Predicción y Análisis de la Deserción Estudiantil en la Facultad de Ingeniería de la Universidad El Bosque mediante Modelos Avanzados de Machine Learning: Identificación de variables más influyentes, Caracterización de Graduados y No Graduados, y Detección Temprana

T. Hernández[*1], D. Useche[**1], O. Fajardo[***1], and E. Rodriguez[****1]

[1] *Maestría en Estadística Aplicada, Universidad el Bosque, Bogotá, Colombia*

Resumen

Este estudio presenta un enfoque mixto para predecir y analizar la deserción estudiantil en la Facultad de Ingeniería de la Universidad El Bosque, integrando técnicas supervisadas y no supervisadas de aprendizaje automático. Se utilizaron datos de 6 054 estudiantes (2008–2025) con variables sociodemográficas (género, estrato socioeconómico, jornada, tipo de estudiante y estado civil) y académicas (edad y semestres aprobados). El preprocesamiento incluyó imputación de valores faltantes mediante KNN, codificación ordinal y selección de atributos.

Se entrenaron y compararon modelos supervisados (árboles de decisión, Random Forest, SVM, XGBoost y LightGBM) con validación cruzada estratificada. LightGBM obtuvo el mejor desempeño (F1-macro = 0,895; AUC = 0,95), destacándose por su bajo número de falsos negativos, ideal para alertas tempranas. La interpretabilidad del modelo se abordó con SHAP, que identificó como predictores clave los semestres aprobados, estrato, jornada, tipo de estudiante y edad.

Complementariamente, se aplicó K-Modes ($k = 7$) para identificar siete perfiles estudiantiles con trayectorias y condiciones heterogéneas. El Análisis de Correspondencias Múltiples reforzó esta segmentación, visualizando asociaciones significativas entre categorías. Los perfiles resultantes incluyen grupos consolidados, como mujeres avanzadas con alto desempeño, y otros en riesgo, como jóvenes de estrato bajo o adultos en jornada nocturna.

En conjunto, este marco metodológico permite anticipar riesgos de deserción, comprender la diversidad estudiantil y orientar decisiones institucionales. Su aplicación puede fortalecer políticas de retención, optimizar recursos y promover una educación más equitativa y basada en datos.

Palabras Clave: Deserción estudiantil, machine learning, lightGBM, SHAP, agrupamiento K-Modes.

Abstract

This study presents a mixed approach to predicting and analyzing student dropout in the School of Engineering at Universidad El Bosque, integrating supervised and unsupervised machine learning techniques. The dataset comprises 6 054 students (2008–2025) with sociodemographic variables (gender, socioeconomic status, study schedule, student type, and marital status) and academic variables (age and approved semesters). Data preprocessing included missing value imputation using KNN, ordinal encoding, and feature selection.

Supervised models (decision trees, Random Forest, SVM, XGBoost, and LightGBM) were trained and evaluated using stratified cross-validation. LightGBM achieved the best performance (F1-macro = 0.895; AUC = 0.95), with a notably low false-negative rate, making it suitable for early dropout alert systems. SHAP was used to interpret the model, identifying the most influential predictors as the number of approved semesters, socioeconomic status, study schedule, student type, and age.

In parallel, K-Modes clustering ($k = 7$) was applied to uncover seven distinct student profiles with heterogeneous academic and demographic trajectories. Multiple Correspondence Analysis (MCA) further supported this segmentation by revealing significant associations among categorical variables. The resulting profiles range from consolidated trajectories—such as high-performing advanced female students—to at-risk groups like low-income freshmen or adult night-shift students.

Overall, this methodological framework enables accurate dropout prediction, nuanced student profiling, and data-driven decision-making. Its application can enhance institutional retention policies, improve resource allocation, and promote a more equitable and evidence-based approach to higher education management.

Keywords: Academic dropout, machine learning, lightGBM, SHAP, K-Modes clustering.

1. Introducción

La deserción estudiantil en la educación superior constituye una de las problemáticas más persistentes y desafiantes en América Latina, especialmente en programas de ingeniería, donde las tasas de abandono suelen superar los promedios nacionales (Ministerio de Educación Nacional de Colombia, 2022; UNESCO, 2023). Este fenómeno no puede ser atribuido a una sola causa, sino que resulta de la interacción compleja entre factores académicos, socioeconómicos, institucionales y personales. En este contexto, comprender de forma oportuna y precisa los elementos que influyen en la permanencia estudiantil se vuelve esencial para el diseño de estrategias de retención efectivas.

Los enfoques estadísticos convencionales han facilitado la identificación de asociaciones generales entre variables, sin embargo, presentan limitaciones al enfrentar relaciones no lineales y estructuras de datos multifactoriales (T. Nguyen, 2020). Esto ha incentivado la incorporación de modelos de aprendizaje automático (*machine learning*) que, por su capacidad de manejar grandes volúmenes de información y descubrir patrones complejos, se perfilan como herramientas potentes para el análisis educativo.

En este estudio se propone una aproximación metodológica integral que combina algoritmos avanzados de *machine learning* con técnicas de interpretabilidad y segmentación. En particular, se centra en la aplicación del modelo SHAP (SHapley Additive exPlanations) como núcleo explicativo, complementado por métodos de agrupamiento como K-Modes y análisis de correspondencias múltiples. Esta integración tiene como finalidad no solo predecir con alta precisión la probabilidad de graduación o abandono, sino también proporcionar explicaciones claras y comprensibles sobre los factores que inciden en dichos resultados, promoviendo así una toma de decisiones más informada, ética y centrada en el estudiante.

El modelo SHAP ha demostrado ser especialmente valioso en entornos educativos, al permitir descomponer cada predicción en contribuciones individuales de las variables. A diferencia de otros enfoques, SHAP se basa en la teoría de juegos cooperativos y garantiza la consistencia y justicia en la distribución de importancia entre variables (Lundberg & Lee, 2017). Su uso ha mejorado la precisión y la transparencia en sistemas de alerta temprana, facilitando intervenciones diferenciadas y fortaleciendo la confianza en el uso institucional de la inteligencia artificial (Dake et al., 2024; Estrada & Rodríguez, 2022).

Esta investigación se apoya en datos institucionales de la Facultad de Ingeniería de la Universidad El Bosque, compuestos por variables académicas, demográficas y administrativas. Se emplean herramientas de código abierto como `Python`, `Scikit-Learn`, `XGBoost` y `LightGBM`, que permiten una implementación técnica eficiente y reproducible.

El objetivo general de este trabajo es desarrollar un modelo predictivo interpretable que permita identificar las variables más influyentes asociadas a la deserción estudiantil. Para ello, se plantean tres objetivos específicos: (i) Diseñar e implementar un modelo de aprendizaje automático para predecir la graduación de estudiantes de programas de pregrado a partir de datos mixtos, integrando SHAP como herramienta de interpretación del modelo.(ii) categorizar a los estudiantes en grupos de riesgo mediante técnicas de *clustering*; y (iii) caracterizar los perfiles de graduados

y no graduados mediante análisis de correspondencias múltiples.

Los resultados preliminares muestran que el modelo propuesto no solo alcanza niveles de precisión competitivos, sino que también permite visualizar con claridad los factores que determinan las trayectorias académicas. Esto aporta evidencia clave para fortalecer las políticas institucionales de permanencia, reducir las tasas de deserción y contribuir a la equidad educativa desde un enfoque analítico y predictivo.

2. Marco Teórico

Deserción estudiantil: definición y contexto

La deserción estudiantil constituye una de las principales problemáticas que enfrenta el sistema educativo a nivel mundial, especialmente en los niveles medio y superior. Este fenómeno se manifiesta cuando los estudiantes interrumpen su trayectoria académica antes de completar el programa educativo en el que se encuentran matriculados, ya sea de manera temporal o definitiva. Desde un enfoque institucional, el Ministerio de Educación Nacional de Colombia (Ministerio de Educación Nacional de Colombia, 2022) define la deserción como la salida de un estudiante del sistema educativo sin haber concluido el ciclo o nivel correspondiente, y propone la Tasa de Deserción Anual (TDA) como indicador clave para su monitoreo.

Diversos organismos han abordado este fenómeno desde múltiples dimensiones. La UNESCO (UNESCO, 2023) destaca que la deserción responde a una interacción compleja de factores estructurales y personales, entre ellos: condiciones socioeconómicas, género, barreras de acceso a recursos educativos, violencia, discriminación, y prácticas pedagógicas poco inclusivas. De manera similar, en el contexto latinoamericano, el Instituto Nacional de Estadística y Geografía (Instituto Nacional de Estadística y Geografía (INEGI), 2023) identifica el abandono como el porcentaje de estudiantes que no continúan sus estudios respecto a la matrícula inicial de un ciclo educativo, subrayando su estrecha relación con la pobreza, el trabajo infantil y la falta de oportunidades.

Estudios recientes han enfatizado que la deserción no solo tiene implicaciones individuales, como la disminución de oportunidades laborales y el incremento del riesgo de exclusión social, sino también consecuencias colectivas para las instituciones educativas y los sistemas de desarrollo económico (Ministerio de Educación del Perú, 2021). En el caso de los estudiantes de secundaria, investigaciones cualitativas han identificado que las dificultades académicas, la falta de motivación y la percepción de baja relevancia de los contenidos escolares pueden ser determinantes en su decisión de abandonar los estudios (Flores-Báez, 2023).

Así, comprender la deserción desde una perspectiva multidimensional resulta fundamental para diseñar estrategias efectivas de intervención y retención. Este artículo se propone, por tanto, contribuir al estudio del fenómeno mediante un enfoque cuantitativo basado en modelos predictivos, integrando variables académicas, demográficas y socioeconómicas.

La deserción estudiantil en la educación superior ha sido ampliamente estudiada en las últimas décadas por su impacto negativo en los sistemas educativos, los proyectos de vida de los estudiantes y la eficiencia institucional (Tinto, 1993; UNESCO, 2023). En América Latina, y particularmente en Colombia, esta problemática se manifiesta de forma más aguda en programas como ingeniería, donde las tasas de abandono superan los promedios nacionales (Ministerio de Educación Nacional de Colombia, 2022). Las causas de la deserción son multicausales e incluyen factores académicos, socioeconómicos, institucionales y personales (Romero & Ventura, 2020), cuya interacción compleja y no lineal requiere enfoques analíticos más sofisticados para ser comprendida integralmente.

Tradicionalmente, los estudios sobre deserción han utilizado modelos estadísticos lineales, como la regresión logística, que si bien permiten identificar asociaciones generales entre variables, presentan limitaciones frente a relaciones más complejas y no capturan adecuadamente la interacción de múltiples factores (T. Nguyen, 2020). Ante ello, en los últimos años ha emergido un enfoque basado en el uso de técnicas de aprendizaje automático (*Machine Learning*), que permite construir modelos predictivos de alto rendimiento a partir de grandes volúmenes de datos (Chen, 2024). Estos modelos han demostrado ser eficaces para anticipar eventos de abandono académico, aunque frecuentemente se les critica por su falta de transparencia e interpretabilidad.

La deserción estudiantil en la educación superior constituye una problemática compleja que ha motivado la aplicación de enfoques avanzados de análisis predictivo. En los últimos años, los modelos de *Machine Learning* (ML) han demostrado ser herramientas eficaces para anticipar este fenómeno y ofrecer a las instituciones información clave para la toma de decisiones oportunas (Hanji, 2023; Khan, 2023).

Modelos como los árboles de decisión, *Random Forest* y máquinas de vectores de soporte (SVM) han sido ampliamente aplicados debido a su capacidad para manejar datos categóricos y numéricos, así como su rendimiento aceptable en entornos de datos educativos. Sin embargo, estudios recientes han evidenciado que algoritmos de *gradient boosting* como XGBoost, CatBoost y LightGBM ofrecen mejoras sustanciales en precisión, especialmente cuando se combinan con técnicas de balanceo como SMOTE o ADASYN (Kabir, 2024; Kaur & Sharma, 2023).

Por ejemplo, XGBoost ha logrado resultados sobresalientes en términos de AUC-ROC (hasta 0.95), permitiendo identificar con alta precisión a estudian-

tes en riesgo de abandono (Hanji, 2023). De manera similar, LightGBM ha sido aplicado con éxito en el análisis de factores académicos y sociodemográficos en programas de ingeniería, revelando variables determinantes como el tiempo de estudio, el rendimiento previo y la edad del estudiante (Sanchez et al., 2023).

2.1. Modelos de Machine Learning Aplicados a la Predicción de la Deserción Estudiantil: Fundamentos Matemáticos y Aplicaciones Recientes

La deserción estudiantil es un fenómeno complejo influenciado por factores académicos, sociales y personales. En los últimos cinco años, los modelos de Machine Learning (ML) se han posicionado como herramientas poderosas para la predicción temprana de este fenómeno, gracias a su capacidad para modelar relaciones no lineales y extraer patrones en grandes volúmenes de datos. Este apartado describe los modelos más utilizados, su fundamento matemático y aplicaciones recientes en el contexto de la deserción.

2.1.1. Árboles de Decisión

Los árboles de decisión (Figura 1)son modelos jerárquicos que segmentan iterativamente el espacio de datos mediante reglas simples basadas en atributos. La selección de atributos se basa en medidas como la entropía y la ganancia de información:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i), \tag{1}$$

$$IG(S, A) = H(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} H(S_v) \tag{2}$$

Estas fórmulas permiten elegir el atributo que más reduce la impureza del conjunto al dividirlo (Quinlan, 1986). Aplicaciones recientes han demostrado su eficacia en programas de ingeniería, alcanzando precisiones superiores al 98 % (Anahua et al., 2025).

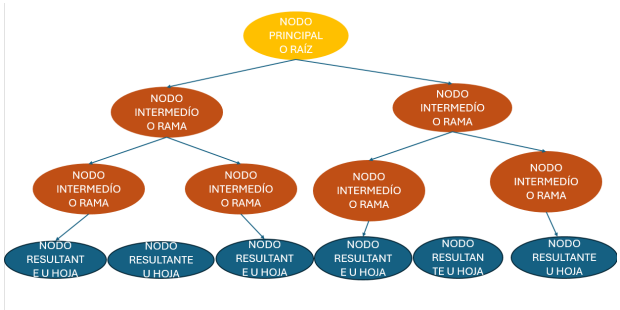

Figura 1: Representación grafica de árbol de decisión. (Torres Acero, 2022)

2.1.2. Random Forest

Random Forest (Figura 2) es un modelo de ensamblado que combina múltiples árboles de decisión construidos con subconjuntos aleatorios de datos y características. La predicción final se obtiene mediante votación mayoritaria:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), ..., h_T(x)) \tag{3}$$

Esta estrategia reduce la varianza del modelo individual y mejora la generalización. (Villar & de Andrade, 2024) demostraron que Random Forest es altamente eficaz en contextos educativos, especialmente cuando se aplican técnicas de balanceo de clases.

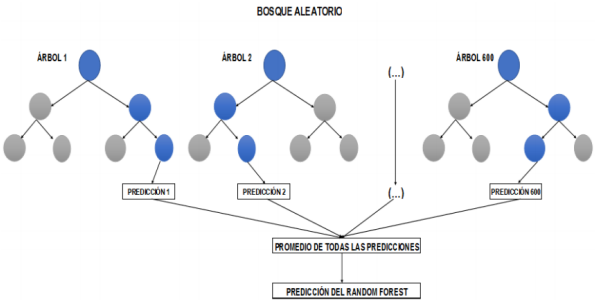

Figura 2: Random Forest. (Torres Acero, 2022)

2.1.3. Support Vector Machines (SVM)

SVM es un modelo de clasificación (Figura 3) que busca el hiperplano que mejor separa las clases, maximizando el margen:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{sujeto a} \quad y_i(w^T x_i + b) \geq 1 \tag{4}$$

Cuando los datos no son linealmente separables, se utilizan funciones kernel para proyectarlos a espacios de mayor dimensionalidad. Recientes estudios han demostrado que SVM ofrece buenos resultados para conjuntos de datos pequeños o balanceados (Boteju et al., 2024).

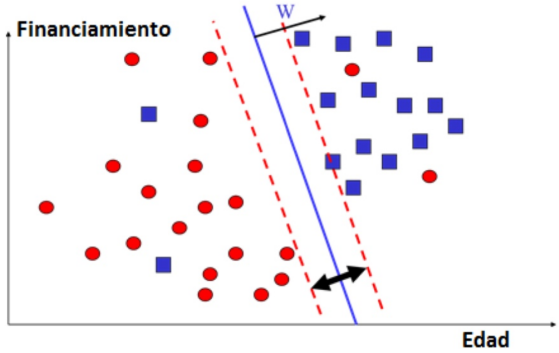

Figura 3: Soporte de máquina vectorial. (Vásquez, 2016)

2.1.4. XGBoost (Extreme Gradient Boosting)

XGBoost (Figura 4) es un algoritmo de boosting que optimiza una función objetivo regularizada combinando modelos aditivos:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{5}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

Donde f_t representa el nuevo árbol añadido en la iteración t. Este enfoque ha sido ampliamente validado en entornos educativos por su alta precisión y eficiencia computacional

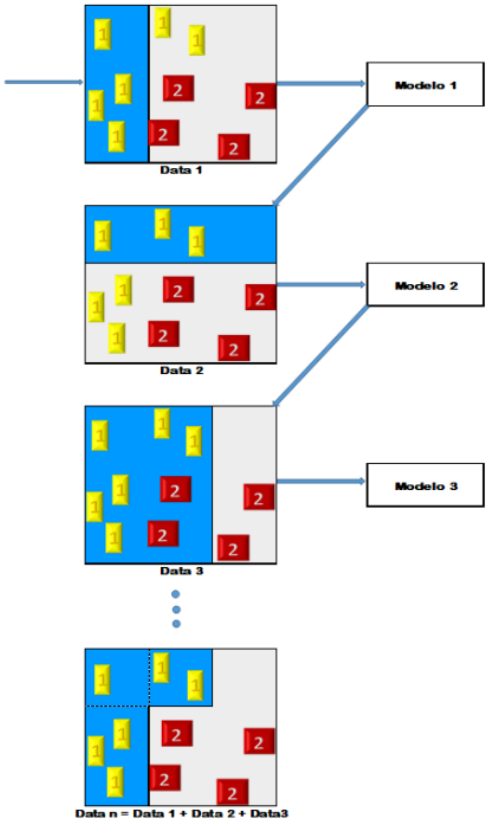

Figura 4: XGBoosting.(Torres Acero, 2022)

2.1.5. LightGBM (Light Gradient Boosting Machine)

LightGBM (Figura 5) es una variante de gradient boosting que utiliza estrategias de crecimiento por hojas (leaf-wise) y optimizaciones como GOSS y EFB para acelerar el entrenamiento:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{7}$$

Se ha destacado por su eficiencia en conjuntos de datos grandes y de alta dimensionalidad, superando en algunos casos a XGBoost en tareas de clasificación educativa (Ke et al., 2017).

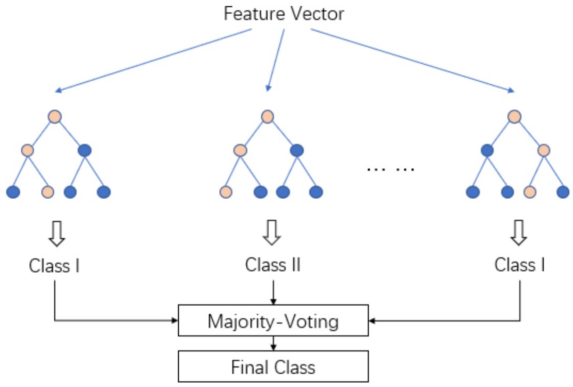

Figura 5: Representación gráfica del modelo LightGBM.

La inclusión de herramientas de interpretabilidad ha fortalecido estos modelos, haciendo posible no solo predecir la deserción, sino también comprender los factores que la explican. Técnicas como SHAP (*SHapley Additive exPlanations*) han permitido descomponer las predicciones de los modelos para asignar importancia a cada característica tanto a nivel individual como global (Liu et al., 2023). Este enfoque resulta especialmente valioso para los responsables institucionales, ya que facilita la creación de intervenciones focalizadas y personalizadas (Jin et al., 2024).

2.2. Definición del modelo SHAP

SHAP (SHapley Additive exPlanations) es un enfoque basado en teoría de juegos para explicar la salida de modelos de *machine learning*. Se fundamenta en los valores de Shapley, propuestos por Lloyd Shapley en 1953, los cuales distribuyen de manera justa la contribución de cada característica a una predicción individual (Lundberg & Lee, 2017).

El objetivo de SHAP es responder a la pregunta: ¿cuánto aporta cada característica a la predicción hecha por el modelo?

La fórmula general del modelo SHAP es:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{8}$$

Donde:

- $f(x)$ es la predicción del modelo.
- ϕ_0 es el valor esperado del modelo (predicción promedio en el conjunto de datos).
- ϕ_i representa el valor SHAP de la característica i, es decir, su contribución a la predicción.

Los valores SHAP cumplen con propiedades de justicia, consistencia y eficiencia, lo que los convierte en una herramienta sólida para explicar decisiones de

modelos complejos como árboles de decisión, XGBoost o redes neuronales (Lundberg & Lee, 2017).

2.2.1. Uso de SHAP en la deserción estudiantil

En el ámbito educativo, especialmente en el estudio de la deserción estudiantil, las herramientas de interpretación de modelos como SHAP se han consolidado como recursos clave para fortalecer la comprensión de los factores que inciden en este fenómeno. SHAP permite interpretar de manera precisa los modelos de predicción, lo cual resulta esencial para diseñar intervenciones informadas por parte de las instituciones. Por ejemplo, facilita la identificación de variables altamente influyentes, como el desempeño académico, la edad o el estrato socioeconómico (Dake et al., 2024; Estrada & Rodríguez, 2022).

Una de las ventajas más destacadas de SHAP es su capacidad para proporcionar interpretaciones tanto locales como globales de los modelos predictivos, ofreciendo así una visión clara de las razones por las cuales un estudiante específico podría estar en riesgo de abandono (Malik & Tüfekci, 2023). Esta capacidad de explicación individualizada fortalece la confianza en los sistemas de inteligencia artificial aplicados en contextos educativos y contribuye a una toma de decisiones más transparente. Diversas investigaciones recientes han demostrado que la incorporación de SHAP en modelos de predicción de deserción mejora sustancialmente la transparencia institucional y la capacidad de respuesta ante el abandono académico (Dake et al., 2024).

Ahora bien, si bien los enfoques supervisados como SHAP permiten comprender mejor los factores de riesgo cuando se dispone de datos etiquetados, en muchos contextos educativos esta información no siempre está disponible. Ante esta limitación, los métodos no supervisados, como las técnicas de *clustering*, han cobrado relevancia al permitir descubrir patrones ocultos en los datos. Estas técnicas agrupan a los estudiantes en perfiles homogéneos que pueden interpretarse posteriormente en función de variables relevantes para el análisis del abandono (Melchor et al., 2025).

Entre los algoritmos no supervisados más adecuados para tratar datos categóricos se encuentra el algoritmo K-Modes. Este método ha demostrado ser particularmente eficiente en el contexto educativo, donde predominan variables cualitativas como género, tipo de institución o modalidad de ingreso. A diferencia de K-Means, que opera con medias y distancias euclidianas, K-Modes emplea una medida de disimilitud basada en la coincidencia de categorías y define los centroides a partir de la moda de los atributos (Dake et al., 2023). Esta característica lo convierte en una técnica especialmente apropiada para trabajar directamente con datos categóricos sin necesidad de codificaciones numéricas.

2.3. El algoritmo K-Modes

El algoritmo K-Modes, propuesto por Huang (1998), es una extensión del algoritmo K-Means diseñada para manejar datos categóricos. En lugar de utilizar medias aritméticas y distancias euclidianas, K-Modes reemplaza estas métricas por modas categóricas y una medida de disimilitud basada en coincidencias. El procedimiento consiste en seleccionar aleatoriamente K centroides iniciales, asignar cada observación al clúster con el centroide más similar, actualizar los centroides mediante la moda de cada atributo y repetir este proceso iterativamente hasta alcanzar la convergencia.

Matemáticamente, el algoritmo busca minimizar la siguiente función de costo:

$$\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(x_{ij}, z_{cj}), \tag{9}$$

donde la función de disimilitud δ se define como:

$$\delta(x_{ij}, z_{cj}) = \begin{cases} 0, & \text{si } x_{ij} = z_{cj} \\ 1, & \text{si } x_{ij} \neq z_{cj} \end{cases}$$

aquí, x_{ij} representa el valor del atributo j de la observación i, y z_{cj} es la moda correspondiente del atributo j en el clúster c ((GeeksforGeeks, 2025)).

2.3.1. Aplicación en la deserción estudiantil

Este algoritmo ha sido utilizado con éxito para identificar subgrupos de estudiantes con características similares, permitiendo clasificar perfiles de riesgo de deserción. Por ejemplo, estudios recientes han empleado K-Modes para analizar bases de datos académicas categóricas, logrando segmentar estudiantes según historial académico, tipo de matrícula, retrasos en el plan de estudios y modalidad de ingreso, lo que facilita el diseño de estrategias de intervención focalizadas ((Darwis et al., 2025; ResearchGate, 2025)).

2.3.2. Selección del número óptimo de clústeres

Una de las decisiones más críticas al aplicar K-Modes es la determinación del número de clústeres K. Esta elección incide directamente en la validez y utilidad práctica de los grupos generados. Para ello, se recomienda emplear una combinación de métodos cuantitativos y juicio experto. El método del codo, basado en la gráfica del costo intra-clúster frente a distintos valores de K, permite identificar el punto donde la mejora marginal disminuye significativamente. Asimismo, pueden aplicarse índices de validación adaptados a datos categóricos, como la silueta ajustada o medidas de entropía ((T. M. Nguyen et al., 2023)).

En estudios aplicados al abandono universitario, también se ha sugerido partir de valores teóricos guiados por conocimiento institucional, como dividir a los

estudiantes en tres grupos: alto, medio y bajo riesgo de deserción.

La aplicación del algoritmo K-Modes ha sido documentada en diversos estudios recientes. Por ejemplo, Melchor et al. (Melchor et al., 2025) evaluaron su eficacia dentro de una propuesta metodológica para detectar estudiantes en riesgo de deserción, utilizando un conjunto de datos mixtos (académicos y sociodemográficos). En su estudio, se aplicaron diferentes técnicas de ingeniería de características (como binning y *encoding*) para transformar los datos, y se validaron los resultados mediante métricas tanto no supervisadas como supervisadas, comparando los grupos formados con etiquetas reales.

Asimismo, Dake et al. (Dake et al., 2023) utilizaron K-Modes para identificar estilos de aprendizaje en estudiantes universitarios. Su investigación permitió formar clústeres interpretables (por ejemplo, visuales, sociales, verbales), que posteriormente se usaron como base para construir modelos de predicción que podrían ser utilizados en procesos de intervención educativa personalizada.

Estas investigaciones destacan tres ventajas clave del uso de K-Modes en el contexto educativo:

- Permite trabajar directamente con datos categóricos sin necesidad de codificaciones que introduzcan sesgos.
- Favorece la identificación de perfiles estudiantiles a partir de patrones latentes, útiles para prevenir el abandono.
- Su aplicación práctica es viable en instituciones con recursos limitados o carencia de bases de datos etiquetadas.

En síntesis, el algoritmo K-Modes representa una alternativa robusta para explorar la estructura interna de los datos categóricos en educación superior, facilitando la toma de decisiones tempranas ante fenómenos críticos como la deserción estudiantil.

2.4. Análisis de Correspondencias Múltiples (ACM)

El Análisis de Correspondencias Múltiples (ACM) se ha consolidado como una herramienta estadística multivariante clave en la exploración de relaciones entre variables categóricas en el ámbito educativo. Su utilidad ha sido ampliamente demostrada en investigaciones orientadas a comprender fenómenos complejos como la deserción estudiantil y la caracterización de perfiles académicos.

En particular, el ACM ha permitido identificar asociaciones latentes entre factores individuales, institucionales y socioeconómicos que inciden en la permanencia o abandono de los estudiantes en la educación superior. Por ejemplo, el estudio de Choque-Soto, Sosa-Jauregui e Ibarra (Choque-Soto et al., 2025) aplicó técnicas de minería de datos combinadas con análisis multivariado para perfilar estudiantes desertores en carreras de informática, revelando patrones específicos como género, estado civil y procedencia regional.

Otro ejemplo significativo es el trabajo de Barros, Silva y Guedes (Barros et al., 2019), donde se empleó el ACM como técnica central para determinar los perfiles de estudiantes desertores. En este estudio, se identificaron características relevantes como la experiencia previa de reprobación, el tipo de institución de procedencia (pública o privada) y el entorno familiar (presencia o no de los padres).

Asimismo, el trabajo de Álvarez-García, Arenas-Parra e Ibar-Alonso (Álvarez-García et al., 2024) en el contexto de PISA 2022 utilizó ACM para construir perfiles estudiantiles basados en variables sociodemográficas, rendimiento académico y percepciones educativas.

Finalmente, el estudio de Olugbara, Letseka y Olugbara (Olugbara et al., 2021) sobre la aceptación de MOOCs empleó el ACM para identificar patrones de asociación entre las creencias, motivaciones y decisiones de los estudiantes respecto al aprendizaje en línea. Esta técnica reveló agrupaciones latentes entre categorías como tipo de curso, modalidad de evaluación y expectativas académicas, lo que permitió generar recomendaciones pedagógicas personalizadas.

Estos hallazgos evidencian que el Análisis de Correspondencias Múltiples no solo permite representar gráficamente estructuras de datos categóricos, sino que también constituye una herramienta poderosa para la toma de decisiones educativas basadas en evidencia empírica.

3. Resultados y Análisis

Para el desarrollo del presente estudio, se utilizó una base de datos compuesta por información de 6 054 estudiantes de la Universidad El Bosque, ubicada en Bogotá, Colombia. El análisis se centró principalmente en los programas adscritos a la Facultad de Ingeniería, considerando tanto aspectos académicos como sociodemográficos.

3.1. Preparación y exploración inicial

La base de datos fue cargada y preprocesada utilizando herramientas del ecosistema Python, específicamente con el uso de bibliotecas como **NumPy** (Harris et al., 2020), **Pandas** (McKinney, 2010), **Matplotlib** (Hunter, 2007), **Seaborn** (Waskom, 2021), **Scikit-learn** (Pedregosa et al., 2011) y **Missingpy** (Boström, 2020), que permitieron el manejo eficiente de estructuras de datos, visualización estadística y técnicas de imputación y aprendizaje automático. Se realizó una exploración inicial para conocer la estructura de los datos, la tipología de las variables y la

existencia de valores faltantes. Las variables incluidas fueron: estado civil, género, localidad, período de ingreso, jornada, tipo de estudiante, institución de egreso, edad, semestre, estrato socioeconómico, fecha de nacimiento, ciudad de nacimiento, programa académico, año de graduación y la variable objetivo *Graduado*.

3.2. Análisis de valores faltantes

Durante la fase inicial de exploración de datos, se identificó la existencia de valores faltantes en algunas variables del conjunto de datos. En particular, las variables categóricas `localidad` e `institucionEgreso` presentaban una cantidad considerable de datos ausentes, con un 35,58 % y 15,64 % de valores nulos, respectivamente. Esta situación requería un tratamiento adecuado para evitar sesgos o distorsiones en los análisis posteriores y en el entrenamiento de modelos predictivos.

Para abordar este problema, se implementó un proceso de imputación específico para variables categóricas, utilizando una combinación de *codificación ordinal* y el algoritmo de aprendizaje supervisado *K-Nearest Neighbors* (KNN). El procedimiento se desarrolló en los siguientes pasos:

1. **Identificación y segmentación de variables:** Se identificaron las columnas categóricas con valores faltantes (`localidad` e `institucionEgreso`) y se seleccionaron las variables numéricas del conjunto de datos para emplearlas como predictores en el modelo KNN.

2. **Codificación ordinal con reconocimiento de categorías desconocidas:** Para convertir las variables categóricas a formato numérico compatible con KNN, se empleó un `OrdinalEncoder` configurado para manejar valores desconocidos mediante la asignación del código `-1`. Esta técnica facilitó la posterior identificación de los registros con datos faltantes.

3. **Entrenamiento del modelo KNN:** Se construyó un modelo de clasificación KNN utilizando los registros que contenían datos completos en las variables categóricas. El modelo fue entrenado con los valores codificados como etiquetas y las variables numéricas como atributos de entrada.

4. **Predicción e imputación de valores faltantes:** Una vez entrenado el modelo, se procedió a predecir los valores correspondientes a las observaciones con código `-1` (es decir, registros con datos faltantes). Las predicciones fueron luego decodificadas a sus respectivas etiquetas categóricas originales.

5. **Verificación posterior a la imputación:** Finalmente, se realizó una revisión para verificar la efectividad del proceso de imputación. Se confirmó que el porcentaje de valores faltantes en ambas variables imputadas se redujo a 0 %, garantizando un conjunto de datos completamente limpio y apto para los análisis posteriores.

Este enfoque permitió conservar la información útil de los registros incompletos sin necesidad de eliminarlos, lo cual es particularmente valioso en contextos donde el tamaño de la muestra puede ser limitado. Además, la estrategia implementada garantiza coherencia con el tipo de variable y aprovecha la similitud entre observaciones, manteniendo la integridad estadística del conjunto de datos.

3.3. Transformación de variables

Posteriormente, se transformaron las variables de tipo fecha (por ejemplo, `periodoIngreso`) a su representación ordinal con el fin de facilitar su tratamiento dentro de los modelos de *Machine Learning*. Asimismo, se eliminaron aquellas variables que no aportaban directamente al análisis predictivo, tales como `codigoEstudiante`, `fechaNacimiento`, `anhoGraduacion` y `CodPrograma`.

3.4. Estadísticas descriptivas

Una vez consolidada la base de datos con 6 054 registros correspondientes a estudiantes de la Universidad El Bosque, se realizó una exploración descriptiva de las principales variables sociodemográficas y académicas. Esta etapa permitió caracterizar de manera general a la población de estudio y sentar las bases para los análisis posteriores.

En cuanto a la edad, se observó que los estudiantes tenían entre 15 y 52 años, con una edad promedio de 22,36 años. La mayoría de los estudiantes (75 %) tenía 24 años o menos, lo que confirma una población predominantemente joven, aunque también se identificaron casos de adultos que cursan programas universitarios, lo cual puede responder a procesos de reintegración académica o formación profesional tardía.

Respecto al semestre más avanzado alcanzado (`semestreMax`), se encontró un promedio de 6,34 semestres, con un rango de 1 a 11. Esto refleja una distribución equilibrada entre estudiantes que apenas inician su formación y otros que están en etapas avanzadas del programa. El 50 % de los estudiantes se encontraba entre los semestres 2 y 10.

El estrato socioeconómico osciló entre 1 (bajo) y 6 (alto), con una moda de 3. La mayoría se ubicaba en los estratos medios, lo cual es coherente con la composición socioeconómica general de estudiantes en instituciones privadas como la Universidad El Bosque.

La variable `periodoIngreso`, que indica la fecha en la que el estudiante ingresó a su programa, abarca desde 1998 hasta 2025. La mediana se ubicó el 1 de

enero de 2019, lo que indica que una gran parte de los estudiantes ingresaron en los últimos años.

En lo que respecta a los programas académicos, se identificó una participación diversa de estudiantes pertenecientes a distintas áreas de la Facultad de Ingeniería, entre ellas: Ingeniería de Sistemas, Ingeniería Ambiental, Ingeniería Electrónica, Ingeniería Industrial y Bioingeniería.

La variable de interés del estudio, *Graduado*, mostró que el 35,28 % de los estudiantes se encontraba registrado como graduado, mientras que el 64,72 % aún no había culminado sus estudios.

Finalmente, en la revisión inicial de los datos se volvieron a identificar valores faltantes en `localidad` (35,58 %) e `institucionEgreso` (15,64 %). Para resolver esta situación, se aplicó un proceso de imputación utilizando el algoritmo KNN, lo que permitió estimar los valores ausentes con base en similitudes entre observaciones y obtener un conjunto de datos sin valores nulos, adecuado para los análisis posteriores.

Para predecir la probabilidad de graduación de estudiantes universitarios, se desarrolló un *pipeline* de modelado basado en técnicas de *Machine Learning*. Se utilizaron cinco modelos: Árbol de Decisión, Random Forest, XGBoost, LightGBM y Máquinas de Vectores de Soporte (SVM). El conjunto de datos fue dividido en entrenamiento y prueba (80/20) con estratificación para conservar la proporción de clases. Las variables predictoras fueron sometidas a preprocesamiento: las variables numéricas fueron estandarizadas con `StandardScaler` y las categóricas codificadas mediante `OneHotEncoder`.

3.5. Modelos de Machine Learning

3.5.1. Entrenamiento y ajuste de modelos

Cada modelo fue entrenado mediante validación cruzada estratificada usando `GridSearchCV`, e incorporó técnicas de balanceo de clases (`class_weight='balanced'` o `scale_pos_weight`) para corregir el desequilibrio entre estudiantes graduados y no graduados.

3.5.2. Evaluación del desempeño de los modelos

Los resultados muestran que los modelos de *boosting* (XGBoost y LightGBM) obtuvieron el mejor desempeño general. En particular, LightGBM alcanzó el mejor *F1-score* macro (0.895) y la mayor precisión (*accuracy* = 0.899), siendo seleccionado para la interpretación de resultados con SHAP. Como se observa en la figura 6

El modelo de **Árbol de Decisión** exhibe la curva ROC más alejada de la esquina superior izquierda, reflejando un desempeño moderado con AUC = 0,86. En contraste, los modelos de ensamble y de gradient boosting (Random Forest, XGBoost y LightGBM) alcanzan AUC superiores a 0,94, lo que indica una

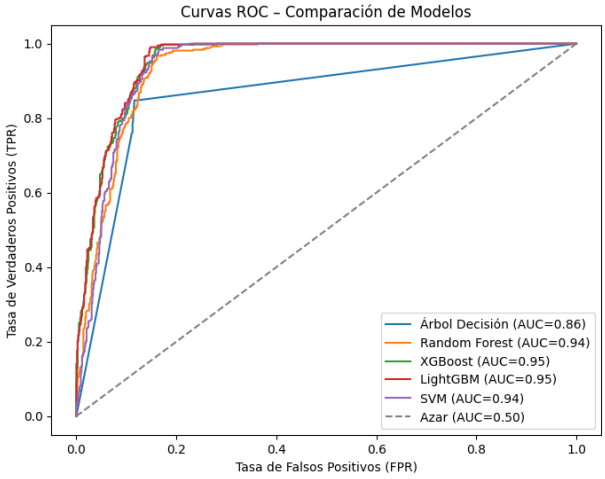

Figura 6: Curva ROC comparación de Modelos de ML.

excelente capacidad para discriminar entre clases positivas y negativas.

XGBoost y LightGBM obtienen la mejor discriminación (AUC = 0,95), con curvas que se acercan al punto ideal (FPR $\approx$ 0, TPR $\approx$ 1).

Random Forest y SVM muestran un rendimiento muy similar (AUC = 0,94), apenas inferior al de los métodos de boosting, y con ligeras diferencias en la pendiente de la curva en la zona de falsos positivos bajos.

El *modelo de azar* (AUC = 0,50) actúa como una línea base o referencia nula. Su curva diagonal indica una clasificación completamente aleatoria, lo que permite confirmar que los modelos evaluados superan significativamente el desempeño esperado por azar, aportando valor predictivo estadísticamente relevante.

Tabla 1: Desempeño de los modelos y su balance óptimo

Modelo	Balance óptimo	Accuracy media CV
Árbol de Decisión	`class_weight='balanced'`	0.861
Random Forest	`class_weight=None`	0.881
XGBoost	`scale_pos_weight=1.83`	0.897
LightGBM	`class_weight='balanced'`	0.897
SVM	`class_weight=None`	0.891

3.5.3. Detalles por modelo

Nota: En las siguientes matrices de confusión, se adopta la convención de que la clase "0" representa **No graduados** y la clase "1" la de **Graduados**, conforme a la codificación establecida en el conjunto de datos analizado.

LightGBM: El modelo LightGBM presentó un excelente desempeño (ver tabla 2)en la tarea de clasificación de estudiantes graduados y no graduados. En

particular, alcanzó un *F1 macro* de 0,895, un área bajo la curva (*AUC*) de 0,95, y una precisión del 99,4 % para la clase "no graduado", con un *recall* del 99,1 % para la clase "graduado". Esto significa que logró identificar correctamente a casi todos los estudiantes que efectivamente se graduaron, minimizando la cantidad de falsos negativos.

La matriz de confusión presentada en la Figura 7 evidencia la capacidad del modelo para diferenciar ambas clases, reflejando una mayor proporción de aciertos para la clase negativa (no graduado), lo cual es clave en contextos de intervención educativa. Por otro lado, los valores detallados en los reportes de clasificación (Tablas 2 y 3) permiten observar la efectividad del modelo en métricas individuales como precisión, *recall* y *F1-score*, tanto a nivel de clases específicas como en sus promedios macro y ponderado.

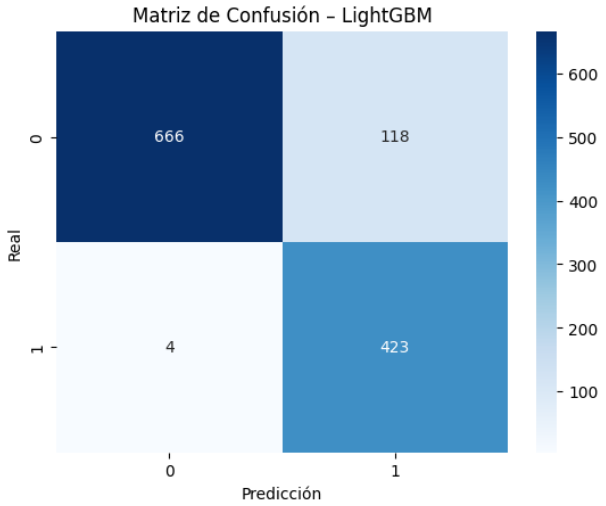

Figura 7: Matriz de confusión de LightGBM.

Tabla 2: Reporte de clasificación — LightGBM

Clase	Precisión	Recall	F1-Score
0	0.995	0.849	0.916
1	0.782	0.991	0.875
Accuracy		**0.899**	

XGBoost: El modelo XGBoost alcanzó un desempeño muy similar al de LightGBM, con un *F1 macro* de 0,885 y un área bajo la curva (*AUC*) de 0,95. Destacó particularmente por su capacidad para minimizar los falsos negativos, registrando únicamente 11 casos en los que estudiantes que efectivamente se graduaron fueron clasificados como no graduados. Esta característica es clave en contextos educativos, ya que evita excluir del acompañamiento a estudiantes con alta probabilidad de éxito académico.

La matriz de confusión presentada en la Figura 8 muestra la distribución de aciertos y errores del modelo, donde se observa una alta sensibilidad hacia la clase positiva (graduado). Además, el desempeño detallado en métricas individuales como precisión, *recall* y *F1-score* se presenta en la Tabla 4, diferenciando claramente entre estudiantes no graduados (clase 0) y graduados (clase 1), así como sus respectivos promedios macro y ponderados.

Tabla 3: Reporte de Clasificación – Modelo LightGBM

Clase	Precisión	Recall	F1-score	Soporte
No graduado (0)	0.994	0.849	0.916	784
Graduado (1)	0.782	0.991	0.874	427
Promedio macro	0.888	0.920	0.895	–
Promedio ponderado	0.901	0.899	0.901	1211
Accuracy global		**0.899**		

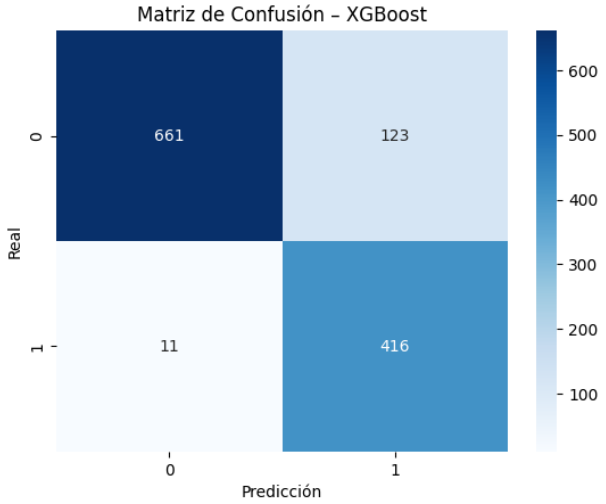

Figura 8: Matriz de confusión de XGBoosting.

Tabla 4: Reporte de clasificación del modelo XGBoost

Clase	Precisión	Recall	F1-score	Soporte
0	0.984	0.843	0.908	784
1	0.772	0.974	0.861	427
Accuracy		**0.889**		1211
Promedio macro	0.878	0.909	0.885	1211
Promedio ponderado	0.909	0.889	0.892	1211

SVM: El modelo de Máquinas de Vectores de Soporte (SVM) también obtuvo buenos resultados en la clasificación de estudiantes. En particular, alcanzó un *F1 macro* de 0,881 y una precisión global del 88,5 %. Mostró un comportamiento robusto al no requerir técnicas de balanceo para alcanzar estos resultados, aunque registró un mayor número de falsos positivos en comparación con otros modelos, lo cual implica una mayor tendencia a clasificar como graduados a estudiantes que finalmente no culminan su programa.

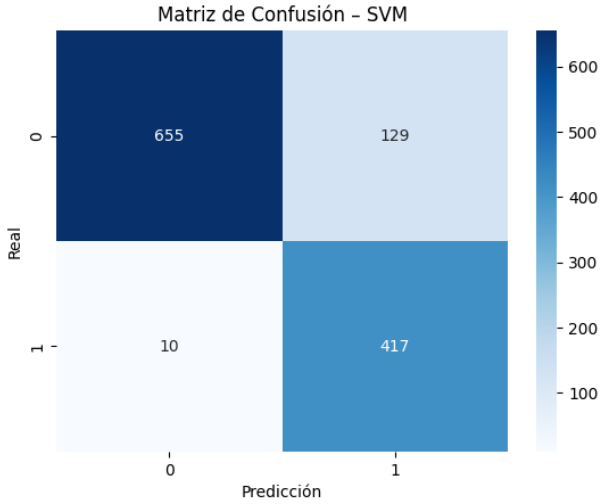

Figura 9: Matriz de confusión de SVM.

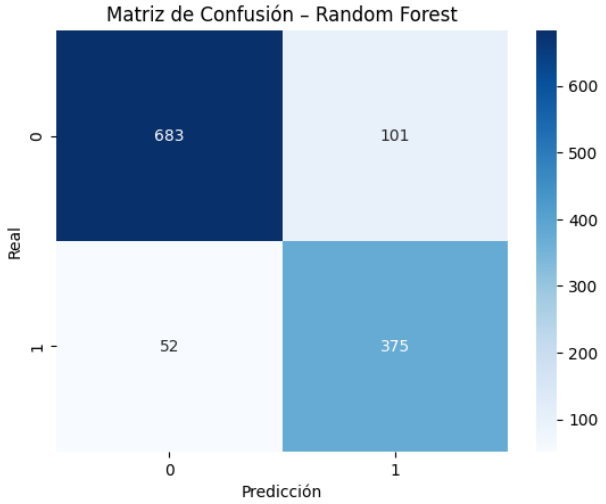

Figura 10: Matriz de confusión de Random Forest.

La Figura 9 presenta la matriz de confusión correspondiente, donde puede observarse el rendimiento del modelo al clasificar las clases 0 (no graduado) y 1 (graduado). Por su parte, la Tabla 5 muestra en detalle las métricas obtenidas, como precisión, *recall* y *F1-score*, así como sus promedios macro y ponderado, con base en un total de 1 211 observaciones del conjunto de prueba.

Tabla 5: Reporte de clasificación del modelo SVM

Clase	Precisión	Recall	F1-Score	Soporte
0	0.985	0.835	0.904	784
1	0.764	0.977	0.857	427
Macro avg	0.874	0.906	0.881	1211
Weighted avg	0.907	0.885	0.888	1211

Random Forest: El modelo Random Forest mostró un buen desempeño general, alcanzando un *F1 macro* de 0,865. Sin embargo, se observó una menor precisión al clasificar correctamente a los estudiantes graduados (clase 1), con un valor de 78,8 %. Este resultado indica que, aunque el modelo es eficaz en términos globales, tiende a generar una mayor cantidad de falsos positivos al predecir la clase "graduado", lo cual puede implicar la asignación de recursos de acompañamiento a estudiantes con bajo riesgo real de deserción.

La Figura 10 presenta la matriz de confusión correspondiente, donde puede visualizarse el rendimiento del modelo para las clases 0 (no graduado) y 1 (graduado).

Árbol de Decisión: El modelo de Árbol de Decisión, si bien es más simple comparado con otros algoritmos, logró mejorar su rendimiento al aplicar un ajuste de pesos con la opción `class_weight=balanced`. Bajo esta configuración, se alcanzó un *F1 macro* de 0,861,

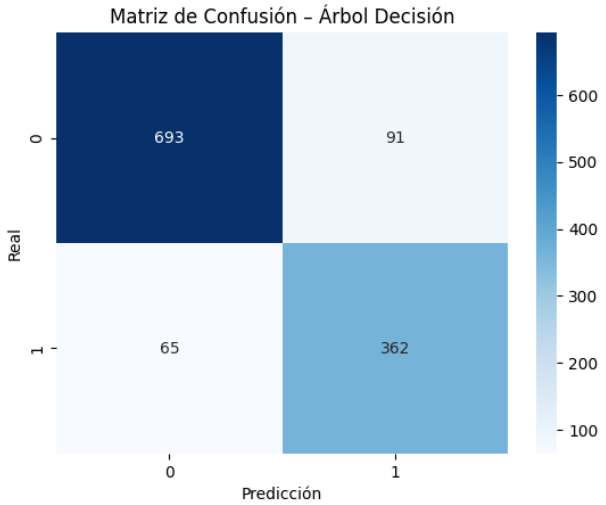

Figura 11: Matriz de confusión de Árbol de Decisión.

lo cual indica una mejora significativa en la equidad del modelo al tratar las clases desbalanceadas.

La Figura 11 muestra la matriz de confusión del modelo, evidenciando su capacidad para clasificar correctamente a estudiantes no graduados (clase 0) y graduados (clase 1), aunque con un rendimiento ligeramente inferior en comparación con modelos más complejos.

3.5.4. Análisis Comparativo entre Clases

El análisis de desempeño por clase reveló que:

- Todos los modelos mostraron un mejor desempeño al clasificar estudiantes no graduados, lo que resulta especialmente valioso para la identificación de perfiles en riesgo de deserción.
- Modelos como XGBoost y LightGBM fueron los más eficaces al clasificar estudiantes graduados (alta sensibilidad), lo cual es crucial en escenarios donde se busca reducir falsos negativos.

- En general, el error más común fue el falso positivo, es decir, clasificar erróneamente a estudiantes como graduados cuando no lo eran.

3.6. Análisis de errores por tipo de clasificación

La Tabla 6 presenta un resumen de los errores de clasificación cometidos por cada modelo de aprendizaje automático, desglosados en verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Este análisis permite evaluar con mayor precisión la capacidad de cada modelo para identificar correctamente tanto a los estudiantes que se graduaron como a aquellos que no lo hicieron.

Verdaderos Positivos (TP) – Graduados bien clasificados. Los modelos con mayor capacidad para identificar correctamente a los estudiantes que efectivamente se graduaron fueron *LightGBM* (423 TP), *SVM* (417 TP) y *XGBoost* (416 TP). Este comportamiento es consistente con los altos valores de *recall* previamente observados para la clase "graduado". La alta tasa de verdaderos positivos sugiere una mayor eficacia en la detección de estudiantes con alta probabilidad de éxito académico.

Verdaderos Negativos (TN) – No graduados bien clasificados. En cuanto a la detección de estudiantes que no se graduaron, el modelo con mejor desempeño fue el *Árbol de Decisión*, que correctamente clasificó a 693 estudiantes como no graduados. Le siguen *Random Forest* (683 TN) y *LightGBM* (666 TN). Este tipo de aciertos es fundamental para alertar tempranamente sobre perfiles en riesgo de deserción.

Falsos Positivos (FP) – No graduados clasificados como graduados. Un alto número de falsos positivos puede implicar una subestimación del riesgo de deserción. En este sentido, el modelo *SVM* presentó el mayor número de FP (129), seguido por *XGBoost* (123) y *LightGBM* (118). Aunque estos modelos destacan en *recall* para graduados, tienden a sobreestimar los casos exitosos. Esto sugiere que, si bien son útiles para identificar a quienes sí se gradúan, deben ser usados con cautela cuando se busca prevenir la deserción.

Falsos Negativos (FN) – Graduados clasificados como no graduados. Este tipo de error es particularmente crítico en contextos educativos, ya que implica no reconocer a un estudiante que finalmente tuvo éxito. En este aspecto, *LightGBM* (4 FN) y *XGBoost* (11 FN) fueron los más eficaces, minimizando este riesgo. En contraste, *Árbol de Decisión* (65 FN) y *Random Forest* (52 FN) presentaron un desempeño inferior, lo que puede limitar su utilidad para acciones de acompañamiento oportuno.

Síntesis crítica. Los resultados evidencian que *LightGBM* logra el mejor equilibrio, combinando una alta tasa de verdaderos positivos con el menor número de falsos negativos. Este comportamiento lo convierte en una herramienta confiable para la identificación tanto de perfiles exitosos como de riesgo. Por su parte, *SVM* y *XGBoost* destacan por su sensibilidad, aunque a costa de una mayor tasa de falsos positivos. *Random Forest* y *Árbol de Decisión* muestran una tendencia más conservadora, con mejor precisión en la clase "no graduado", pero con menor capacidad para identificar estudiantes que sí culminaron su formación.

3.7. Interpretabilidad del Modelo con SHAP

Con el objetivo de interpretar el comportamiento del modelo de mejor desempeño (**LightGBM**), se aplicó la técnica de descomposición *SHAP* (SHapley Additive exPlanations), la cual permite cuantificar el impacto individual de cada variable en las predicciones del modelo. Esta herramienta resulta esencial para transformar modelos de caja negra en sistemas comprensibles, facilitando su aplicación en contextos educativos y de toma de decisiones institucionales.

Procedimiento de interpretación El procedimiento consistió en:

1. Extraer el modelo y el preprocesador del pipeline final de LightGBM.
2. Transformar los datos de entrenamiento con el preprocesador correspondiente.
3. Calcular los valores SHAP sobre el conjunto de datos transformado.
4. Generar visualizaciones globales (importancia media de las variables) y locales (explicaciones de predicciones individuales) para interpretar los efectos de las variables más influyentes.

3.8. Resultados globales: variables más influyentes

La visualización global generada por SHAP (ver figura 12) muestra las variables ordenadas por importancia. Las características con mayor impacto en la predicción de graduación fueron:

`num__semestreMax` Variable más influyente. A mayor cantidad de semestres aprobados, mayor probabilidad de graduación.

`num__estrato` Los estudiantes pertenecientes a estratos socioeconómicos altos tienden a presentar mejores tasas de graduación.

`cat__jornada_Diurna` Estudiar en jornada diurna favorece significativamente el éxito académico.

`cat__tipoEstudiante_Antiguo` Los estudiantes antiguos presentaron una mayor tasa de graduación que los nuevos.

Tabla 6: Comparativo de clasificaciones por modelo

Tipo de Clasificación	LightGBM	Random Forest	Árbol de Decisión	SVM	XGBoost
TP (Graduados bien clasificados)	423	375	362	417	416
TN (No graduados bien clasificados)	666	683	693	655	661
FP (No graduados clasificados como graduados)	118	101	91	129	123
FN (Graduados clasificados como no graduados)	4	52	65	10	11

`num__edadMax` Comportamiento no lineal: impacto negativo en estudiantes de mayor edad con bajo avance académico.

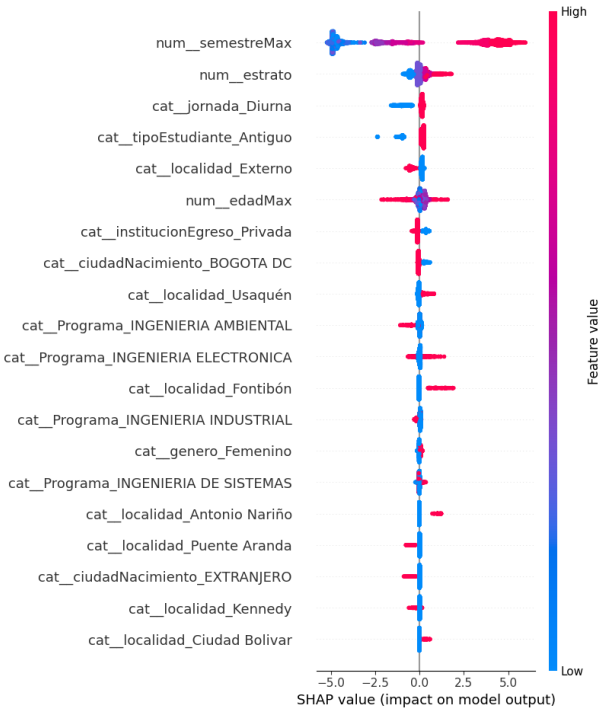

Figura 12: SHAP summary plot – Importancia global de variables.

3.9. Resultados locales: análisis de `edadMax` y su interacción

Se realizó una visualización específica de la variable `edadMax` (ver figura 13), con el fin de analizar su interacción con otras variables. Se evidenció lo siguiente:

- Para edades bajas, el impacto SHAP tiende a ser **positivo**, incrementando la probabilidad de graduación.
- A medida que la edad aumenta, el impacto en la probabilidad de graduación se torna **negativo**, especialmente si el estudiante ha cursado pocos semestres.
- Existe una interacción clara con `semestreMax`:
  - Estudiantes mayores con pocos semestres aprobados presentan alto riesgo de no graduarse.
  - Si tienen un avance académico significativo (semestres altos), la edad pierde relevancia negativa.

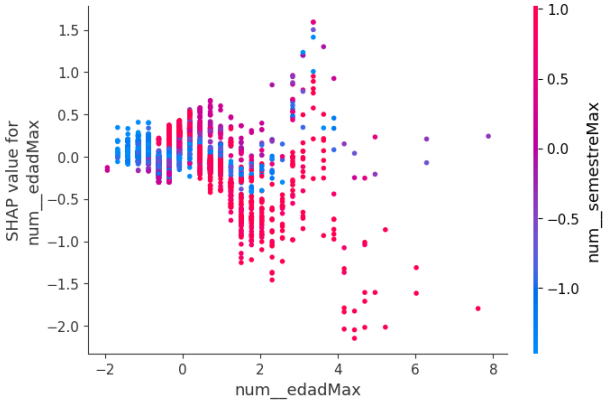

Figura 13: SHAP dependence plot – Relación entre edad y semestre aprobado.

El uso de SHAP permitió no solo identificar las variables más relevantes para la predicción de graduación, sino también entender cómo estas interactúan (Tabla 7). La visualización detallada aporta evidencia empírica que puede ser utilizada por las instituciones para diseñar estrategias de intervención dirigidas. Por ejemplo, estudiantes mayores con bajo avance curricular podrían beneficiarse de programas de acompañamiento intensivo.

La interpretación basada en SHAP transforma el modelo de LightGBM de una herramienta altamente precisa a una herramienta altamente explicativa, alineándose con los principios de equidad, comprensión y toma de decisiones informadas en el ámbito educativo.

3.9.1. Agrupamiento de estudiantes según perfil de riesgo de deserción

Con el fin de categorizar a los estudiantes según su perfil de riesgo de deserción, se implementó un análisis de agrupamiento utilizando el algoritmo `K-Modes`, debido a su idoneidad para manejar variables categóricas. El conjunto de datos estuvo compuesto por 6 054 estudiantes de programas de ingeniería, con variables sociodemográficas y académicas, incluyendo estado

Tabla 7: Variables más influyentes y su interpretación (SHAP global)

Variable	Interpretación
`num__semestreMax`	Es la más influyente. Semestres altos (rojo) impactan **positivamente** en la predicción de graduación.
`num__estrato`	Estratos más altos (rojo) suelen asociarse a mayor probabilidad de graduación.
`cat__jornada_Diurna`	Estudiar en jornada diurna tiene un **impacto positivo** (predice graduación).
`cat__tipoEstudiante_Antiguo`	Los estudiantes antiguos tienen mayor probabilidad de graduarse.
`cat__localidad_Externo`	Posiblemente estudiantes fuera de Bogotá muestran un patrón distinto.
`num__edadMax`	Edad más alta tiende a impactar **negativamente** en la probabilidad de graduación.
Programas académicos (Ingeniería Ambiental, Electrónica, Industrial o de Sistemas)	Aparecen con diferente peso, lo que sugiere variaciones en tasas de graduación según el programa.
Género femenino	Tiene un efecto moderado y parece variar según el caso.

civil, género, jornada, tipo de estudiante, institución de egreso, semestre cursado, estrato socioeconómico y edad categorizada mediante *binning* supervisado.

3.10. Selección del número óptimo de clústeres

La curva de costo generada al variar el número de clústeres entre 2 y 10 mostró una disminución pronunciada del costo hasta $k = 6$, punto a partir del cual la pendiente comenzó a aplanarse. Sin embargo, se observó que al llegar a $k = 7$ aún existía una reducción apreciable en el costo, lo que sugiere que este valor puede capturar una estructura adicional relevante dentro de los datos.

La selección del número óptimo de grupos (k) para el algoritmo K-Modes se realizó mediante el *método del codo*. Como se muestra en la Figura 14, al incrementar el número de clusters de 2 a 6 se evidencia una reducción significativa en la función de costo (inercia intra-cluster), reflejando una mejora considerable en la cohesión de los grupos. A partir de $k = 6$, si bien la curva comienza a estabilizarse, el valor de $k = 7$ presenta una disminución adicional no despreciable en el costo, lo que justifica su consideración como punto óptimo. Esta elección permite capturar una mayor heterogeneidad en los datos sin sacrificar la interpretabilidad del modelo, ofreciendo un equilibrio adecuado entre complejidad y calidad de segmentación.

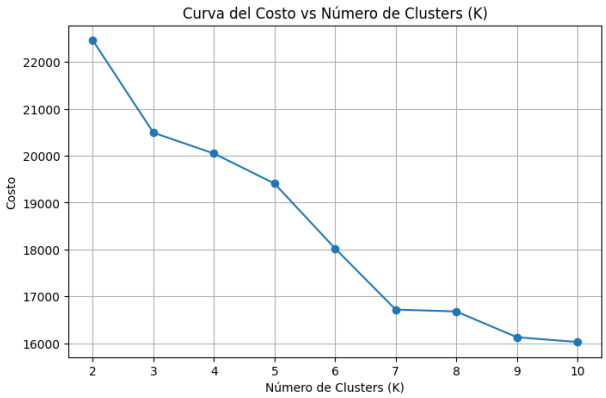

Figura 14: Curva de costo vs número de clústeres (K).

3.10.1. Distribución de estudiantes por clúster

El modelo con $k = 7$ generó los siguientes tamaños de clúster:

La Tabla 8 y la Figura 15 presenta la distribución de los estudiantes agrupados mediante el algoritmo K-Modes con $k = 7$. Se observa una marcada heterogeneidad en la frecuencia de los grupos, lo que sugiere

Tabla 8: Distribución de estudiantes por clúster (K-Modes)

Cluster	Frecuencia	Porcentaje (%)
0	1873	30.94
1	931	15.38
2	1308	21.61
3	368	6.08
4	342	5.65
5	694	11.46
6	538	8.89

la existencia de perfiles claramente diferenciados en la población analizada.

El **clúster 0** concentra la mayor proporción de individuos, con un 30.94 % del total, lo cual indica que este grupo representa un perfil dominante o recurrente en la muestra. Este clúster podría estar compuesto por estudiantes con características mayoritarias o patrones sociodemográficos comunes.

Le siguen en tamaño los clústeres **2** (21.61 %) y **1** (15.38 %), que también agrupan una porción significativa de la población. Estos grupos podrían representar segmentos intermedios con ciertas similitudes al clúster dominante, pero con variaciones importantes en variables clave que justifican su diferenciación.

En contraste, los clústeres **3** (6.08 %), **4** (5.65 %) y **6** (8.89 %) presentan una menor representación. Estos pueden estar capturando perfiles más específicos, posiblemente asociados a estudiantes en condiciones particulares o con trayectorias atípicas dentro del sistema educativo.

Finalmente, el **clúster 5** agrupa al 11.46 % de los estudiantes, ubicándose como un subgrupo de tamaño medio, potencialmente relevante en términos de políticas de intervención focalizada.

En conjunto, esta distribución refuerza la validez del modelo al evidenciar una segmentación que logra identificar tanto grupos mayoritarios como minoritarios, permitiendo un análisis más granular de los distintos perfiles de estudiantes en función de sus características compartidas.

Caracterización general de los clústeres

El análisis de segmentación mediante K-Modes permitió identificar siete grupos de estudiantes con perfiles diferenciados. A continuación, se presenta una caracterización general basada en las principales variables sociodemográficas y académicas:

- **Clúster 0:** Corresponde al grupo más numeroso. Está compuesto predominantemente por estudiantes **masculinos (83.66 %)**, pertenecientes al programa de **Ingeniería de Sistemas (51.36 %)**, de **estrato 3 (67.75 %)**, en modalidad **diurna (83.5 %)**, y con una **edad entre 21 y 26 años (72.61 %)**. La mayoría son **estudiantes antiguos (96.95 %)** y egresados de **instituciones privadas (88.68 %)**. Este clúster representa un perfil tradicional en términos académicos y sociodemográficos.

- **Clúster 1:** Se caracteriza por una mayor proporción de estudiantes provenientes de **instituciones públicas (97.95 %)**, y una mayor presencia de **estudiantes nuevos (21.58 %)**, lo que sugiere un perfil reciente de ingreso. La mayoría son **masculinos (71.21 %)**, con alta concentración en **Ingeniería de Sistemas (50.91 %)**, y edades jóvenes, principalmente entre **15 y 20 años (64.87 %)**.

- **Clúster 2:** Está altamente compuesto por **mujeres (76.22 %)** y concentra un porcentaje importante de estudiantes de **Ingeniería Industrial (29.28 %)** e **Ingeniería Ambiental (51.14 %)**. La mayoría tiene entre **21 y 26 años (79.28 %)**, es de **estratos 2 y 4**, y estudia en jornada **diurna (97.17 %)**. La gran mayoría son **estudiantes antiguos** y egresados de **colegios privados**.

- **Clúster 3:** Representa un grupo menor, pero distintivo, con alta proporción de estudiantes **masculinos (88.31 %)**, **en jornada nocturna (93.2 %)**, **mayores de 26 años (75 %)**, y predominantemente del programa de **Ingeniería Electrónica (51.35 %)**. Además, se concentra en estudiantes de **estrato 3**, lo que sugiere un perfil trabajador-estudiante.

- **Clúster 4:** Se destaca por una fuerte concentración en **Ingeniería Industrial (77.48 %)**, en jornada **diurna (96.2 %)**, con edades **entre 21 y 26 años (38.01 %)** y mayores de 26 años (47.36 %). La mayoría son **estudiantes antiguos**, provenientes de **instituciones privadas**, y pertenecen a **estratos 2 y 3**.

- **Clúster 5:** Incluye una proporción considerable de estudiantes **nocturnos (8.78 %)**, con predominio de **hombres (89.05 %)**, y una importante presencia de **estudiantes nuevos (30.11 %)**, lo cual sugiere un grupo en transición. Se encuentran repartidos en distintos programas, destacando **Ingeniería de Sistemas** y **Bioingeniería**.

- **Clúster 6:** Tiene un perfil marcadamente **femenino (87.36 %)**, con alta representación de estudiantes de **Bioingeniería (63.94 %)**, pertenecientes en su mayoría a **estratos 2 y 3**, en jornada **diurna (98.7 %)**, y en su mayoría egresados de **instituciones privadas (86.6 %)**. Las edades se concentran en el rango de **15 a 21 años**.

Esta caracterización evidencia que los clústeres identificados reflejan patrones consistentes de agrupamiento basados en combinaciones significativas de variables demográficas y académicas, lo cual valida la utilidad del enfoque no supervisado para segmentar perfiles de estudiantes y orientar estrategias diferenciadas de acompañamiento institucional.

La distribución observada en los clústeres revela un **alto grado de variabilidad**, lo cual implica que algunos grupos concentran una proporción considerablemente mayor de individuos en comparación con otros. En particular, el clúster 0 agrupa al 30,94 % de la población total, mientras que clústeres como el 4 y el 3 apenas representan el 5,65 % y 6,08 %, respectivamente.

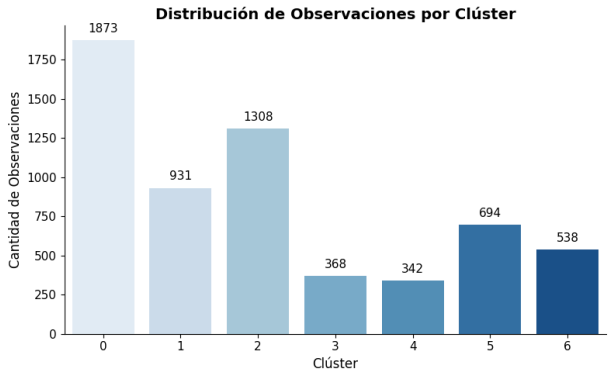

Figura 15: Distribución de observaciones por clúster.

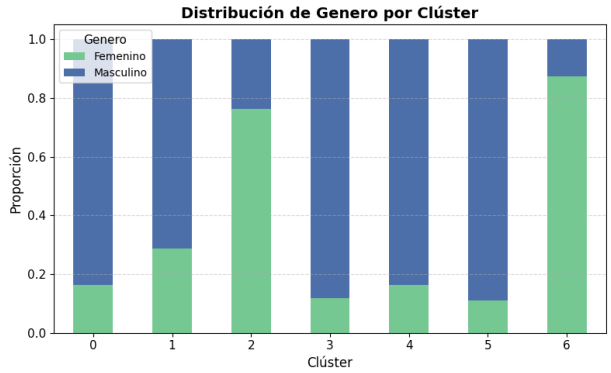

Figura 16: Distribución de género por clúster.

3.10.1. Análisis de Variables Sociodemográficas

La Figura 16 presenta la distribución proporcional del género en cada uno de los siete clústeres identificados mediante el algoritmo K-Modes. La gráfica evidencia una clara heterogeneidad en la composición por género entre los distintos grupos, lo cual refuerza la validez del modelo de segmentación en la captura de patrones sociodemográficos diferenciados.

En términos generales, se observa un predominio masculino en la mayoría de los clústeres, con excepción del **clúster 2** y el **clúster 6**, donde las mujeres representan el **76.2 %** y el **87.4 %** respectivamente. Estos grupos destacan como segmentos fuertemente feminizados, posiblemente asociados a programas como Bioingeniería o a cohortes más jóvenes.

En contraste, los clústeres **0, 3, 4** y **5** presentan una marcada mayoría de hombres, con proporciones que oscilan entre el 83 % y el 89 %, lo cual puede estar relacionado con carreras tradicionalmente masculinizadas como Ingeniería de Sistemas o Electrónica, y modalidades nocturnas de estudio.

El **clúster 1** muestra una distribución más equilibrada en comparación con los demás, aunque sigue predominando el género masculino (71.2 %).

Este patrón de distribución por género no solo refleja la composición real de la muestra, sino que también ofrece oportunidades para el diseño de intervenciones específicas por subgrupo, considerando dinámicas de género que puedan incidir en la permanencia y el rendimiento académico.

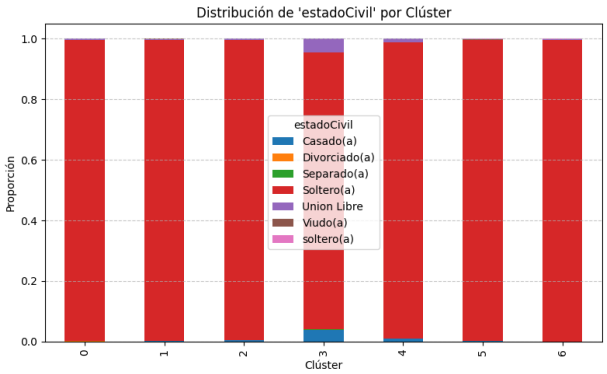

Figura 17: Distribución del estado civil por clúster.

La Figura 17 muestra la proporción de los distintos estados civiles dentro de cada uno de los siete clústeres identificados mediante K-Modes. La variable estado civil evidencia un patrón altamente homogéneo, donde el estado **soltero(a)** predomina de forma casi absoluta en todos los grupos.

Más específicamente, más del 90 % de los estudiantes en todos los clústeres se identifican como solteros(as), siendo esta proporción superior al 99 % en la mayoría de los casos. Esto sugiere que el estado civil no representa una variable de alta variabilidad dentro del conjunto de datos, aunque puede estar correlacionada con otras características como la edad o el tipo de jornada.

Excepcionalmente, el **clúster 3** presenta una mayor diversidad en los estados civiles, incluyendo pequeñas proporciones de estudiantes en *unión libre* (4.6 %), *casados(as)* (3.8 %) y *separados(as)* (0.27 %). Esta diversidad coincide con las características previamente descritas de este grupo, donde predominan estudiantes de jornada nocturna y de mayor edad (mayores de 26 años), lo cual podría reflejar situaciones familiares distintas a las de los estudiantes de jornada diurna y menor edad.

En suma, si bien el estado civil no permite una discriminación robusta entre la mayoría de los clústeres, sí aporta valor para la caracterización de grupos particulares con trayectorias educativas no convencionales, como es el caso del clúster 3.

La Figura 18 presenta la distribución proporcional de los estratos socioeconómicos de los estudiantes agrupados por clúster. Esta variable permite identificar diferencias significativas en el perfil económico de los grupos, aportando información relevante para el diseño de políticas de equidad y acompañamiento académico.

Se observa que el **estrato 3** es predominante en la mayoría de los clústeres, lo cual es coherente con la composición socioeconómica general de las cohortes universitarias de estratos medios. En particular, los clústeres **0**, **2**, **3** y **6** presentan proporciones que superan el 60 % en dicho estrato.

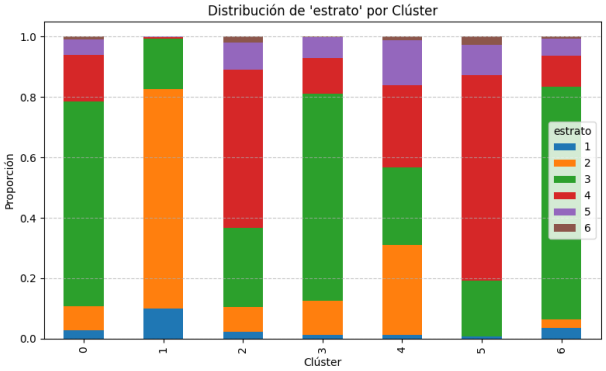

Figura 18: Distribución del estrato socioeconómico por clúster.

El **clúster 1** destaca por concentrar una proporción significativa de estudiantes del **estrato 2**, lo que sugiere una mayor representación de perfiles socioeconómicamente vulnerables. Este mismo grupo tiene también una presencia más marcada de estudiantes del **estrato 1**, lo cual puede requerir medidas específicas de acompañamiento académico y financiero.

Por su parte, los clústeres **4** y **5** exhiben una mayor heterogeneidad, incluyendo una combinación más equilibrada de estudiantes de los estratos 3, 4 y 5, y en menor proporción del estrato 6. Esta diversidad podría indicar condiciones económicas intermedias y mayor dispersión en términos de acceso a recursos.

En suma, la variable estrato permite enriquecer la comprensión de las condiciones sociales de los grupos formados por el algoritmo, lo cual resulta útil para orientar estrategias institucionales que promuevan la permanencia estudiantil con criterios de equidad y focalización.

3.10.2. Análisis Académico y de Trayectoria

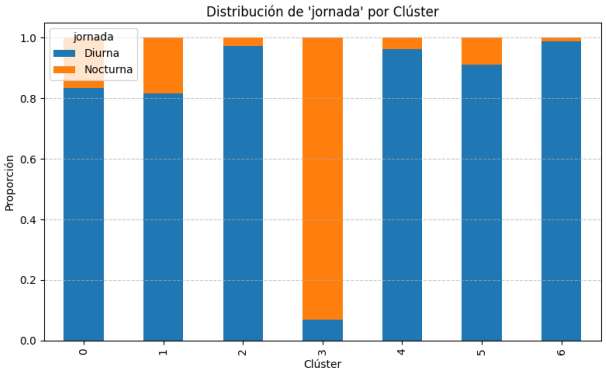

Figura 19: Distribución de jornada académica por clúster.

La Figura 19 muestra la distribución de los estudiantes según la jornada académica (diurna o nocturna) en cada uno de los clústeres identificados por el modelo K-Modes. Esta variable permite observar patrones relevantes en relación con la modalidad horaria de estudio, frecuentemente asociada a características socioeconómicas, laborales o de edad.

Se evidencia que la gran mayoría de los clústeres están compuestos principalmente por estudiantes de **jornada diurna**. En particular, los clústeres **0**, **1**, **2**, **4**, **5** y **6** tienen proporciones superiores al 80 % de estudiantes en esta modalidad, lo que refleja una fuerte prevalencia de la jornada tradicional entre los grupos identificados.

No obstante, el **clúster 3** destaca de forma significativa por su alta concentración de estudiantes de **jornada nocturna**, superando el 90 %. Este patrón es consistente con otras características previamente observadas en este grupo, como la mayor edad promedio y una mayor diversidad en el estado civil, lo cual sugiere que este clúster está conformado por estudiantes que probablemente combinan estudio con trabajo u otras responsabilidades.

Este contraste en la distribución horaria refuerza la utilidad del análisis de clústeres para identificar perfiles diferenciados y potencialmente vulnerables dentro de la población estudiantil, brindando insumos clave para el diseño de estrategias de acompañamiento específicas según las condiciones de cursado.

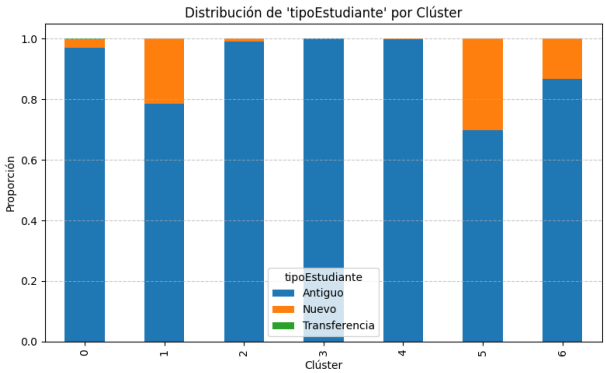

Figura 20: Distribución del tipo de estudiante por clúster.

La Figura 20 presenta la proporción de estudiantes antiguos, nuevos y por transferencia dentro de cada uno de los clústeres generados por el algoritmo K-Modes. Esta variable resulta particularmente útil para identificar grupos con diferentes niveles de integración y experiencia en el sistema universitario.

En términos generales, se observa un claro predominio de estudiantes **antiguos** en casi todos los clústeres. Los clústeres **2**, **3** y **4** están compuestos casi exclusivamente por este tipo de estudiantes, superando el 99 % de representación. Esto sugiere que estos clústeres agrupan perfiles con trayectorias más consolidadas dentro de la universidad.

Por su parte, los clústeres **1**, **5** y **6** presentan una mayor proporción de **estudiantes nuevos**, especialmente el clúster 5, donde cerca del 30 % de los integrantes pertenecen a esta categoría. Este hallazgo podría estar relacionado con etapas tempranas del

proceso de adaptación al entorno universitario y con factores de riesgo académico o de deserción.

El **clúster 0** también muestra una ligera presencia de estudiantes por **transferencia**, aunque en proporciones muy reducidas, lo cual indica que este subgrupo no representa un patrón dominante en el conjunto de datos.

En síntesis, esta variable permite diferenciar a los estudiantes según su nivel de permanencia institucional, siendo clave para el diseño de estrategias diferenciadas de apoyo académico que consideren el momento del ciclo de vida estudiantil en el que se encuentra cada grupo.

Figura 21: Distribución de rangos de edad por clúster.

La Figura 21 ilustra la distribución de los estudiantes agrupados por clúster según rangos etarios. Esta variable es de especial interés debido a su asociación con trayectorias académicas diferenciadas, condiciones laborales y responsabilidades extracurriculares.

Los clústeres **0**, **1**, **2**, **5** y **6** presentan una fuerte concentración de estudiantes en los rangos de **15 a 26 años**, lo cual es consistente con una población universitaria en formación inicial. Particularmente, los clústeres **0**, **2** y **6** tienen una representación notable del grupo entre **21 y 26 años**, lo que puede asociarse a estudiantes en niveles intermedios o avanzados de sus carreras.

Por contraste, el **clúster 4** presenta una proporción destacada de estudiantes en el rango de **26 a 52 años**, siendo el único grupo donde este rango supera al 40 % de su composición. Este hallazgo sugiere que este clúster puede representar un perfil de estudiantes no tradicionales, posiblemente con trayectorias académicas interrumpidas, reincorporaciones o estudios en paralelo con actividades laborales y familiares.

El **clúster 3**, por su parte, muestra una distribución más equilibrada, con una proporción relevante tanto de estudiantes jóvenes como adultos. Esto podría indicar un grupo de transición con dinámicas mixtas entre juventud y experiencia.

En síntesis, la edad emerge como una variable relevante para comprender la heterogeneidad de los perfiles estudiantiles identificados mediante el análisis de clústeres, permitiendo distinguir entre estudiantes tradicionales y no tradicionales, y aportando insumos para el diseño de políticas diferenciadas de acompañamiento académico.

3.11. Análisis de Correspondencias Múltiples (ACM) para la Caracterización de Estudiantes

La Figura 22 muestra la distribución bidimensional de los estudiantes proyectados sobre los dos primeros ejes factoriales obtenidos mediante Análisis de Correspondencias Múltiples (ACM). Esta técnica permite reducir la dimensionalidad de datos categóricos, identificando estructuras latentes y agrupamientos potenciales a partir de combinaciones de modalidades compartidas.

En la gráfica se aprecia una elevada concentración de puntos cerca del origen, lo que indica que la mayoría de los estudiantes no presentan perfiles extremos en relación con las variables incluidas en el análisis. La forma de "cono horizontal" sugiere una mayor variabilidad explicada sobre el eje **Dimensión 1**, mientras que el eje **Dimensión 2** muestra menor dispersión, indicando una menor capacidad de discriminación.

Asimismo, se observa una dispersión moderada hacia ambos lados del eje horizontal, lo que puede reflejar la existencia de subgrupos diferenciados con características categóricas contrastantes (por ejemplo, en género, jornada, programa académico, etc.). La presencia de algunos puntos alejados del centro en la parte superior de la gráfica puede corresponder a perfiles atípicos o menos frecuentes dentro de la muestra.

En conjunto, esta representación sugiere que el ACM ha permitido condensar patrones estructurales relevantes en los dos primeros ejes, facilitando el análisis visual de perfiles estudiantiles y constituyendo una base adecuada para técnicas de agrupamiento o interpretación de categorías activas.

Representación general de los individuos mediante ACM

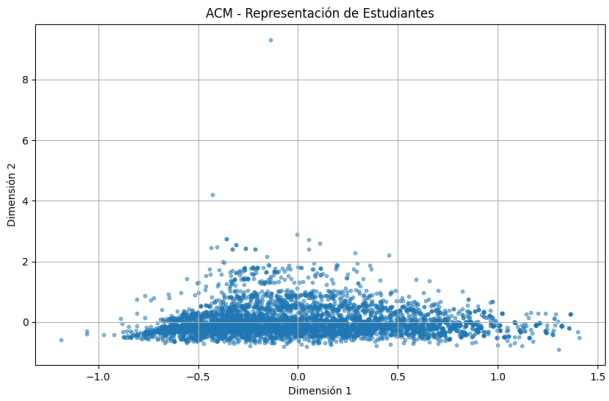

Figura 22: ACM – Representación de Estudiantes

La Figura 23 muestra la proyección de los estudiantes sobre los dos primeros ejes factoriales obtenidos mediante Análisis de Correspondencias Múltiples (ACM), con la incorporación de la segmentación generada por el algoritmo K-Modes. Esta visualización permite evaluar gráficamente el grado de separación, solapamiento y cohesión de los grupos formados a partir de variables categóricas.

En general, los clústeres presentan una distribución parcialmente solapada, aunque es posible identificar ciertos patrones de concentración espacial. El **clúster 2** (en verde) se ubica predominantemente en la parte izquierda del plano, con baja dispersión en ambas dimensiones, lo que indica un grupo homogéneo y bien definido. En contraste, el **clúster 1** (en naranja) se proyecta principalmente hacia la derecha, lo que sugiere un perfil diferenciado a lo largo de la Dimensión 1.

El **clúster 3** (en rojo) muestra una dispersión considerable en ambas dimensiones, lo que podría reflejar una mayor heterogeneidad interna o la presencia de estudiantes con combinaciones de características menos frecuentes. Por otro lado, el **clúster 6** (rosado) y el **clúster 0** (azul) se concentran cerca del origen, lo que sugiere que comparten perfiles moderados o comunes respecto a las variables consideradas.

Esta representación confirma que el modelo de *clustering* ha logrado identificar grupos con cierto grado de separación estructural en el espacio de las variables categóricas, especialmente a lo largo del primer eje factorial. No obstante, también evidencia áreas de solapamiento entre algunos clústeres, lo que puede ser indicativo de transiciones suaves entre perfiles estudiantiles o de una naturaleza continua en las combinaciones de categorías.

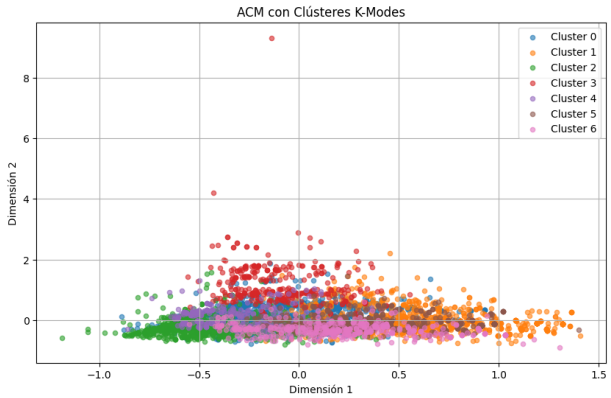

Figura 23: ACM con Clústeres K-Modes

3.11.1. ACM: representación conjunta de individuos y modalidades más influyentes

La Figura 24 presenta la proyección conjunta de los individuos y las 50 modalidades más representativas sobre los dos primeros ejes factoriales del Análisis de Correspondencias Múltiples (ACM). Este tipo de visualización permite interpretar el posicionamiento relativo de los estudiantes en función de las categorías que más contribuyen a la varianza explicada.

Los puntos azules representan a los individuos, quienes se concentran fuertemente alrededor del origen, lo cual indica que la mayoría de ellos poseen perfiles intermedios respecto a las variables categóricas consideradas. Esta densidad sugiere una estructura poblacional homogénea con escasos perfiles extremos.

Por otro lado, las modalidades (representadas por triángulos y etiquetas) se ubican en las periferias del plano factorial, indicando una mayor contribución a la variabilidad explicada. Entre las más influyentes se destacan categorías de la variable **ciudad de nacimiento**, que aparecen ampliamente dispersas a lo largo de ambos ejes, reflejando su capacidad discriminante dentro de la estructura de datos.

Se identifican también algunas categorías específicas de otras variables como *estado civil*, *estrato* y *tipo de estudiante*, aunque en menor proporción, lo cual sugiere que la **variable geográfica tiene un peso significativo** en la segmentación factorial.

En conjunto, esta representación permite visualizar no solo la distribución de los perfiles individuales, sino también qué modalidades particulares se asocian a posiciones extremas en el espacio factorial, facilitando así la interpretación cualitativa de los componentes latentes del ACM y la comprensión de las variables que estructuran la heterogeneidad de la población estudiada.

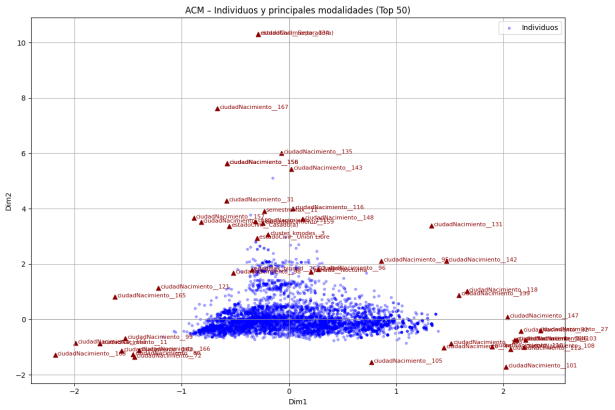

Figura 24: Individuos y modalidades – ACM

La Figura 25 muestra las 50 modalidades categóricas con mayor contribución a la varianza explicada en los dos primeros ejes del Análisis de Correspondencias Múltiples (ACM). El color y el tamaño de los puntos indican la magnitud de dicha contribución, medida mediante el *coseno al cuadrado* ($\cos^2$), donde valores más altos representan mayor relevancia en la definición del plano factorial.

Se destaca que la modalidad **estadoCivil_Separado(a)** presenta la mayor contribución, ubicada en una posición extrema del gráfico y resaltada en color rojo intenso. Este comportamiento

indica que dicha categoría tiene una capacidad altamente discriminante y que está asociada a perfiles estudiantiles específicos con características particulares.

Adicionalmente, varias modalidades de la variable **ciudadNacimiento** se distribuyen ampliamente sobre el plano, lo que sugiere una influencia considerable de la dimensión geográfica en la configuración factorial. Entre ellas, las ciudades identificadas con los códigos 167, 135, 158 y 143 muestran una posición periférica y colores más intensos, lo cual indica una contribución significativa en la diferenciación de los estudiantes.

Otras modalidades relevantes incluyen combinaciones asociadas al **estrato**, **estado civil** y **tipo de jornada**, lo cual evidencia que estas variables también aportan información estructural al análisis.

En conjunto, esta visualización permite identificar cuáles categorías específicas estructuran el espacio factorial y, por ende, cuáles deben ser consideradas prioritarias en la interpretación y segmentación de perfiles estudiantiles.

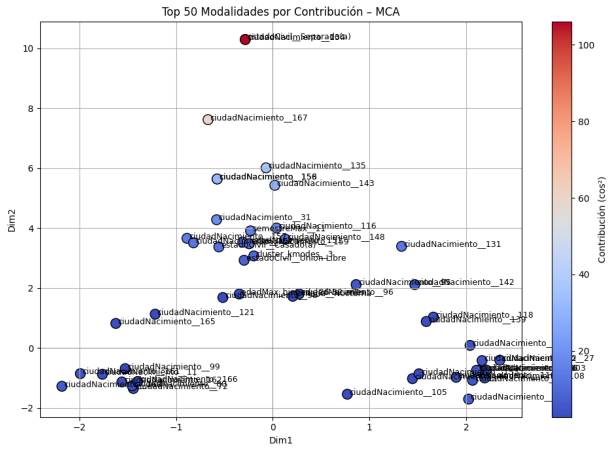

Figura 25: Contribución de variables – ACM (cos²)

La Figura 26 representa las 40 modalidades categóricas más representativas del Análisis de Correspondencias Múltiples (ACM), proyectadas sobre los dos primeros ejes factoriales. Esta visualización destaca aquellas categorías cuya contribución y calidad de representación en el espacio reducido es más alta, permitiendo identificar las dimensiones que estructuran la variabilidad de los datos.

Se observa que varias de las modalidades mejor representadas pertenecen a la variable **ciudad de nacimiento**, distribuidas en posiciones periféricas del plano factorial. Este patrón sugiere que la dimensión geográfica es una de las principales fuentes de variabilidad dentro de la población estudiantil analizada, y que ciertas ciudades tienen un peso diferenciador en la segmentación de perfiles.

Entre las modalidades más extremas se encuentran *estadoCivil_Separado(a)* y diversas ciudades como *ciudadNacimiento_167* y *ciudadNacimiento_158*, las cuales se ubican lejos del origen. Esta distancia indica una fuerte contribución a la configuración de los ejes y un perfil atípico en relación con las demás categorías.

Asimismo, categorías asociadas a variables como *estrato*, *tipo de jornada*, *edad* y *tipo de estudiante* también figuran en el plano, lo que confirma que estas variables influyen en la estructuración de los datos aunque con menor dispersión.

En conjunto, esta nube de categorías permite identificar los elementos que explican la diferenciación entre los individuos en el espacio factorial, siendo clave para la interpretación cualitativa de los ejes y la elaboración de perfiles estudiantiles con base en las combinaciones de modalidades más influyentes.

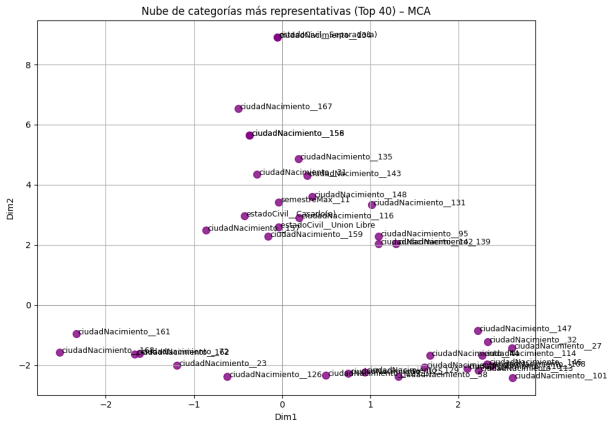

Figura 26: Nube de categorías – ACM

3.11.2. Clasificación de Perfiles mediante Árbol de Decisión y Pruebas de Independencia

La Figura 27 representa un árbol de decisión diseñado para clasificar a los estudiantes en función de los perfiles obtenidos mediante el algoritmo de agrupamiento K-Modes. Este modelo supervisado proporciona reglas interpretables basadas en variables sociodemográficas y académicas, permitiendo comprender las condiciones que conducen a cada grupo o clúster.

En el nodo raíz, la variable **tipo de institución de egreso** (pública o privada) emerge como el principal discriminador, lo que subraya el rol del origen educativo previo en la estructuración de perfiles académicos. A partir de este nodo, el árbol se divide en dos grandes ramas:

- La primera rama agrupa a estudiantes egresados de **instituciones privadas**, donde se identifican trayectorias académicas mixtas. Dentro de esta rama, la progresión se explica en función de variables como el **semestre máximo cursado**, la **edad** y el **género**. Por ejemplo, se diferencian perfiles de *estudiantes mujeres avanzadas con alto desempeño*, así como *ingresantes de alto estrato* y estudiantes con trayectorias iniciales.

- La segunda rama incluye a estudiantes de **instituciones públicas**, donde la edad (particularmente si se encuentran en el rango de 21–26 años) y el **estrato socioeconómico** adquieren protagonismo. Esta rama permite identificar perfiles como *estudiantes jóvenes de estrato bajo con bajo avance académico* o *adultos nocturnos de bajo estrato*.

El árbol incluye subdivisiones como se muestra en la Figura 27 que incorporan interacciones relevantes entre variables. Por ejemplo, dentro de los estudiantes jóvenes de instituciones privadas, el género y el estrato permiten identificar subperfiles asociados al rendimiento académico y la permanencia. En la rama pública, el *semestre máximo cursado* ayuda a diferenciar entre estudiantes con trayectorias truncadas o prolongadas.

En términos generales, el modelo evidencia una capacidad explicativa sólida, proporcionando descripciones precisas de grupos estudiantiles que integran tanto factores estructurales (estrato, institución, jornada) como factores de trayectoria (semestre, edad), lo que lo convierte en una herramienta valiosa para la toma de decisiones en políticas de retención y acompañamiento institucional.

3.11.3. Análisis multivariado y asociación con el estado de graduación

La Tabla 9 muestra un resumen multivariado de los perfiles dominantes por clúster obtenidos mediante el algoritmo K-Modes. Cada fila representa el valor modal de las variables categóricas dentro del clúster respectivo. Los resultados evidencian una fuerte homogeneidad interna y permiten etiquetar los grupos según características académicas y sociodemográficas recurrentes.

3.11.4. Análisis de asociación mediante prueba de Chi-cuadrado

Se realizaron pruebas de independencia Chi-cuadrado entre diversas variables categóricas y el estado de graduación. Los resultados se resumen a continuación:

- **Género**: $\chi^2 = 135{,}70$, $p < 0{,}001 \Rightarrow$ Hay una relación significativa entre género y graduación.
- **Programa académico**: $\chi^2 = 670{,}30$, $p < 0{,}001 \Rightarrow$ Existe una fuerte asociación entre el programa cursado y el estado de graduación.
- **Jornada académica**: $\chi^2 = 36{,}61$, $p < 0{,}001 \Rightarrow$ La jornada también influye significativamente en la probabilidad de graduación.
- **Estado civil**: $\chi^2 = 8{,}61$, $p = 0{,}197 \Rightarrow$ No se evidenció una asociación significativa entre estado civil y graduación.

Estos resultados permiten concluir que variables como el género, el programa académico y la jornada están asociadas de forma significativa con el resultado académico final. Por el contrario, variables como el estado civil no mostraron diferencias significativas.

3.11.5. Implicaciones

Los hallazgos refuerzan la pertinencia de los perfiles detectados por los clústeres. Perfiles como los estudiantes nocturnos adultos de bajo estrato, o las mujeres avanzadas con alto desempeño, tienen fundamentos estadísticos que justifican su diferenciación. Además, la integración del análisis de perfiles con pruebas de asociación fortalece la base empírica para diseñar intervenciones académicas focalizadas.

4. Conclusiones

1. **Desempeño de los modelos predictivos.** Los algoritmos de *gradient boosting*, en particular LightGBM, demostraron el mejor equilibrio entre precisión y sensibilidad, alcanzando un F1-macro de 0.895 y un AUC de 0.95 en la tarea de predecir la probabilidad de graduación de los estudiantes. Este nivel de exactitud, acompañado de un muy bajo número de falsos negativos, convierte a LightGBM en una herramienta confiable para sistemas de alerta temprana de deserción académica.

2. **Identificación de factores de riesgo clave.** El análisis de interpretabilidad global (SHAP) reveló que las variables más influyentes fueron el número de semestres aprobados (`semestreMax`), el estrato socioeconómico, la jornada de estudio (diurna vs. nocturna), el tipo de estudiante (nuevo vs. antiguo) y la edad. En particular, a mayor avance curricular y estrato socioeconómico, aumenta la probabilidad de graduación, mientras que la edad elevada sin progreso académico paralelo incrementa el riesgo de abandono.

3. La segmentación estudiantil mediante el algoritmo K-Modes permitió identificar siete clústeres claramente diferenciados, los cuales reflejan trayectorias, condiciones sociodemográficas y niveles de riesgo académico heterogéneos. Estos perfiles no solo evidencian patrones comunes de ingreso y progresión académica, sino también desigualdades estructurales asociadas al género, el estrato socioeconómico, el tipo de jornada y la institución de egreso.

   Los resultados destacan la existencia de grupos consolidados con trayectorias positivas, como el de *mujeres avanzadas con alto desempeño*, y otros con señales de vulnerabilidad, como los *estudiantes nocturnos adultos* o los *jóvenes de estrato bajo en semestres iniciales*. Esta diversidad de perfiles

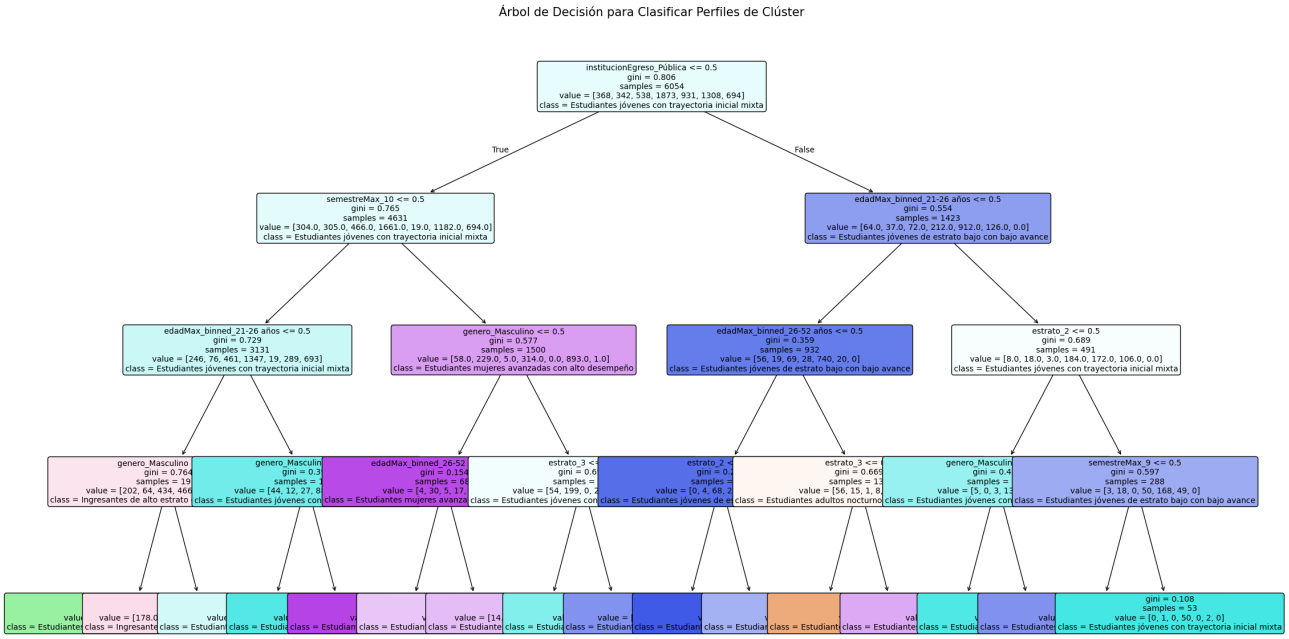

Figura 27: Árbol de Decisión para Clasificar Perfiles de Clúster

Tabla 9: Perfiles dominantes por clúster (modalidad más frecuente)

Clúster	Estado Civil	Género	Jornada	Tipo Est.	Inst. Egreso	Semestre	Estrato	Edad	Programa	Perfil nombrado
0	Soltero(a)	Masculino	Diurna	Antiguo	Privada	9	3	21–26	Ing. Sistemas	Jóvenes con trayectoria inicial mixta
1	Soltero(a)	Masculino	Diurna	Antiguo	Pública	1	2	15–20	Ing. Sistemas	Jóvenes de estrato bajo con bajo avance
2	Soltero(a)	Femenino	Diurna	Antiguo	Privada	10	4	21–26	Ing. Ambiental	Mujeres avanzadas con alto desempeño
3	Soltero(a)	Masculino	Nocturna	Antiguo	Privada	11	3	26–52	Ing. Electrónica	Adultos nocturnos con progresión lenta
4	Soltero(a)	Masculino	Diurna	Antiguo	Privada	10	2	26–52	Ing. Industrial	Estudiantes avanzados de bajo estrato
5	Soltero(a)	Masculino	Diurna	Antiguo	Privada	1	4	15–20	Ing. Sistemas	Ingresantes de alto estrato con riesgo
6	Soltero(a)	Femenino	Diurna	Antiguo	Privada	1	3	15–20	Bioingeniería	Femeninas jóvenes en transición

demuestra que el riesgo de deserción o rezago no responde a un solo factor, sino a una interacción compleja de dimensiones académicas, sociales y personales.

En este sentido, los clústeres no solo ofrecen un panorama descriptivo del estudiantado, sino que se configuran como una herramienta estratégica para el diseño de políticas de retención diferenciadas y con enfoque preventivo. La caracterización multivariada facilita el diseño de intervenciones focalizadas, mientras que el uso de modelos explicativos como árboles de decisión y pruebas estadísticas permite validar empíricamente las relaciones entre variables clave y el éxito académico.

4. El **Análisis de Correspondencias Múltiples (ACM)** evidenció que variables como el programa académico, el estrato, la jornada y el género contribuyen de manera significativa a la estructuración del espacio factorial, reforzando la validez de los clústeres formados.

5. La **clasificación mediante árbol de decisión** demostró que variables como institución de egreso, edad, género y semestre cursado son determinantes para asignar a los estudiantes a sus respectivos perfiles, lo que aporta una herramienta explicativa e interpretable para la toma de decisiones institucionales.

6. Las **pruebas de independencia de Chi-cuadrado** revelaron que existen asociaciones estadísticamente significativas entre el estado de graduación y variables como el género, el programa académico y la jornada. Sin embargo, no se evidenció una asociación significativa con el estado civil.

7. Se constató la presencia de **perfiles vulnerables**, como estudiantes de estrato bajo, jornada nocturna o provenientes de instituciones públicas, quienes presentan mayor probabilidad de no culminar sus estudios, lo que resalta la necesidad de acciones diferenciales de acompañamiento.

5. Recomendaciones

- **Desarrollar intervenciones focalizadas a partir de los factores de riesgo clave.** Las variables identificadas por SHAP como altamente influyentes (semestre aprobado, estrato, jornada, tipo de estudiante y edad) deben ser utilizadas

como criterios prioritarios para diseñar acciones de acompañamiento académico, financiero y psicosocial. En especial, se debe prestar atención a estudiantes mayores con bajo avance curricular y a quienes se encuentran en modalidades nocturnas.

- **Utilizar la segmentación por clústeres como base para políticas diferenciadas.** Los siete perfiles estudiantiles identificados mediante K-Modes ofrecen una tipología útil para dirigir recursos institucionales de forma más eficiente. Por ejemplo, se recomienda priorizar a los estudiantes de clústeres en riesgo (como los jóvenes de estrato bajo en semestres iniciales o los adultos de jornada nocturna) con programas específicos de tutoría, monitoreo personalizado y apoyo financiero.

- **Aprovechar el ACM para análisis diagnósticos y toma de decisiones visual.** Dado que variables como programa académico, estrato, jornada y género fueron determinantes en el espacio factorial, el ACM puede ser usado periódicamente como una herramienta visual y exploratoria para monitorear la evolución de la composición estudiantil y detectar cambios estructurales en los perfiles de riesgo.

- **Incorporar árboles de decisión como instrumentos explicativos para orientar políticas.** Los árboles construidos permiten comprender cómo interactúan variables clave para segmentar a los estudiantes. Su simplicidad e interpretabilidad los hacen ideales para que directivos y personal de apoyo diseñen estrategias basadas en reglas claras y replicables.

- **Establecer mecanismos de seguimiento diferencial por programa académico y jornada.** Las asociaciones significativas encontradas mediante pruebas de Chi-cuadrado indican que ciertos programas y modalidades están más expuestos al riesgo de deserción. Se sugiere implementar estrategias de retención adaptadas al contexto de cada facultad o modalidad (diurna/nocturna).

- **Promover políticas institucionales con enfoque de equidad.** Dado que la vulnerabilidad académica está fuertemente ligada a condiciones estructurales (estrato bajo, egreso de institución pública, jornada nocturna), las estrategias institucionales deben ir más allá de lo académico, incorporando criterios de justicia social y apoyo integral que contribuyan a la permanencia con equidad.

Referencias

Álvarez-García, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Uncovering student profiles: An explainable cluster analysis approach to PISA 2022. *Computers & Education*, *223*(100), 105166. https://doi.org/10.1016/j.compedu.2024.105166

Anahua, T., Yana, V., & ReynosoD, A. (2025). Machine Learning Algorithms for Predicting Student Dropout in Engineering Programs. *Information Management and Big Data: 11th Annual International Conference, SIMBig 2024, Ilo, Peru, November 20–22, 2024, Proceedings*, *2496*, 68.

Barros, T. M., Silva, I., & Guedes, L. A. (2019). Determination of dropout student profile based on correspondence analysis technique. *IEEE Latin America Transactions*, *17*(9), 1517-1523. https://doi.org/10.1109/TLA.2019.8931146

Boström, H. (2020). MissingPy: Imputation and missing data analysis in Python.

Boteju, G., Tang, L., & Brown, M. S. (2024). Support Vector Machine for Predicting Student Dropout Under Different Normalization Methods. *2024 IEEE International Conference on Big Data (BigData)*, 8633-8636.

Chen, L. (2024). Model-driven early alert systems for student retention. *Educational Data Science*, *2*(1), 1-15.

Choque-Soto, V. M., Sosa-Jauregui, V. D., & Ibarra, W. (2025). Characterization of the dropout student profile using data mining techniques. *Revista de Gestão Social e Ambiental*. https://www.researchgate.net/publication/389099375

Dake, D. K., Bada, G. K., & Techie-Menson, H. (2023). Using Machine Learning to Cluster and Predict the Learning Pattern of University Students. *Telematique*, *22*(1), 77-84. https://www.researchgate.net/publication/367190613

Dake, D. K., Bada, G. K., & Techie-Menson, H. (2024). Model interpretability on private-safe oriented student dropout prediction. *PLOS ONE*, *19*(5), e0317726. https://doi.org/10.1371/journal.pone.0317726

Darwis, M., Hasibuan, L. H., & Firmansyah, M. (2025). An analysis of students' academic performance using K-means clustering algorithm. *JISA (Jurnal Informatika dan Sains)*. https://www.researchgate.net/publication/359547536

Estrada, J. C., & Rodríguez, A. (2022). *Using SHAP Values to Explain Dropout Prediction Models in Higher Education* [Tesis doctoral, Tecnológico de Monterrey]. https://repositorio.tec.mx/bitstreams/f2e9ebda-ed86-4fd3-a949-4ee8b9ba26f2/download

Flores-Báez, M. A. (2023). Dificultades en la asignatura de Formación Cívica y Ética como causa de deserción escolar en secundaria. *Revista Latinoamericana de Educación*, *16*(2), 35-52. https://alumnieditora.com/index.php/ojs/article/view/175

GeeksforGeeks. (2025). K-mode clustering in Python. https://www.geeksforgeeks.org/k-mode-clustering-in-python/

Hanji, S. (2023). Predicting Student Dropout Using Machine Learning.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., & Cournapeau, D. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Instituto Nacional de Estadística y Geografía (INEGI). (2023). Estadísticas a propósito del día internacional de la educación [Consultado el 23 de mayo de 2025]. https://www.inegi.org.mx

Jin, W., Li, X., & Zhang, H. (2024). From Data to Decision: Machine Learning and Explainable AI in Student Dropout Prediction. *Journal of Education and Learning*. https://ibimapublishing.com/articles/JELHE/2024/246301/

Kabir, R. (2024). Student Dropout Prediction Through Machine Learning Optimization. *Scientific Reports*, *14*, 10032. https://www.nature.com/articles/s41598-025-93918-1

Kaur, H., & Sharma, V. (2023). A Study on Dropout Prediction for University Students Using Machine Learning.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Khan, R. A. (2023). Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study. *International Journal of Data Science and Analytics*, *7*(1), 45-59. https://link.springer.com/article/10.1007/s44163-023-00079-z

Liu, Y., Wang, H., & Xu, Y. (2023). Who Will Dropout from University? Academic Risk Prediction Based on Interpretable Machine Learning.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4768-4777.

Malik, S., & Tüfekci, A. (2023). An Explainable Machine Learning Approach to Predicting and Understanding Dropouts in MOOCs. *Educational Data Mining Conference*. https://open.metu.edu.tr/bitstream/handle/11511/102701/10.24106-kefdergi.1246458-2933723.pdf

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51-56. https://doi.org/10.25080/Majora-92bf1922-00a

Melchor, F., Conejero, J. M., Fernández-García, A. J., Sánchez-Figueroa, F., & Rodríguez-Echeverría, R. (2025). An empirical evaluation of clustering processes for early detection of university dropout. https://doi.org/10.21203/rs.3.rs-6146415/v1

Ministerio de Educación del Perú. (2021). Análisis de la deserción escolar en el nivel primario durante la pandemia por COVID-19 [Consultado el 23 de mayo de 2025]. https://www.inei.gob.pe/media/MenuRecursivo/investigaciones/desercion-escolar.pdf

Ministerio de Educación Nacional de Colombia. (2022). Indicadores de eficiencia interna del sistema educativo [Consultado el 23 de mayo de 2025]. https://www.mineducacion.gov.co

Nguyen, T. (2020). Predicting student dropout using machine learning: A review. *IEEE Access*, *8*, 20036-20054.

Nguyen, T. M., Doan, Q. H., & Pham, V. T. (2023). A comparative study of clustering algorithms for categorical data: Evaluation and applications. *Expert Systems with Applications*, *213*, 118957. https://doi.org/10.1016/j.eswa.2022.118957

Olugbara, C. T., Letseka, M., & Olugbara, O. O. (2021). Multiple correspondence analysis of factors influencing student acceptance of massive open online courses. *Sustainability*, *13*(23), 13451. https://doi.org/10.3390/su132313451

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Blondel, M. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106.

ResearchGate. (2025). An empirical evaluation of clustering processes for early detection of university dropout. https://www.researchgate.net/publication/389606807

Romero, L., & Ventura, J. (2020). Factores asociados a la deserción en ingeniería. *Revista Educación y Desarrollo*, *18*, 45-60.

Sanchez, D., Martínez, G., & Torres, M. (2023). College Dropout Factors: An Analysis with LightGBM and Shapley's Cooperative Game Theory.

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press.

Torres Acero, N. (2022). *Modelos para la predicción de deserción universitaria de estudiantes de Psicología de la Universidad El Bosque* [Trabajo de grado]. Universidad El Bosque.

UNESCO. (2023). Why are boys at risk of dropping out of school? [Consultado el 23 de mayo de 2025]. https://www.unesco.org/es/gender-equality/education/boys

Vásquez, J. (2016). *Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior* [Tesis de magíster]. Universidad de Chile, Facultad de Economía y Negocios.

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2.

Waskom, M. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021