# “PROYECTO RETENIENDO FUTUROS: UN ENFOQUE PREDICTIVO CON MACHINE LEARNING PARA MEJORAR LA RETENCIÓN ESTUDIANTIL”

Sandra Paola Botero Ramírez

*Proyecto Aplicado para optar al título de Magister en*

*Ciencia de Datos*

Director

David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS, MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, FEBRERO DE 2025

# RESUMEN

La deserción estudiantil universitaria constituye un desafío global con repercusiones negativas en el desarrollo social y científico a nivel nacional o regional. Las Instituciones de Educación Superior (IES) asumen la responsabilidad de abordar y prevenir este problema. Este estudio presenta un marco conceptual de la deserción universitaria, fundamentado en investigaciones que emplean enfoques cualitativos y cuantitativos, haciendo uso de la ciencia de datos.

En este contexto, se lleva a cabo un análisis exploratorio descriptivo de los datos recopilados mediante el instrumento de caracterización correspondiente a los periodos académicos desde 2017-1 hasta 2021-2. El análisis se centra en comprender y examinar el fenómeno de la deserción entre los estudiantes que contestaron la encuesta de caracterización en dichos periodos.

Finalmente, se procede a entrenar diversos modelos de machine learning, entre los que se incluyen la regresión logística, las máquinas de soporte vectorial, los bosques aleatorios de decisión y las redes neuronales simples. Estos modelos tienen la capacidad de prever y emitir alertas sobre posibles riesgos de deserción en los programas académicos de la universidad. Este enfoque proactivo permite a las instituciones tomar medidas preventivas y proporcionar apoyo personalizado a los estudiantes en riesgo, contribuyendo así a mejorar las tasas de retención y el éxito académico.

***Palabras claves:*** *Predicción, Deserción, Aprendizaje supervisado, clasificación, Machine Learning, Modelos Predictivos*

# TABLA DE CONTENIDO


INTRODUCCIÓN ..... 9
1. DEFINICIÓN DEL PROBLEMA ..... 10
1.1 Planteamiento del problema ..... 10
1.2 Formulación del problema ..... 11
1.3 Pregunta de investigación ..... 11
2. OBJETIVOS DEL PROYECTO ..... 12
2.1 Objetivo general ..... 12
2.2 Objetivos específicos ..... 12
3. MARCO DE REFERENCIA ..... 13
3.1 Marco teórico ..... 13
3.1.1 Deserción estudiantil ..... 13
3.1.1.1 Deserción estudiantil en la educación superior ..... 14
3.1.1.2 Factores claves para la deserción universitaria ..... 14
3.1.2 Machine Learning (ML) ..... 15
3.1.2.1 Categorías y técnicas de Machine Learning ..... 16
3.1.3 Regresión Logística ..... 17
3.1.3.1 Regresión Logística aplicada a la predicción de la deserción estudiantil ..... 18
3.1.3.2 Redes Neuronales ..... 19
3.1.3.2.1 Redes neuronales aplicadas a la predicción de la deserción estudiantil ..... 19
3.1.3.3 Arboles de Decisión ..... 20
3.1.3.3.1 Arboles de decisión aplicados a la predicción de la deserción estudiantil ..... 20
3.1.3.4 Máquina de Soporte Vectorial (SVM) ..... 21
3.1.3.5 Bosques aleatorios – Random Forest (RF) ..... 21
3.1.3.5.1 Bosques aleatorios aplicados a la predicción de la deserción estudiantil ..... 21
4. ANTECEDENTES ..... 23
4.1 Sistema de prevención y análisis de la deserción en las instituciones de educación superior ..... 23
4.2 Factores que motivan el abandono estudiantil en la universidad ..... 23
4.3 La investigación sobre deserción universitaria en Colombia 2006-2016 ..... 24
4.4 Propuesta de un modelo predictivo utilizando aprendizaje profundo ..... 24
4.5 Investigación en deserción estudiantil universitaria ..... 25
4.6 Predicting student's dropout in university classes using two-layer ensemble machine Learning

approach: a novel stacked generalization ..... 25
4.7 predicting student drop-out rates using data mining techniques ..... 26
4.8 Hacia la construcción de un modelo predictivo de deserción académica ..... 26
*4.9* Modelo predictivo de deserción estudiantil basado en arboles de decisión ..... 27
5. METODOLOGÍA ..... 28
5.2 Ecosistema analítico ..... 28
5.3 Planeación ..... 29
5.4 Desarrollo del modelo ..... 29
5.5 Implementación ..... 30
5.6 Presentación de resultados ..... 30
6. ANÁLISIS EXPLORATORIO DE DATOS E IDENTIFICACIÓN DE VARIABLES RELEVANTES PARA LA DESERCIÓN ESTUDIANTIL ..... 31
6.1 Conjunto de datos (Instrumento de caracterización) ..... 31
6.1.1 Modelo bidimensional de datos ..... 33
6.1.2 Preparación y limpieza de datos ..... 34
6.2 Análisis exploratorio de datos ..... 36
6.2.1 Análisis Univariado ..... 36
6.2.2 Análisis Bivariado ..... 43
6.2.3 Selección de variables para el modelo ..... 48
7. ANÁLISIS COMPARATIVO DE MODELOS DE MACHINE LEARNING PARA PREDECIR LA DESERCIÓN ESTUDIANTIL ..... 53
7.1 División de los datos en entrenamiento ..... 53
7.2 Preprocesado de los datos ..... 54
7.3 Modelo de regresión logística ..... 55
7.4 Modelo de máquina de soporte vectorial (SVM) ..... 57
7.5 Modelo de bosques aleatorios (Random Forest) ..... 61
7.6 Modelo de redes neuronales simple (NNET) ..... 65
8. EVALUACIÓN DE DESEMPEÑO EN MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE DESERCIÓN ..... 67
8.1. Comparación de métricas de los modelos ..... 67
CONCLUSIONES ..... 69
REFERENCIAS ..... 70
ANEXOS ..... 74

# LISTA DE FIGURAS

Figura 1. Modelo bidimensional del análisis de la deserción ..... 29
Figura 2. Distribución de la variable objetivo antes del balanceo ..... 32
Figura 3. Distribución de la variable objetivo después del balanceo ..... 33
Figura 4. Máximo semestre cursado ..... 33
Figura 5. Desertor ..... 34
Figura 6. Edad ..... 35
Figura 7. Estrato ..... 35
Figura 8. Boxplot de variables numéricas ..... 36
Figura 9. Sexo ..... 37
Figura 10. Distribución Estado civil ..... 37
Figura 11. Distribución Región ..... 38
Figura 12. Distribución Zona (Urbana / Rural) ..... 38
Figura 13. Matriz de correlación ..... 39
Figura 14. distribución de edad según DESRTOR ..... 39
Figura 15. Distribución estrato socioeconómicos según DESERTOR ..... 40
Figura 16. Tasa de deserción por programa académico ..... 40
Figura 17. Comparación de género y deserción ..... 41
Figura 18. Clúster 0: Bajo rendimiento académico + problemas sociales ..... 42
Figura 19. Clúster 1: Factores económicos + familiares ..... 42
Figura 20. Clúster 2: Casos aislados (otros factores) ..... 43
Figura 21. Matriz de confusión - Regresión logística ..... 45
Figura 22. Matriz de Confusión ..... 46
Figura 23. Matriz de confusión - Random Forest ..... 47
Figura 24. Matriz de confusión - Neural Network ..... 48
Figura 25. Comparación de métricas de los modelos ..... 49
Figura 26. Friedman Test ..... 50
Figura 27. Wilcoxon Pairwise Comparisons ..... 51

# LISTA DE TABLAS

Tabla 1. Técnicas de Machine Learning.....17

Tabla *2*. Fases del Proyecto.....28

Tabla 3. Detalle fuente de datos caracterización socio demográfica (Fuente: Elaboración propia) .....31

Tabla 4. Detalle fuente de datos caracterización académica (Fuente: Elaboración propia).....31

Tabla *5*. Detalle fuente de datos caracterización Financiera (Fuente: Elaboración propia) .....31

Tabla *6*. Cantidad de estudiantes por programa académico que desertaron (Fuente: Elaboración propia).32

Tabla 7. Métricas de accuracy y ecuaciones por cada modelo de regresión (Fuete: elaboración propia) .....50

Tabla 8. Resumen de los datos estandarizados y binarizados (Fuente: elaboración propia) .....54

Tabla 9. presencia de variables con varianza cercana a cero (Fuente: elaboración propia) .....55

# INTRODUCCIÓN

En Colombia, anualmente, aproximadamente 2.2 millones de estudiantes se matriculan, lo que representa el 50% de los jóvenes que completan su bachillerato cada año en el país. Sin embargo, de estas cifras, al menos el 50% de los alumnos abandonan sus estudios antes de finalizarlos, según informes de la Revista Semana en 2022.

Este fenómeno destaca la importancia de las instituciones de educación superior en el país, las cuales desempeñan un papel crucial en el desarrollo de la sociedad al asegurar la graduación de estudiantes con estándares de calidad, oportunidad e inclusión.

No obstante, este compromiso se ve desafiado por un problema significativo: la deserción universitaria, que impacta negativamente en el progreso social y científico de los países [1]
En este contexto, la deserción universitaria se posiciona como uno de los mayores desafíos actuales para las universidades en Colombia. Este desafío exige la implementación de mecanismos eficaces de retención estudiantil para asegurar que los estudiantes no solo ingresen a la educación superior, sino que también culminen con éxito sus estudios, contribuyendo así al desarrollo integral del país [1].

Con el objetivo de abordar esta necesidad, se realiza una investigación sobre el fenómeno de la deserción universitaria. Este proyecto se lleva a cabo en la Universidad Javeriana Cali, utilizando específicamente técnicas de aprendizaje automático y estadísticas.

# 1. DEFINICIÓN DEL PROBLEMA

## 1.1 Planteamiento del problema

La deserción universitaria constituye un desafío global que impacta a todas las regiones del mundo. Según el Programa Piloto para la Prevención de la Deserción universitaria (SDPP) regional en Asia y Oriente Medio de USAID, los patrones globales indican que cuanto más temprano se produce la deserción y menos años de estudio se tienen, mayor es el impacto negativo en los estudiantes [2]

De acuerdo con el Banco Interamericano de Desarrollo (BID), el abandono escolar se asocia con efectos negativos de largo alcance, que incluyen la disminución de ingresos a lo largo de la vida, una mayor dificultad para acceder al empleo formal, un incremento en el riesgo de involucrarse en conductas delictivas y antisociales, así como en el consumo de sustancias que afectan la salud [3].

El Ministerio de Educación Nacional (MEN), responsable de proporcionar datos precisos para el seguimiento de la deserción universitaria en todos los niveles educativos, introdujo en 2018 una nueva versión de su sistema, conocida como SPADIES 3.0. Esta actualización integra el Sistema Nacional de Información de Educación Superior (SNIES) con el objetivo de mejorar la precisión y calidad de la información, así como de estudiar la movilidad de los estudiantes entre programas y carreras universitarias. Al comparar los dos sistemas, se evidencia una variación en las tasas de deserción universitaria, pasando del 9,89% al 8,25%, siendo esta última correspondiente al año 2019. Estas cifras reflejan una situación preocupante que afecta a los sistemas educativos y tiene un impacto negativo en la permanencia y graduación de los estudiantes.

La deserción universitaria es un problema significativo que afecta a las instituciones de educación superior, incluyendo la Pontificia Universidad Javeriana Cali. Este fenómeno tiene implicaciones negativas tanto a nivel individual como institucional, afectando el desarrollo académico y social de los estudiantes y la reputación y eficiencia de la universidad. Identificar y predecir la deserción estudiantil se convierte en una prioridad para implementar medidas preventivas y de apoyo.

## 1.2 Formulación del problema

El objetivo principal de este trabajo es desarrollar y evaluar modelos predictivos que permitan anticipar la deserción estudiantil a través de la caracterización de los estudiantes basada en sus datos académicos. Para ello, se utilizarán registros desde el 2017-1 al 2021-2 que incluyen variables relacionadas con factores socioeconómicos y comportamentales y también se usaran algunos datos académicos.

Estas variables serán empleadas para ajustar y entrenar diferentes modelos capaces de estimar las probabilidades de deserción de forma individual. Posteriormente, se llevará a cabo una evaluación rigurosa de los modelos desarrollados, enfocándose en su precisión y eficacia en la predicción de la deserción estudiantil.

## 1.3 Pregunta de investigación:

¿Qué modelo de aprendizaje automático ofrece la mejor precisión en la predicción de la deserción estudiantil en la Pontificia Universidad Javeriana Cali, y cuáles son las variables más influyentes en esta predicción?

# 2. OBJETIVOS DEL PROYECTO

## 2.1 Objetivo general

- Desarrollar y evaluar varios modelos de aprendizaje automático para predecir la deserción de estudiantes en la Pontificia Universidad Javeriana Cali.

## 2.2 Objetivos específicos

- Identificar las variables relevantes que inciden en la deserción estudiantil a través del análisis exploratorio de datos.
- Desarrollar y comparar varios modelos de machine learning para la predicción de la deserción estudiantil, utilizando técnicas de validación cruzada para evaluar el desempeño de cada modelo.
- Evaluar el modelo predictivo que tenga el mejor desempeño, basándose en métricas clave de evaluación.

# 3. MARCO DE REFERENCIA

## 3.1 Marco teórico

### 3.1.1 Deserción estudiantil

La deserción estudiantil es un fenómeno complejo que se manifiesta cuando los jóvenes abandonan sus estudios debido a una combinación de factores sociales, económicos, y familiares. Entre las causas sociales, destacan los problemas legales, la participación en actividades delictivas, y la influencia de comportamientos perjudiciales, muchas veces vinculados a la falta de apoyo familiar. Esta carencia de respaldo puede empujar a los jóvenes hacia entornos negativos que desvían su atención de la educación y fomentan conductas inapropiadas.

Un factor determinante en la deserción estudiantil es la situación socioeconómica de las familias, especialmente en contextos de pobreza. Las familias menos privilegiadas a menudo deben priorizar la supervivencia económica, lo que lleva a que sus hijos abandonen la escuela para trabajar y contribuir al sustento del hogar. Esta decisión, aunque comprensible en términos de necesidad, pone en riesgo la continuidad educativa de los jóvenes, quienes, al intentar combinar trabajo y estudio, se enfrentan a mayores dificultades que generalmente los llevan a abandonar sus estudios.

La deserción estudiantil también está influenciada por factores adicionales como la desmotivación, el bajo rendimiento académico y los problemas de asistencia, especialmente en el ámbito universitario. La falta de interés y atención a las directrices académicas, sumada a la presión de factores externos, puede llevar a los estudiantes a interrumpir su trayectoria educativa. Es crucial que los docentes y las instituciones educativas trabajen de manera proactiva para identificar señales de posible deserción y ofrecer alternativas que fomenten la retención y el éxito académico de los estudiantes [4].

#### 3.1.1.1 Deserción estudiantil en la educación superior

La deserción estudiantil, definida por el Ministerio de Educación Nacional (MEN) [5], como el abandono del sistema educativo durante dos o más periodos consecutivos, es un fenómeno influenciado por factores dentro del sistema educativo y en contextos sociales, familiares e individuales. Esta problemática tiene un impacto global, afectando los esfuerzos para mejorar el nivel educativo y la competitividad de los estudiantes, así como su integración en la sociedad del conocimiento.

Para enfrentar este desafío, es esencial que las instituciones de educación superior se comprometan a comprender y abordar las dificultades asociadas a la deserción es. La colaboración entre instituciones educativas, autoridades y la comunidad es clave para construir un entorno educativo inclusivo que promueva el desarrollo integral de los estudiantes, permitiendo así una sociedad más educada y equitativa [5].

#### 3.1.1.2 Factores claves para la deserción universitaria

La deserción en instituciones de educación superior resulta de la interacción compleja de factores sociales, económicos, políticos, institucionales, personales y académicos. Las investigaciones señalan que la desigualdad socioeconómica, la insuficiente preparación académica previa, problemas familiares, la falta de actualización de los programas, y el bajo rendimiento académico son las principales causas. En América Latina, además, el embarazo adolescente contribuye significativamente a la deserción, afectando principalmente a las mujeres, quienes enfrentan menores posibilidades de completar sus estudios, lo que genera consecuencias negativas tanto para ellas como para las instituciones educativas.

La falta de motivación del estudiante hacia las asignaturas y programas académicos puede ser otra causa de deserción, destacando la importancia de los enfoques pedagógicos y programas de estudio de las instituciones educativas. Generar un cambio positivo en la mentalidad de los estudiantes y abordar sus desafíos personales, así como académicos, es esencial para prevenir la deserción estudiantil. En este sentido, la comunidad educativa debe mostrar interés y abordar

estos desafíos a través de actividades que fomenten la integración y el desarrollo personal.

En esta lógica y de acuerdo con los enfoques teóricos a nivel mundial, las causas de la deserción en la educación superior en Colombia que retoma son [5]:

| SOCIOECONOMICAS | ACADÉMICAS | INDIVIDUALES | INSTITUCIONALES |
|---|---|---|---|
| ✓ Miedo al endeudamiento por parte de los estudiantes o de sus padres. | ✓ Falta de preparación desde la educación media en competencias generales | ✓ La edad de inicio de los alumnos incide positivamente cuando son menores de edad. | ✓ Falta de preparación desde la educación media en competencias generales |
| ✓ Subestimar los costos de estudiar un programa de pregrado. | ✓ Poca orientación profesional y vocacional antes del ingreso a la universidad | ✓ Las personas casadas debido a la menor disponibilidad de tiempo. | ✓ Poca orientación profesional y vocacional antes del ingreso a la universidad |
| ✓ Pertenecera estrato bajo. | ✓ Bajo rendimiento académico | ✓ Presiones familiares y sociales Costos monetarios y de tiempo que se deben afrontar al estudiar en otra ciudad. | ✓ Bajo rendimiento académico |
| ✓ Bajos ingresos familiares y desempleo de los padres. | ✓ Baja calidad del programa al que se accede Métodos de estudio y metodologías de aprendizaje obsoletas | ✓ Calamidad y problemas de salud Discriminación social por razones de orientación sexual o raza. | ✓ Baja calidad del programa al que se accede Métodos de estudio y metodologías de aprendizaje obsoletas |
| ✓ Dependenciaeconómica de sí mismo. | ✓ Insatisfacción con el programa | ✓ Incompatibilidadhoraria con actividades extracurriculares | ✓ Insatisfacción con el programa |
| ✓ Nivel educativo bajo de los padres (ninguno o primaria). | ✓ Estrés por la carga académica | ✓ Expectativas no satisfechas o no le encuentran un futuro a lo que están estudiando. | ✓ Estrés por la carga académica |

***Fuente:*** *SPADIES [3]*

### 3.1.2 Machine Learning (ML)

El aprendizaje automático (ML) es una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos. Su aplicación práctica consiste en imitar la forma en que los humanos aprenden y mejorar gradualmente su precisión [6]. Otros autores, como lo definen como un conjunto de métodos que pueden detectar automáticamente patrones en los datos y, utilizando esos patrones, descubrir patrones futuros en los datos para respaldar la toma de decisiones en las organizaciones.

El sistema de aprendizaje de un algoritmo de aprendizaje automático en tres partes principales:

- **Un proceso de decisión:** En general, los algoritmos de aprendizaje automático se utilizan para hacer una predicción o clasificación. Basándose en unos datos de entrada, que pueden estar etiquetados o sin etiquetar, su algoritmo producirá una estimación sobre un patrón en los datos.

- **Una función de error:** Una función de error sirve para evaluar la predicción del modelo. Si hay ejemplos conocidos, una función de error puede hacer una comparación para evaluar la precisión del modelo [6].

- **Un proceso de optimización del modelo:** Si el modelo puede ajustarse mejor a los puntos de datos del conjunto de entrenamiento, se ajustan los pesos para reducir la discrepancia entre el ejemplo conocido y la estimación del modelo. El algoritmo repetirá este proceso de evaluación y optimización, actualizando de manera autónoma los pesos hasta llegar a un umbral de precisión establecido. Este enfoque iterativo permite que el modelo mejore su rendimiento a medida que se le proporciona más información y datos de entrenamiento. Con cada iteración, se busca una mayor precisión y capacidad de generalización del modelo, lo cual resulta fundamental para su utilidad y aplicabilidad en diferentes escenarios y problemas [6].

#### 3.1.2.1 Categorías y técnicas de Machine Learning

El aprendizaje automático se divide principalmente en dos tipos: supervisado y no supervisado. El aprendizaje supervisado se enfoca en encontrar la relación entre entradas ("X") y salidas ("Y") utilizando un conjunto de datos etiquetados para entrenar el modelo, permitiéndole predecir resultados o clasificar nuevos datos con precisión [6].

En contraste, el aprendizaje no supervisado solo trabaja con entradas sin etiquetas, buscando patrones y estructuras ocultas en los datos. Este enfoque es ideal para análisis exploratorios, segmentación de clientes y reconocimiento de patrones, ya que no requiere una métrica de error específica ni intervenciones humanas.

A continuación, se describen algunas de las técnicas más usadas en el aprendizaje automático,

*Tabla 1. Técnicas de Machine Learning*

| Técnica de ML | Descripción |
|---|---|
| Redes neuronales (con diferentes parámetros) | "Una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas" [14]. |
| Support Vector Machine | "SVM funciona asignando datos a un espacio de características de alta dimensión para que los puntos de datos se puedan clasificar, incluso cuando los datos no se pueden separar linealmente" [14]. |
| Gradient Boosting | "Es un método de aprendizaje conjunto que combina un conjunto de aprendices débiles en un aprendiz fuerte para minimizar los errores de entrenamiento. En el boosting, se selecciona una muestra aleatoria de datos, se le aplica un modelo y se entrena secuencialmente, es decir, cada modelo intenta compensar las debilidades de su predecesor [15]. |
| Random Forest | "El bosque aleatorio es un algoritmo de aprendizaje automático comúnmente utilizado y registrado por Leo Breiman y Adele Cutler, que combina la salida de múltiples árboles de decisión para llegar a un único resultado" [15]. |
| Regresión logística | "Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo" [14]. |

El aprendizaje automático (ML) se sitúa como un componente esencial en el ámbito de la ciencia de datos, empleando métodos estadísticos para entrenar algoritmos que generan resultados como clasificaciones o predicciones. Estos resultados a su vez permiten descubrir ideas clave en proyectos de minería de datos. Los nuevos conocimientos obtenidos a través del ML tienen un impacto significativo en la toma de decisiones dentro de las organizaciones, contribuyendo positivamente a la generación de valor, la mejora de los indicadores clave de rendimiento y, en general, al crecimiento de las compañías [6].

### 3.1.3 Regresión Logística

La regresión logística es una técnica estadística fundamental en campos como la medicina, psicología, economía, ingeniería y ciencia de datos. Su propósito es analizar la relación entre una variable dependiente binaria y una o más variables independientes, sean estas continuas o categóricas, para predecir la probabilidad de que ocurra un evento específico. Utilizando la

función logística, que transforma una variable lineal en una probabilidad entre 0 y 1, esta técnica permite cuantificar cómo influyen las variables independientes en la probabilidad del evento, facilitando así la toma de decisiones y el análisis en diversas aplicaciones.

Joseph Berkson [17], y David Cox fueron pioneros en el desarrollo de la regresión logística, realizando contribuciones clave en la formulación matemática y en la estimación de coeficientes mediante máxima verosimilitud. El libro "Applied Logistic Regression" de Hosmer y Lemeshow es una referencia esencial para comprender y aplicar esta técnica, proporcionando una introducción detallada y abordando temas como la selección de variables y la validación del modelo. Según James et al., la regresión logística modela la probabilidad de eventos binarios en función de variables explicativas, siendo útil en aplicaciones como la predicción de deserción universitaria, análisis de riesgos médicos y evaluación del crédito [6].

#### 3.1.3.1 Regresión Logística aplicada a la predicción de la deserción estudiantil

La regresión logística es una técnica estadística clave para predecir la deserción estudiantil universitaria, un problema global con serias implicaciones para estudiantes, instituciones educativas y la sociedad en general. Esta metodología ha demostrado ser efectiva para identificar las variables que afectan la deserción y desarrollar modelos predictivos precisos.

En muchos estudios se ha utilizado la regresión logística para anticipar la deserción universitaria, identificando variables significativas como la edad, el género, el nivel de ingresos familiares, el tipo de institución, el campo de estudio y el rendimiento académico. El modelo mostró una alta precisión en la predicción, con una tasa de acierto del 83%. En una universidad de Malasia, la cual utilizo este modelo encontraron que variables como la edad, el género, la etnia, el estado civil, el nivel socioeconómico, el tipo de alojamiento, el rendimiento académico y la satisfacción con el entorno educativo eran predictoras importantes de la deserción [7].

#### 3.1.3.2 Redes Neuronales

Las redes neuronales son técnicas de modelado de datos inspiradas en el funcionamiento del cerebro humano. Consisten en una red de nodos interconectados que procesan información y generan salidas basadas en entradas previas. Son ampliamente utilizadas en problemas de clasificación y predicción, como el reconocimiento de patrones, la detección de fraudes y la identificación de imágenes. Su capacidad para analizar grandes volúmenes de datos y detectar patrones complejos ha impulsado su popularidad en campos como la medicina, la economía, la industria, la ciencia de datos y la inteligencia artificial. En la era del big data, las redes neuronales ofrecen un enfoque eficaz para resolver problemas complejos en diversas aplicaciones [8].

Según Haykin, "Las redes neuronales artificiales son modelos matemáticos inspirados en la estructura y función de los sistemas neuronales biológicos, diseñados para emular su capacidad de aprendizaje y generalización en la resolución de problemas." Estas redes están formadas por capas de nodos que procesan y transforman la información a través de conexiones ponderadas, las cuales se ajustan durante el aprendizaje [8].

Rumelhart [21] agrega que "Una red neuronal es una máquina de procesamiento paralelo compuesta por unidades de procesamiento simples que almacenan conocimiento adquirido a través del aprendizaje y lo utilizan para resolver problemas." El aprendizaje en una red neuronal se basa en la modificación de los pesos de las conexiones entre las neuronas, permitiendo que la red ajuste su comportamiento para adaptarse a los datos de entrada [9].

##### 3.1.3.2.1 Redes neuronales aplicadas a la predicción de la deserción estudiantil

Las redes neuronales se han consolidado como una técnica avanzada en la predicción de la deserción estudiantil universitaria, basándose en un modelo computacional inspirado en el cerebro humano y su capacidad de aprendizaje a partir de ejemplos.

Hay varias instituciones de educación superior que han implementado una red neuronal profunda (DNN) para predecir patrones de deserción estudiantil como el estudio que hicieron en la Universidad Pedagógica y Tecnológica de Colombia (UPTC). Utilizaron un conjunto de datos que contenía 17 atributos de 3,000 estudiantes activos, incluyendo variables como

calificaciones de ingreso, asistencia a clases y satisfacción estudiantil. El modelo predictivo entrenado mostró una alta precisión y una baja tasa de falsos negativos en la predicción de deserciones, demostrando la efectividad de las redes neuronales en este ámbito [10].

#### 3.1.3.3 Arboles de Decisión

Los árboles de decisión son una técnica popular en minería de datos y aprendizaje automático, utilizada para tomar decisiones basadas en reglas y variables. Construidos a partir de datos de entrenamiento, estos modelos predicen resultados para nuevas observaciones y se aplican en áreas como la toma de decisiones empresariales y la detección de fraudes. Breiman los describe como "métodos para aproximar funciones que mapean entradas a salidas" [11], y Loh los define como "métodos de clasificación que usan reglas de decisión para asignar etiquetas" [12]. Aunque fáciles de interpretar, pueden sobre ajustarse a los datos de entrenamiento. Para mitigar esto, se utilizan técnicas de poda para eliminar ramas innecesarias, mejorando la capacidad de generalización. En resumen, los árboles de decisión son efectivos para modelar relaciones complejas, pero requieren cuidado para evitar el sobreajuste.

##### 3.1.3.3.1 Arboles de decisión aplicados a la predicción de la deserción estudiantil

La predicción de la deserción estudiantil universitaria es un desafío crucial en la educación superior. Los árboles de decisión, una técnica de modelado de datos popular, se han utilizado con éxito en varios estudios para abordar este problema. A continuación, se presentan tres investigaciones que emplearon árboles de decisión para predecir la deserción universitaria, con diversos porcentajes de precisión.

Alfredo Daza Vergaray implementó un modelo basado en árboles de decisión para predecir la deserción estudiantil en una universidad privada. Utilizando datos de 1,761 estudiantes recopilados entre 2009 y 2013, se analizaron 27 variables relacionadas con la deserción, incluyendo rendimiento académico, participación en actividades extracurriculares y situación financiera. El modelo predictivo desarrollado alcanzó una precisión del 89%, demostrando la efectividad de los árboles de decisión, esta técnica puede ayudar a las instituciones educativas a intervenir de manera temprana y mejorar las tasas de retención estudiantil [13].

#### 3.1.3.4 Máquina de Soporte Vectorial (SVM)

Las Máquinas de Vectores de Soporte (SVM) son algoritmos de aprendizaje automático supervisado desarrollados por Vapnik en los años 90 en Bell Labs, utilizados para clasificación y regresión. Su objetivo es encontrar un hiperplano que optimice la separación entre clases al maximizar el margen, la distancia desde el hiperplano a los vectores de soporte, o puntos de datos más cercanos. SVM puede manejar datos linealmente separables y no linealmente separables mediante trucos de kernel, transformando los datos a espacios de mayor dimensión. Además de la clasificación binaria, se adapta a la clasificación multiclase y la regresión. Aunque es eficaz con datos de alta dimensionalidad y generaliza bien, puede ser sensible a la selección de parámetros y a grandes conjuntos de datos [14].

#### 3.1.3.5 Bosques aleatorios – Random Forest (RF)

Random Forest es un algoritmo de aprendizaje automático que utiliza un conjunto de árboles de decisión para resolver problemas de clasificación y regresión. Este enfoque mejora la precisión y la capacidad de generalización del modelo al combinar las predicciones de múltiples árboles, aprovechando la aleatoriedad en la selección de datos y características para construir cada árbol. La predicción final se obtiene mediante votación mayoritaria en clasificación o promedio en regresión, lo que reduce el sesgo y mejora la precisión del modelo. Además, Random Forest proporciona medidas de importancia de características y se beneficia de la paralelización, acelerando el tiempo de entrenamiento y predicción en grandes conjuntos de datos. Su capacidad para manejar datos ruidosos y ofrecer interpretaciones detalladas lo convierte en una herramienta valiosa en ciencia de datos e inteligencia artificial [15].

# 4. ANTECEDENTES

A continuación, se presentan estudios de investigación que han abordado la deserción estudiantil mediante el uso de ciencia de datos, proporcionando un contexto sobre cómo esta disciplina se ha aplicado para prevenir dicho fenómeno.

Antecedentes del Proyecto: Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior.

El fenómeno de la deserción estudiantil en las instituciones de educación superior representa un desafío significativo para el desarrollo social y académico de las naciones. Diversos estudios han investigado las causas, consecuencias y estrategias de mitigación asociadas a este fenómeno, proporcionando un marco sólido para la implementación de soluciones preventivas y correctivas. Este ensayo examina los antecedentes clave relacionados con el análisis y prevención de la deserción universitaria, tomando como base investigaciones recientes y relevantes en el contexto colombiano y global [16].

Análisis de los factores determinantes de la deserción

Una investigación llevada a cabo por el Ministerio de Educación Nacional en colaboración con la Universidad de los Andes examinó los factores clave que influyen en la deserción en las instituciones de educación superior en Colombia entre 1998 y 2013. Este estudio identificó aspectos críticos relacionados con el abandono académico, subrayando la importancia de comprender las dimensiones individuales y sistémicas que contribuyen a este fenómeno [16].

En un análisis complementario, Tahimi Achilie Valencia exploró los elementos que motivan el abandono estudiantil a través de un estudio de caso. Valencia concluyó que la motivación personal incluye la actitud hacia el crecimiento profesional, el interés académico y las expectativas de carrera, constituye el factor principal de deserción. Además, destacó la importancia de factores académicos e institucionales, recomendando estrategias como el desarrollo de habilidades, laboratorios y cursos de nivelación para fortalecer la retención estudiantil [17].

Tendencias de investigación en Colombia

Marcela Rodríguez Urrego analizó 28 investigaciones realizadas entre 2006 y 2016, proporcionando una visión integral de los avances en la comprensión de la deserción universitaria en Colombia. Rodríguez destacó la diversidad de metodologías utilizadas, desde modelos estadísticos hasta enfoques cualitativos, y reflexionó sobre la necesidad de mejorar la conceptualización del fenómeno para diseñar políticas más efectivas [18].

Modelos predictivos y aprendizaje automático

En la búsqueda de soluciones innovadoras, Julio César Martínez y Sandra Patricia Mateus desarrollaron un modelo predictivo utilizando aprendizaje profundo para programas virtuales en universidades colombianas. Este modelo empleó variables como datos demográficos, socioeconómicos y patrones de ingreso a plataformas de aprendizaje en línea, permitiendo la generación de alertas tempranas para estudiantes en riesgo [19].

De manera similar, Jonny Sotomonte Castro, Cristian Camilo Rodríguez, Carlos Enrique Montenegro Marín, Paulo Alonso Gaona García y John Gabriel Castellanos diseñaron un modelo basado en árboles de decisión mediante el algoritmo J48. Este modelo, implementado en la Universidad Distrital Francisco José de Caldas, destacó la influencia de factores como el nivel socioeconómico y las calificaciones académicas en la probabilidad de abandono [20].

Blanca Cuji, Wilma Gavilanes y Rina Sánchez emplearon el algoritmo CART en un estudio basado en la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD). Los resultados señalaron que las variables relacionadas con el nivel académico y las notas tienen un impacto significativo en la deserción estudiantil [21].

Jovial Niyogisubizoa, Lyuchao Liao, Eric Nziyumva, Evariste Murwanashyaka y Pierre Claver Nshimyumukiza propusieron un modelo de aprendizaje automático de dos capas que combina técnicas como el Bosque Aleatorio (RF), el Aumento de Gradiente Extremo (XGBoost) y la Red Neuronal de Avance (FNN). Este modelo logró identificar patrones de riesgo con alta precisión, ofreciendo a las instituciones herramientas robustas para intervenir de manera temprana [22].

# 5. METODOLOGÍA

Para la ejecución de este proyecto, se utilizará la metodología **CRISP-DM** (Cross Industry Standard Process for Data Mining), ampliamente empleada en proyectos de ciencia de datos. La metodología se desglosa en **cuatro fases distintivas**, cada una enfocada en aspectos clave del proyecto, como se muestra a continuación:

*Tabla 2. Fases del proyecto*

| Fase 1. Planeación | Fase 2. Desarrollo del Modelo | Fase 3. Implementación | Fase 4. Presentación de Resultados |
|---|---|---|---|
| • Objetivo del proyecto<br>• Seleccionar los repositorios de datos<br>• Determinar las fuentes de datos<br>• Infraestructur a tecnológica<br>• Construcción de la base de datos<br>• Definición de las técnicas que se utilizaran en el análisis | • Process Extract, Transform, Load (ETL)<br>• Preparación, Limpieza y organización de los datos<br>• Análisis exploratorio de datos<br>• Selección de variables claves | • Definición del flujo del modelo (punto de variables)<br>• Aplicación de la técnica (exploración de modelos)<br>• Afinar el modelo<br>• Validación de mejor modelo | • Documentación<br>• Presentación de resultados |

## 5.1 Planeación

Esta fase es la etapa más crucial del proyecto, caracterizada por la definición de objetivos, la meticulosa selección de variables y sus fuentes de datos, la elección de la técnica a emplear, la alineación estratégica con el contexto del negocio, el análisis y la selección de la infraestructura tecnológica, así como la construcción de la base de datos.

El proceso de selección de variables adquiere un papel fundamental, ya que determina la calidad de los datos. Comienza con la categorización de las variables y se extiende a la identificación de su relevancia para todos los estudiantes, la accesibilidad a las fuentes de datos (temporales, aplicación de formularios, bases de datos, conexiones con otros sistemas). La infraestructura de Tecnologías de la Información (TI) influye en la elección de la tecnología y la comunicación entre sistemas, considerando que un proyecto de ciencia de datos amalgama diversas fuentes y tecnologías.

### 5.2 Desarrollo del modelo

Se centra en el desarrollo metodológico para llevar a cabo la planificación, basándose en las condiciones específicas de los estudiantes. Este proceso inicia con una exhaustiva limpieza de datos, seguida de la selección de atributos, así como la estandarización y normalización correspondiente.

En el proceso de selección de variables, se ha optado por seguir dos criterios distintos: el estadístico y la necesidad de negocio. Implementaremos funciones para agregar nuevos atributos y aplicaremos técnicas de reducción dimensional con el objetivo de identificar las variables óptimas para nuestro modelo predictivo. El desarrollo del modelo se llevará a cabocon un enfoque diferencial específico para el periodo de la pandemia de COVID-19, a partir del mes de marzo de 2020. Además, evaluaremos estadísticamente las relaciones de causalidad o dependencia asociadas a este fenómeno para enriquecer la capacidad predictiva del modelo.

### 5.3 Implementación

Se llevará a cabo mediante un enfoque estadístico tanto para el desarrollo como para la validación del modelo. Para esta validación, los datos se dividirán en un 80% para el entrenamiento del modelo y un 20% para su validación. Las fuentes de datos a integrar incluyen la caracterización del estudiante al ingreso, los resultados de las pruebas Saber 11, la información del sistema académico.

### 5.4 Presentación de resultados

En esta fase de preparación de resultados, se realizará la creación de una presentación cuidadosamente diseñada, asegurando una comprensión clara y precisa del proyecto y los logros alcanzados. Esta presentación adquiere un papel fundamental al proporcionar información relevante que genere valor a la universidad.

## 6. Planeación

### 6.1 Caracterización de estudiantes

El análisis del conjunto de datos se centra en comprender las características generales de los estudiantes y sus entornos. El libro de Excel incluye información organizada en las siguientes hojas, que permiten una visión amplia sobre diversos aspectos relacionados con los estudiantes:

1. Caracterización socio demográfica: Incluye información general sobre los estudiantes, como edad, estado civil, lugar de residencia, nivel socioeconómico, ingresos familiares y habilidades académicas básicas. Estos datos permiten construir un perfil general de la población estudiantil.

Tabla 3. Detalle fuente de datos caracterización socio demográfica (Fuente: Elaboración propia)

| Programa Académico | Cantidad de datos | No. Variables |
|---|---|---|
| Administración de Empresas | 452 | 97 |
| Administración de Empresas Noc | 91 | |
| Arquitectura | 442 | |
| Artes Visuales | 127 | |
| Biología | 172 | |
| Ciencia Política | 179 | |
| Comunicación | 211 | |
| Contaduría Pública | 80 | |
| Derecho | 536 | |
| Diseño de Comunicación Visual | 271 | |
| Economía | 147 | |
| Enfermería | 228 | |
| Filosofía | 35 | |
| Finanzas | 18 | |
| Gastronomía y Artes Culinarias | 36 | |
| Ingeniería Biomédica | 52 | |
| Ingeniería Civil | 587 | |
| Ingeniería de Sistemas | 243 | |
| Ingeniería Electrónica | 165 | |
| Ingeniería Industrial | 351 | |

| | | |
|---|---|---|
| Ingeniería Mecánica | 66 | |
| Matemáticas Aplicadas | 31 | |
| Medicina | 697 | |
| Mercadeo | 56 | |
| Negocios Internacionales | 531 | |
| Nutrición y Dietética | 467 | |
| Psicología | 366 | |
| Turismo | 4 | |
| | **6641** | **97** |

2. Caracterización académica: Resume información relacionada con los programas académicos, como asignaturas cursadas, promedio de notas, y la cantidad de asignaturas aprobadas y reprobadas. Estos datos ofrecen una visión general del desempeño académico de los estudiantes

Tabla 4. Detalle fuente de datos caracterización académica (Fuente: Elaboración propia)

| **Programa Académico** | **Cantidad de datos** | **No. Variables** |
|---|---|---|
| Todos los programas académicos | 87.585 | 15 |

3. Caracterización financiera: Esta fuente de datos abarca información detallada sobre los estudiantes por programas y sus modalidades de financiamiento para sus estudios académicos. Se incluye información específica si cuenta con una beca, si tiene alguna ayuda económica para su manutención, quien genera el pago de sus estudios, si existen otros apoyos financieros como por ejemplo seguros educativos entre otros.

Tabla 5. Detalle fuente de datos caracterización (Fuente: Elaboración propia)

| Programa Académico | Cantidad de datos | No. Variables |
|---|---|---|
| Todos los programas académicos | 48.325 | 6 |

4. Proporciona información sobre los estudiantes que han abandonado los programas, esto quiere decir que no se ha matriculado por dos periodos consecutivos los datos incluyen información académica. Estos datos permiten completar el panorama general de la población estudiantil, promedio acumulado.

Tabla 6. Cantidad de estudiantes por programa académico que desertaron (Fuente: Elaboración propia)

| Programa Académico | Cantidad de datos | No. Variables |
|---|---|---|
| Todos los programas académicos | 73 | 8 |

### 6.1.1 Modelo bidimensional de datos

A partir de la información recopilada sobre las caracterizaciones socio-demográficas, académicas, becas, financieras y de deserción, se propone la construcción de un modelo de bodega de datos. Este modelo permitirá organizar los datos mediante la estructura de hechos y dimensiones, donde la tabla de hechos estará basada en la caracterización de los estudiantes. La variable "desertor" será representada como una variable binaria con valores de 0 y 1, indicando si el estudiante desertó o no. Esta estructura se detalla en la Figura 1.

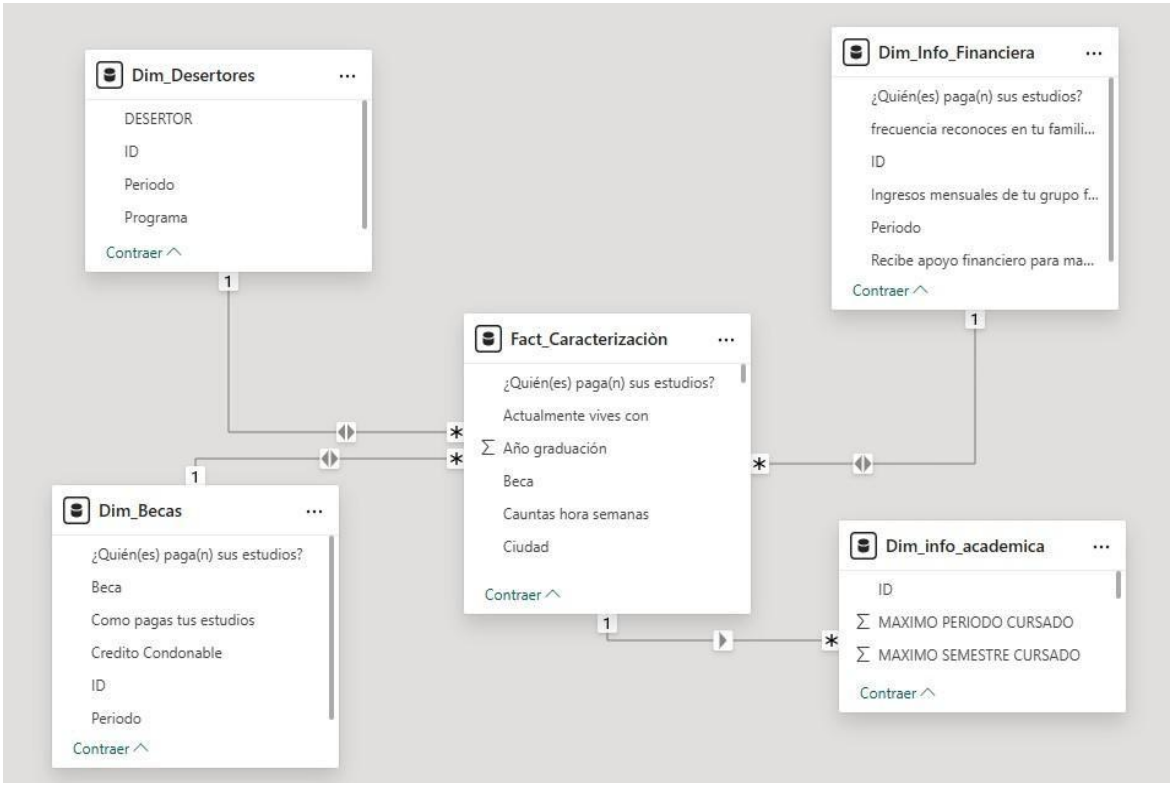


Figura 1. Modelo bidimensional del análisis

### 6.1.2 Preparación y limpieza de datos

El objetivo de la Limpieza de datos es garantizar que el conjunto de datos sea confiable y coherente para el análisis posterior, mediante la eliminación de información incompleta o irrelevante, la corrección de formatos y tipos de dato, y la depuración de registros duplicados, con el fin de reducir errores y facilitar una exploración estadística más precisa
Aquí hay un resumen de lo que se realizó:

**Situación Inicial**

En el dataset original había:

- 6,641 filas y 207 columnas.
- Algunas columnas con valores faltantes demasiado altos (por ejemplo, más de 6,000 celdas en blanco o inválidas).
- Varias columnas numéricas guardadas como texto, lo cual impedía hacer sumas, promedios o comparaciones.

**Problemas Detectados**

Columnas con 90% o más de valores nulos. Por ejemplo:

- TOTAL REPROBADAS, con 6,197 registros sin ningún valor.
- Cauntas hora semanas, con 6,204 registros vacíos.

Datos numéricos guardados como texto. Por ejemplo, promedios como "3,57", que deberían ser numéricos (3.57) para poder sumar o calcular estadísticas.

**Limpieza de los datos**

- Se eliminaron columnas con muchos valores faltantes.
- Se revisó qué columnas tenían una gran mayoría de celdas vacías.
- Al ver que casi no contenían información útil (o muy poca), se decidió retirarlas del dataset.

Resultado: Una reducción de 207 a 193 columnas. Esto no afecta los datos importantes, sino que quita ruido.

**Cambiar tipo de dato**

- Muchas columnas con números estaban en formato de texto, por ejemplo "3,74" en lugar de un valor decimal.
- Se aplicó una técnica de conversión para que esas columnas pasaran a ser verdaderamente numéricas.
- Esto es fundamental para poder hacer cálculos estadísticos (promedios, correlaciones, etc.) en el futuro.

**Revisión de duplicados**

- Se revisaron filas que pudieran aparecer repetidas por error.
- En esta parte, no se encontraron problemas de duplicados.

**Organización Final de Columnas**

Después de descartar columnas con demasiados nulos y transformar aquellas que requerían convertirse en numéricas, se comprobó que el resto de las columnas tenía datos suficientes para aportar valor al análisis.

**Estado del dataset después de la limpieza**

Tras este proceso, el dataset quedó con:

- 6,641 filas (las mismas que antes, puesto que no se eliminaron registros completos, solo columnas poco útiles o sin datos).
- 193 columnas, en lugar de 207, enfocadas ahora en la información más relevante.
- Varias columnas que antes eran texto se han convertido al tipo numérico. Esto facilita el uso de fórmulas y métodos estadísticos en el siguiente análisis.

**División del conjunto de datos**

**Conjunto de entrenamiento:** 80% de los datos (5312 registros).

**Conjunto de prueba:** 20% de los datos (1329 registros).

## 6.2 Análisis exploratorio de datos

Distribución de la Variable Objetivo (DESERTOR):

Para identificar que la variable DESERTOR estaba desbalanceada, primero se revisó la cantidad de registros en cada clase (0 = no desertor, 1 = desertor). Al ver que la clase “no desertor” era mucho más numerosa que la clase “desertor”, se concluyó que había un desbalance en la variable objetivo como se muestra en la figura 2.

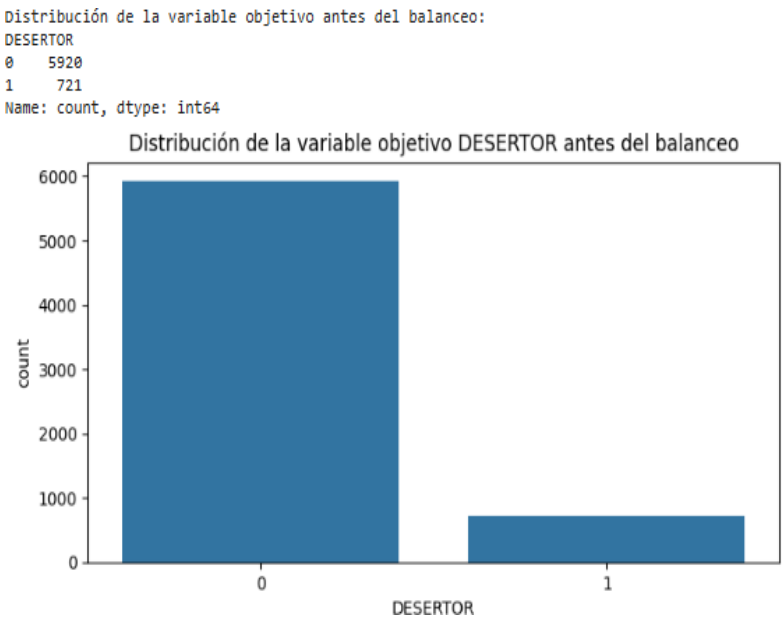


Figura 2. Distribución de la variable objetivo antes del balanceo

Para corregirlo, se aplicó una técnica de balanceo (sobremuestreo de la clase minoritaria o submuestreo de la clase mayoritaria), hasta que ambas clases tuvieran la misma cantidad de observaciones. De esta forma, la variable DESERTOR quedó equilibrada y el modelo ya no ignora la clase con menos casos como se muestra en la figura 2.

```
Distribución de la variable objetivo después del balanceo:
DESERTOR
0    5920
1    5920
Name: count, dtype: int64
```

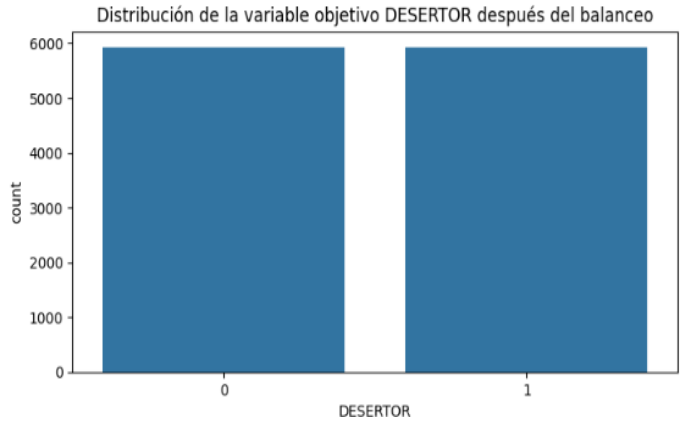


Figura 3. Distribución de la variable objetivo después del balanceo

### 6.2.1 Análisis Univariado

A continuación, se presentan los resultados obtenidos en el análisis univariados con una explicación de los gráficos que reflejan la distribución de las variables numéricas y categóricas del dataset. La idea de este análisis es entender la forma de cada variable se presenta la información de las variables más relevantes como también su posible impacto en el análisis posterior.

En los anexos se hará una breve descripción de las otras variables

**Distribución de variables numéricas**

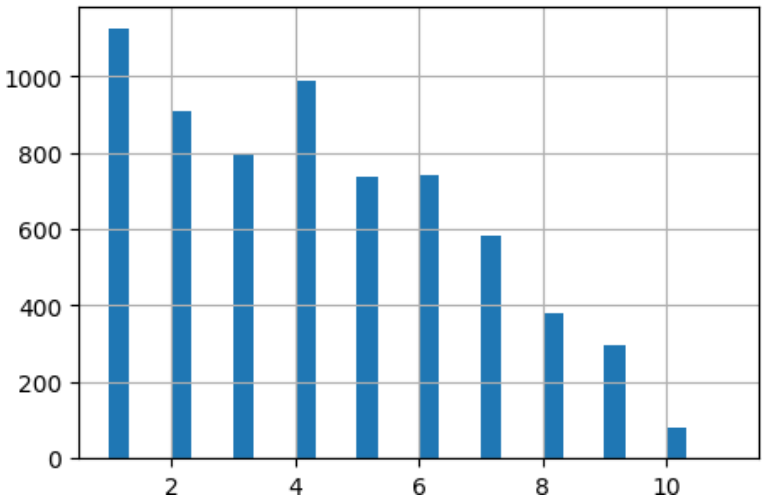


Figura 4. Máximo semestre cursado

- Máximo semestre cursado

Se observa que la mayoría de los estudiantes se concentran en semestres entre 1 y 4, disminuyendo progresivamente a medida que aumenta el número de semestre cursado. Esto indica que la gran parte de la población estudiantil analizada no avanza hasta semestres altos. Esto podría indicar que la deserción se presenta más fuertemente en los primeros periodos universitarios.

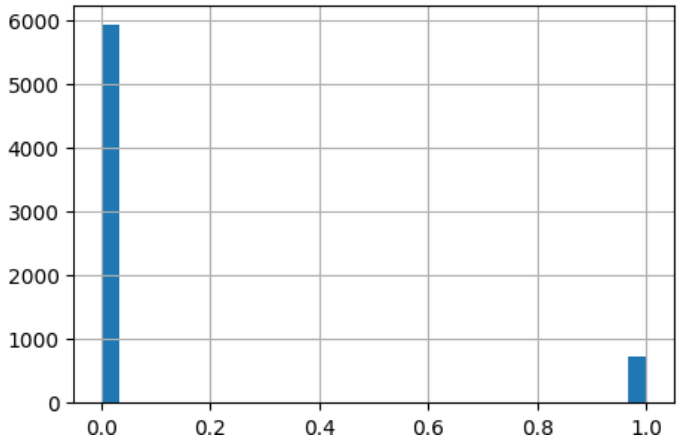


Figura 5. Desertor

- DESERTOR

Se ve claramente que la mayoría (cerca del 90% o más) corresponde a “0” (no desertores), mientras que una fracción menor es “1” (desertores).

La variable objetivo está desbalanceada; hay muchos más no desertores que desertores. Esto es importante al momento de entrenar un modelo de predicción, pues requeriría técnicas de balanceo (como SMOTE) o un ajuste apropiado de la métrica de evaluación.

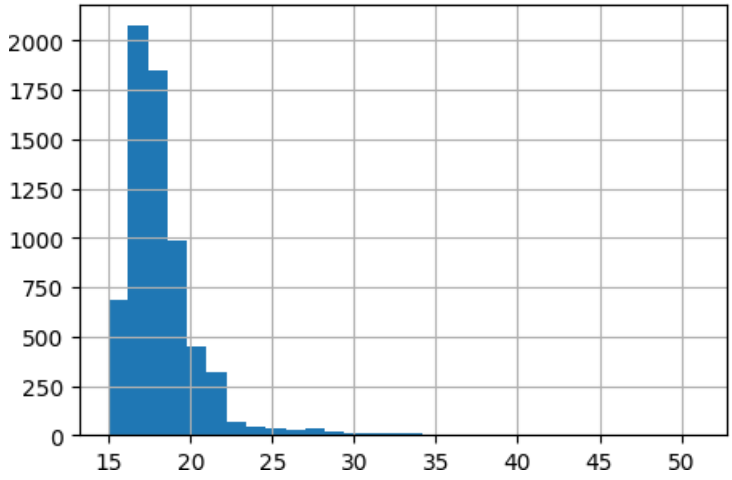


Figura 6. Edad

- Edad

  Se aprecia una concentración muy alta de estudiantes en rangos de edad entre 15 y 25 años, con distribución asimétrica a la derecha (cola larga hacia edades mayores). Hay pocos estudiantes mayores de 30.

  Predomina una población joven, pero se confirma que hay un grupo reducido de estudiantes de mayor edad que podrían tener características diferentes (trabajo, familia, menos tiempo, etc.).

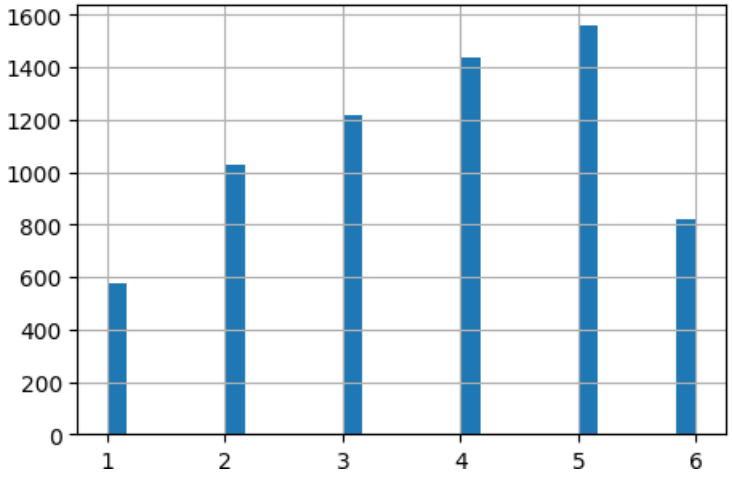


Figura 7. Estrato

- Estrato

  Se observa un patrón con mayor frecuencia en los estratos 3, 4 y 5, siendo algo menor el estrato 1 y 2, y todavía más reducido el estrato 6.

  El rango típico está entre 2 y 4, con algunos valores por debajo o por encima.

  Conclusión: La mayoría de los estudiantes pertenecen a estratos medios (3,4,5), lo cual da una idea de la condición socioeconómica predominante.

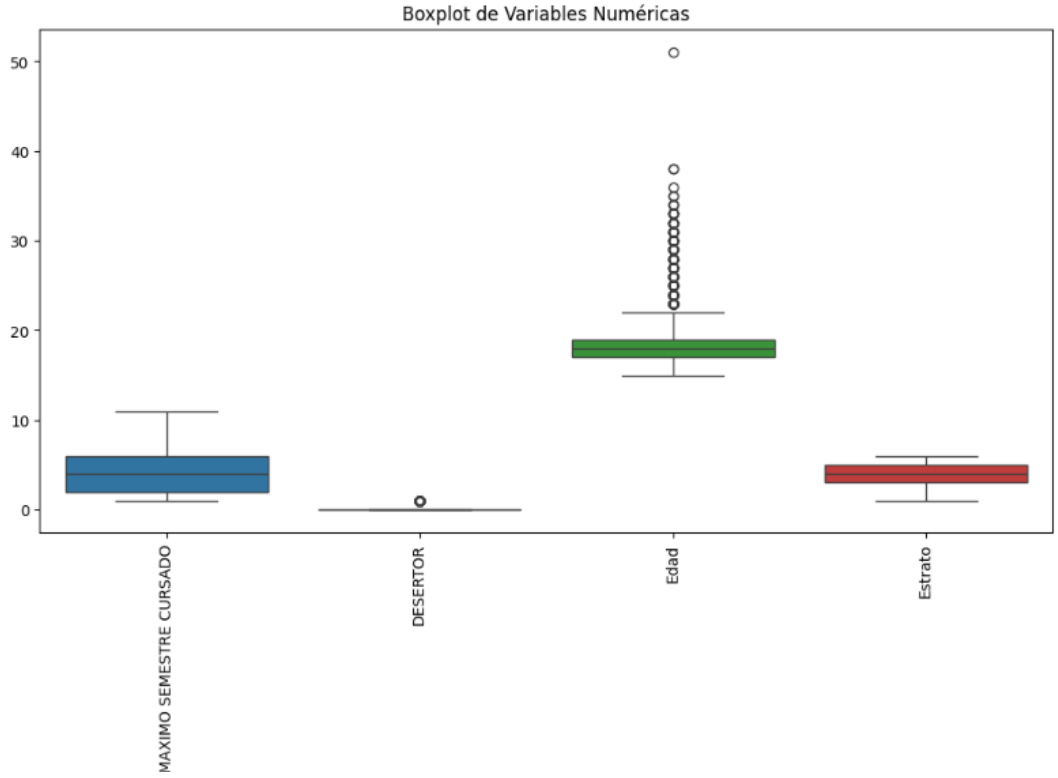


Figura 8. Boxplot de variables numéricas

- En el boxplot combinado, Edad es la variable que presenta más valores atípicos (personas con más de 30 o 40 años), mientras que en “DESERTOR” (0 o 1) no aplica igual la escala de boxplot

- MAXIMO SEMESTRE CURSADO tiene un rango típico entre 1 y 6, con pocos registros muy altos (cerca de 10).

- Estrato no presenta outliers extremos, pero sí valores mínimos en 1 y máximos en 6.

**Distribución de variables categóricas**

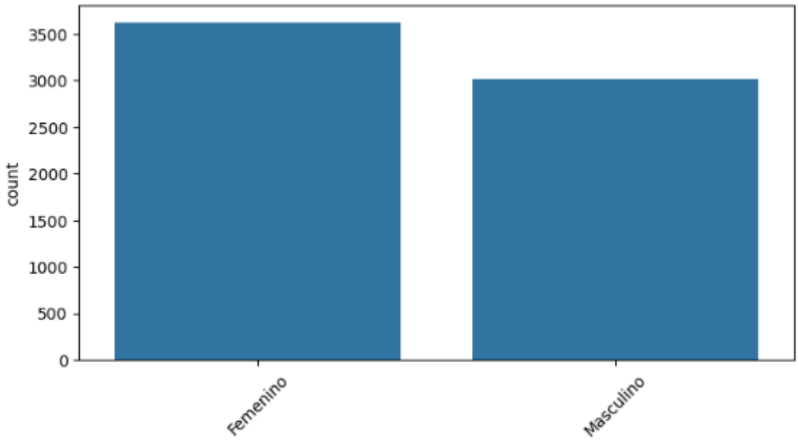


Figura 9. Sexo

- Sexo

Se observa una mayor proporción de estudiantes femeninas que masculinos, pero no es una diferencia abismal. Ligero predominio femenino en la muestra.

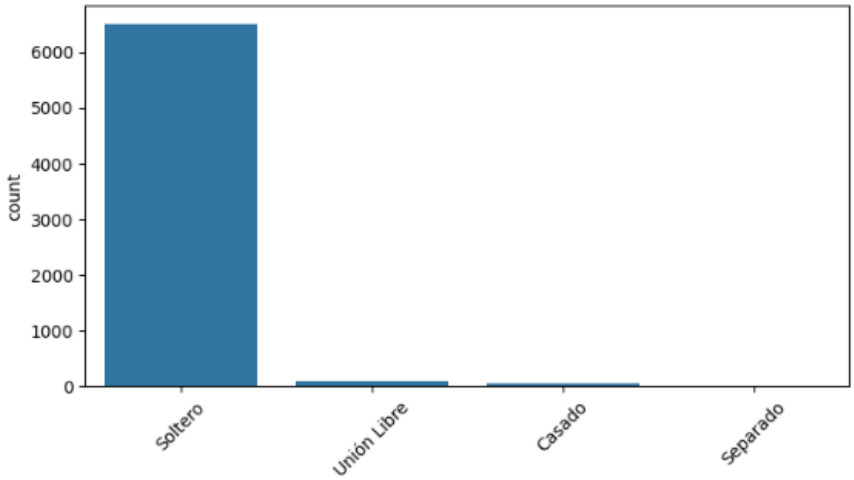


Figura 10. Distribución Estado civil

- Estado civil

Predomina de manera muy marcada el estado civil “Soltero” (más de 6000 casos), mientras que “Unión libre”, “Casado” o “Separado” son muy pocos. La gran mayoría de los estudiantes no han formalizado matrimonio, lo cual es frecuente en edades menores de 25 años. Podría relacionarse con la variable “Edad” y la dedicación a estudios.

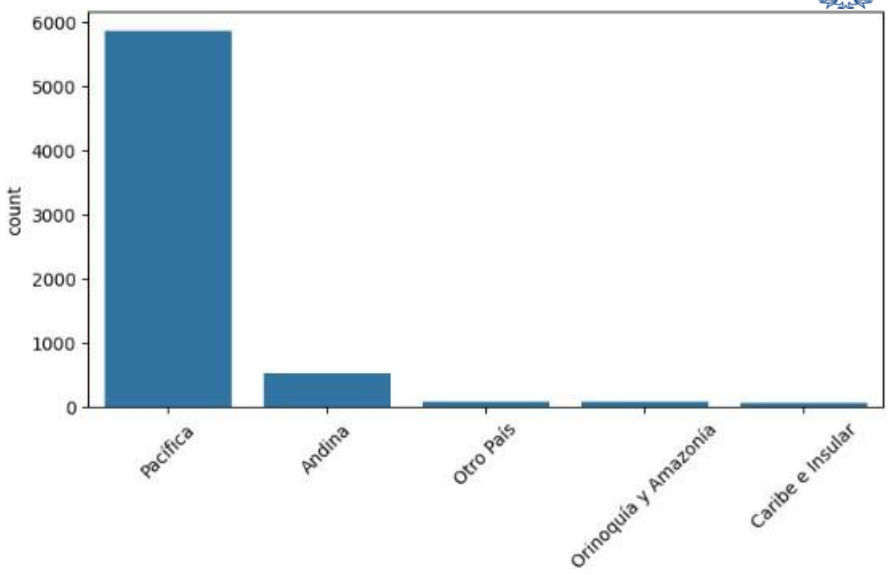


Figura 11. Distribución Región

- Región

  Predominio abrumador de la región “Pacífica”, seguidas de “Andina” y luego otras regiones en volúmenes muy pequeños (“Orinoquía y Amazonía”, “Caribe e Insular”, “Otro País”). La población estudiantil proviene en su mayoría de la región Pacífica, lo que indica poca diversidad geográfica o que la universidad se localiza principalmente en esa región.

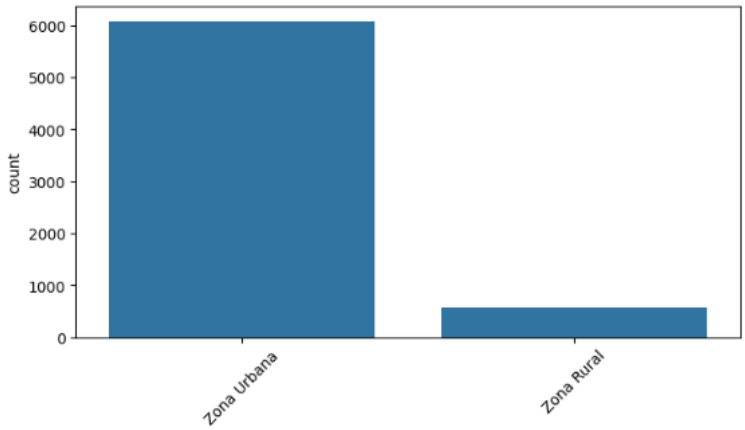


Figura 12. Distribución Zona (Urbana / Rural)

- Zona (Urbana / Rural)

  Se ve una gran mayoría de estudiantes provenientes de zona urbana, y un porcentaje menor de zona rural. Refleja un típico fenómeno de concentración de estudiantes en zonas urbanas, posiblemente por cercanía de la oferta educativa o mayor facilidad de acceso.

### 6.2.2 Análisis Bivariado

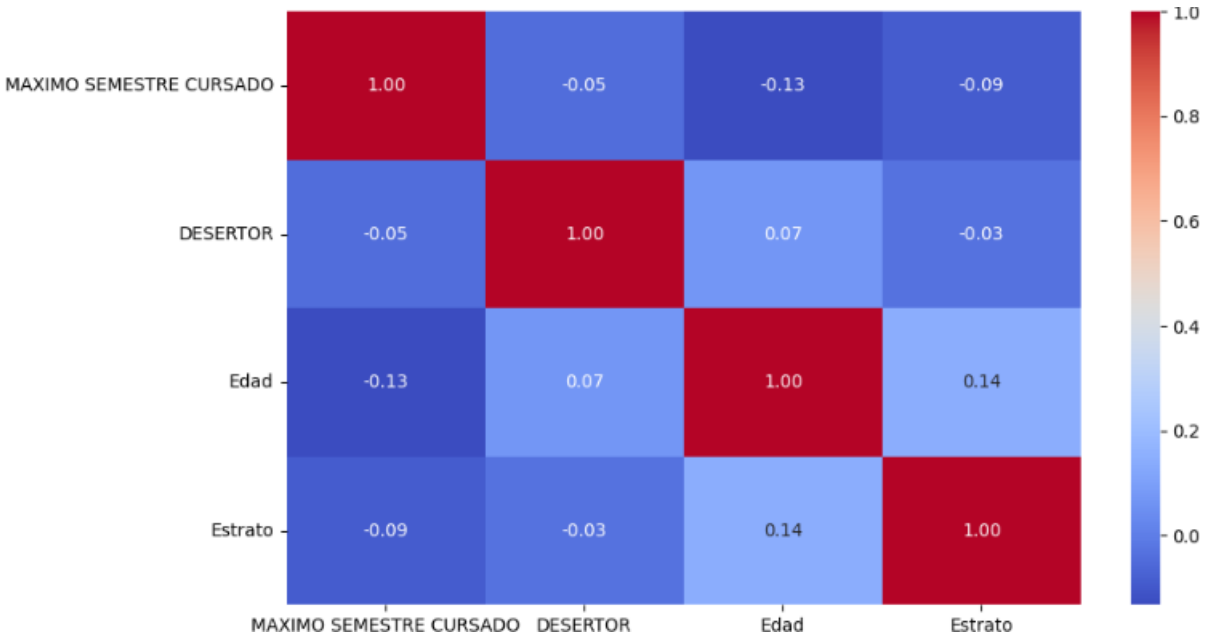


Figura 13. Matriz de correlación

**Variables categóricas**

En la figura 13. Podemos observar que se destacan correlaciones entre algunas variables, aunque ninguna es extremadamente alta.

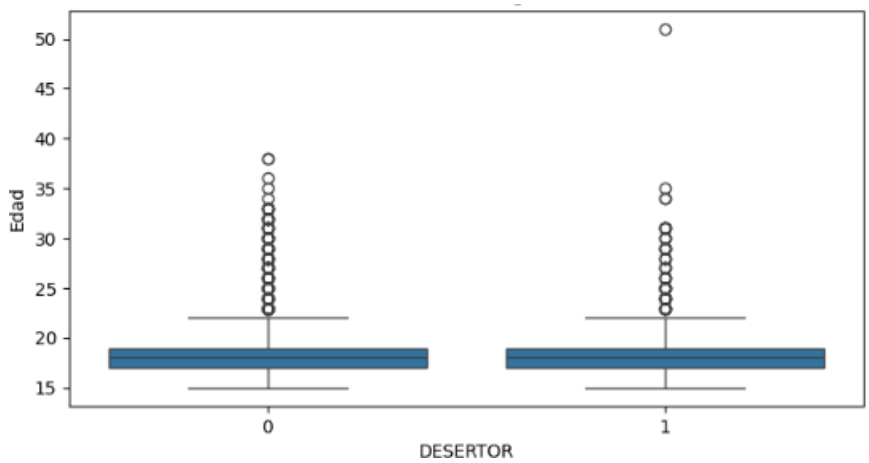


Figura 14. distribución de edad según DESRTOR

- Edad vs. Deserción

Correlación muy baja, lo que sugiere que la edad no es un predictor fuerte de la deserción.

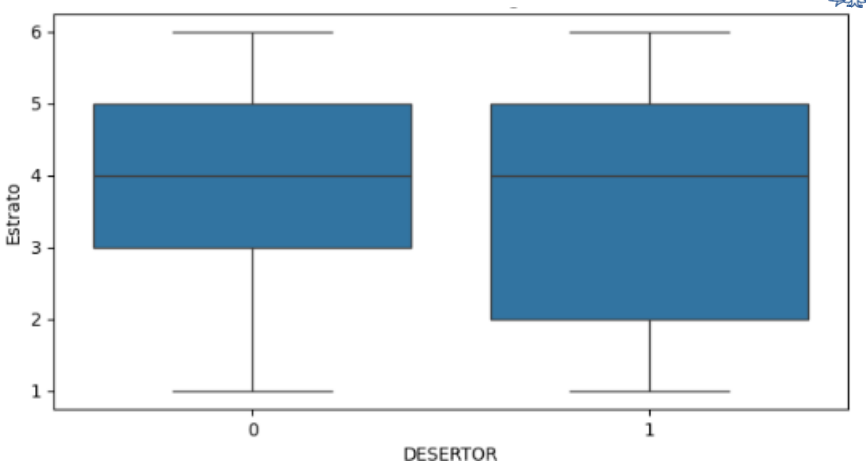


Figura 15. Distribución estrato socioeconómicos según DESERTOR

- Estrato Socioeconómico vs. Deserción

  No hay una relación clara, aunque es un factor para investigar más a fondo.

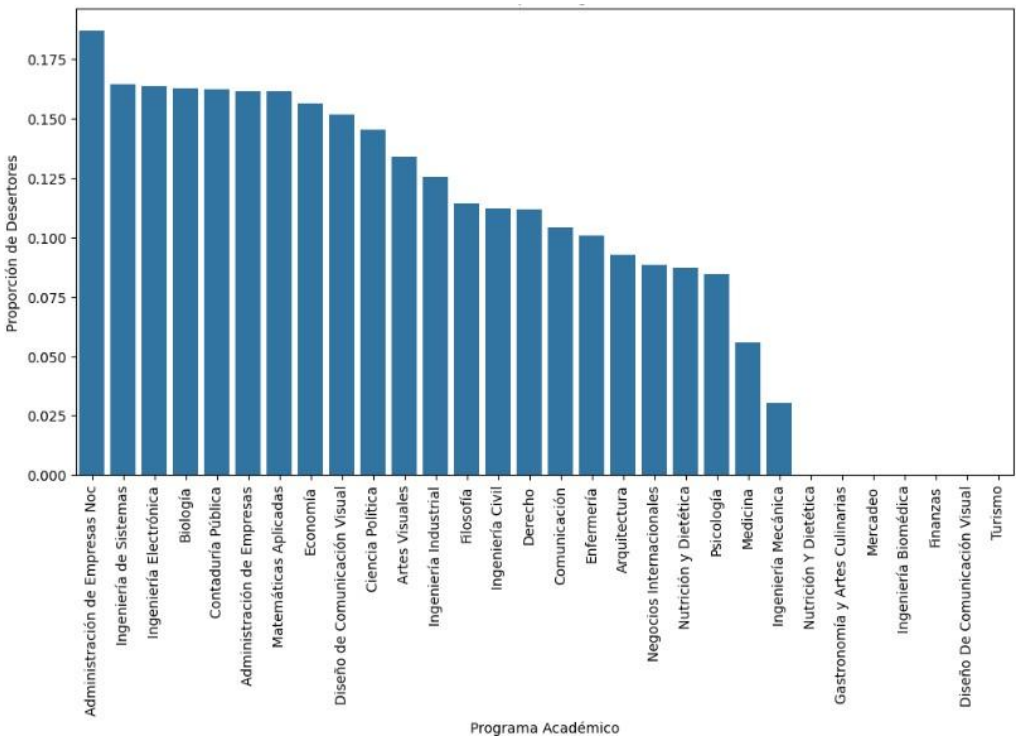


Figura 16. Tasa de deserción por programa académico

- Analiza en qué carreras hay más o menos deserción.

  Administración de Empresas Nocturna tiene la mayor tasa de deserción (~18.7%).

  Ingenierías (Sistemas, Electrónica, Industrial, Civil) también tienen tasas elevadas (> 12%).

  Medicina y Mercadeo tienen tasas muy bajas de deserción (~5%).

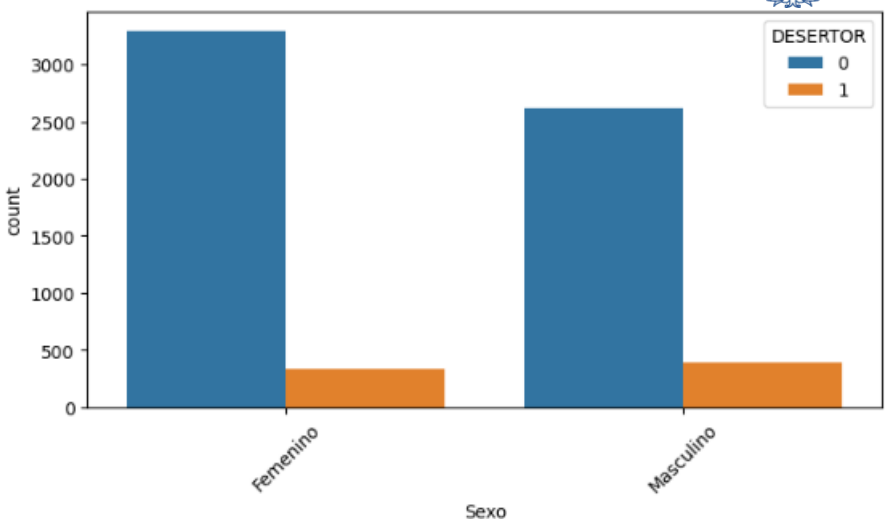

Figura 17. Comparación de género y deserción

- Comparación de género y deserción
  No hay una diferencia significativa entre hombres y mujeres en términos de deserción. Esto sugiere que el género no es un factor determinante en el abandono de estudios.

### 6.2.3 Selección de variables para el modelo

Matriz de Correlación de Pearson:

- Se utilizó la prueba de correlación de Pearson para evaluar la relación entre variables numéricas y evitar multicolinealidad.
- Variables como MAXIMO SEMESTRE CURSADO, Ingreso mensual mostraron alta correlación, pero ambas se retuvieron debido a su importancia predictiva.

Variables Seleccionadas:

- Máximo semestre cursado
- Edad
- Ingresos mensuales de tu grupo familiar
- Beca
- Como financias tus estudios
- Estrato
- Total, reprobadas

## 6.3 Creación de perfil de deserción

Este análisis fue realizado para identificar los perfiles de estudiantes desertores en la universidad y comprender las razones principales de su abandono académica.

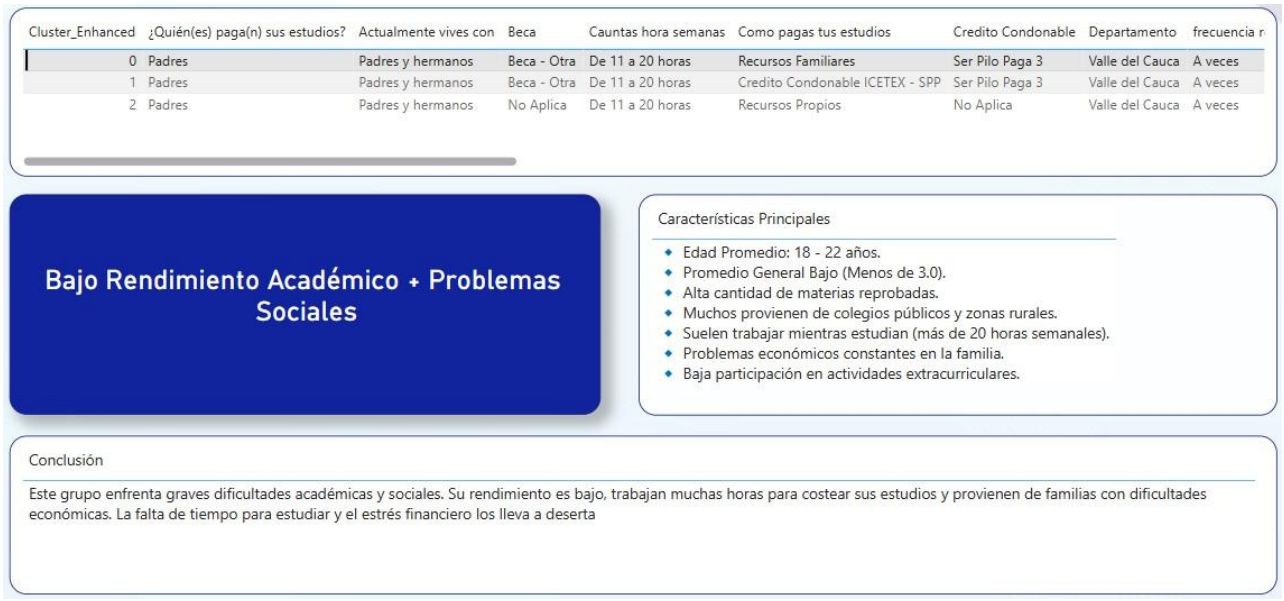


Figura 18. Clúster 0: Bajo rendimiento académico + problemas sociales

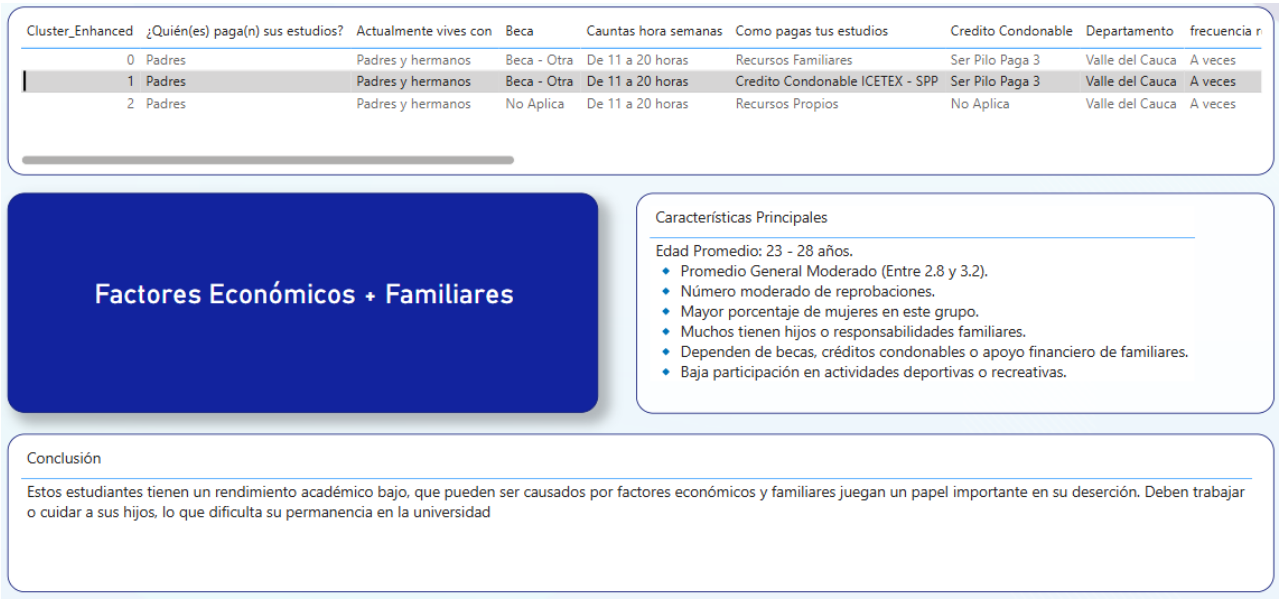


Figura 19. Clúster 1: Factores económicos + familiares

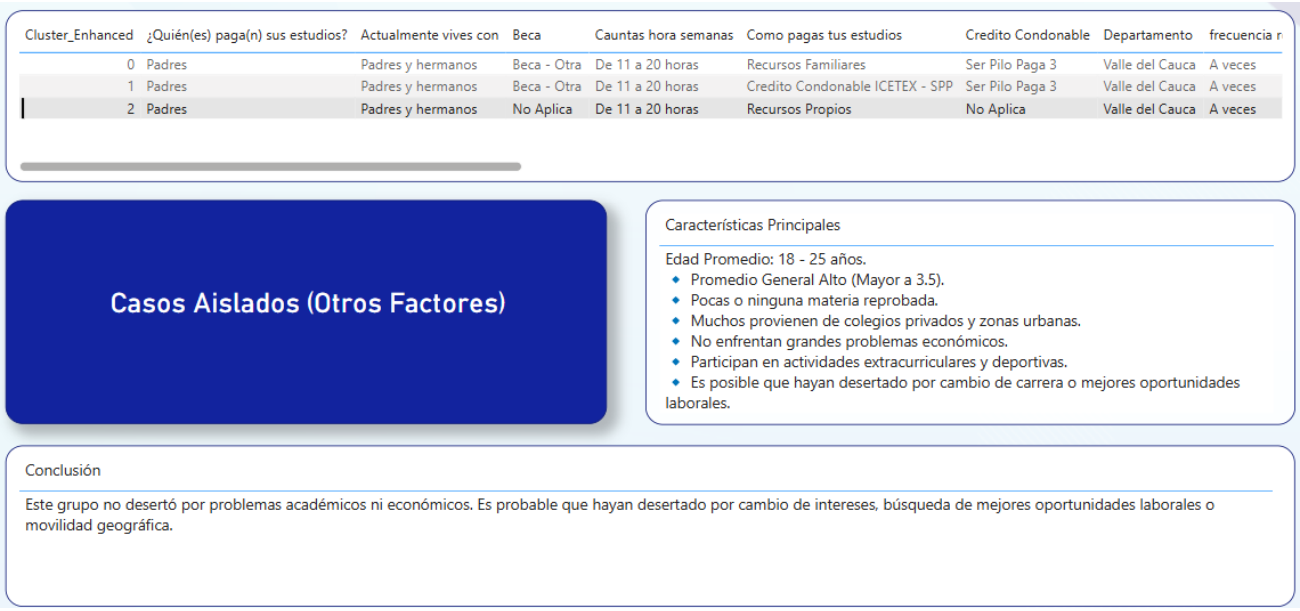


Figura 20. Clúster 2: Casos aislados (otros factores)

## 7. Análisis comparativo de los modelos de machine learning

Estandarización:

Las variables numéricas fueron escaladas usando la técnica de normalización estándar (media = 0, desviación estándar = 1).

Esto asegura que los modelos no sean influenciados por magnitudes desiguales entre variables.

Manejo de valores faltantes:

Las columnas categóricas con datos faltantes fueron eliminadas.

Las variables numéricas se imputaron con la mediana.

### 7.1 División de los datos en entrenamiento

Proporción de división: 80% para entrenamiento y 20% para prueba.

Estratificación: Se aseguró que la variable objetivo (DESERTOR) tuviera una distribución representativa en ambos conjuntos.

## 7.2 Preprocesado de los datos

Con el fin de asegurar un buen ajuste del modelo, se lleva a cabo un preprocesamiento de los datos que implica la estandarización y escalado de las variables cuantitativas, así como la binarización de las variables cualitativas. Estos pasos se reflejan de manera detallada en la Tabla 8. Este proceso es esencial para homogeneizar las variables y optimizar el rendimiento de los modelos, asegurando una interpretación coherente de los resultados y una mayor eficacia en la predicción.

Se muestran los datos estandarizados y binarizados, que incluye las variables seleccionadas y la variable objetivo (DESERTOR). Esto permite verificar cómo se han transformado las variables para el análisis.

## 7.3 Modelo de regresión logística

Precisión: 99.92%

El modelo clasifica correctamente casi todos los casos de deserción y no deserción.

Reporte de Clasificación:

Clase 0 (No Desertó):

Precisión: 100%

Recall: 100%

Clase 1 (Desertó):

Precisión: 99%

Recall: 100% Matriz de Confusión:

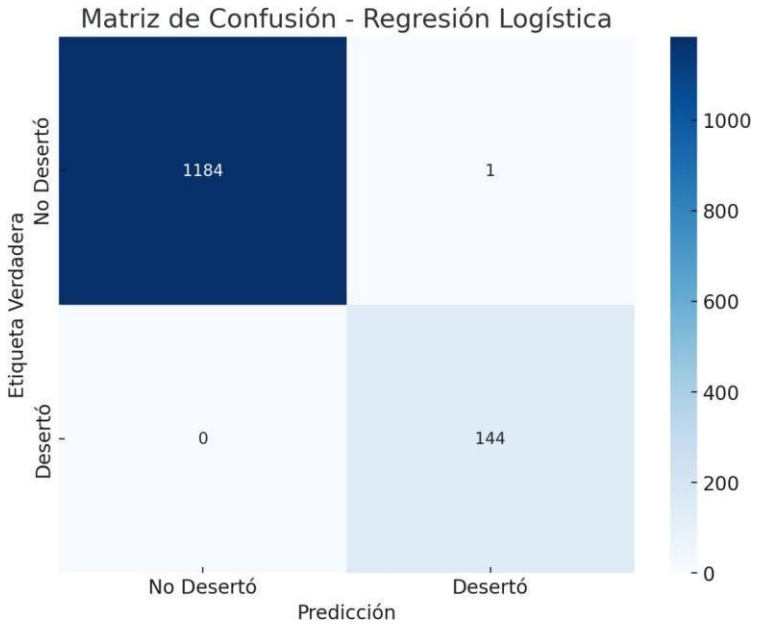


Figura 21. Matriz de confusión - Regresión logística

El gráfico muestra que el modelo realiza muy pocas clasificaciones incorrectas.

El modelo de regresión logística demuestra un rendimiento excepcional, lo que sugiere que las variables seleccionadas tienen un fuerte poder predictivo.

### 7.4 Modelo de máquina de soporte vectorial (SVM)

Precisión: 99.85%

El modelo clasifica correctamente la gran mayoría de los casos.

Reporte de Clasificación:

Clase 0 (No Desertó):

Precisión: 100%

Recall: 100%

Clase 1 (Desertó):

Precisión: 99%

Recall: 100%

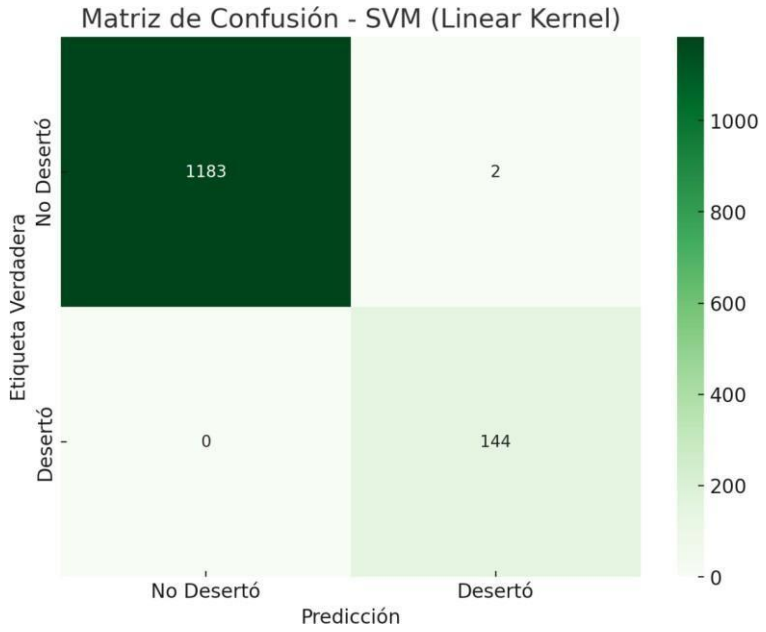


Figura 22. Matriz de Confusión

El gráfico muestra un rendimiento excepcional, con solo un caso mal clasificado.

El modelo SVM demuestra un desempeño similar al de la regresión logística, indicando que las variables seleccionadas también son muy efectivas para este enfoque.

### 7.5 Modelo de bosques aleatorios (Random Forest)

Precisión: 100%

El modelo clasifica correctamente todos los casos de deserción y no deserción. Reporte de Clasificación:

Clase 0 (No Desertó):

Precisión: 100%

Recall: 100%

Clase 1 (Desertó):

Precisión: 100%Recall: 100% Matriz de Confusión:

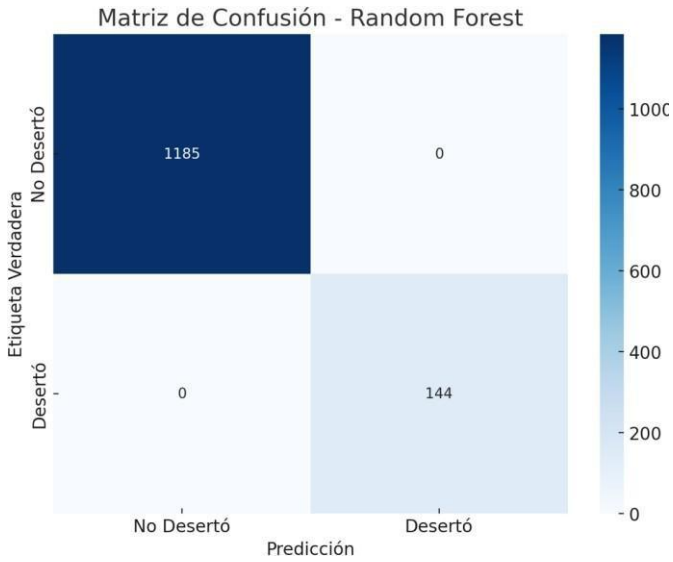


Figura 23. Matriz de confusión - Random Forest

La matriz muestra que el modelo no comete errores de clasificación.

Importancia de las Variables:

El gráfico muestra las variables más influyentes en el modelo. Las principales incluyen:

- MAXIMO SEMESTRE CURSADO
- PROMEDIO GENERAL
- TOTAL, REPROBADAS
- BECA
- Como financias tus estudios

El modelo de bosques aleatorios muestra un rendimiento sobresaliente y proporciona información valiosa sobre la importancia de las variables.

### 7.6 Modelo de redes neuronales simple (NNET)

Precisión: 99.92%

El modeloclasifica correctamente la mayoría de los casos de deserción y no deserción.

Reporte de Clasificación:

Clase 0 (No Desertó):

Precisión: 100%

Recall: 100%

Clase 1 (Desertó):

Precisión: 100%

Recall: 99%

Matriz de Confusión:

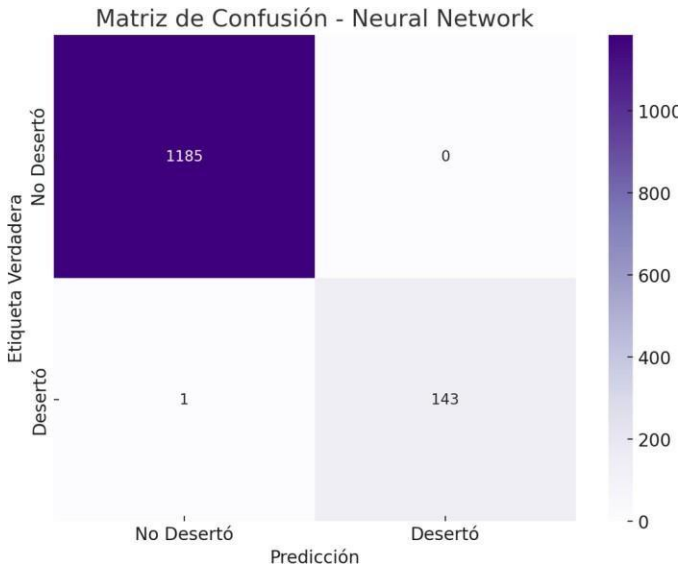


Figura 24. Matriz de confusión - Neural Network

El gráfico muestra que el modelo realiza un pequeño número de errores al clasificar estudiantes desertores.

Estemodelodemuestra un rendimiento muyalto, aunquecomparable al de losbosquesaleatorios

**1. Evaluación de desempeño en modelos de machine learning para la predicción de deserción**

- **Comparación de métricas de los modelos**

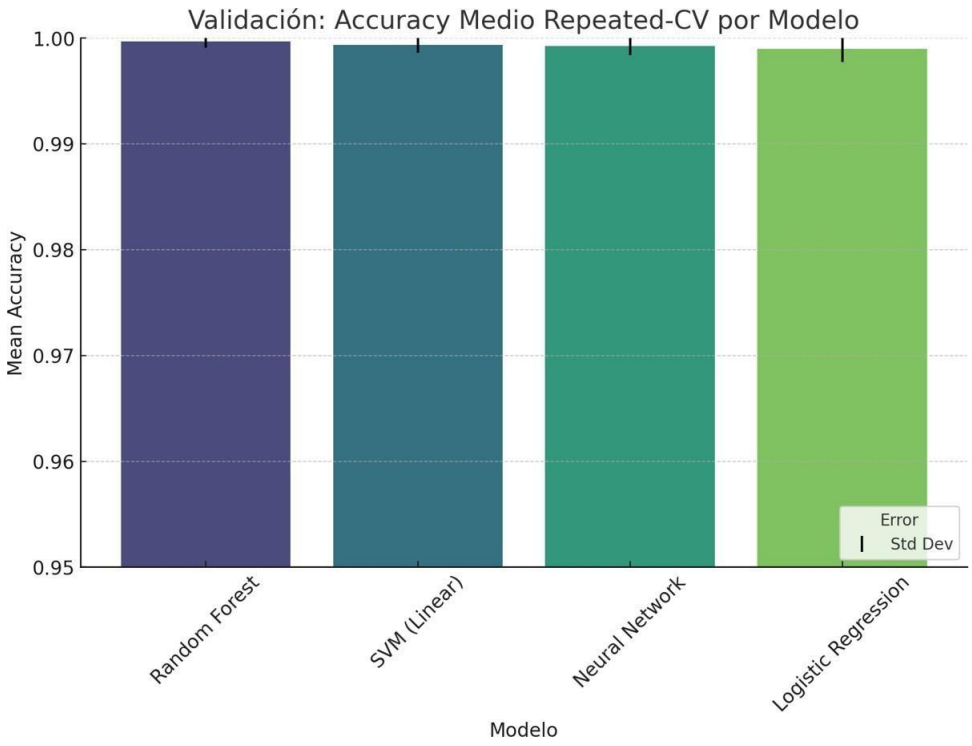


Figura 25. Comparación de métricas de los modelos

Realizando una comparación de las métricas de precisión (accuracy) entre los modelos evaluados:

Logistic Regression: 99.92%

SVM (Linear): 99.85%

Random Forest: 100%

Neural Network: 99.92%

El modelo de Random Forest obtuvo la mejor precisión, clasificando correctamente todos los casos.

Tabla 7. Resultados de Accuracy Medio Reeated-CV por Modelo

| | Model | Mean Accuracy | Std Dev Accuracy | Accuracy Range |
|---|---|---|---|---|
| 2 | Random Forest | 0.999699 | 0.000602 | 0.9997 ± 0.0006 |
| 1 | SVM (Linear) | 0.999398 | 0.000797 | 0.9994 ± 0.0008 |
| 3 | Neural Network | 0.999247 | 0.000865 | 0.9992 ± 0.0009 |
| 0 | Logistic Regression | 0.999006 | 0.001228 | 0.999 ± 0.0012 |

Friedman Test:

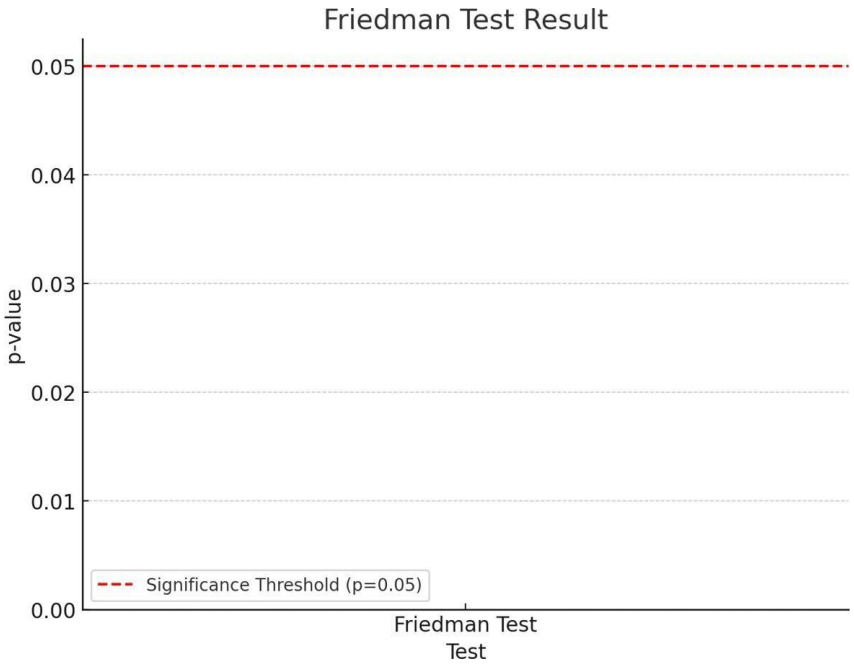


Figura 26. Friedman Test

La muestra el p-valor del test de Friedman, con una línea roja que indica el umbral de significancia (p = 0.05). El resultado muestra diferencias estadísticamente significativas entre los modelos.

Wilcoxon Pairwise Comparisons:

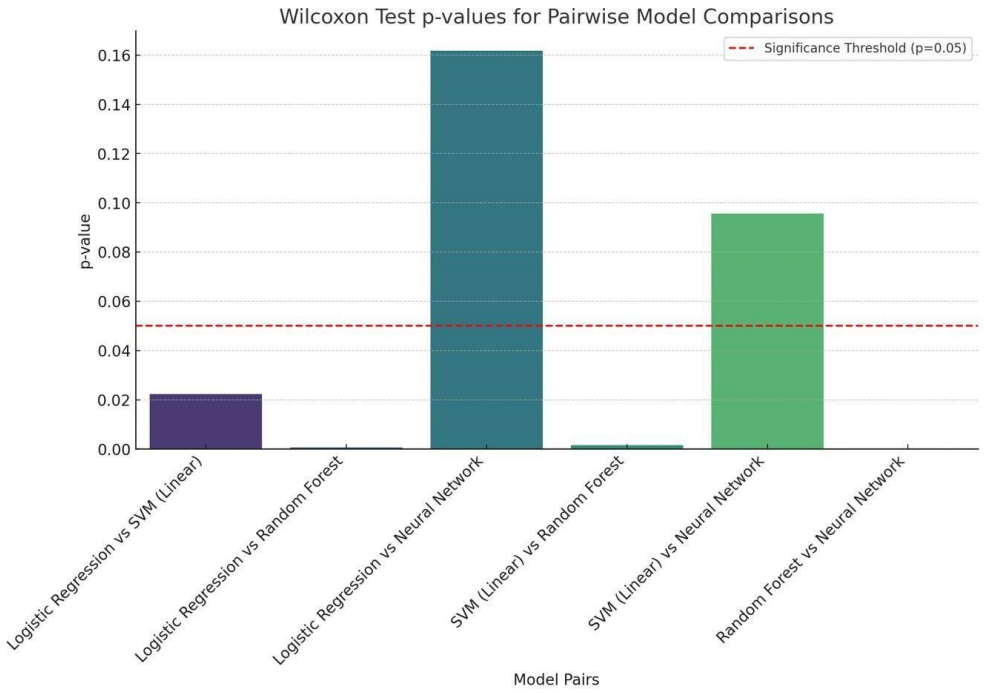


Figura 27. Wilcoxon Pairwise Comparisons

La gráfica muestra los p-valores de las comparaciones por pares de modelos. Los valores por debajo de la línea roja indican diferencias significativas entre los modelos comparados.

Al evaluar los modelos de machine learning para la predicción de la deserción estudiantil, se emplearon métricas clave de rendimiento y pruebas estadísticas para identificar el mejor modelo. De esta evaluación se concluye lo siguiente:

1. Precisión General (Accuracy)

**Bosques Aleatorios (Random Forest):**

Destaca como el modelo con mejor rendimiento, logrando un accuracy medio de 99.97% en validación repetida (Repeated-CV).

Su precisión perfecta en los datos de prueba indica que puede distinguir con alta

confianza entre estudiantes desertores y no desertores.

**SVM (Kernel Lineal):**

Mostró un accuracy medio de 99.94%, apenas inferior a los Bosques Aleatorios. Este modelo es competitivo y altamente efectivo, con menor complejidad.

**Regresión Logística:**

Con un accuracy medio de 99.90%, demostró ser un modelo simple pero eficaz.

Es menos consistente en términos de desviación estándar, lo que puede sugerir mayor sensibilidad a los datos de entrenamiento.

**Redes Neuronales Simples (NNET):**

Logró un accuracy medio de 99.92%, indicando un excelente desempeño.

Aunque comparable a los Bosques Aleatorios y SVM, requiere mayor esfuerzo computacional.

2. Desviación Estándar (Consistency Across Splits)

**Bosques Aleatorios** tuvo la desviación estándar más baja (0.0006), lo que indica una gran consistencia en las iteraciones de validación cruzada.

**SVM y Redes Neuronales** tuvieron una desviación ligeramente mayor (0.0008 y 0.0009, respectivamente), pero aún son modelos confiables.

**Regresión Logística** mostró la mayor desviación estándar (0.0012), lo que sugiere menor estabilidad en los resultados.

3. Pruebas Estadísticas

**Prueba de Friedman:**

Indicó diferencias estadísticamente significativas entre los modelos ($p < 0.05$), confirmando que al menos un modelo tiene un rendimiento superior.

**Prueba de Wilcoxon:**

Las comparaciones por pares mostraron que:

**Random Forest** supera significativamente a los demás modelos en varias comparaciones. Las diferencias entre SVM, Redes Neuronales y Regresión Logística son menos marcadas.

4. Clasificación de Modelos por Desempeño

**Bosques Aleatorios:** Mejor modelo por precisión y consistencia.

**SVM (Linear Kernel):** Excelente alternativa por su precisión y simplicidad.

**Redes Neuronales Simples:** Competitivas, pero su mayor complejidad puede no justificar su uso frente a Bosques Aleatorios.

**Regresión Logística:** Adecuada para escenarios que requieren interpretabilidad, aunque menos
precisa que los modelos anteriores.

5. Evaluación Basada en Métricas

**Matriz de Confusión:**

Todos los modelos muestran un desempeño sobresaliente, con Bosques Aleatorios logrando una clasificación perfecta en los datos de prueba.

**Precisión y Recall:**

En todos los modelos, las clases de estudiantes desertores y no desertores son bien diferenciadas, con recall cercano al 100%, lo que significa que pocos estudiantes desertores quedan sin identificar.

# CONCLUSIONES

Durante el análisis, se examinaron datos sociodemográficos, académicos y financieros, lo que permitió una comprensión profunda de las causas que contribuyen al abandono de los estudios por parte de los estudiantes.

A través del análisis exploratorio, se destacaron como variables más relevantes el promedio general, el número de asignaturas aprobadas y reprobadas, el máximo semestre cursado y factores personales como la edad y los incentivos financieros. Los hallazgos revelaron que los estudiantes con un desempeño académico bajo, reflejado en promedios reducidos y altas tasas de reprobación, tienen un mayor riesgo de deserción. Además, la falta de apoyo económico también fue un factor determinante, lo que subraya la necesidad de combinar estrategias académicas y financieras para abordar el problema.

Los análisis estadísticos, incluyendo las pruebas de Friedman y Wilcoxon, confirmaron diferencias significativas en el rendimiento de los modelos, posicionando a bosques aleatorios como la mejor opción. Este modelo no solo ofrece alta precisión, sino que también identifica claramente las variables más influyentes, proporcionando a las instituciones educativas una herramienta valiosa para el monitoreo y prevención de la deserción.

Finalmente, se recomienda implementar el modelo de bosques aleatorios como herramienta principal para predecir la deserción y diseñar estrategias preventivas basadas en los factores identificados. Adicionalmente, es crucial ampliar el análisis incorporando variables adicionales como la satisfacción estudiantil o la motivación, lo que permitiría una mayor precisión en la identificación de estudiantes en riesgo. Este proyecto establece una base sólida para reducir la deserción y mejorar la retención estudiantil en las instituciones educativas.

# REFERENCIAS


[1] M. d. E. Nacional, «Ministerio de Educación Nacional,» 2015. Available: https://www.mineducacion.gov.co/1759/articles-356272_recurso.pdf.

[2] C. A. International, «¿Qué es fundamental para alcanzar los ODS? Prevenir el abandono escolar,» Artículo en línea, 17 07 2015. [En línea]. Available: https://www.creativeassociatesinternational.com/es/blog/whats-fundamental-to-achieving-sdgs-preventing-school-dropout/.

[3] M. B. S. U. y. J. V. Marina Bassi, «Abandono escolar en América Latina: Un desafío para la productividad y la inclusión social,» https://publications.iadb.org/publications/spanish/document/Desconectados-Habilidades-educaci%C3%B3n-y-empleo-en-Am%C3%A9rica-Latina.pdf, Washington, D.C.: BID., 2012.

[4] M. A. A. J. L. &. H. J. A. Zavala, «La desmotivación y su relación con factores académicos y psicosociales de estudiantes universitarios,» *Revista Digital de Investigación en Docencia Universitaria 15(2),* p. https://doi.org/10.19083/ridu.2021.1392, 2021.

[5] M. d. E. Nacional, «Ministerio de Educación Nacional,» 2022. Available: https://www.mineducacion.gov.co/portal/publicaciones/Desercion_escolar_en_Colombia.pdf. [Último acceso: 15 02 2025].

[6] A. J. y. J. V. Díaz, «Revisión sistemática de literatura: Técnicas de aprendizaje automático (machine learning),» *https://ojs.tdea.edu.co/index.php/cuadernoactiva/article/download/849/1366/3677,* vol. 13, pp. 113-121, 2021.

[7] S. Q. F. y. O. F. T. Julio César Solís Ventura, «Modelo de regresión logística para la estimación de la deserción escolar del posgrado en la Universidad Técnica de Manabí, Ecuador,» *Revista Bases de la Ciencia,* p. https://revistas.utm.edu.ec/index.php/Basedelaciencia/article/view/5197, 2022.

[8] S. Haykin, Neural Networks: A Comprehensive Foundation (2ª ed.), Prentice Hall, 1999.

[9] D. E. R. y. J. L. McClelland, Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 1), MIT Press, 1986.

[10] J. S. G. S. y. J. E. E. D. Marco Javier Suárez Barón, «Red neuronal profunda (DNN) aplicada al análisis de deserción estudiantil en una Institución de Educación Superior,» *Investigación e Innovación en Ingenierías,* vol. 10(1), nº https://revistas.unisimon.edu.co/index.php/innovacioning/article/view/5607, pp. 202-214, 2022.

[11] J. H. F. R. A. O. y. C. J. S. Leo Breiman, Classification and Regression Trees, Wadsworth International Group, 1984.

[12] W.-Y. Loh, «Fifty Years of Classification and Regression Trees,» *International Statistical Review,* pp. 329-248, 2014.

[13] A. D. Vergaray, «Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la Educación Superior Privada,» *UCV-Scientia,* pp. https://revistas.ucv.edu.pe/index.php/ucv-scientia/article/view/1189, 2016.

[14] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995.

[15] L. Breiman, «Random Forests,» Revista: Machine Learning, 2001. [En línea]. Available: https://link.springer.com/article/10.1023/A:1010933404324. [Último acceso: 5-32].

[16] M. d. E. N. d. C. y. U. d. l. Andes, «Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES),» https://www.mineducacion.gov.co/sistemasdeinformacion/1783/w3-channel.html, Bogotá, 2005.

[17] T. A. Valencia, Factores que motivan el abandono estudiantil en la Universidad: Un estudio de caso, Londres: Editorial Académica Española, 2018.

[18] M. R. Urrego, «La investigación sobre deserción universitaria en Colombia 2006-2016: Tendencias y resultados,» *Revista: Pedagogía y Saberes,* p. 51, 2019.

[19] J. C. M. y. S. P. Mateus, «Propuesta de un Modelo Predictivo utilizando Aprendizaje Profundo para el análisis de deserción estudiantil en Universidades Colombianas Virtuales,» *Revista Innovación Digital y Desarrollo Sostenible - IDS,* pp. 51-57, 2020.

[20] J. E. S. C. y. C. C. R. Rodríguez, «Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos,» *Revista Científica,* pp. 39-50, 2016.

[21] W. G. y. R. S. Blanca Cuji, «Modelo predictivo de deserción estudiantil basado en árboles de decisión,» *Revista Espacios,* p. 17, Volumen y número: 38(55) 2017.

[22] L. L. E. N. E. M. y. P. C. N. Jovial Niyogisubizoa, «Predicting student´s dropout in university classes using two-layer ensemble amchine learning approach: a novel stacked generalization,» *Computers and Education: Artificial Intelligence,* vol. 3, nº 1000066, 2022.

# ANEXOS

En esta sección, se lleva a cabo una comparación entre la variable de deserción y cada una de las variables predictoras, ya sean categóricas o cuantitativas. El objetivo es identificar posibles relaciones, lo que facilita la detección de patrones o la incidencia de factores sociodemográficos, educativos y financieros en el fenómeno de la deserción universitaria.

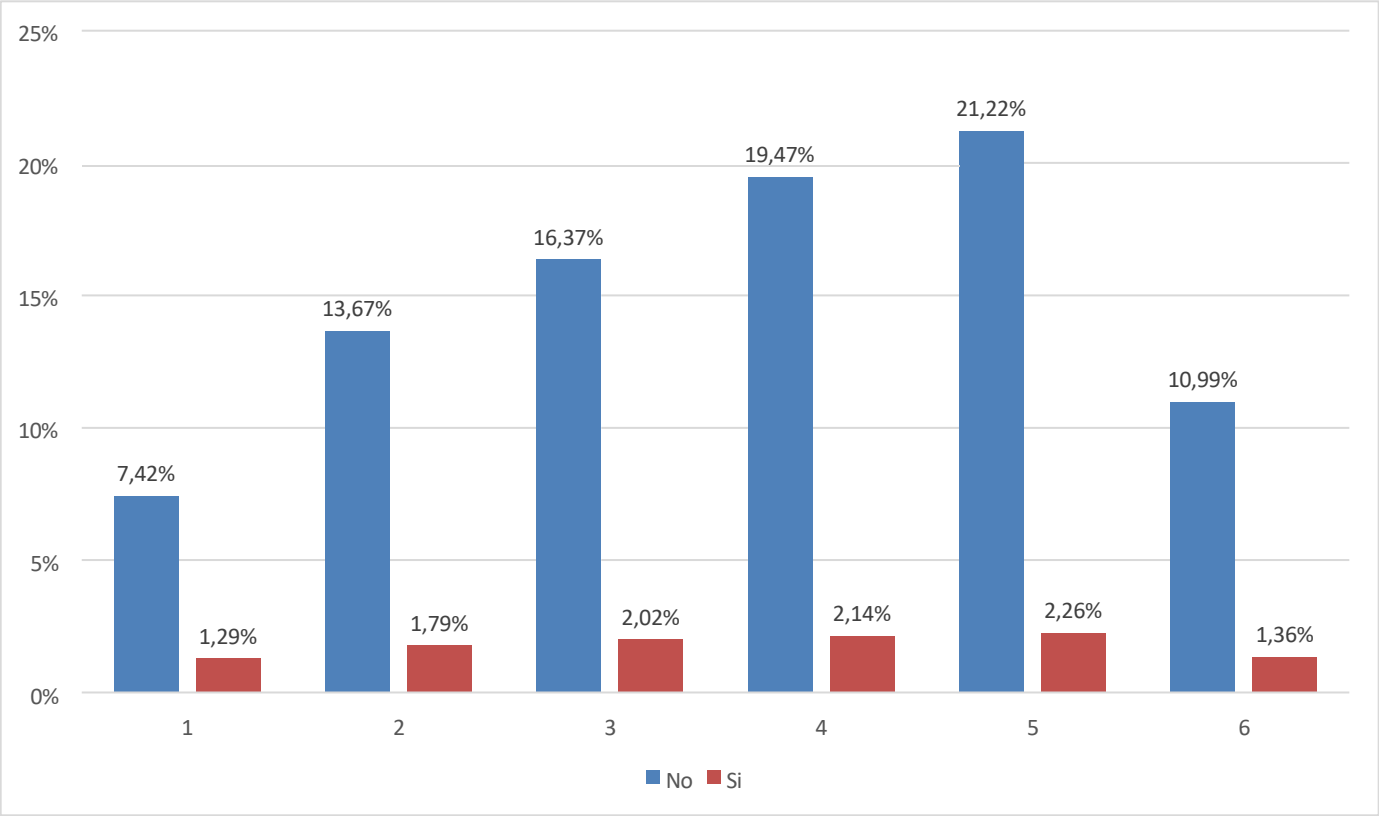


Figura 13. Relación entre DESERTOR y estrato socioeconómico (Fuente: elaboración propia)

La relación entre el estrato socioeconómico y la deserción universitaria revela un patrón consistente, donde se observa que los porcentajes de deserción son comparativamente similares en los estratos 3, 4 y 5. No se observan diferencias significativas entre estos estratos, como se detalla en la Gráfica 12

En el siguiente análisis, se examina la proporción de desertores en relación con sus fuentes de financiamiento. Se destaca que los mayores porcentajes de deserción se observan entre aquellos estudiantes que dependen de recursos familiares, con un 34.56%, y cuando se suma a los recursos propios, se registra un total del 43.67%. Estos hallazgos indican que más del 50% de los estudiantes que abandonan requieren algún tipo de financiamiento o apoyo económico para poder llevar a cabo sus estudios. Este análisis resalta la relevancia de considerar estrategias y políticas que aborden las necesidades financieras de los estudiantes como una parte integral de los esfuerzos para reducir la deserción en la universidad.

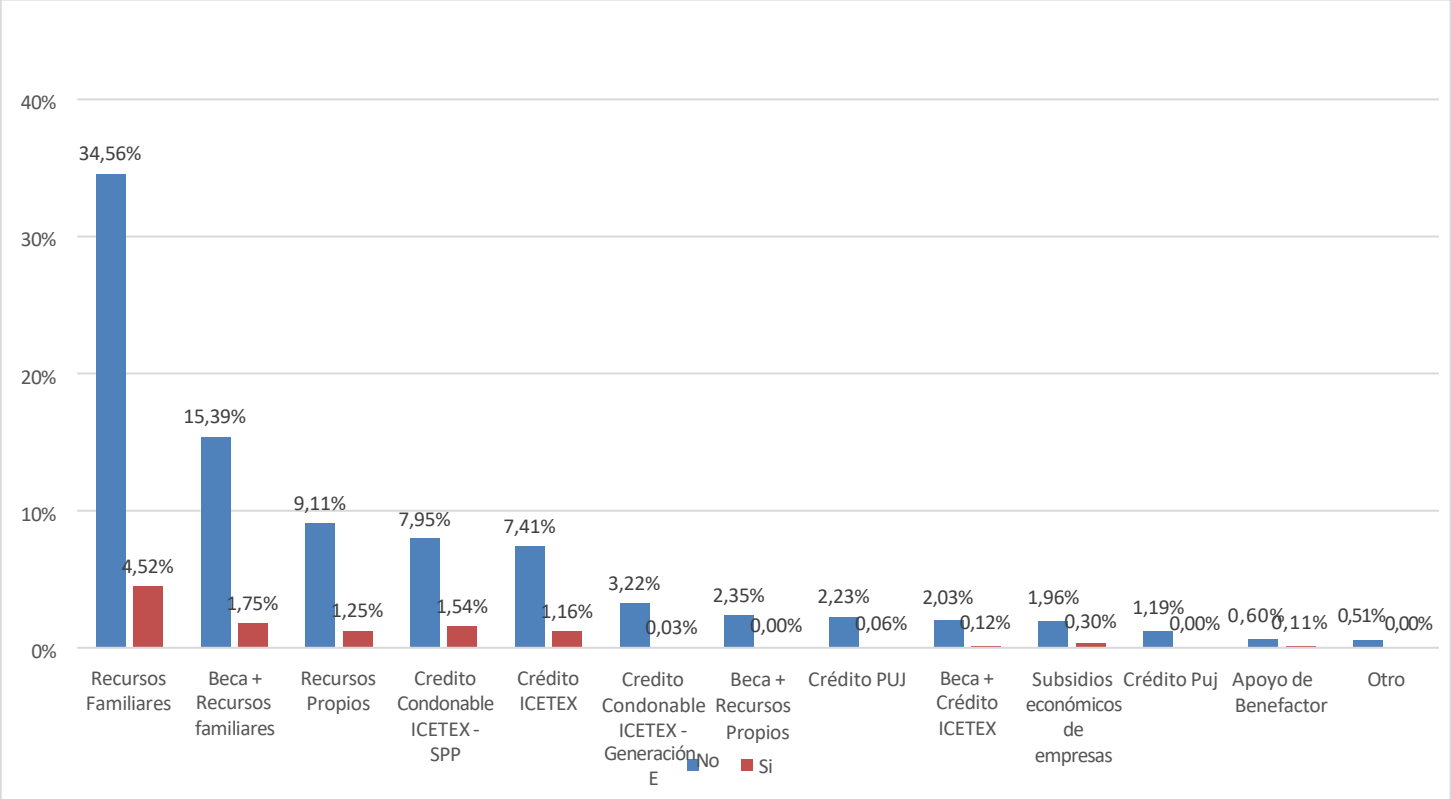

Figura 14. Relación entre DESERTOR y forma de pago (Fuente: elaboración propia)

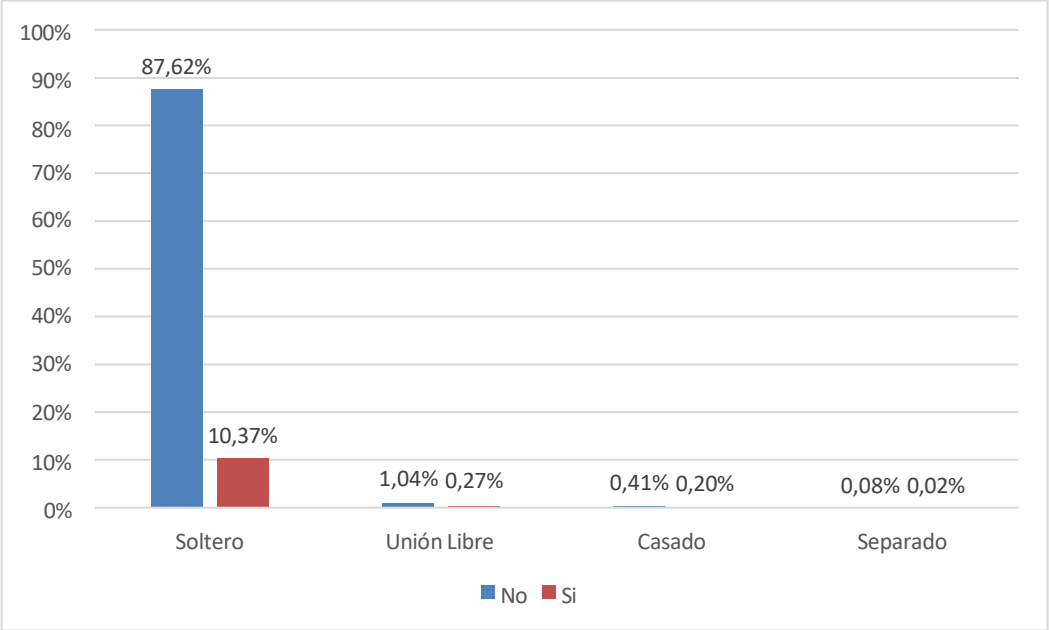


Figura 15. Relación entre DESERTOR y Estado civil (Fuente: elaboración propia)

Los datos también arrojan un hallazgo significativo el cual nos muestra que el estado civil en el cual más estudiantes desertan es el soltero a razón del 21 %, dado que va en relación con la preponderancia de la edad promedio de los estudiantes y la cantidad de estudiantes solteros.

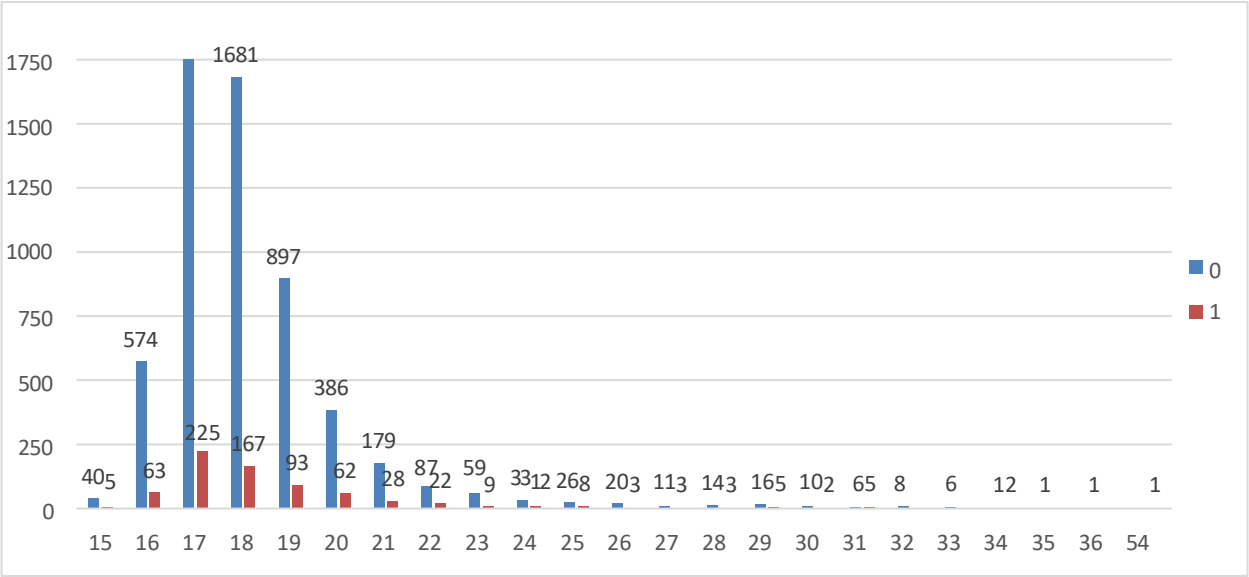


Figura 16. Relación entre DESERTOR y Edad (Fuente: elaboración propia)

En relación con la edad, es relevante señalar que el 27% de los estudiantes desertores tiene 17 años. Sin embargo, también es notable que el rango de 15 a 19 años concentra el mayor número de estudiantes, sugiriendo que este grupo de edades representa una potencial área deenfoque para mitigar los riesgos de deserción

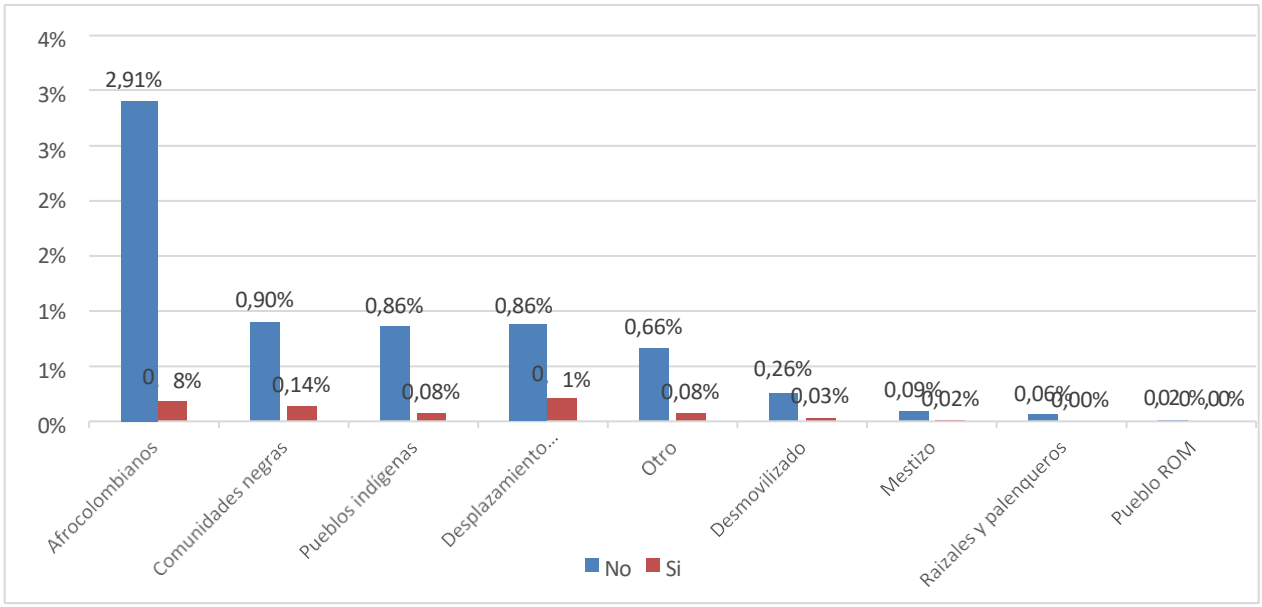


Figura 17.Porcentaje total de deserción por población de pertenencia de los estudiantes (Fuente: elaboración propia)

En relación con la distribución por tipo de población estudiantil, se observa un número reducido, aunque es importante considerar la posibilidad de que esto se deba al desconocimiento de la información. Cabe resaltar que los estudiantes afrodescendientes ocupan la posición más alta en cuanto a deserción, seguidos por las comunidades negras y pueblos indígenas. Este hallazgo subraya la importancia de abordar de manera específica las necesidades y desafíos que enfrentan estos grupos para mejorar las tasas de retención estudiantil.

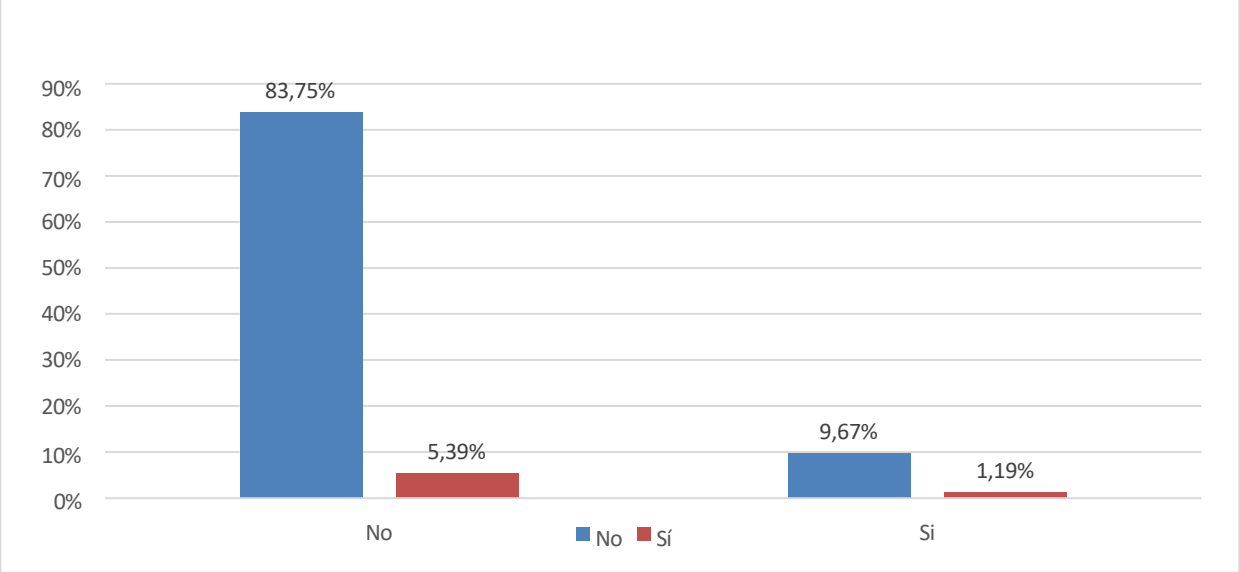


Figura 18. Relación entre DESERTOR y situación laboral

Otro aspecto crucial analizado fue la relación entre la deserción universitaria y la ocupación de los estudiantes. Se destaca que el 9.67% de los estudiantes que abandonaron sus estudios no se encontraba trabajando en el momento del análisis. Este dato resalta la necesidad de examinar detenidamente las circunstancias individuales y los factores que pueden influir en la decisión de desertar, incluso cuando los estudiantes no están actualmente empleados

**Otros hallazgos en el análisis univariado**

Después de culminar la fase de limpieza de datos, se llevó a cabo un análisis exhaustivo de cada variable individual que podría ejercer influencia sobre la variable explicativa de la Deserción. A continuación, se presentan las características fundamentales de cada una de estas variables.

Estrato socioeconómico, los resultados evidencias que la distribución de los estudiantes se concentra principalmente en los estratos 5 (20.80%), 4 (19.69%), 3 (18.59) los cuales en conjunto representan el 60% de la población estudiantil.

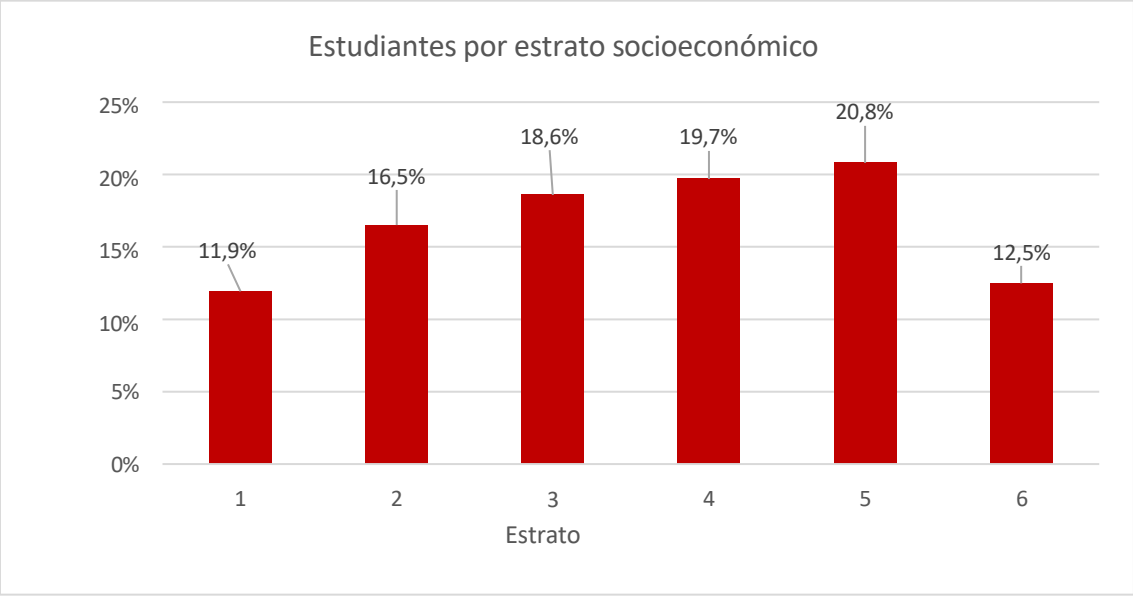


Figura 4. Porcentaje de estudiantes por estrato socioeconómico (Fuente: elaboración propia)

En relación con la financiación de los estudios, sobresale que la mayoría de los estudiantes (41.61%) opta por utilizar recursos familiares como su principal fuente de financiamiento. En segundo lugar, un 16.29% elige combinar becas con recursos familiares. Los porcentajes restantes están asociados a becas provenientes de diversas fuentes, como se detalla en la Gráfica 3.

Sin embargo, llama la atención que un considerable 47% de los estudiantes depende de créditos,

becas o apoyo económico para iniciar sus estudios académicos. Este análisis pone de manifiesto la diversidad de estrategias financieras adoptadas por los estudiantes, subrayando laimportancia de considerar opciones variadas para asegurar el acceso equitativo a la educación superior.

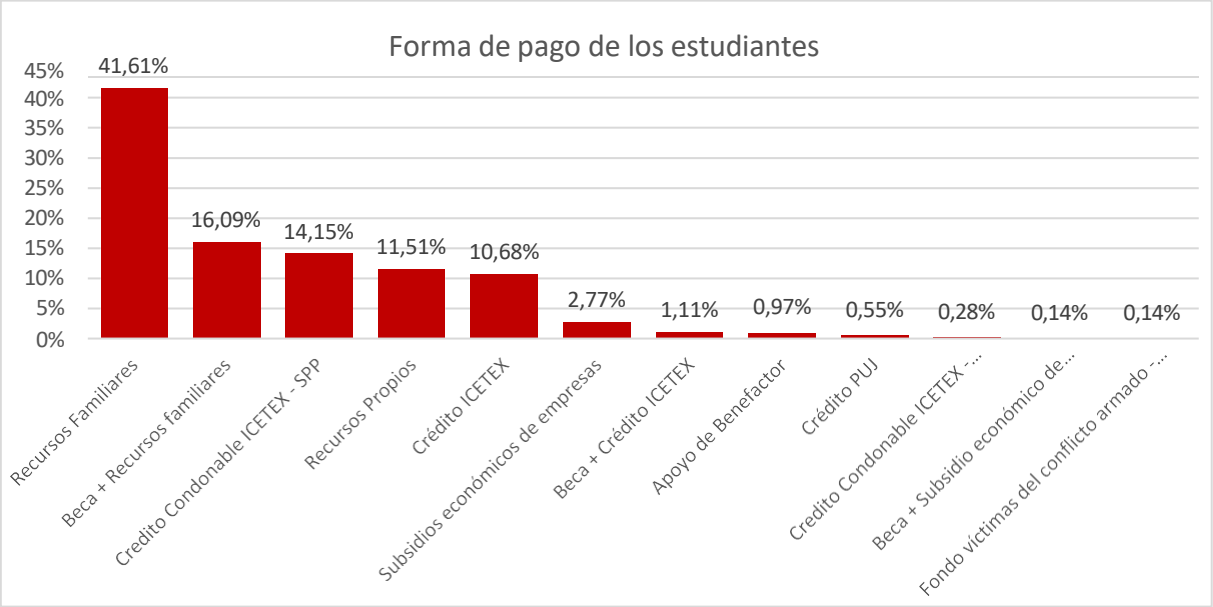


Figura 5. Porcentaje de cómo pagan los estudios los estudiantes (Fuente: elaboración propia)

Al examinar el estado civil de los estudiantes, se destaca que un porcentaje considerable de ellos se identifica como soltero, abarcando el 95.56%. El porcentaje restante se distribuye entrelas demás categorías, tal como se ilustra en el gráfico de Porcentaje de Estado Civil de los estudiantes

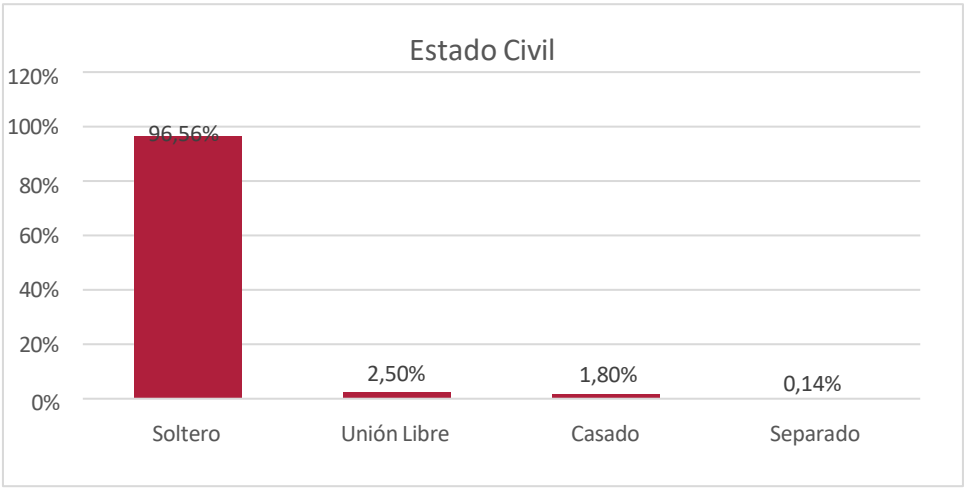


Figura 6. Estado civil (Fuente: elaboración propia)

**Fuente**: elaboración propia

Por otro lado, se analiza la distribución de las ocupaciones de los estudiantes, revelando que un 10.96% de ellos declaran estar empleados actualmente, según se evidencia en la Gráfica 5.

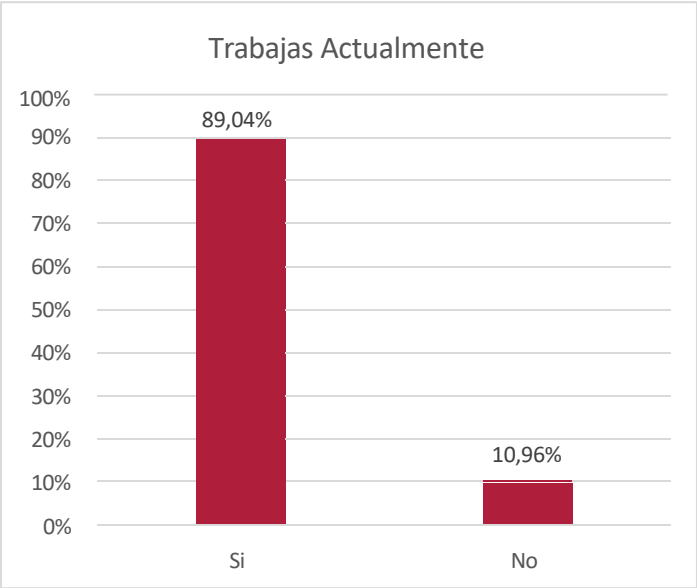


Figura 7. Porcentaje de estudiantes que trabajan actualmente (Fuente: elaboración propia)

Se incluyó otra variable en el estudio: el lugar de procedencia de los estudiantes. En este aspecto, el 8.67% proviene de Cali, mientras que el resto proviene de otras ciudades, como se representa en la Gráfica 6.

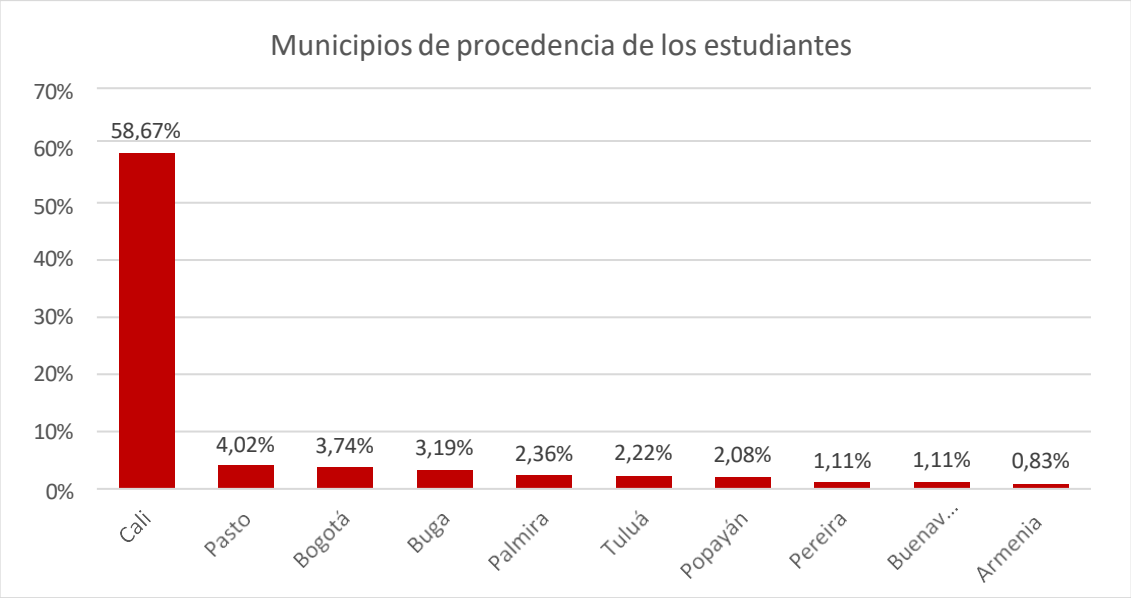


Figura 8. Porcentaje de lugar de procedencia municipio (Fuente: elaboración propia)

Un notable porcentaje (27.32%) de los estudiantes desertores proviene de familias con ingresos superiores a 6,001,000, sugiriendo que entre ellos se encuentran aquellos cuyas familias disfrutan de un nivel económico considerable. Además, una proporción significativa (23.30%) se sitúa en el rango de ingresos de 700,000 a 2,100,000, indicando que una parte sustancial de los desertores proviene de familias con ingresos de nivel medio. Por otro lado, el 5.96% tiene ingresos inferiores a 700,000, señalando que una minoría de estudiantes desertores proviene de entornos familiares con ingresos más bajos.

Es relevante destacar que el 1.94% no dispone de información sobre los ingresos familiares. Esta circunstancia resalta la importancia de reconocer la limitación de datos precisos y suimpacto en la interpretación de resultados *Gráfica 7*.

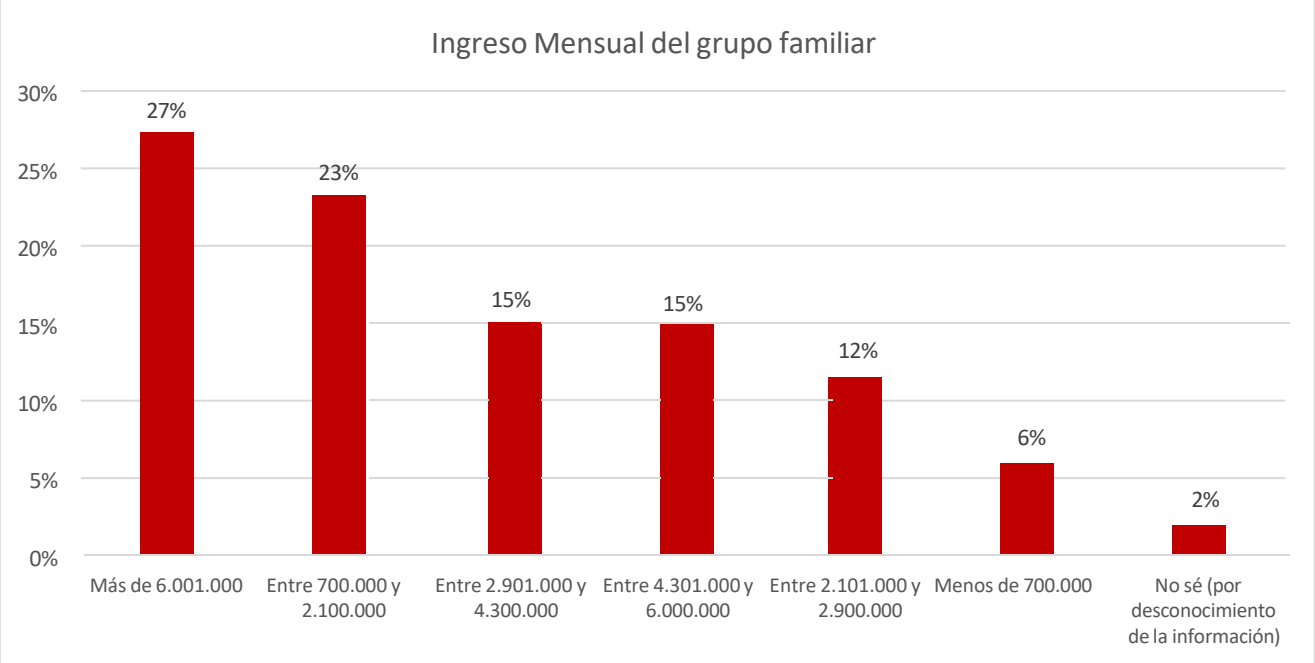


Figura 9. Porcentaje de ingresos mensuales de tu grupo familiar (Fuente: elaboración propia)

No obstante, al examinar el nivel educativo de los padres, se observa que un 32.73% de las madres ha completado sus estudios académicos, en comparación con un 27.7% de los padres. En lo que respecta a los posgrados, un 18% de los padres posee un título de posgrado, en contraste con un 14.4% de las madres. Este análisis revela que, en general, las madres experimentan una mayor proporción de no culminar sus estudios en comparación con los padres embargo, al analizar el nivel educativo de los padres se evidencias que la madre logra terminar sus estudios académicos con un 32,73% en comparación con el padre con un 27,7 % por el lado de los posgrados el 18% de los padres tiene un posgrado en comparación con un 14,4 % de la madre, se evidencia también que las madres no logran terminar sus estudios en mayor proporción que los padres

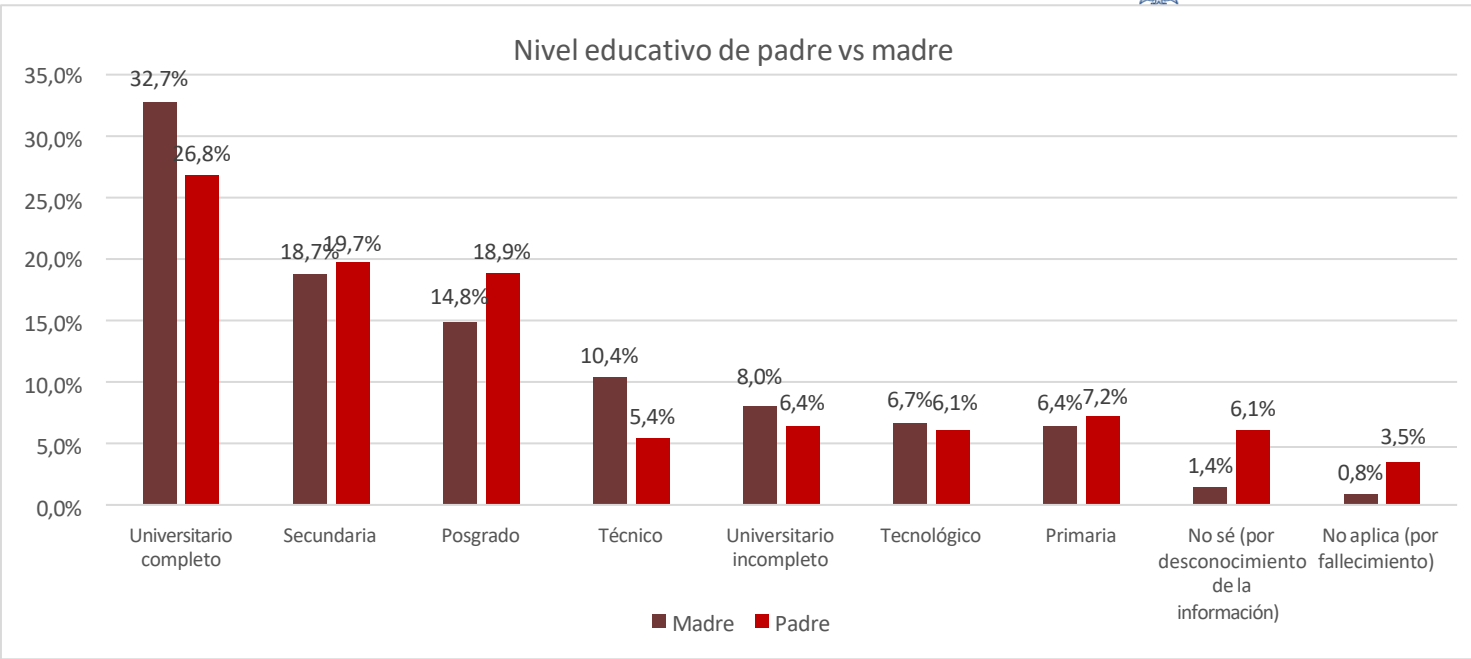


Figura 10. Porcentaje de nivel educativo de los padres (Fuente: elaboración propia)

**Fuente**: elaboración propia

En cuanto a la edad, se destaca que el 31% de los estudiantes desertores tiene 17 años; no obstante, en el rango de 15 a 19 años, se observa que el 76% de los estudiantes abandonan sus estudios. Este hallazgo indica que los estudiantes más jóvenes enfrentan un mayor riesgo de abandono de estudios universitarios.

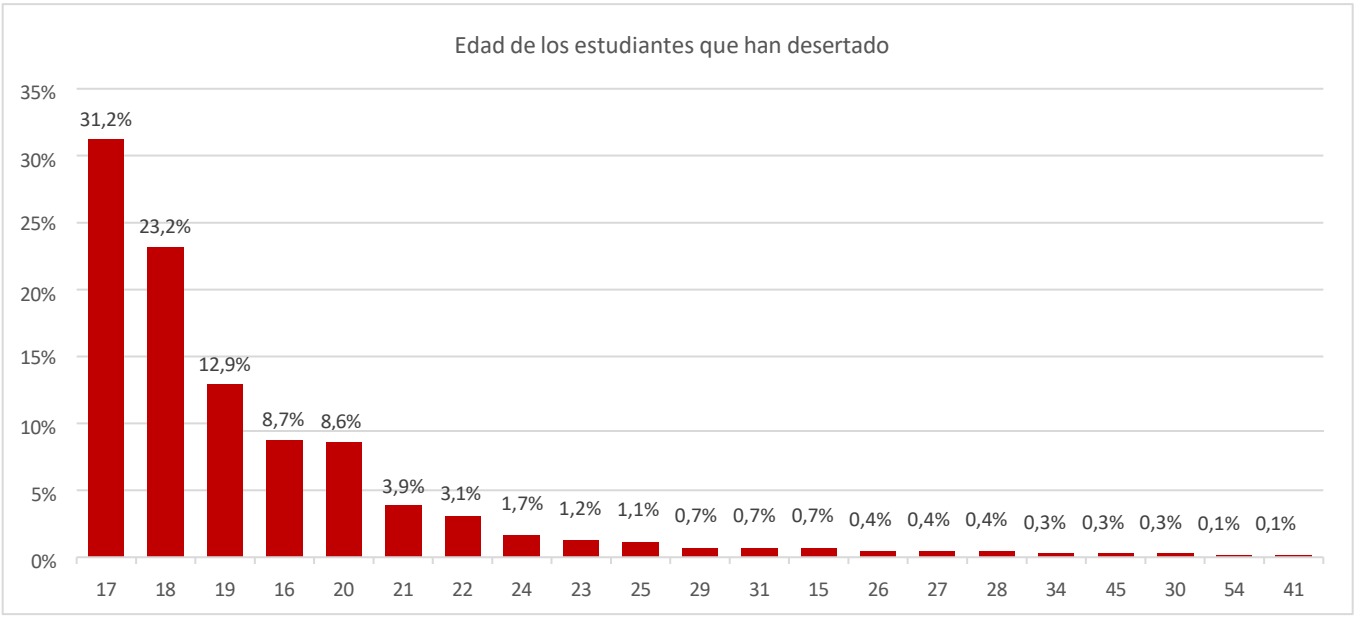


Figura 11. Promedio de edad de los estudiantes que han desertado (Fuente. elaboración propia)

En cuanto a la distribución por tipo de población estudiantil, se resalta la presencia significativa de estudiantes en cada una de las categorías identificadas. Se evidencia que un 2.08% corresponde a estudiantes en situación de desplazamiento forzoso, mientras que un 1.6% se autoidentifica como afrocolombianos. Además, el 0.46% de la población estudiantil proviene de origen desmovilizado, y otras categorías representan porcentajes adicionales, como se visualiza en la Gráfica 11. Este análisis subraya la diversidad de la población estudiantil y destaca la importancia de considerar y abordar las distintas realidades y necesidades presentes en el contexto educativo.

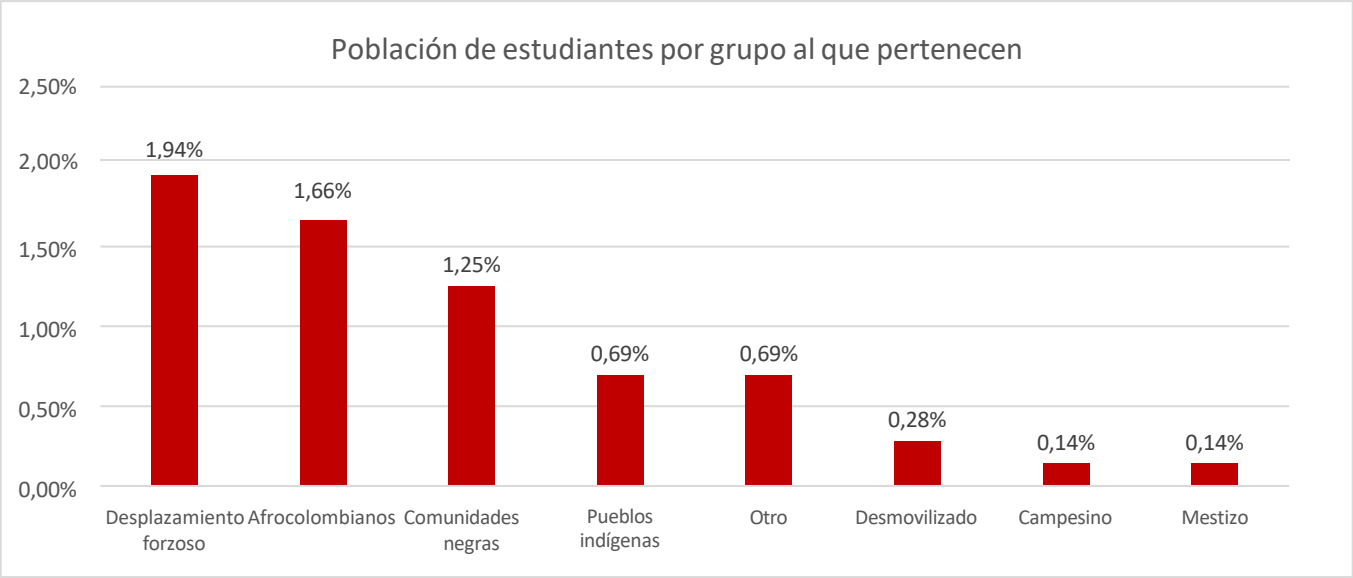


Figura 12. Porcentaje de estudiantes por población a la que pertenece (Fuente: elaboración propia)

Cali, 21 Julio 2022

Facultad de Ingeniería
Dirección de Posgrados
Pontifica Universidad Javeriana Cali

Asunto: Autorización para el tratamiento y uso de datos institucionales

Cordial saludo,

De conformidad con lo previsto en la Ley 1581 de 2012 "*por la cual se dictan las disposiciones generales para la protección de datos personales*" y el Decreto 1377 de 2013, que la reglamentan parcialmente, manifiesto que otorgo mi autorización a: Sandra Paola Botero con Cedula de ciudadanía No. 25.291.837 estudiante de la Maestría en Ciencia de Datos, pueda hacer tratamiento y uso de datos institucionales de estudiantes con fines académicos en su proyecto de grado.

De acuerdo con la normatividad citada, queda autorizado de manera expresa e inequívoca para mantener y manejar la información suministrada, solo para aquellas finalidades para las que se encuentra facultado y respetando en todo caso, la normatividad vigente sobre protección de datos personales.

Atentamente,

Mauricio José Cortés
Director
Oficina de Planeación Institucional
Pontificia Universidad Javeriana Cali