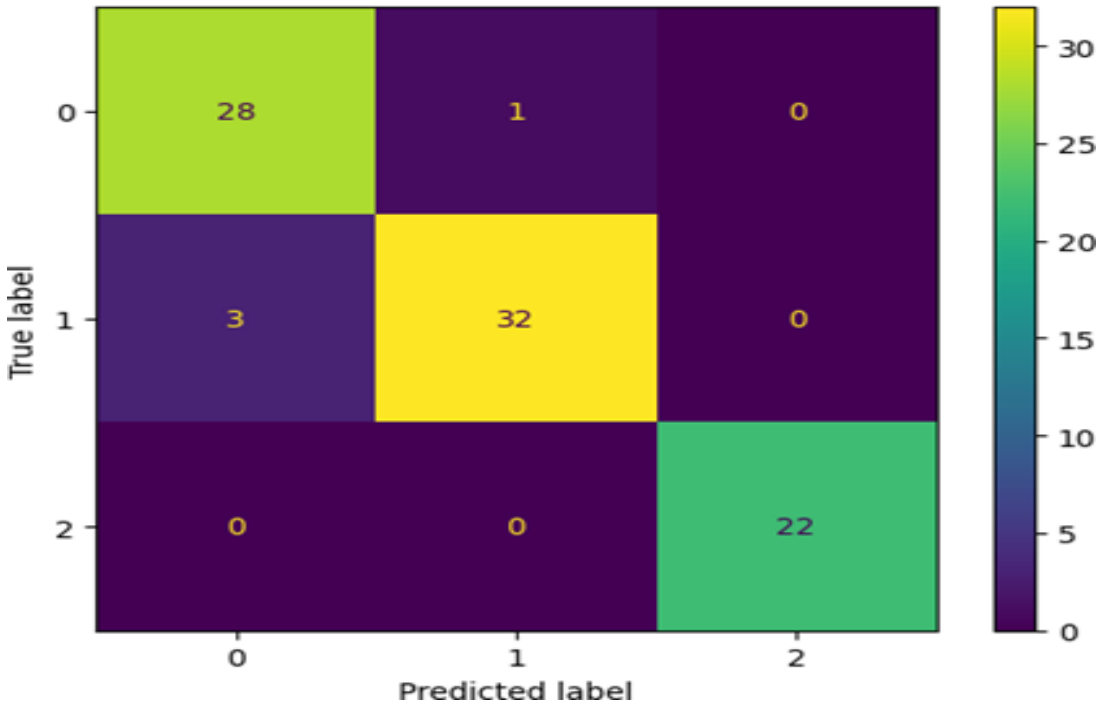


Figura 5-11. Matriz de confusión del algoritmo Redes Neuronales (Perceptrón Multicapa).

5.4 Interpretación de resultados

El análisis de datos académicos, utilizando algoritmos de aprendizaje no supervisado como K-means en combinación con la técnica de Reducción de Dimensionalidad (PCA), y validado mediante el método del codo y el coeficiente de silueta promedio, ha determinado que la configuración óptima para la agrupación de los estudiantes se establece en el patrón 0, patrón 1 y patrón 2.

Se ha identificado que dos de estos patrones están compuestos por estudiantes que manifiestan indicadores robustos de madurez académica, un rendimiento escolar consistentemente sobresaliente, un elevado nivel de apoyo familiar y una percepción institucional marcadamente favorable. Este hallazgo sugiere que estos segmentos estudiantiles poseen una base sólida para el éxito de sus trayectorias académicas.

Contrariamente, el tercer patrón (patrón 1) presenta un perfil de riesgo, caracterizado por patrones que sugieren una potencial desvinculación emocional o académica. Esta detección temprana subraya la urgencia de implementar estrategias de intervención focalizadas. Se recomienda encarecidamente la articulación de programas de orientación psicológica especializados, el establecimiento de esquemas de mentoría personalizada y un acompañamiento institucional continuo y proactivo. Estas acciones están diseñadas para mitigar el riesgo de deserción y fomentar la reinserción efectiva de los estudiantes en el entorno académico, garantizando su bienestar integral y éxito formativo.

Para caracterizar los patrones identificados, se realizó un análisis cualitativo complementado con el cálculo de la moda por variable dentro de cada clúster, lo que permitió interpretar las variables dominantes que describen los perfiles estudiantiles.

Posteriormente, se implementaron modelos de aprendizaje supervisado con el propósito de estimar la pertenencia de estudiantes que no fueron incluidos en los conjuntos de entrenamiento ni de prueba a alguno de los grupos previamente definidos. Los resultados obtenidos evidencian la viabilidad de generar predicciones precisas respecto al patrón de asignación de nuevos estudiantes con características similares. Esta capacidad predictiva contribuye a la identificación temprana de posibles situaciones de riesgo académico o emocional, lo que a su vez favorece la toma de decisiones informadas y la ejecución de estrategias de intervención oportunas por parte de las instituciones educativas.

Para lograr una comprensión profunda de los patrones estudiantiles previamente identificados, se implementó un análisis cualitativo riguroso, complementado con el cálculo de la moda por variable dentro de cada clúster. Esta aproximación metodológica dual permitió dilucidar las características dominantes que definen los perfiles de los estudiantes en cada segmento, ofreciendo una base sólida para la interpretación de su comportamiento y necesidades.

Consecuentemente, se procedió al entrenamiento de modelos de aprendizaje supervisado. El objetivo principal de esta fase fue desarrollar la capacidad predictiva para asignar nuevos estudiantes (excluidos de los conjuntos de entrenamiento y prueba originales) a uno de los patrones caracterizados. Los resultados empíricos derivados de este proceso han validado la factibilidad y fiabilidad de realizar predicciones sobre la pertenencia patronal de nuevos ingresos con perfiles de características similares.

Esta capacidad predictiva representa un avance significativo, ya que facilita la detección temprana de potenciales situaciones de riesgo académico o emocional. En un contexto institucional, esta información es invaluable para la toma de decisiones proactiva, permitiendo la implementación de intervenciones educativas y de apoyo oportunas y personalizadas. En última instancia, esta metodología no solo optimiza la comprensión de la población estudiantil, sino que también refuerza la capacidad de las instituciones para fomentar la retención y el éxito académico.

Capítulo 6

Conclusiones y trabajos futuros

La tesis proporciona un aporte significativo a la Minería de Datos Educativa a través de un estudio metodológicamente riguroso, cuyos principales hallazgos y conclusiones son:

- Metodología Estructurada Validada: Se aplicó una metodología de cuatro fases (recolección, preprocesamiento, modelado e interpretación) que garantiza la fiabilidad y consistencia del proceso. La validación de los datos primarios mediante V de Aiken y Alfa de Cronbach confirma la pertinencia y consistencia interna de las variables utilizadas.

- Muestra Representativa: La investigación se basó en una muestra significativa de 430 estudiantes de diversos niveles académicos de una Institución de Educación Superior adscrita al Tecnológico Nacional de México.

- Reducción de Dimensionalidad Eficiente: Se implementó el Análisis de Componentes Principales (PCA), lo que permitió la reducción exitosa de la dimensionalidad del conjunto de datos original, preservando la mayor varianza y optimizando las etapas analíticas subsiguientes.

- Segmentación Robusta de la Población: El algoritmo K-means se utilizó para el agrupamiento de estudiantes. La determinación del número óptimo de componentes y grupos se realizó mediante criterios estadísticos robustos (método del codo y coeficiente de silueta promedio), asegurando una segmentación conceptualmente y estadísticamente pertinente.

-

- Identificación de Riesgo de Desvinculación: El análisis de los datos reveló la existencia de un segmento estudiantil con claros indicios de desvinculación emocional y/o académica.

- Necesidad de Intervención Proactiva: Este hallazgo crítico subraya la imperiosa necesidad de implementar estrategias de intervención proactivas, incluyendo programas de orientación psicológica, esquemas de mentoría personalizada y servicios de asesoramiento vocacional adaptados.

- Modelos Predictivos de Alto Rendimiento: Se realizó una evaluación comparativa de cuatro modelos de aprendizaje supervisado (MLP, SVM, Regresión Logística y Random Forest). Los modelos Perceptrón Multicapa (MLP), Máquinas de Soporte Vectorial (SVM) y Regresión Logística (LR) demostraron un rendimiento superior en la tarea de clasificación.

- Validación Rigurosa de Modelos: La eficacia de los modelos fue validada rigurosamente utilizando matrices de confusión y un conjunto de métricas estándar de desempeño (exactitud, precisión, sensibilidad*/recall y F1-score*), garantizando una comprensión exhaustiva de su capacidad predictiva.

6.1 Trabajos futuros

Es crucial replicar el estudio en otras instituciones, regiones o niveles educativos para confirmar si los hallazgos son aplicables a contextos más amplios. Esto valida la robustez del modelo y evita que las conclusiones sean específicas de una sola muestra.

Un trabajo futuro puede centrarse en explorar algoritmos de aprendizaje automático más avanzados (como redes neuronales profundas) para mejorar la precisión predictiva. Además, se pueden incorporar nuevas variables (factores socioeconómicos, interacciones sociales) que no se consideraron en el estudio inicial para enriquecer el análisis.

El siguiente paso es trascender la predicción y enfocarse en la prevención. Para ello, se propone el desarrollo de un dashboard interactivo que identifique en tiempo real a los estudiantes en riesgo. Esta herramienta permitirá a los administradores educativos implementar intervenciones proactivas y personalizadas, como mentorías y apoyo psicológico, para combatir la deserción.

Referencias

Castillo Aráuz, D., & Martínez, J. J. (2023). Predicción Del Rendimiento Académico En La UNADECA Por Medio De Sistemas De Clasificación. *Una Ciencia Revista De Estudios E Investigaciones*, *16*(31), 17–35. Https://Doi.Org/10.35997/Unaciencia.V16i31.738

Cotrina-Teatino, M. A., Riquelme Sandoval, A. I., Guartan Medina, J. A., & Marquina Araujo, J. J. (2025). Machine Learning Aplicado A La Exploración Minera Usando Matriz De Confusión. *Revista Ciencia Y Tecnología*, *21*(1), 63–74. Https://Doi.Org/10.17268/Rev.Cyt.2025.01.06

David, Y., Islas, G., Ramos Ojeda, E., & Guzmán Zazueta, A. (2018). Abandono Escolar Guia (Gestión Universitaria Integral Del Abandono) Prototype Of Data Mining In The Timely Detection Of Students At Risk Of Dropping Out Of School Guidance (Comprehensive University Management Of Abandonment). In *Tecnológico Nacional De México En Celaya Pistas Educativas* (Vol. 39, Issue 129). Http://Itcelaya.Edu.Mx/Ojs/Index.Php/Pistas

De, F., De, C., De, T., Información, L. A., & De Titulación, T. (2020). *Universidad Estatal Península De Santa Elena*.

Del Carmen Del Amo Chicharro, M., Anguita Acero, J. M., & González Olivares, Á. L. (2025). Design And Validation Of The Neurodiedu Questionnaire To Assess The Implementation Of Neurodidactics In Pre-School And Primary School Classrooms. *Aula Abierta*, *54*(1), 51–63. Https://Doi.Org/10.17811/Rifie.20944

De, T., & -Ecuador, G. (2022). *Maestría En Opción De Titulación: Artículo Profesional De Alto Nivel*.

Escurra M, L. M. (2020). Diseño Y Validación Del Cuestionario Decompetencias Socioemocionales. *Universidad De Perú*, 1–95.

Hernández G, C. L. , R. R. J. E. (2019). *Preprocesamiento De Datos Estructurados*. 1–22.

Herrera Flóres, J. E. (2025). Análisis De La Predicción Temprana De Los Factores Determinantes En La Deserción De La Educación Superior Desde El Enfoque De Machine Learning - Un Análisis Desde El Contexto Colombiano. *Universidad Nacional Abierta Y A Distancia*, 1–94.

López Pedraza, F. J., Macias González, Ma. D. C., & Sandoval García, E. R. (2019). Minería De Datos: Identificando Causas De Deserción En Las Instituciones Públicas De Educación Superior De México. *TIES, Revista De Tecnología E Innovación En Educación Superior*, *2*, 1–14. Https://Doi.Org/10.22201/Dgtic.26832968e.2019.2.4

Maćkiewicz, A., & Ratajczak, W. (1993). Principal Components Analysis (PCA). *Computers & Geosciences*, *19*(3), 303–342. Https://Doi.Org/10.1016/0098-3004(93)90090-R

Mazurett González, M. V. (2025). *Modelos-Predictivos-De-Desercion-Estudiantil-Incorporando-Variables-De-Salud-Mental-En-La-Universidad-De-Chile-Para-Elaborar-Estrategias-De-Intervencion-Temprana*. 1–25.

Pérez-Gutiérrez, B. R. (2020). Comparación De Técnicas De Minería De Datos Para Identificar Indicios De Deserción Estudiantil, A Partir Del Desempeño Académico. *Revista UIS Ingenierías*, *19*(1), 193–204. Https://Doi.Org/10.18273/Revuin.V19n1-2020018

Pérez Niño, J. L., Gualdrón Guerrero, O. E., & Barrera Oliveros, D. J. (2024). Modelos De Inteligencia Artificial En Minería De Datos Educativos Para Predecir La Deserción En Educación Superior: Una Revisión Integral. *Tecnura*, *28*(82), 134–155. Https://Doi.Org/10.14483/22487638.23670

Sulla Torres, J. A. (2021). *Universidad Católica Santa María*.

Urbina-Nájera, A. B., Camino-Hampshire, J. C., & Cruz Barbosa, R. (2020). University Dropout: Prevention Patterns Through The Application Of Educational Data Mining. *Relieve - Revista Electronica De Investigacion Y Evaluacion Educativa*, *26*(1), 1–19. Https://Doi.Org/10.7203/Relieve.26.1.16061

Urbina-Nájera, A. B., Téllez-Velázquez, A., & Barbosa, R. C. (2021). Patterns To Identify Dropout University Students With Educational Data Mining. *Revista Electronica De Investigacion Educativa*, *23*. Https://Doi.Org/10.24320/REDIE.2021.23.E29.3918

Villarroel-Henríquez, V., & Gallardo-Aguayo, C. (2025). Educational Management Using Learning Analytics And Big Data In Higher Education. *Revista De Investigacion En Educacion*, *23*(1), 76–92. Https://Doi.Org/10.35869/Reined.V23i1.6110

Anexo A

El objetivo principal de la Figura A-1. Diseño del instrumento de medición, al desarrollar un instrumento de medición, concebido como un diccionario de variables, radica en establecer una estructura metodológica rigurosa para la recolección y análisis de datos. En este contexto, cada pregunta se define como una variable, a la que se le asignan especificaciones precisas que incluyen su tipo de escala (ej., nominal, ordinal, de intervalo, de razón) y las respuestas correspondientes siendo este el ANEXO A.

Diseño del instrumento de medida: Diccionario (Dataset)			
VARIABLE	**DESCRIPCIÓN**	**TIPO DE VARIABLE**	**VALORES**
Q1_cual_edad	¿Cuál es tu edad en años?	Categórica	R1 estudiante de 18 años R2 estudiante de 19 años R3 estudiante de 20 años R4 estudiante de 21 años R5 Estudiante de más de 21 años R6 Egresado, de más de 21 años R7 Sin egresar, de más de 21 años
Q2_cual_sexo	¿Cúal es tu género?	Dicotómica	R1 Femenino R2 Masculino
Q3_ing_cursas	¿En qué carrera estas inscrito?	Categórica	R1 Civil R2 Gestión Empresarial R3 Informática
Q4_semetre_actual	Selecciona, ¿Qué semestre cursas actualmente?	Categórica	R1 Segundo semestre R2 Tercer semestre R3 Cuarto semestre R4 Quinto semestre R5 Octavo semestre R6 Noveno semestre R7 Egresado/egresada R8 Sin egresar
Q5_promedio	¿Cuál es tu promedio académico general?	Likert	R1 5 o Menos:Reprobado R2 6:Suficiente R3 7:Aprobado R4 8:Bien R5 9:Notable R6 10:Sobresaliente
Q6_promedio_semestre	¿Cuál fué tu promedio en el último semestre?	Likert	R1 5 o Menos:Reprobado R2 6:Suficiente R3 7:Aprobado R4 8:Bien R5 9:Notable R6 10:Sobresaliente
Q7_materias_no aprobadas	¿Cuantas materias no aprobadas presentas hasta ahora?	Categórica	R1 Ninguna materia R2 1 Materia reprobada R3 2 Materias reprobadas R4 3 Materias reprobadas R5 4 Materias reprobadas
Q8_nivel_compren_contenido	¿Cómo calificarías tu nivel de comprensión de los contenidos en tu carrera?	Likert	R1 Totalmente en desacuerdo R2 En desacuerdo R3 Neutral R4 De acuerdo R5 Totalmente de acuerdo

Q9_satis_ense-univer	¿Cuál es el grado de satisfacción con la enseñanza en la universidad?	Likert	R1 Muy insatisfecho R2 Insatisfecho R3 Neutral R4 Satisfecho R5 Muy satisfecho
Q10_satis_met_ense	¿Cuál es tu nivel de satisfacción con los métodos de enseñanza?	Likert	R1 Muy insatisfecho R2 Insatisfecho R3 Neutral R4 Satisfecho R5 Muy satisfecho
Q11_comunica_prof_est_efectiva	¿Consideras que la comunicación entre profesores y estudiantes es efectiva?	Likert	R1 Totalmente en desacuerdo R2 En desacuerdo R3 Neutral R4 De acuerdo R5 Totalmente de acuerdo
Q12_prof_capa_enseñar	¿Sientes que tus profesores están capacitados para enseñar y fomentar la sostenibilidad	Likert	R1 Totalmente en desacuerdo R2 En desacuerdo R3 Neutral R4 De acuerdo R5 Totalmente de acuerdo
Q13_presionado_carga_acad	¿Qué tan presionado te sientes por la carga académica?	Likert	R1 Muy presionado R2 Algo presionado R3 Neutral R4 Poco presionado R5 Nada precionado
Q14_carga_acad_excesiva	¿Consideras que la carga académica es excesiva?	Likert	R1 Muy en desacuerdo R2 En desacuerdo R3 Neutral R4 De acuerdo R5 Muy de acuerdo
Q15_comunicar_capaz_prof_compa	¿Te sientes capaz de comunicar efectivamente con tus compañeros y profesores sobre tus	Likert	R1 Totalmente incapaz R2 Incapaz R3 Neutral R4 Capaz R5 Totalmente capaz
Q16_horas_estudio_fuera	¿Cuántas horas de estudio adicionales dedicas fuera de tu horario académico?	Likert	R1 Ninguna hora R2 Menos de una hora R3 1-2 horas R4 3-4 horas R5 más de 4 horas
Q17_abandonar_estudios	¿Has pensado en abandonar tus estudios académicos?	Likert	R1 Sí, definitivamente R2 Neutral R3 Posiblemente R4 No, nunca
Q18_oport_talen_habili	¿Consideras que la escuela te brinda oportunidades para desarrollar tus talentos y habilidades?	Likert	R1 Totalmente en desacuerdo R2 En desacuerdo R3 Neutral R4 De acuerdo R5 Totalmente de acuerdo
Q19_ingreso_familiar	¿Cuál es el ingreso mensual de tu familia?	Categórica	R1 Menos de $5,000 R2 De $5,000 a $10,000 R3 De 10,000 a $20,000 R4 Más de $20,000
Q20_beca_apoyo_financiamiento	¿Cuentas con beca o apoyo financiero?	Likert	R1 Sin apoyo R2 Apoyo gubernamental R3 Beca cultural R4 Beca deportiva R5 Beca académica

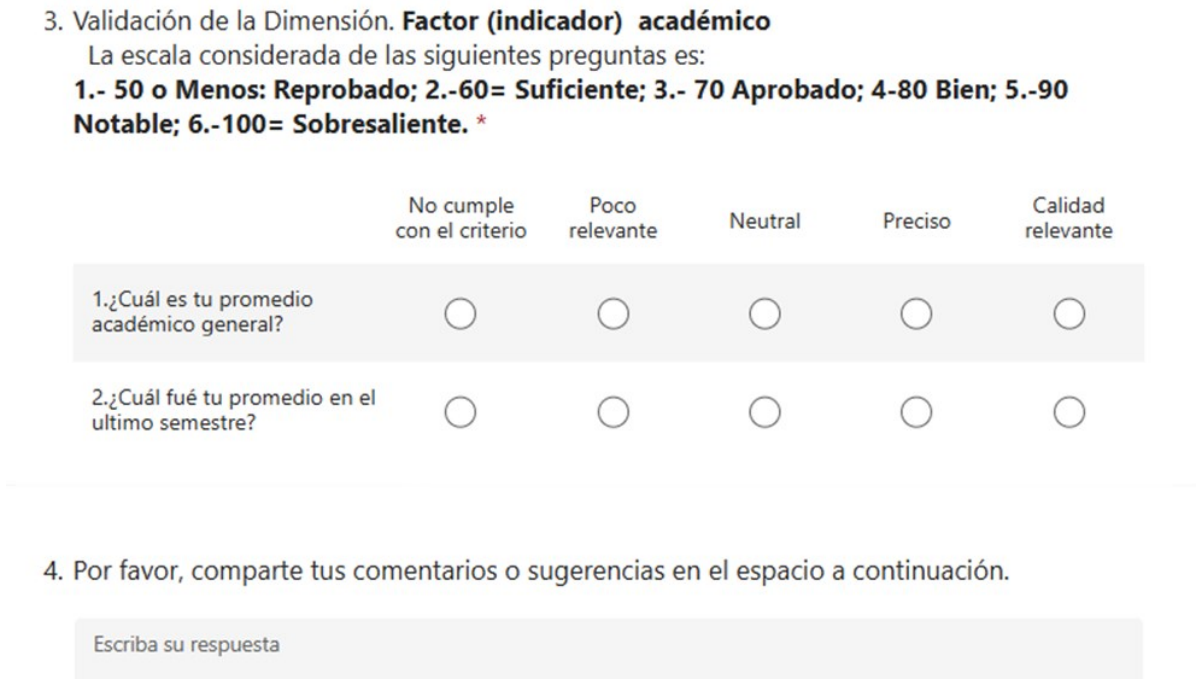
Q21_financiar_estudios_futuro	¿En que medida te preocupa financiar tus estudios en el futuro?	Likert	R1 Nada preocupado R2 Poco preocupado R3 Neutral R4 Algo preocupado R5 Muy preocupado
Q22_tienes_compu_personal	¿Tienes acceso a una computadora personal para estudiar?	Dicotómica	R1 Sí R2 No
Q23_tienes_internet_en_casa	¿Tienes acceso a Internet en casa?	Dicotómica	R1 Sí R2 No
Q24_trabajas_estudias	¿Trabajas mientras estudias?	Dicotómica	R1 Sí R2 No
Q25_horas_laborables_semanal	Si trabajas, ¿cuántas horas laboras por semana?	Categórica	R1 No trabajo R2 Menos de 20 horas R3 20-30 horas R4 31-40 horas R5 Más de 40 horas
Q26_problemas_salud_rendimiento	¿Ha tenido problemas de salud que afectan su rendimiento académico?	Categórica	R1 Problemas graves R2 Problemas moderados R3 Neutral R4 Problemas leves R5 No tengo problemas
Q27_ansiedad_depresion_por_estudios	¿Ha sentido ansiedad o depresión relacionada con sus estudios?	Likert	R1 Nunca R2 Rara vez R3 A veces R4 A menudo R5 Siempre
Q28_manejar_estres_relacionado_escuela	¿Te sientes capaz de manejar el estrés relacionado con la escuela?	Likert	R1 Totalmente incapaz R2 Muy incapaz R3 Incapaz R4 Neutral R5 Capaz R6 Muy capaz
Q29_discriminado_excluido_escuela	¿Te sientes discriminado (a) o excluido(a) en la escuela?	Likert	R1 Nunca R2 Rara vez R3 Algunas veces R4 Frecuentemente R5 Siempre
Q30_apoyo_familiar_seguir_estudiando	¿Cuentas con apoyo familiar para seguir estudiando?	Likert	R1 Nunca R2 Rara vez R3 Algunas veces R4 Frecuentemente R5 Siempre
Q31_motivado_terminar_carrera	¿Qué tan motivado te sientes para terminar tu carrera?	Likert	R1 Nada motivado R2 Poco motivado R3 Neutral R4 Algo motivado R5 Muy motivado

Figura A-0-1.Diseño del instrumento de medición

ANEXO B

Se muestran en la Figura B-1 las preguntas que forman parte del cuestionario a ser evaluado, se observan las preguntas que forman parte del cuestionario y que será validado por un grupo de expertos en el área de psicología y muestra una sección del instrumento después de su validación por jueces expertos. Esta estructura permite una evaluación matizada y profesional de cada elemento, reflejando el juicio experto sobre su adecuación y calidad. Este instrumento utiliza una escala de valoración discreta y categórica, diseñada para evaluar criterios específicos. Las opciones de respuesta disponibles en esta escala son:

- No cumple con el criterio
- Poco relevante
- Neutral
- Preciso
- Calidad relevante



3. Validación de la Dimensión. **Factor (indicador) académico**
La escala considerada de las siguientes preguntas es:
1.- 50 o Menos: Reprobado; 2.-60= Suficiente; 3.- 70 Aprobado; 4-80 Bien; 5.-90 Notable; 6.-100= Sobresaliente. *

	No cumple con el criterio	Poco relevante	Neutral	Preciso	Calidad relevante
1.¿Cuál es tu promedio académico general?	○	○	○	○	○
2.¿Cuál fué tu promedio en el ultimo semestre?	○	○	○	○	○

4. Por favor, comparte tus comentarios o sugerencias en el espacio a continuación.

Escriba su respuesta

Figura A-2 1.Preguntas que forman parte del cuestionario a ser evaluado.

ANEXO C

Se muestra en la Tabla C-1 un ejemplo de la Base de datos que evaluaron los jueces. Se puede observar la base de datos que evaluaron los jueces o personas expertas en temas de Minería de Datos Educativa, solo se muestran las preguntas del 5 al 10 como ejemplo del contenido, ya que las primeras contienen datos personales como ¿Cuál es tu nombre?, ¿Cuál es tu edad?, ¿Cuál es tu correo electrónico?, etc.

Tabla C-1. Ejemplo de la Base de datos que evaluaron los jueces.

Q5 ¿Cuál es tu promedio académico general?	Q6 ¿Cuál fué tu promedio en el ultimo semestre?	Q7¿Cuantas materias no aprobadas presentas hasta ahora?	Q8 ¿Cómo calificarías tu nivel de comprensión de los contenidos en tu carrera?	Q9 ¿Cuál es el grado de satisfacción con la enseñanza en la Universidad?	Q10 ¿Cuál es tu grado de satisfacción con los métodos de enseñanza?
Neutral	Neutral	Calidad relevante	Calidad relevante	Neutral	Preciso
Preciso	Preciso	Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante
Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante
Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante	No cumple con el criterio	Calidad relevante
Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante	Calidad relevante
Preciso	Preciso	Preciso	Calidad relevante	Calidad relevante	Calidad relevante
Preciso	Preciso	Preciso	Neutral	Calidad relevante	Preciso
Neutral	Poco relevante	Preciso	Calidad relevante	Calidad relevante	Calidad relevante

El proceso de validación por expertos generalmente implicó los siguientes pasos:

- Selección de expertos: Se identifican y seleccionan expertos que cumplan con las características deseadas. Para este proyecto se solicitó la ayuda de las y los siguientes expertos:

- Dra. Alicia Martínez Rebollar
- Dra. Ana María Peña
- Dra. Blanca Dina Valenzuela Robles
- Dr. David Luviano
- Dra. Eunice Paola Gallegos
- Dra. Hilda del Carmen Lee

- Dr. Manuel Juárez
- Mtra. Verónica Sotelo

Para optimizar la calidad del cuestionario, se convocó a ocho expertos en psicología. La participación de estos especialistas es crucial, ya que su conocimiento profundo del área nos permitirá validar la pertinencia y eficacia de cada ítem.

Se compartido el formulario con los especialistas vía correo electrónico, adjuntando el enlace directo y detallando la finalidad de su evaluación: asegurando que el instrumento sea claro, relevante y representativo de los constructos que deseamos medir. Sus aportaciones son invaluables para garantizar la validez de contenido y la fiabilidad de los datos que se obtendrán, pilares fundamentales para una investigación rigurosa.

ANEXO D

Se muestra en la Tabla D-1 la pertinencia al área y claridad de redacción, se muestran los resultados de la evaluación inicial del cuestionario realizado por ocho jueces expertos, que indican un alto grado de pertinencia temática y claridad en la redacción. Tras un análisis descriptivo de la varianza de las valoraciones de los expertos, se obtuvo un coeficiente de 0.74, lo cual valida la idoneidad del instrumento para proseguir con la investigación. Este consenso refuerza la fiabilidad del cuestionario en cuanto a la relevancia y comprensión de cada ítem.

Tabla D-1. Pertinencia al área y claridad de la redacción.

Pertinencia al área y claridad de redacción

	D. académico			D. Satisfacción con la enseñanza		D. Interacción y apoyo	D. Compromiso y hábitos de estudio	D. Motivación y permanencia	D. Desarrollo de habilidades e	D. Factor socieconómico				D. Recursos tecnólogicos y acceso a internet		D. Trabajo y dedicación al estudio		Factores psicologicos y motivacionales				D. Apoyo familiar y social	D. Motivación y metas	D. Experiencia académica				D. Sostenibilidad									
Jueces	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Q13**	**Q14**	**Q15**	**Q16**	**Q17**	**Q18**	**Q19**	**Q20**	**Q21**	**Q22**	**Q23**	**Q24**	**Q25**	**Q26**	**Q27**	**Q28**	**Q29**	**Q30**	**Q31**	**Q32**	**Q33**	**Q34**	**Q35**	**Q36**	**Q37**	**Q38**	**Q39**	**Q40**
Juez 1	0.50	0.50	0.25	1.00	0.50	0.75	0.75	0.75	0.50	0.00	1.00	0.50	0.00	0.50	0.50	1.00	0.00	1.00	1.00	1.00	0.00	0.00	0.25	1.00	1.00	0.00	0.00	1.00	0.75	0.75	0.00	1.00	1.00	0.00	1.00	0.00
Juez 2	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.50	0.75	0.75	1.00	0.75	1.00	1.00	0.00	0.75	1.00	1.00	1.00	0.75	1.00	1.00	1.00	0.75	0.75	0.50	0.75	0.50	0.75	0.75	0.50	0.75	0.50	1.00
Juez 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.50	0.50	0.50	0.50	0.00
Juez 4	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00
Juez 5	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.75	0.50	1.00	0.50	1.00	0.50	0.75	1.00	1.00	0.75	1.00	1.00	1.00	0.25	0.50	0.75	0.50	0.75	1.00	0.75	0.25	0.75	1.00	0.50	1.00	0.25	0.00	0.25	0.75
Juez 6	0.75	0.75	0.75	1.00	1.00	1.00	1.00	1.00	0.50	0.75	1.00	0.50	0.75	1.00	0.50	0.75	0.75	0.75	0.75	0.50	0.50	0.75	0.75	0.75	0.75	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.75	0.75	0.75
Juez 7	0.75	0.75	0.75	0.50	1.00	0.75	1.00	0.75	1.00	0.75	1.00	0.25	1.00	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.75	1.00	1.00	1.00	1.00	1.00	0.75
Juez 8	0.50	0.25	0.75	1.00	1.00	1.00	0.75	1.00	0.75	0.50	0.75	0.75	1.00	0.75	0.75	1.00	1.00	0.75	1.00	0.50	0.25	1.00	1.00	0.75	1.00	0.75	1.00	0.75	0.75	0.75	0.50	0.25	0.75	0.50	1.00	0.25
Total	**0.78**	**0.75**	**0.81**	**0.94**	**0.81**	**0.94**	**0.91**	**0.91**	**0.72**	**0.56**	**0.75**	**0.72**	**0.78**	**0.72**	**0.69**	**0.97**	**0.63**	**0.91**	**0.97**	**0.88**	**0.63**	**0.63**	**0.84**	**0.75**	**0.81**	**0.63**	**0.81**	**0.56**	**0.59**	**0.72**	**0.59**	**0.69**	**0.59**	**0.44**	**0.63**	**0.44**
	0.78			0.88		0.94	0.91	0.91	0.72	0.68			0.75		0.83		0.84				0.63	0.63	0.76				0.61									
	0.74																																			

Menor valor 1
Número de categorias 13

ANEXO E

A continuación, se muestran los resultados y comentarios realizados por las personas expertas en el área de psicología, se les pidió analizar y valorar críticamente cada uno de los ítems en relación con categorías de factores de información general, factores académicos, factores socioeconómicos, factores psicológicos y motivacionales, factores de sostenibilidad académica, factores institucionales, interacción con la tecnología y metodología de la enseñanza y de factor sostenibilidad, utilizando en la mayoría una escala Likert. También se les solicitó señalar posibles aspectos en la medición del constructo que no hubieran sido considerados por los investigadores. A continuación, se describen algunas de las principales valoraciones:

Juez ID 1. En la dimensión factores académicos, comentó que las preguntase están acordes a lo que deseamos buscar en los alumnos.

Juez ID 2. En la dimensión carga académica y dificultades, en el Ítem Q15 ¿Te sientes capaz de comunicar efectivamente con tus compañeros y profesores sobre tus necesidades académicas? Propuso que es importante saberse comunicar para expresar lo que piensas y que los oyentes entiendan la idea a transmitir con una redacción alternativa como "¿Te sientes cómodo al expresar tus necesidades académicas a compañeros y profesores?".

También sugiere en el Ítem Q22 ¿Tienes acceso a una computadora personal para estudiar? Donde el factor socioeconómico en la dimensión ingreso familiar y situación económica, hace mención que "La computadora y el acceso al internet son básicos para poder hacer investigación y cumplir con los compromisos educativos".

El Juez ID 3. En la dimensión da factores académicos, en el Ítem Q8 ¿Cómo calificarías tu nivel de comprensión de los contenidos en tu carrera? Menciona que, es interesante, un ejercicio de autorreflexión, pero muy subjetivo y sin forma de comprobar la veracidad de la respuesta. En este caso el alumno sabría como contestar mediante una respuesta confiable como: "Considero que mi nivel de comprensión de los contenidos de mi carrera es alto. Me baso en mi capacidad para aplicar los conocimientos en tareas y proyectos, así como en mi

desempeño en evaluaciones y discusiones en clase." "Evalúo mi comprensión de los contenidos de mi carrera como satisfactoria. Constantemente busco conectar los conceptos teóricos con situaciones prácticas y me siento capaz de explicar los temas a mis compañeros.", "Desde mi perspectiva, mi nivel de comprensión de los contenidos de mi carrera es bueno.

En la dimensión experiencia académica el Ítem Q10 ¿Cuál es tu grado de satisfacción con los métodos de enseñanza? Y Q11 ¿Consideras que la comunicación entre profesores y estudiantes es efectiva? Menciona: Las dos preguntas requerirían aclaración, ¿cómo sabes que un estudiante sabe qué es sostenibilidad? En este caso el alumno al contestar la respuesta deberá saber que su grado de satisfacción con los métodos de enseñanza es alto. Porque el valora la claridad con la que los profesores explican los conceptos y el uso de ejemplos prácticos que facilitan la comprensión. La segunda pregunta no se corresponde con la primera ni con la pregunta 10, esta ¿a qué dimensión pertenece? En la Q10 pertenece a la dimensión de Proceso Enseñanza-Aprendizaje y la Q11 es a la dimensión de Interacción y comunicación.

En el Ítem Q15 ¿Consideras que la escuela ofrece suficiente apoyo para los estudiantes con dificultades académicas? Menciona "Nuevamente depende de cada estudiante, ¿cómo podría reformularse esta pregunta?" Enfocada en la conciencia de los servicios quedaría "¿Estás al tanto de los servicios de apoyo académico que ofrece la escuela (por ejemplo, tutorías, asesorías personalizadas, grupos de estudio asistidos)?" o "¿Conoces los recursos disponibles en la escuela para estudiantes que tienen dificultades en alguna materia?"

Dentro del factor de dedicación al estudio y a la dimensión de tiempo de estudio Individual en el Ítem Q16 ¿Cuántas horas de estudio adicionales dedicas fuera de tu horario académico? Menciona, En 1 solo debe decir "ninguna", ya que la escala contempla las opciones de (Ninguna hora, Menos de una hora, 1-2 horas, 3-4 horas y Más de 4 horas).

Dentro del factor académico en la dimensión desarrollo de habilidades e intereses, en el Ítem

Q18 ¿Consideras que la escuela te brinda oportunidades para desarrollar tus talentos y habilidades? Menciona la pregunta es Subjetiva e imposible de verificar, para resolver esto se propone replantear la pregunta a "¿Has participado en alguna actividad o programa de la escuela diseñado para desarrollar tus talentos o habilidades?"

En el factor socioeconómico la dimensión ingreso familiar y situación económica, en el ítem Q19 ¿Cúal es el ingreso mensual de tu familia? y la Q23 ¿Tienes acceso a Internet en casa? Hace mención que estas preguntas deberían ir al inicio, por motivos de sensibilidad al colocarla al principio, los encuestados aún no han invertido mucho tiempo en la encuesta, por lo que, si se sienten incómodos respondiendo, la "pérdida" de tiempo sería menor. Sin embargo, esto también podría generar desconfianza al inicio.

Dentro del factor socioeconómico en la dimensión recursos tecnológicos y acceso a internet, en el Ítem Q22 ¿Tienes acceso a una computadora personal para estudiar? Menciona “no podría estar aquí, si es un Likert, debería respetar el formato de 5 respuestas o colocar tal vez al inicio del cuestionario” porque este tipo de escala mide el grado de acuerdo o desacuerdo con una afirmación, o la frecuencia de algo. Si se aplicara escala Likert ninguna de esas opciones se ajusta a la naturaleza binaria de la pregunta sobre el acceso a una computadora. Una persona tiene acceso o no tiene acceso; no hay un espectro de "acuerdo" o "frecuencia" aplicable en este caso.

Dentro del factor socioeconómico en la dimensión trabajo y dedicación al estudio el Ítem Q24 ¿Trabajas mientras estudias? Menciona: que debería formularse de otra manera: Trabajas y estudias o buscar formular opciones (100% dedicado al estudio, 75% estudio y 25 % trabajo ...) la opción de formulación incluyendo los porcentajes recomendados, quedaría de la siguiente manera: "¿Cuál de las siguientes opciones describe mejor tu situación actual?"

- 100% dedicado al estudio
- Mayormente dedicado al estudio (ej. 75% estudio, 25% trabajo)
- Dedicación equitativa (ej. 50% estudio, 50% trabajo)
- Mayormente dedicado al trabajo (ej. 25% estudio, 75% trabajo

- 100% dedicado al trabajo (y actualmente no estudio)
- No trabajo y no estudio actualmente

Dentro del factor psicológico y motivacional en la dimensión bienestar y salud en el ítem Q27 ¿Ha sentido ansiedad o depresión relacionada con sus estudios?, Menciona que aquí debería decir "causada u ocasionada por sus estudios", quedando de la siguiente manera: "¿Ha sentido ansiedad o depresión causada por sus estudios?"

En el factor psicológico y motivacional en la dimensión apoyo familiar y social, en la

Q30 ¿Cuentas con apoyo familiar para seguir estudiando? Menciona que “aquí la respuesta es sí o no, la pregunta tendría que cambiar a "¿cuándo te apoya tu familia para seguir estudiando?" no se realizaron cambios, porque la pregunta es confusa.

El juez 4, Menciona que en el factor académico en la dimensión satisfacción con la enseñanza en el Ítem Q10 ¿Cuál es tu grado de satisfacción con los métodos de enseñanza? ¿Comenta, considero la pregunta es ambigua, que significa satisfacción con la enseñanza? me caen bien los maestros, mis compañeros, el lugar etc. Con la finalidad de no ser ambigua la pregunta se replantea de la siguiente manera:"¿En qué medida consideras que los métodos de enseñanza utilizados por tus profesores te ayudan a comprender los contenidos de las materias?" y de esta manera ya no es ambigua.

En la dimensión de carga académica y carga académica con dificultades, siendo los Ítems Q13 ¿Qué tan presionado te sientes por la carga académica? y el Q14 ¿Consideras que la carga académica es excesiva? el juez menciona: A mi parecer esta pregunta es reiterativa con la anterior, además se considera excesiva la carga académica es apreciativa, los planes y programas están diseñados para que "un sujeto común" pueda desarrollarlos**,** sin embargo se considera que los planes de estudio están diseñados**,** personalmente ya que la cantidad de trabajo requerido en algunas asignaturas es considerable**.** Esto a veces genera niveles de estrés que dificultan el proceso de aprendizaje en los alumnos**."**

En el factor socioeconómico en la dimensión ingreso familiar y situación económica, en el Ítem Q19 ¿Cúal es el ingreso mensual de tu familia? Menciona:

Pregunta que debería ir enfocada a ¿“el ingreso familiar es suficiente para cubrir las necesidades de cada uno de sus miembros? la cantidad es poco informativo, sugiere replantear la pregunta mencionada anteriormente.

En el factor socioeconómico en la dimensión recursos tecnológicos y acceso a internet, en el Ítem Q22 ¿Tienes acceso a una computadora personal para estudiar? Menciona que es una herramienta base para cualquier estudiante, desde los estudios de licenciatura. Se considera imprescindible la herramienta.

En el factor institucional en el Ítem Q40 ¿Recibes retroalimentación frecuentemente sobre tu desempeño académico? Menciona: La retroalimentación no tendría que ser frecuente sino más que nada que estuviera a la disposición del alumno cuando fuera necesario, tal vez el termino seria periódica en lugar de frecuente.

El juez 5, no comenta ninguna observación.

El juez ID 6, en el factor académico, en el Ítem Q8 ¿Cómo calificarías tu nivel de comprensión de los contenidos en tu carrera? Y Q10 ¿Cuál es tu grado de satisfacción con los métodos de enseñanza? Menciona: Considero importante la

atención a la respuesta que se obtenga de estas preguntas ya que es de suma importancia en el proceso de aprendizaje de los alumnos, así como de los métodos y tipos de enseñanza a implementar para que los conocimientos adquiridos sean favorables.

Muchas veces la sobrecarga, el estrés y la carga académica excesiva puede generar altos niveles de estrés, ansiedad y agotamiento en los estudiantes. Esto puede disminuir su motivación, concentración y capacidad de retención, afectando negativamente su rendimiento académico general, por eso es importe saber que tanto lo comprenden.

En el factor académico, en la dimensión experiencia académica, en el Ítem Q11 ¿Consideras que la comunicación entre profesores y estudiantes es efectiva? Menciona que es de suma importancia. Con esto se puede valorar si fuera necesario la toma de cursos o talleres (a profesores), con el fin de mejorar la calidad de enseñanza, así como la dinámica que se crea en cada grupo, entre profesores y alumnos. Es una buena recomendación para mejora la forma de enseñanza - aprendizaje de los docentes hacia los alumnos con ello garantizando calidad en la educación.

En el factor académico, en la dimensión carga académica y dificultades, en el Ítem Q14 ¿Consideras que la carga académica es excesiva? Menciona: La pregunta da oportunidad de saber precisamente si existe algún problema y consideran que es excesivo.

En el factor académico, en la dimensión interacción y apoyo en el Ítem Q15 ¿Consideras que la escuela ofrece suficiente apoyo para los estudiantes con dificultades académicas? Menciona: De nuevo la respuesta obtenida es importante para la creación de estrategias, técnicas o dinámicas que promuevan ambientes saludables, de empatía, respeto y atención a sus necesidades alumno- profesor o profesora. Es importante para mejorar la calidad de la educación y compromiso del Tecnológico.

En el factor dedicación al estudio, en la dimensión tiempo de estudio individual en el Ítem Q16 ¿Cuántas horas de estudio adicionales dedicas fuera de tu horario académico? Menciona: No considero que tenga mucha relevancia.

En el factor académico, en la dimensión desarrollo de habilidades e intereses, en el Ítem Q18 ¿Consideras que la escuela te brinda oportunidades para desarrollar tus talentos y habilidades? Menciona: Es relevante saber si los alumnos se sienten acompañados y motivados a desarrollarse.

En el factor socioeconómico, en la dimensión recursos tecnológicos y acceso a internet en el Ítem Q22 ¿Tienes acceso a una computadora personal para estudiar? Menciona: Es preciso saber con qué herramientas sé que cuenta.

En el factor socioeconómico en la dimensión trabajo y dedicación al estudio en los Ítems Q25 Si trabajas, ¿cuántas horas laboras por semana? Menciona: Las preguntas 37-40 a mi parecer no representan motivo para poder evaluar el aprendizaje, hay chicos en esa situación que tienen un buen desempeño académico.

En el factor psicológico y motivacional, en la dimensión bienestar y salud en el Ítem Q26 ¿Ha tenido problemas de salud que afectan su rendimiento académico? Menciona: Es importante saber si hay algún tema médico, para informar a los profesores y si se tienen que tomar medidas especiales para tomar sus clases o justificar faltas (más no actividades) por algún tema médico con la finalidad de no perjudicarlos.

En el factor psicológico y motivacional, en la dimensión bienestar y salud en el Ítem Q27 ¿Ha sentido ansiedad o depresión relacionada con sus estudios? Menciona:

Si considero importante la pregunta ya que he visto casos de estrés sobre todo en periodos de exámenes o evaluaciones.

El Juez ID 7 comenta que, en el factor académico, en el Ítem Q6 ¿Cuál fué tu promedio en el último semestre? Menciona Si está dirigido a alumnos del TecNM considerar que la mínima aprobatoria es de 70, por tanto, podría fusionar la respuesta 1 y 2. Se hace la observación ya que se tienen como respuestas en escala Likert "5 o Menos: Reprobado; 6: Suficiente; 7: Aprobado; 8: Bien; 9: Notable y 10: Sobresaliente"

En el factor de Interacción y comunicación en la dimensión experiencia académica en el Ítem Q15 Q11 ¿Te sientes capaz de comunicar efectivamente con tus compañeros y profesores sobre tus necesidades académicas? recomienda replantearla de esta manera ¿Te sientes capaz de comunicarte efectivamente con tus compañeros y profesores sobre tus necesidades académicas?

En el factor socioeconómico en la dimensión recursos tecnológicos y acceso a internet en el Ítem Q22 ¿Tienes acceso a una computadora personal para estudiar? Menciona: La respuesta es dicotómicas, ¿no afecta que las otras sean de 5 opciones? Además, muchos alumnos

utilizan sus celulares o tabletas para elaborar sus trabajos una ellos se les está excluyendo.

El juez ID 8 comenta que, en el factor académico el ítem Q6 ¿Cuál fué tu promedio en el último semestre? Menciona: los promedios deben ser considerados si ya cursaron un semestre anterior, ya que, si son de nuevo ingreso, no habría comparativa. Para esta pregunta, se están considerando alumnos a partir del segundo semestre.

En el factor académico en la dimensión motivación y permanencia, en el Ítem Q17 ¿Has pensado en abandonar tus estudios académicos? Menciona. Aquí se deben considerar las tutorías como relevantes. Es una manera de detectar a tiempo la deserción en algún alumno.

En el factor socioeconómico en la dimensión recursos tecnológicos y acceso a internet en el Ítem Q22 ¿Tienes acceso a una computadora personal para estudiar? Menciona: No necesariamente se requiere de computadoras para llevar a cabo las actividades escolares, ya que existen otros tipos de dispositivos. Esta respuesta concuerda con el Juez ID 7

En el factor psicológico y motivacional en la dimensión bienestar y salud en el Ítem Q28 ¿Te sientes capaz de manejar el estrés relacionado con la escuela? Menciona: Quizás la pregunta podría enfocarse como: ¿Te sientes capaz de manejar el estrés relacionado con tus actividades académicas o escolares?

En el factor institucional. El ítem Q40 ¿Recibes retroalimentación frecuentemente sobre tu desempeño académico? Menciona: La pregunta podría ser: ¿La retroalimentación que recibes impacta en tu desempeño académico?

Para que un instrumento sea idóneo y que se pueda utilizar con la confianza requerida tiene que cumplir dos requisitos: fiabilidad y validez (González, 2008).

ANEXO F

Se presentan los requerimientos funcionales para asegurar el óptimo funcionamiento y los resultados deseados del sistema de programación. Se detallan a continuación los requisitos funcionales asociados a cada librería:

- *pandas* (Importado como pd): para gestión y manipulación de datos: Capacidad para cargar, estructurar y transformar conjuntos de datos tabulares. Esto incluye operaciones como lectura de archivos (CSV, Excel), filtrado, agrupación, fusión y pivoteo de datos.

- *sklearn.preprocessing.OneHotEncoder*: para codificación de variables categóricas: Funcionalidad para transformar características categóricas (texto o etiquetas discretas) en un formato numérico binario (*one-hot encoding*), adecuado para el procesamiento por algoritmos de aprendizaje automático.
- *sklearn.decomposition.PCA*: para la reducción de dimensionalidad: Posibilidad de aplicar el Análisis de Componentes Principales para reducir el número de características de un conjunto de datos, manteniendo la mayor parte de su varianza, lo que optimiza el rendimiento computacional y facilita la visualización.
- *sklearn.cluster.KMeans*: para el agrupamiento (*Clustering*) de datos: Implementación del algoritmo K-Means para agrupar observaciones similares en clústeres, basándose en la distancia entre los puntos de datos.
- *sklearn.metrics (con calinski_harabasz_score, silhouette_score, silhouette_samples*): en evaluación de modelos de agrupamiento: Provisión de métricas como el índice Calinski-Harabasz y el coeficiente de silueta para cuantificar la calidad y coherencia de los clústeres generados.

- *matplotlib.pyplot* (Importado como plt): para la visualización de datos: Habilidad para generar gráficos estáticos, interactivos y animados, incluyendo diagramas de dispersión, histogramas, gráficos de barras y de líneas, para la exploración y presentación de datos.

- *numpy* (Importado como np): en computación numérica de alto rendimiento: Soporte para operaciones avanzadas con arreglos y matrices multidimensionales, fundamental para cálculos matemáticos y estadísticos eficientes.

- *sklearn.model_selection.train_test_split*: para la división de conjuntos de datos: Capacidad para dividir un conjunto de datos en subconjuntos de entrenamiento y prueba, lo cual es esencial para evaluar la generalización de un modelo de aprendizaje automático.
- *sklearn.model_selection.RandomizedSearchCV*: para la optimización de hiperparámetros (búsqueda aleatoria): Funcionalidad para realizar una búsqueda aleatoria de los mejores hiperparámetros para un modelo, explorando un espacio de parámetros definido y mejorando el rendimiento del modelo.

- *sklearn.ensemble.RandomForestClassifier*: para la clasificación predictiva (Random Forest): Implementación de un clasificador basado en el algoritmo Random Forest, capaz de manejar problemas de clasificación complejos y ofrecer alta precisión.

- *scipy.stats.randint*: para la generación de distribuciones discretas: Provisión de funcionalidades para generar números enteros aleatorios a partir de una distribución uniforme discreta, útil en la definición de rangos de parámetros para búsquedas aleatorias.

- *sklearn.metrics (con confusion_matrix, ConfusionMatrixDisplay, accuracy_score, precision_score, recall_score, f1_score, classification_report*): para la evaluación exhaustiva de modelos de clasificación: Conjunto de herramientas para calcular y visualizar métricas clave de rendimiento para modelos de clasificación, incluyendo matriz de confusión, precisión, exhaustividad *(recall*) y puntaje F1.

- *sklearn.model_selection.GridSearchCV*: para la optimización de hiperparámetros (búsqueda en cuadrícula): Capacidad para realizar una búsqueda exhaustiva en una cuadrícula predefinida de hiperparámetros, con el fin de identificar la combinación óptima para un modelo.

- *sklearn.svm.SVC*: para la clasificación predictiva (Máquinas de Vectores de Soporte): Implementación de Máquinas de Vectores de Soporte para tareas de clasificación, destacando por su eficacia en espacios de alta dimensionalidad.

- *sklearn.neural_network.MLPClassifier*: para la clasificación predictiva (Redes Neuronales): Provisión de un clasificador basado en una red neuronal multicapa (MLP), apto para aprender patrones complejos y realizar clasificaciones.

Estos requisitos delinean las capacidades que el sistema debe tener al utilizar estas herramientas, permitiendo desde la preparación de datos hasta la construcción, optimización y evaluación de modelos de aprendizaje automático.