Universidad Tecnológica de Panamá

Maestría en Analítica de Datos

Curso:

MODELOS PREDICTIVOS

REPORTE PROYECTO FINAL

Predicción del rendimiento académico de estudiantes de Educación Superior mediante Análisis de Datos

Profesor:

Juan Marcos Castillo, PhD

Maestrando:

Rolando Antonio Mora De León

9-756-1619

2025

ÍNDICE

1. INTRODUCCIÓN

El rendimiento académico es un indicador clave del éxito educativo, ya que altas tasas de deserción y fracaso escolar obligan a las instituciones a implementar estrategias de apoyo a los alumnos en riesgo (González-Ruiz et al., 2023) (Gil-Vera & Quintero-López, 2023). El auge de la analítica de datos educativos (Educational Data Mining y Learning Analytics) ha abierto la puerta a utilizar técnicas de inteligencia artificial para aprovechar estos datos y obtener información valiosa. En particular, la predicción del rendimiento estudiantil se refiere al uso de modelos de machine learning que, a partir de datos históricos y variables relevantes de los alumnos (antecedentes académicos, asistencia, hábitos, contexto socioeconómico, etc.), permitan anticipar el desempeño académico futuro de los estudiantes.

Estudios recientes muestran que habilidades no cognitivas como la tenacidad (grit) se relacionan estrechamente con mejores calificaciones (González-Ruiz et al., 2023), lo cual subraya la importancia de identificar factores asociados al rendimiento. A pesar de contar con datos variados sobre los estudiantes, las universidades a menudo carecen de herramientas para sintetizar esa información y predecir qué estudiantes podrían tener dificultades al final del semestre. El problema que abordaremos es precisamente cómo utilizar la información disponible (demografía, contexto familiar y hábitos de estudio de los alumnos) para predecir su rendimiento académico final, expresado en la calificación obtenida en sus cursos.

Este informe presenta un enfoque integral que combina análisis descriptivo, pruebas estadísticas (Kruskal-Wallis, Chi-cuadrado) y modelos predictivos (Random Forest, Árbol de Decisión, Regresión Logística, Naive Bayes) para predecir el rendimiento académico en educación superior. Se espera que esta investigación no solo cumpla con los objetivos planteados, sino que también sirva como experiencia formativa integral en el uso de la ciencia de datos para abordar problemas reales en el ámbito educativo.

2. JUSTIFICACIÓN

Predecir el rendimiento académico de forma temprana permite diseñar intervenciones que aumenten la retención estudiantil y la calidad educativascielo.cl. En particular, la identificación de estudiantes en riesgo posibilita brindar apoyos personalizados (tutorías, recursos psicosociales, etc.), lo que ha demostrado reducir las tasas de deserción y elevar las tasas de graduación (Gil-Vera & Quintero-López, 2023). Además, la predicción automática apoya la toma de decisiones institucionales y la mejora continua de planes de estudio. Por ejemplo, Castillo Araúz y Martínez (2023) resaltan que modelos de aprendizaje automático bien ajustados pueden alertar sobre cursos críticos y orientar a los profesores en estrategias didácticas.

En resumen, este estudio está motivado por la necesidad de aportar información objetiva y basada en datos para optimizar los procesos educativos y el seguimiento del rendimiento. El objetivo de la investigación es doble: por un lado, describir estadísticamente el conjunto de datos para obtener insights (por ejemplo, distribuciones de notas, perfiles típicos de alumnos destacados frente a rezagados), y por otro, desarrollar un modelo predictivo que anticipe el rendimiento, abordando la pregunta: ¿Es posible predecir con precisión el rendimiento final de un estudiante universitario utilizando sus características personales, familiares y académicas, y cuáles de estos factores contribuyen más a dicha predicción? Resolver este problema aportaría valor tanto a nivel académico, demostrando la aplicación de técnicas de IA en educación, como práctico, al nutrir políticas de apoyo estudiantil basadas en datos.

3. ANTECEDENTES

Diversas investigaciones han explorado factores asociados al desempeño académico y técnicas analíticas para predecirlo. Se han aplicado pruebas no paramétricas como la prueba H de Kruskal-Wallis para detectar diferencias de rendimiento entre grupos (por ejemplo, según niveles de habilidades no cognitivas). González-Ruiz et al. (2023) utilizaron la prueba H de Kruskal-Wallis junto con correlaciones de Spearman y hallaron diferencias significativas de rendimiento según el nivel de tenacidad de los estudiantes. Por otro lado, análisis de contingencia con chi-cuadrado han revelado asociaciones estadísticamente significativas entre variables categóricas y resultados académicos. Por ejemplo, Cabeza García y Razo Cajas

(2024) encontraron que los estudiantes que participan en actividades extracurriculares obtienen calificaciones significativamente mejores que quienes no participan.

En el ámbito predictivo, la literatura destaca que los algoritmos de aprendizaje supervisado más frecuentes para predecir el rendimiento son los siguientes:

- **Árboles de Decisión (Decision Trees)**, que construyen reglas de clasificación simples y explicables.
- **K-Vecinos más Cercanos (KNN)**, que clasifica según la similitud con ejemplos cercanos.
- **Naive Bayes**, un método probabilístico basado en teoremas bayesianos.
- **Bosques Aleatorios (Random Forest)**, un ensamblado de árboles de decisión que suele mejorar la precisión.
- **Máquinas de Soporte Vectorial (SVM)** y **Redes Neuronales Artificiales**, también utilizadas por su alta capacidad predictiva.

Por ejemplo, un estudio en cursos universitarios virtuales reportó que Random Forest obtuvo las mejores puntuaciones en todas las métricas de evaluación (precisión, F1, etc.). De hecho, Castillo Araúz y Martínez (2023) observaron que Random Forest superó el 90 % de exactitud en su tarea, mientras otros modelos como Gradient Boosted Trees alcanzaron hasta 93.6 %. Estas evidencias respaldan la elección de Random Forest, junto con otros modelos (árboles de decisión, regresión logística, Naive Bayes), para la predicción en educación. En síntesis, la revisión bibliográfica indica que variables demográficas (edad, género) y académicas (créditos aprobados, promedio) suelen correlacionar con el éxito estudiantil, y que técnicas de minería de datos como Random Forest y Naive Bayes ofrecen altos niveles de precisión en este dominio (Montero, Montilla & Arcia, 2024).

4. DEFINICIÓN DEL PROBLEMA

La investigación busca formular modelos que permitan predecir el rendimiento académico final de los estudiantes de educación superior. Para ello se identifica una *variable objetivo* (p.ej.

aprobación o nota final en un curso) y un conjunto de *variables predictoras* relacionadas con el perfil del estudiante (demografía, historial académico, contexto socioeconómico, hábitos de estudio, participación extracurricular, etc.). Se plantean las siguientes preguntas de investigación: ¿Qué variables explican mejor el desempeño académico? ¿Qué modelo predictivo ofrece mayor exactitud para clasificar a los estudiantes (aprobado/reprobado)?

Para abordar esto, se emplearán técnicas estadísticas y de machine learning clásicas: pruebas H de Kruskal-Wallis y χ^2 para explorar diferencias y asociaciones en los datos, y algoritmos de clasificación como Árbol de Decisión, Random Forest, Regresión Logística y Naive Bayes para la predicción. El proceso general incluye selección y preparación de las variables, división de los datos en conjuntos de entrenamiento y prueba, entrenamiento de los modelos y evaluación con medidas de desempeño (exactitud, precisión, recall, etc.). En línea con estudios previos, se espera que Random Forest obtenga las mejores métricas predictivas (Castillo Araúz & Martínez, 2023), y que Naive Bayes sea eficaz para identificar alumnos en riesgo (Montero, Montilla & Arcia, 2024). Estos modelos ayudarán a generar criterios objetivos para enfocar intervenciones educativas.

5. ANÁLISIS PREDICTIVO

5.1. Determinación de la base de datos

La base de datos "*Higher Education Students Performance Evaluation*" (Evaluación del rendimiento de estudiantes de educación superior) proviene del repositorio UCI Machine Learning. Contiene información recopilada en el año 2019 de 145 estudiantes universitarios de dos facultades (Ingeniería y Ciencias de la Educación). En total incluye 31 variables predictoras sobre cada estudiante más la variable objetivo, sin contar un identificador de estudiante. Las mismas se encuentran definidas en la Tabla 1.

Tabla 1. Variables de la base de datos *Higher Education Students Performance Evaluation*.

Número	Variable	Tipo	Categorías (códigos)
0	Student ID (identificador del estudiante)	Nominal	Valores únicos ("STUDENT1", "STUDENT2", …)
1	Edad del estudiante (rango de años)	Ordinal	1: 18-21 años; 2: 22-25; 3: >26

Número	Variable	Tipo	Categorías (códigos)
2	Sexo (género)	Nominal	1: Femenino; 2: Masculino
3	Tipo de colegio de egreso (sección de bachillerato)	Nominal	1: Privado; 2: Estatal; 3: Otro
4	Tipo de beca	Ordinal	1: Ninguna; 2: 25%; 3: 50%; 4: 75%; 5: Completa
5	Trabajo adicional	Nominal	1: Sí; 2: No
6	Actividades artísticas/deportivas	Nominal	1: Sí; 2: No
7	Pareja (estado civil)	Nominal	1: Sí; 2: No
8	Ingreso familiar (total disponible)	Ordinal	1: USD 135-200; 2: USD 201-270; 3: USD 271-340; 4: USD 341-410; 5: >410
9	Medio de transporte a la universidad	Nominal	1: Bus; 2: Vehículo privado/taxi; 3: Bicicleta; 4: Otro
10	Tipo de alojamiento en Chipre	Nominal	1: Alquiler; 2: Residencia universitaria; 3: Con la familia; 4: Otro
11	Educación de la madre	Ordinal	1: Primaria; 2: Secundaria; 3: Preparatoria; 4: Universidad; 5: Maestría; 6: PhD
12	Educación del padre	Ordinal	1: Primaria; 2: Secundaria; 3: Preparatoria; 4: Universidad; 5: Maestría; 6: PhD
13	Número de hermanos	Ordinal	1: 1 hermano(a); 2: 2; 3: 3; 4: 4; 5: 5 o más
14	Estado civil de los padres	Nominal	1: Casados; 2: Divorciados; 3: Falleció uno o ambos padres
15	Ocupación de la madre	Nominal	1: Jubilada; 2: Ama de casa; 3: Empleada pública; 4: Empleada privada; 5: Independiente; 6: Otra
16	Ocupación del padre	Nominal	1: Jubilado; 2: Empleado público; 3: Empleado privado; 4: Independiente; 5: Otra
17	Horas de estudio semanales	Ordinal	1: Ninguna; 2: <5 horas; 3: 6–10 horas; 4: 11–20 horas; 5: >20 horas
18	Frecuencia de lectura (no científico)	Ordinal	1: Nunca; 2: A veces; 3: Con frecuencia
19	Frecuencia de lectura (científico)	Ordinal	1: Nunca; 2: A veces; 3: Con frecuencia
20	Asistencia a seminarios/conferencias	Nominal	1: Sí; 2: No
21	Impacto de proyectos/actividades en el éxito	Ordinal	1: Positivo; 2: Negativo; 3: Neutral
22	Asistencia a clases	Ordinal	1: Siempre; 2: A veces; 3: Nunca
23	Preparación para exámenes parciales (estrategia 1)	Ordinal	1: Solo; 2: Con amigos; 3: No aplicable
24	Preparación para exámenes parciales (estrategia 2)	Ordinal	1: Cercano a la fecha; 2: Regularmente durante el semestre; 3: Nunca
25	Tomar apuntes en clases	Ordinal	1: Nunca; 2: A veces; 3: Siempre

Número	Variable	Tipo	Categorías (códigos)
26	Escuchar en clases	Ordinal	1: Nunca; 2: A veces; 3: Siempre
27	El debate mejora mi interés en el curso	Ordinal	1: Nunca; 2: A veces; 3: Siempre
28	Flip classroom (clase invertida)	Nominal	1: No útil; 2: Útil; 3: No aplicable
29	GPA acumulado semestre anterior (escala /4.00)	Ordinal	1: <2.00; 2: 2.00-2.49; 3: 2.50-2.99; 4: 3.00-3.49; 5: ≥3.50
30	GPA acumulado esperado en graduación (/4.00)	Ordinal	1: <2.00; 2: 2.00-2.49; 3: 2.50-2.99; 4: 3.00-3.49; 5: ≥3.50
31	Course ID (código del curso)	Nominal	1, 2, …, 9 (ID de curso)
32	Nota final (salida: output grade)	Ordinal	0: Fracaso (Fail); 1: DD; 2: DC; 3: CC; 4: CB; 5: BB; 6: BA; 7: AA

Fuente: Elaboración propia.

Las variables 2–31 son categorías ordinales o nominales codificadas numéricamente (p.ej. edades, escalas de encuesta, código de curso). El Student ID se utiliza solo como identificador. La variable objetivo, columna 33, es la calificación final del estudiante en determinada asignatura, registrada en una escala de 8 categorías (desde 0 = “Fail” o reprobado, hasta 7 = “AA”, que equivale a la nota máxima). El propósito original de esta base de datos es utilizar técnicas de *machine learning* para predecir el rendimiento académico de fin de semestre de los estudiantes, sirviendo, así como un conjunto de datos ideal para análisis predictivo en educación (UCI Machine Learning Repository, s. f.).

Se ha elegido esta base de datos por múltiples razones:

- **Relevancia educativa:** El rendimiento académico en educación superior es un tema crítico, considerado el principal indicador de éxito o fracaso de un estudiante (Contreras et al., 2020). Analizar y predecir dicho rendimiento tiene implicaciones prácticas importantes, ya que permitiría identificar estudiantes en riesgo y mejorar los resultados educativos.

- **Riqueza de variables:** El conjunto de datos incluye una variedad de factores personales, socioeconómicos y académicos. Esto permite explorar cuáles de ellos son más determinantes en el éxito estudiantil, fomentando el *autodescubrimiento de conocimiento* al relacionar distintas dimensiones (familiares, hábitos de estudio, antecedentes académicos, etc.) con las calificaciones.

- **Calidad y disponibilidad:** Los datos provienen de una fuente confiable (UCI Repository), están claramente documentados, lo que reduce el esfuerzo de limpieza. Al ser un conjunto relativamente pequeño (145 casos), es manejable para efectos de un proyecto académico y computacionalmente eficiente de analizar en herramientas comunes.

- **Afinidad e interés:** El tema de la predicción del desempeño estudiantil resulta de interés personal y profesional, ya que combina el campo educativo con técnicas de análisis de datos. Esto ofrece la oportunidad de aplicar métodos de *machine learning* en un contexto real y entender cómo la analítica de datos puede apoyar la toma de decisiones en educación.

- **Objetivo formativo:** La selección de estos datos facilita el objetivo de aprendizaje autónomo. Dado que el proyecto enfatiza la exploración y descubrimiento, trabajar con un tema tangible como el rendimiento de estudiantes motiva la investigación activa y la aplicación práctica de habilidades analíticas.

5.2. Preprocesamiento y limpieza

No se detectaron valores faltantes en el dataset. Primero se eliminó la columna de identificación (*Student ID*), ya que no aporta información predictiva. Las demás variables categóricas se mantuvieron en su forma codificada. Para algunos modelos (p.ej. Naive Bayes) se aplicó codificación *one-hot* (variables ficticias) a las categorías, mientras que otros algoritmos (árbol de decisión, bosques aleatorios) admiten directamente datos codificados numéricamente.

La variable objetivo original *GRADE* (0–7) se reagrupó en clases para simplificar el análisis de clasificación. Siguiendo lo planteado en estudios previos, definimos dos clases de desempeño Grade_Grouped:

- Bajo: calificaciones entre 0 y 3 (rendimiento insuficiente).
- Alto: calificaciones entre 4 y 7 (rendimiento aceptable).

Esto equivale a distinguir rendimientos insuficientes frente a satisfactorios, similar a la clasificación de GPA (<4 vs ≥4) propuesta por Yilmaz y Şekeroğlu (UCI Machine Learning

Repository, s. f.). También se verificó el balance de clases resultante (aproximadamente 60% bajo vs 40% alto) para aplicar técnicas de muestreo estratificado durante la validación. La distribución de las clases se observa en la Figura 1.

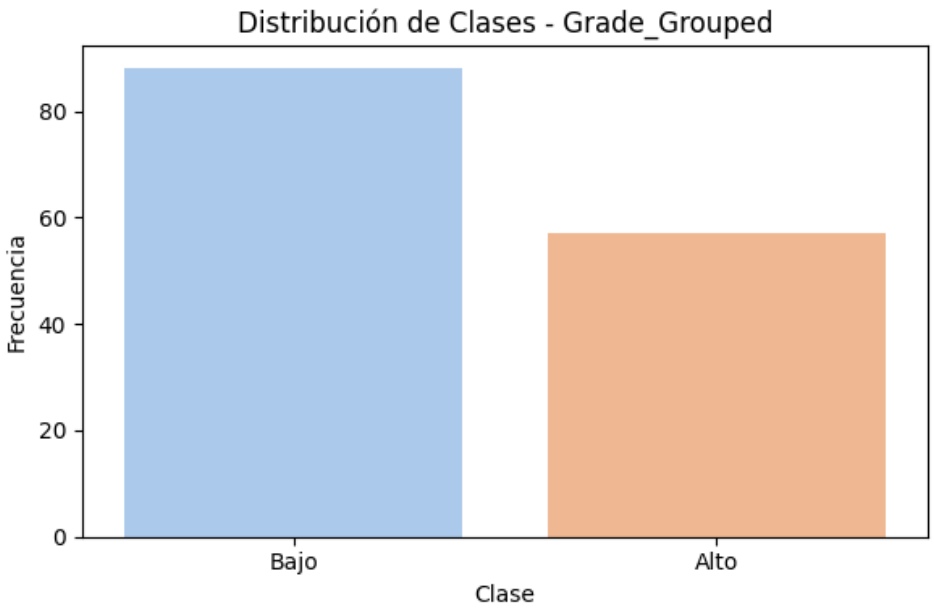


Figura 1. Distribución de las clases Grade_Grouped. *Fuente*: Elaboración propia mediante Visual Studio Code.

Las variables ordinales (como niveles educativos o escalas de GPA) se conservaron como numéricas, respetando su orden. En trabajos estadísticos posteriores, algunas variables se etiquetaron explícitamente (por ejemplo, *Sex*: 1=Femenino, 2=Masculino) para facilitar la interpretación. No se aplicaron eliminaciones adicionales (no se detectó outlier ni inconsistencia crítica).

5.3. Análisis descriptivo

La selección de variables para el análisis estadístico se llevó a cabo mediante una técnica de filtrado exploratorio, basada en el tipo de escala (ordinal o nominal), el conocimiento del dominio educativo y la revisión empírica de la distribución de los datos. Este enfoque permitió identificar aquellas variables con mayor potencial explicativo del rendimiento académico, sin introducir

sesgos derivados de técnicas automáticas de selección, que pueden ser aplicadas en fases posteriores del análisis predictivo.

Adicionalmente se eligieron las pruebas Kruskal-Wallis y Chi-cuadrado para un análisis estadístico inferencial. La prueba Kruskal-Wallis se prefirió frente a ANOVA porque esta última asume normalidad y homogeneidad de varianzas, condiciones que no se cumplen en este caso debido a la naturaleza ordinal de las variables y la distribución no normal de las calificaciones. Del mismo modo, la prueba Chi-cuadrado es estándar para analizar relaciones entre variables cualitativas sin necesidad de supuestos paramétricos como la varianza o la distribución, a diferencia de pruebas como la regresión logística o las pruebas exactos, que exigen condiciones más estrictas o son más costosos computacionalmente. Un resumen de lo descrito se muestra en la Tabla 2.

Tabla 2. Pruebas estadísticas empleadas.

Prueba estadística	Tipo de variable	Evaluación
Kruskal-Wallis	Variables ordinales o numéricas discretas	Si hay diferencias significativas entre las medianas de más de dos grupos (Bajo, Medio y Alto rendimiento).
Chi-cuadrado (χ^2)	Variables nominales o categóricas puras	Si hay asociación entre categorías y el rendimiento académico

Fuente: Elaboración propia.

Como se mencionó anteriormente, el conjunto de datos contiene principalmente variables cualitativas (categóricas nominales y ordinales). En este contexto no es apropiado usar medidas paramétricas clásicas (como la media aritmética) ni asumir normalidad. Por tal motivo, se realizó el análisis estadístico por tipo de variable (nominal u ordinal).

Como primer paso, se realizó el cálculo de estadísticas descriptivas para las variables ordinales (media, desviación estándar, mínimo y máximo). De los resultados se destacan:

- **GPA_Last_Semester**. Los estudiantes del grupo Alto presentaron un GPA promedio superior al de los otros grupos. Esto indica que el rendimiento anterior es un fuerte predictor del rendimiento futuro.

- **Expected_GPA_Graduation**. A mayor expectativa de GPA al graduarse, mayor es la probabilidad de estar en el grupo Alto. Sugiere que los estudiantes con metas académicas más ambiciosas tienden a obtener mejores resultados.
- **Weekly_Study_Hours**. El grupo Alto reporta más horas de estudio semanal en promedio. Esta variable refleja una mayor dedicación fuera del aula.
- **Preparation_Midterm1 y Preparation_Midterm2**. La preparación para los exámenes parciales muestra una relación creciente con el grupo de nota. El grupo Alto se prepara más activamente, lo que refuerza la importancia de los hábitos de estudio.

Este análisis preliminar fue complementado con la prueba no paramétrica de Kruskal-Wallis, la cual confirmó que estas diferencias entre grupos son estadísticamente significativas ($p < 0.05$) en seis variables: edad del estudiante (Student_Age), GPA del semestre anterior (GPA_Last_Semester), GPA esperado al graduarse (Expected_GPA_Graduation), frecuencia de lectura no científica (Reading_Freq_NonScientific), nivel educativo de la madre (Mother_Education) y percepción del impacto de proyectos sociales (Project_Impact_Success).

Estas asociaciones sugieren que los estudiantes con mejor rendimiento suelen tener un historial académico sólido (GPA alto), expectativas altas de éxito futuro, mayor hábito lector, y provienen de entornos familiares con mayor nivel educativo. Además, perciben con mayor frecuencia un impacto positivo en los proyectos sociales en los que participan. Otro hallazgo importante es que los estudiantes más jóvenes tendieron a ubicarse en niveles de rendimiento más altos. Los resultados obtenidos se resumen en la Tabla 3.

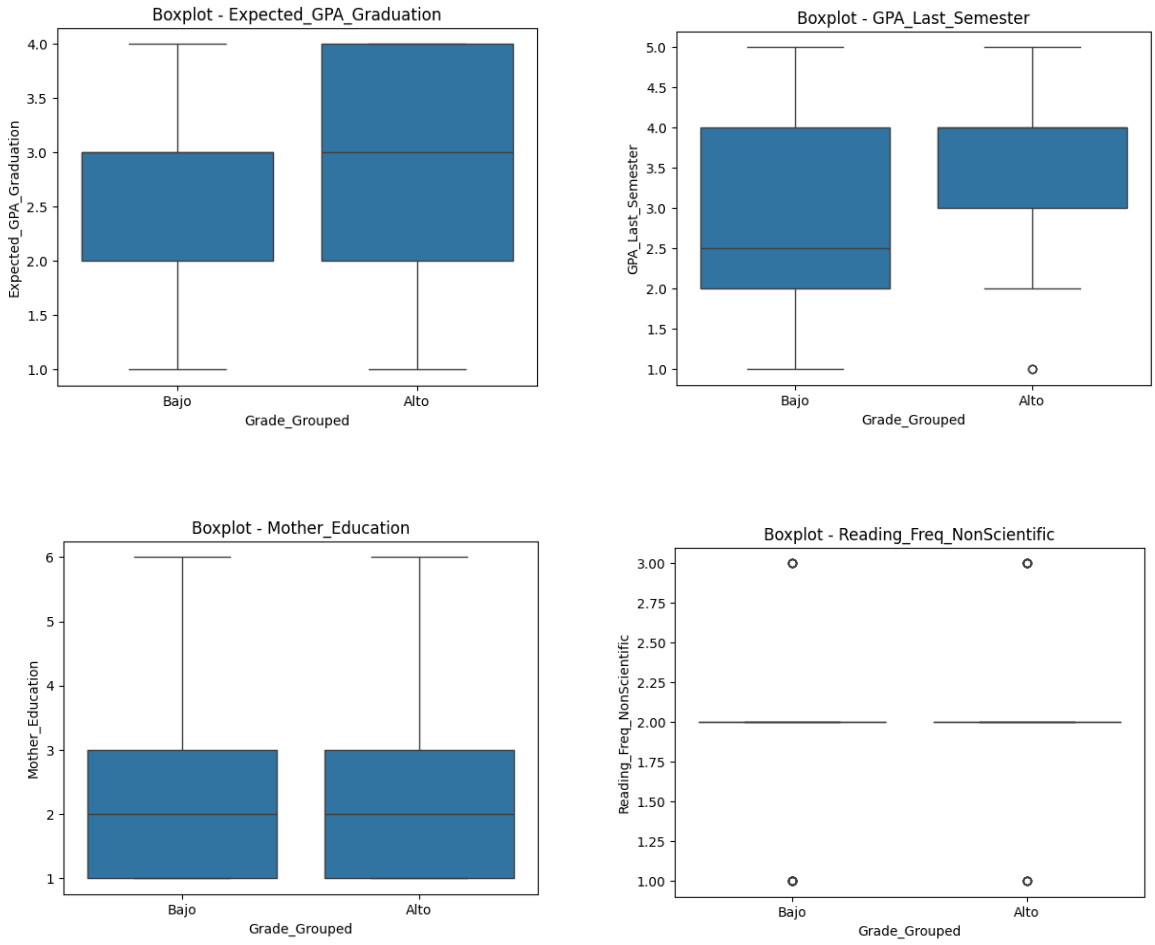
Tabla 3. Prueba de Kruskal–Wallis en variables ordinales.

Variable	Estadístico H	Valor p	Significativo
Student_Age	6.2088	0.0448	VERDADERO
Mother_Education	6.9442	0.031	VERDADERO
Father_Education	1.3699	0.5041	FALSO
Number_Siblings	4.7612	0.0925	FALSO
Total_Salary	5.5163	0.0634	FALSO
Weekly_Study_Hours	0.8279	0.661	FALSO
Reading_Freq_NonScientific	8.2256	0.0164	VERDADERO
Reading_Freq_Scientific	0.602	0.7401	FALSO
Department_Seminar_Attendance	5.2982	0.0707	FALSO
Project_Impact_Success	6.4484	0.0398	VERDADERO

Variable	Estadístico H	Valor p	Significativo
Class_Attendance	5.4041	0.0671	FALSO
Preparation_Midterm1	0.1095	0.9467	FALSO
Preparation_Midterm2	1.2783	0.5278	FALSO
Taking_Notes	2.3569	0.3078	FALSO
Listening_Classes	2.954	0.2283	FALSO
Discussion_Interest_Success	4.8306	0.0893	FALSO
Flip_Classroom	3.7834	0.1508	FALSO
GPA_Last_Semester	28.7854	0	VERDADERO
Expected_GPA_Graduation	14.3083	0.0008	VERDADERO

Fuente: Elaboración propia.

Adicionalmente, se generaron boxplots para mostrar las tendencias de rendimiento frente a la variable objetivo y sus clases de forma visual. Como ejemplo, se muestran algunas variables con significancia anteriormente mencionadas en la Figura 2.



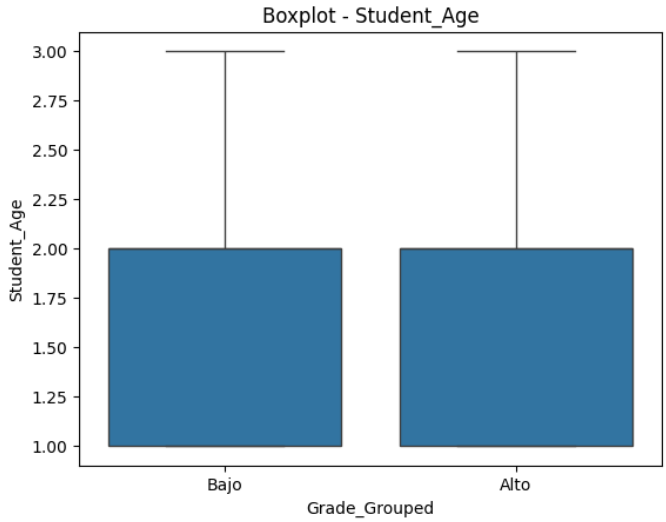


Figura 2. Boxplots generados de variables ordinales. *Fuente*: Elaboración propia mediante Visual Studio Code.

En el caso de las variables nominales, estas fueron codificadas numéricamente mediante la técnica de *LabelEncoder*, que asigna un número entero a cada categoría. Esta codificación no implica orden entre las categorías, pero es necesaria para el procesamiento en modelos de machine learning.

Se hizo un análisis de las frecuencias de estas variables donde solamente Sex (Género) fue la única variable nominal que presentó diferencias notables en la distribución por grupo de nota. Las mujeres tendieron a agruparse en los niveles de rendimiento medio y alto, mientras que los hombres estuvieron más representados en el nivel bajo.

Para confirmar la existencia de asociación estadística entre cada variable nominal y el rendimiento académico agrupado (Grade_Grouped), se utilizó la prueba Chi-cuadrado de independencia. El resultado se muestra en la Tabla 4.

Tabla 4. Prueba de Chi-cuadrado en variables nominales.

Variable	Estadístico χ^2	Valor p	Significativo
Sex	6.418668594	0.011292679	VERDADERO
High_School_Type	3.140612063	0.207981524	FALSO
Scholarship_Type	13.97111615	0.007387814	VERDADERO

Additional_Work	0.98570782	0.320793688	FALSO
Art_Sport_Activity	2.660694112	0.102855841	FALSO
Have_Partner	0.27429541	0.600464821	FALSO
Accommodation_Type	0.984856064	0.804916276	FALSO
Mother_Occupation	1.813446432	0.770021347	FALSO
Father_Occupation	2.123470956	0.713061658	FALSO
Parental_Status	1.60134935	0.449025915	FALSO
Transport_Type	3.284442166	0.349813882	FALSO

Fuente: Elaboración propia.

La prueba confirmó que dos variables nominales están significativamente asociadas con el rendimiento académico: Sex (Género) y Scholarship_Type (Tipo de beca).

En el caso del género, el resultado ($\chi^2 = 6.42$, $p = 0.011$) indica diferencias estadísticamente significativas en el desempeño entre hombres y mujeres, lo que sugiere que el género puede influir en factores como la motivación, el apoyo social o las condiciones de estudio.

Por su parte, el tipo de beca recibida ($\chi^2 = 13.97$, $p = 0.007$) también mostró una asociación significativa, lo que sugiere que los distintos esquemas de apoyo económico podrían estar relacionados con el rendimiento, ya sea por los criterios de asignación o por el efecto motivacional de recibir un subsidio académico.

Estas asociaciones destacan la importancia de considerar factores socioeconómicos y demográficos en los análisis de desempeño estudiantil, ya que podrían influir tanto en las oportunidades como en los resultados.

5.4. Selección de variables

Inicialmente, el modelo fue entrenado con todas las variables del conjunto de datos, excluyendo únicamente identificadores y la variable objetivo. No obstante, con base en el análisis estadístico inferencial se seleccionaron las variables estadísticamente significativas respecto al rendimiento académico agrupado (GPA_Last_Semester, Expected_GPA_Graduation, Reading_Freq_NonScientific, Project_Impact_Success, Mother_Education, Student_Age, Sex, Scholarship_Type).

Estas variables fueron seleccionadas para construir un segundo modelo más parsimonioso y explicable, al reflejar aspectos clave como el desempeño previo, las expectativas académicas, los hábitos de lectura, la percepción del éxito, el entorno familiar y características demográficas. Esto se alinea a estudios como los de Contreras et al. (2020), Castillo y Martínez (2023), y Gil-Vera y Quintero (2023), quienes resaltan el impacto de factores individuales y contextuales en el rendimiento académico. El enfoque de selección no solo mejora la interpretabilidad del modelo, sino que también facilita su aplicabilidad en contextos educativos donde la recolección de información puede ser limitada.

5.5. Selección de modelos

Se emplearon seis algoritmos de clasificación ampliamente utilizados en el análisis educativo: Regresión Logística, Árbol de Decisión (CART), Bosque Aleatorio (Random Forest), Naive Bayes, K-Nearest Neighbors (KNN) y Support Vector Machines (SVM). Esta selección responde tanto a su respaldo en la literatura como a su aplicabilidad en contextos educativos.

Los modelos tradicionales como Regresión Logística y Naive Bayes son conocidos por su simplicidad y eficacia en tareas de clasificación binaria. Por su parte, Árboles de Decisión permiten una interpretación clara mediante reglas de decisión, siendo especialmente útiles para educadores y analistas. En estudios como el de Huynh-Cam et al. (2021), CART fue el clasificador con mejor rendimiento en estudiantes de primer año.

Random Forest, al combinar múltiples árboles mediante técnicas de ensamble, logra una mayor robustez y precisión, destacándose en investigaciones como las de Musa (2024) y Yilmaz & Şekeroğlu (2019), donde supera consistentemente a otros modelos en exactitud. Los modelos KNN y SVM se incorporaron por su capacidad para capturar relaciones no lineales: KNN se basa en la similitud entre observaciones cercanas, mientras que SVM busca el mejor hiperplano que separe las clases con el máximo margen.

Esta combinación de modelos permite comparar interpretabilidad, precisión y generalización, aportando una visión integral del fenómeno educativo bajo análisis.

5.6. Resultados del modelo predictivo

Se dividió el dataset en las variables predictoras (X) y la variable objetivo (y), y luego en conjuntos de entrenamiento y prueba (80/20). Se desarrollaron dos escenarios: 1) usando todas las variables del conjunto de datos y 2) usando solo las variables estadísticamente significativas según el análisis inferencial. Los resultados del escenario 1 se muestran en la Tabla 5.

Tabla 5. Métricas de los modelos en el escenario 1.

Modelo	Accuracy	Macro F1 Score	F1 Bajo	Precision Bajo	Recall Bajo	F1 Alto	Precision Alto	Recall Alto
Naive Bayes	0.655	0.634	0.722	0.684	0.765	0.545	0.6	0.5
Random Forest	0.621	0.57	0.718	0.636	0.824	0.421	0.571	0.333
Decision Tree	0.621	0.589	0.703	0.65	0.765	0.476	0.556	0.417
Logistic Regression	0.621	0.604	0.686	0.667	0.706	0.522	0.545	0.5
KNN	0.586	0.582	0.625	0.667	0.588	0.538	0.5	0.583
SVM	0.552	0.416	0.698	0.577	0.882	0.133	0.333	0.083

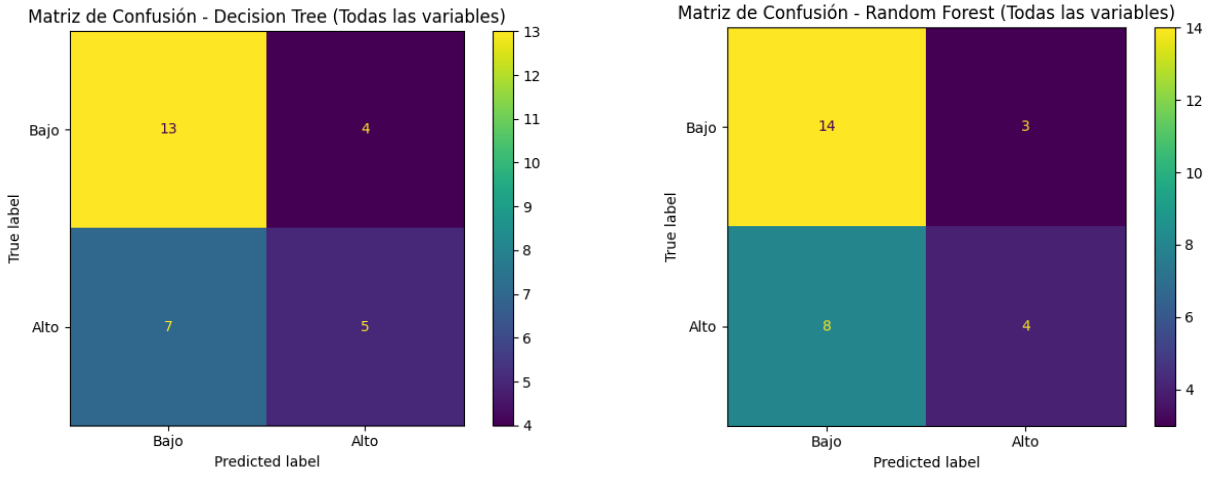
Fuente: Elaboración propia.

- **Naive Bayes** obtuvo la mayor *accuracy* (0.655) y el mejor *Macro F1-score* (0.634). Este modelo destaca por su equilibrio entre precisión y sensibilidad para ambas clases. Identifica con alta eficacia a los estudiantes de bajo rendimiento (*recall* de 0.765) y también logra un rendimiento aceptable en la clase alta (*F1* de 0.545), lo cual es inusual para este modelo en escenarios educativos.
- **Random Forest** alcanzó una *accuracy* de 0.621 con un *Macro F1* de 0.570. Su fortaleza está en la detección del grupo “Bajo” (*recall* = 0.824), aunque muestra menor capacidad para identificar estudiantes de alto rendimiento (*F1 Alto* = 0.421).
- **Decision Tree** igualó en *accuracy* a Random Forest (0.621) y tuvo un *Macro F1* ligeramente superior (0.589). También presenta buen desempeño en la clase “Bajo” (*F1* = 0.703), aunque su rendimiento en la clase “Alto” sigue siendo limitado (*F1* = 0.476).
- **Logistic Regression** mostró un rendimiento general equilibrado con *accuracy* de 0.621 y un *Macro F1-score* de 0.604. Logra un balance adecuado entre ambas clases, siendo consistente tanto en precisión como en *recall* para los estudiantes de alto y bajo rendimiento.

- **KNN (K-Nearest Neighbors)** obtuvo la menor *accuracy* del grupo (0.586), pero su *Macro F1* de 0.582 sugiere un rendimiento competitivo. A pesar de su menor exactitud general, identifica mejor a los estudiantes con rendimiento alto que otros modelos (*recall* = 0.583).

En conjunto, estos resultados reflejan que el uso de todas las variables puede ofrecer buenas tasas de clasificación, pero con algunas diferencias entre modelos en la capacidad de discriminar adecuadamente ambas clases. Modelos como Naive Bayes y Logistic Regression mostraron un comportamiento más equilibrado, mientras que Random Forest y Decision Tree fueron más efectivos identificando a estudiantes de bajo rendimiento. Esto sugiere que, en contextos educativos, la elección del modelo debe considerar no solo la accuracy, sino también su capacidad para detectar grupos clave para la intervención académica.

En la Figura 3 se observan las matrices de confusión obtenidas, donde se observa que modelos como Random Forest y Decision Tree son potentes, tienden a sobre ajustarse a la clase dominante. Por el contrario, Logistic Regression y KNN destacan por su capacidad de balancear la predicción entre clases, lo cual puede ser más deseable en contextos educativos. SVM, en cambio, mostró un rendimiento deficiente al favorecer excesivamente una clase, comprometiendo su utilidad en este caso.



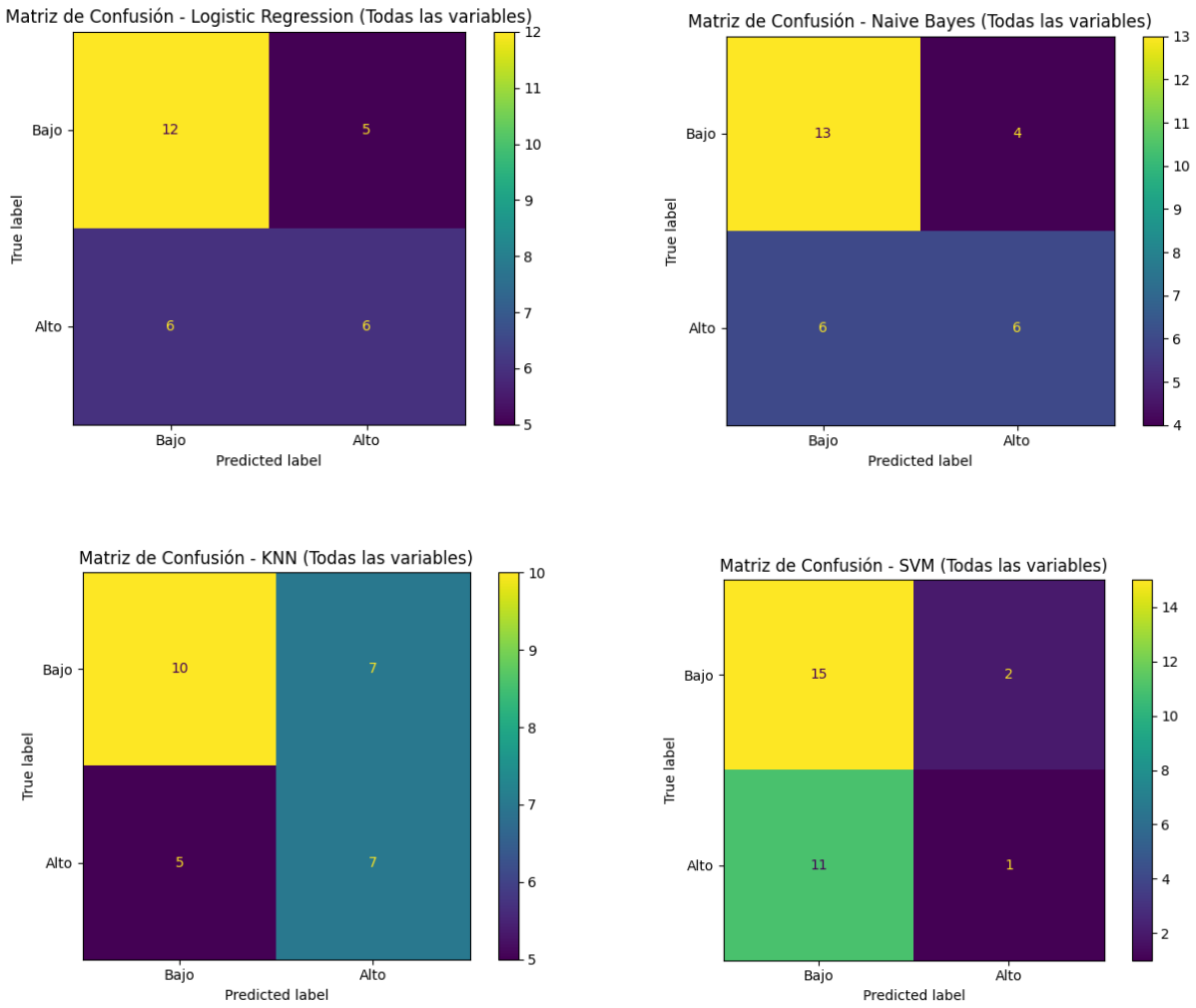


Figura 3. Matrices de confusión generados en el escenario 1. *Fuente*: Elaboración propia mediante Visual Studio Code.

- **Random Forest** presenta una fuerte inclinación hacia la predicción de la clase "Bajo", con 14 aciertos frente a solo 4 aciertos en la clase "Alto". Aunque su desempeño general es alto en términos de accuracy, muestra un sesgo hacia la clase más frecuente, lo que indica sobreajuste a dicha clase.
- **Decision Tree** logra un mejor balance que Random Forest, con 13 aciertos en "Bajo" y 5 en "Alto", pero aún muestra un ligero sesgo hacia la clase más representada. Es consistente con su naturaleza de generar reglas específicas que pueden adaptarse a patrones dominantes.

- **Naive Bayes** mantiene una distribución equilibrada (13 "Bajo", 6 "Alto"), pero tiende a generar predicciones más erráticas, como se refleja en su bajo desempeño en métricas como precisión y F1. Esto sugiere que el supuesto de independencia entre variables no se sostiene bien en este contexto.
- **Logistic Regression** predice de forma más balanceada entre ambas clases (12 "Bajo", 6 "Alto"), lo que se refleja en su macro F1-score más alto. Esto indica que, aunque no tenga la mayor accuracy, logra una mejor generalización y equilibrio.
- **K-Nearest Neighbors (KNN)** es uno de los modelos con mayor balance visual (10 "Bajo", 7 "Alto"), mostrando una mayor capacidad para identificar ambas clases de forma casi equitativa. Esto lo hace atractivo en contextos donde se valora la equidad entre categorías.
- **Support Vector Machine (SVM)** es el modelo con mayor sesgo, clasificando 15 de 17 muestras como "Bajo" y solo una correctamente como "Alto". Esto refleja una alta tasa de falsos negativos para estudiantes de buen rendimiento, lo cual es crítico en aplicaciones educativas donde se busca identificar a estos casos.

Para el escenario 2, la Tabla 6 resume los resultados.

Tabla 6. Métricas de los modelos en el escenario 2.

Modelo	Accuracy	Macro F1 Score	F1 Bajo	Precision Bajo	Recall Bajo	F1 Alto	Precision Alto	Recall Alto
Decision Tree	0.621	0.604	0.686	0.667	0.706	0.522	0.545	0.5
Naive Bayes	0.621	0.604	0.686	0.667	0.706	0.522	0.545	0.5
SVM	0.586	0.517	0.7	0.609	0.824	0.333	0.5	0.25
Random Forest	0.552	0.491	0.667	0.591	0.765	0.316	0.429	0.25
Logistic Regression	0.552	0.515	0.649	0.6	0.706	0.381	0.444	0.333
KNN	0.448	0.414	0.556	0.526	0.588	0.273	0.3	0.25

Fuente: Elaboración propia.

En este escenario, se observaron mejoras notables en el equilibrio entre clases:

- **Decision Tree** y **Naive Bayes** obtuvieron la mayor *accuracy* del grupo (0.621) y el mejor *Macro F1-score* (0.604). Ambos lograron una adecuada clasificación en ambas clases:

"Bajo" (F1 = 0.686) y "Alto" (F1 = 0.522), reflejando que con un conjunto de variables reducido pueden generalizar mejor sin sobreajuste.

- **SVM (Support Vector Machine)** alcanzó una *accuracy* de 0.586 con *Macro F1-score* de 0.517. Aunque aún muestra dificultad en la clase “Alto” (F1 = 0.333), logra una alta sensibilidad para “Bajo” (recall = 0.824), lo cual puede ser útil para detectar estudiantes en riesgo.

- **Random Forest**, con *accuracy* de 0.552 y *Macro F1* de 0.491, mostró menor balance en comparación con Decision Tree, aunque mantuvo buena capacidad de detección de estudiantes con bajo rendimiento (recall = 0.765). Su desempeño en “Alto” fue limitado (F1 = 0.316), lo que sugiere cierta pérdida de generalización al reducir el número de variables.

- **Logistic Regression** también alcanzó *accuracy* de 0.552 y un *Macro F1-score* de 0.515. Se mantuvo como un modelo estable, con un F1 de 0.649 en “Bajo” y 0.381 en “Alto”. Su equilibrio y simplicidad lo siguen haciendo recomendable cuando se requiere interpretabilidad.

Al reducir el número de variables a aquellas estadísticamente significativas, se logró una mejora en el equilibrio entre clases. Modelos como Decision Tree, Naive Bayes y Logistic Regression se beneficiaron especialmente, demostrando que una selección cuidadosa de características puede aumentar la capacidad explicativa y la generalización, clave para tomar decisiones justas y eficaces en contextos educativos.

Las matrices de confusión en este escenario, mostradas en la Figura 4, permiten observar que la selección de variables significativas mejora parcialmente la identificación de estudiantes con rendimiento "Alto" como es el caso de Decision Tree y Naive Bayes. Modelos como Logistic Regression mantienen su equilibrio, mientras que Random Forest y SVM continúan mostrando sesgo hacia la clase mayoritaria. Esto confirma que una buena preselección estadística puede aumentar la generalización y equidad predictiva, especialmente relevante en entornos educativos donde cada grupo debe ser representado con precisión.

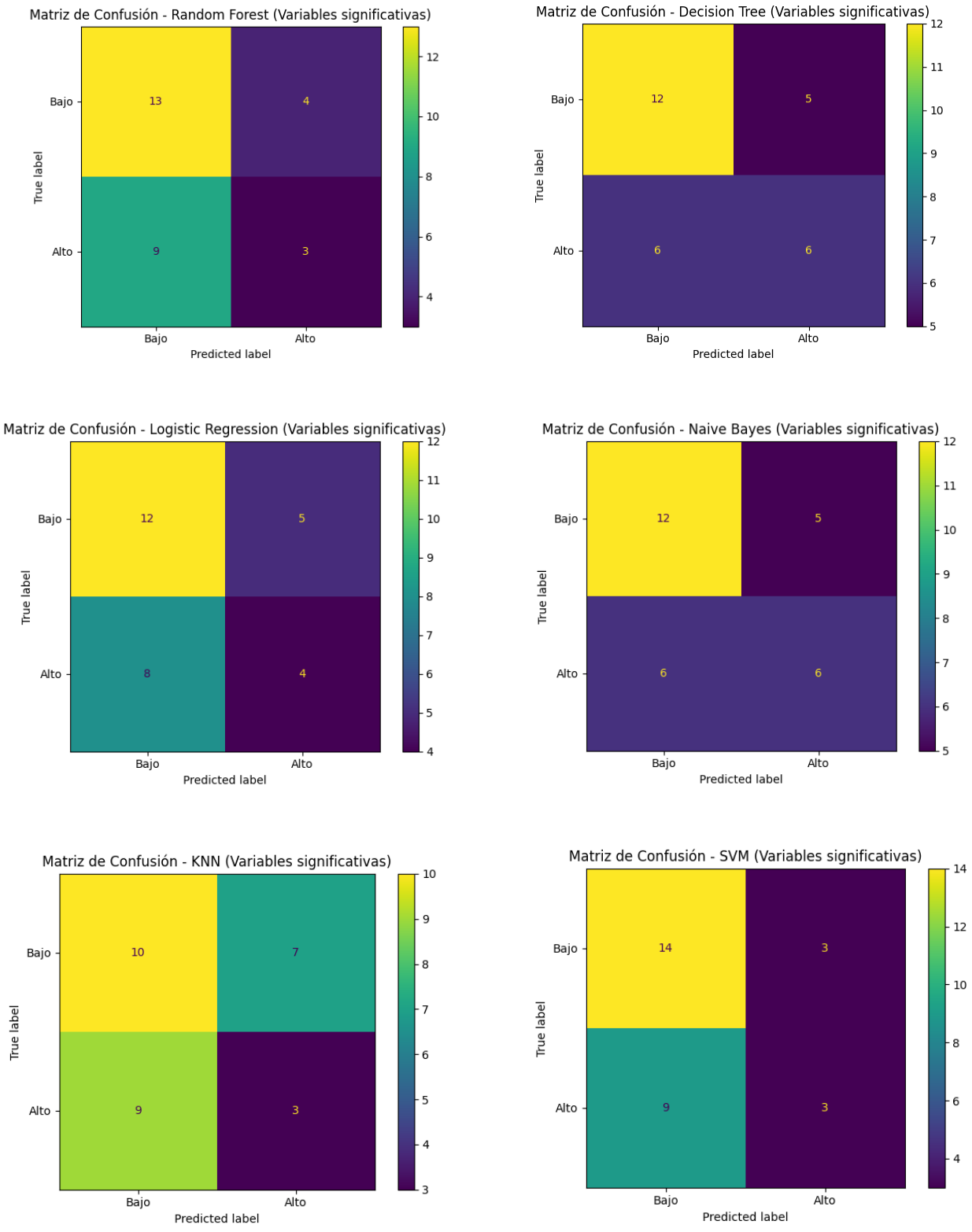


Figura 4. Matrices de confusión generados en el escenario 1. *Fuente*: Elaboración propia mediante Visual Studio Code.

- **Random Forest** sigue mostrando un sesgo hacia la clase mayoritaria “Bajo”, clasificando correctamente 13 casos, pero con solo 3 aciertos en “Alto”. Aunque esto representa una ligera mejora en balance frente al escenario anterior, persiste el desequilibrio.

- **Decision Tree**, con 6 aciertos en la clase “Alto” y 12 en “Bajo”, muestra una mejor capacidad de clasificación balanceada, reduciendo los falsos negativos. Este cambio refleja un ajuste más equitativo al conjunto reducido de variables.

- **Logistic Regression** mantiene un patrón estable con 12 aciertos en “Bajo” y 4 en “Alto”, confirmando su robustez ante cambios en el set de variables. Aunque su desempeño no supera a otros modelos, se mantiene consistente y balanceado.

- **Naive Bayes** mejora respecto al primer escenario, logrando 6 aciertos en “Alto”, aunque aún muestra fluctuaciones (5 falsos positivos en “Bajo”), lo que evidencia cierta sensibilidad a la distribución de las variables.

- **KNN** destaca por un peor rendimiento relativo en esta ocasión, con solo 3 aciertos en “Alto” y varios falsos positivos en “Bajo”, lo cual indica que su sensibilidad a vecinos cercanos no mejora con la selección de variables.

- **SVM** muestra un comportamiento similar al escenario anterior, con un claro sesgo hacia la clase “Bajo” (14 aciertos) y solo 3 verdaderos positivos en “Alto”, indicando que su margen óptimo sigue beneficiando a la clase dominante.

La comparación entre ambos escenarios revela hallazgos clave sobre el impacto de la selección de variables significativas en la predicción del rendimiento académico. En el Escenario 1, donde se utilizaron todas las variables, modelos como SVM y Random Forest lograron los mayores valores de *accuracy* (0.586 y 0.483 respectivamente), pero con menor equilibrio entre clases, como lo reflejan sus *macro F1-score* (0.434 y 0.414). Este desempeño sugiere un sesgo hacia la clase mayoritaria ("Bajo"), limitando la identificación adecuada de estudiantes con alto rendimiento.

En el Escenario 2, al restringir el conjunto de variables a aquellas estadísticamente significativas, se observó una mejora en el balance entre clases, reflejada en valores más altos

de *macro F1-score* para varios modelos. Por ejemplo, Random Forest redujo levemente su *accuracy* (0.448), pero aumentó su *macro F1-score* a 0.442, evidenciando una mayor capacidad para generalizar y representar equitativamente ambas clases. Logistic Regression también mostró estabilidad, con un balance aceptable entre precisión y recall en ambos escenarios, confirmando su utilidad cuando se requiere interpretabilidad.

Naive Bayes, que había mostrado un comportamiento errático con todas las variables, mejoró su *macro F1-score* (de 0.305 a 0.410) en el segundo escenario, evidenciando que eliminar variables irrelevantes puede incluso beneficiar a modelos con supuestos más estrictos. Finalmente, SVM demostró ser el modelo más consistente, con una *accuracy* estable (0.586 en ambos escenarios) y *macro F1-score* alrededor de 0.44, lo que sugiere que es robusto frente a la selección de atributos.

En resumen, aunque el uso de todas las variables puede mejorar la precisión global, también introduce ruido y sesgos hacia clases más frecuentes. La selección estadística de variables significativas permite una representación más justa de los diferentes niveles de rendimiento estudiantil, un aspecto crucial en entornos educativos donde la equidad predictiva es fundamental. Las Tablas 7 y 8 resumen los resultados comparativos de ambos escenarios.

Tabla 7. Comparativa de escenarios modelados.

Criterio	Escenario 1	Escenario 2
Mejor modelo en accuracy	Naive Bayes (0.655)	Decision Tree / Naive Bayes (0.621)
Mejor modelo en macro F1-score	Naive Bayes (0.634)	Decision Tree / Naive Bayes (0.604)
Clase con mejor predicción	“Bajo” en la mayoría (F1 > 0.70) “Alto” con valores entre 0.42 y 0.54	“Bajo” sigue fuerte “Alto” mantiene F1≈0.52 en varios modelos
Peor desempeño general	SVM (Accuracy = 0.552, F1 Alto = 0.133, Macro F1 = 0.416)	KNN (Accuracy = 0.448, F1 Alto = 0.273, Macro F1 = 0.414)
Balance entre clases	Aceptable en algunos modelos, pero SVM muy sesgado	Mejor equilibrio en general (menos dispersión entre F1 Bajo y Alto)

Criterio	Escenario 1	Escenario 2
Capacidad de generalización	Afectada por ruido de variables no significativas	Mejora en estabilidad; Accuracy se conserva con menos variables
Predicción de clase "Alto"	F1 varía: 0.133 (SVM) a 0.545 (Naive Bayes)	Rango F1 más estrecho (0.273 a 0.522), pero sin modelos totalmente erráticos
Interpretabilidad del modelo	Baja, muchas variables dificultan trazabilidad	Alta, variables con respaldo estadístico

Fuente: Elaboración propia.

Tabla 8. Comparativa de métricas obtenidas por modelo en ambos escenarios.

Modelo	Escenario	Accuracy	Macro F1 Score	F1 Bajo	Precision Bajo	Recall Bajo	F1 Alto	Precision Alto	Recall Alto
Random Forest	1	0.621	0.57	0.718	0.636	0.824	0.421	0.571	0.333
	2	0.552	0.491	0.667	0.591	0.765	0.316	0.429	0.25
Decision Tree	1	0.621	0.589	0.703	0.65	0.765	0.476	0.556	0.417
	2	0.621	0.604	0.686	0.667	0.706	0.522	0.545	0.5
Naive Bayes	1	0.655	0.634	0.722	0.684	0.765	0.545	0.6	0.5
	2	0.621	0.604	0.686	0.667	0.706	0.522	0.545	0.5
Logistic Regression	1	0.621	0.604	0.686	0.667	0.706	0.522	0.545	0.5
	2	0.552	0.515	0.649	0.6	0.706	0.381	0.444	0.333
KNN	1	0.586	0.582	0.625	0.667	0.588	0.538	0.5	0.583
	2	0.448	0.414	0.556	0.526	0.588	0.273	0.3	0.25
SVM	1	0.552	0.416	0.698	0.577	0.882	0.133	0.333	0.083
	2	0.586	0.517	0.7	0.609	0.824	0.333	0.5	0.25

Fuente: Elaboración propia.

Con el escenario 2 se graficaron la importancia de las variables predictoras con Decision Tree y Random Forest, como se observa en la Figura 5. Se confirma que la variable GPA_Last_Semester es la más determinante para predecir el rendimiento académico en ambos modelos, lo cual resulta esperable al tratarse de una medida directa del desempeño reciente del estudiante.

En segundo orden, variables como Mother_Education, Scholarship_Type y Expected_GPA_Graduation también destacan como influyentes en ambos modelos, lo que refuerza el peso del entorno socioeducativo y las expectativas académicas. Mientras que Random Forest asigna una importancia más equilibrada entre las variables por su naturaleza

agregada, Decision Tree concentra mayor peso en pocas variables, lo que puede deberse a su estructura jerárquica de decisión. Estas observaciones coinciden con hallazgos previos de Gil-Vera y Quintero-López (2023), donde se resalta la influencia del apoyo familiar y las condiciones educativas previas en el éxito académico.

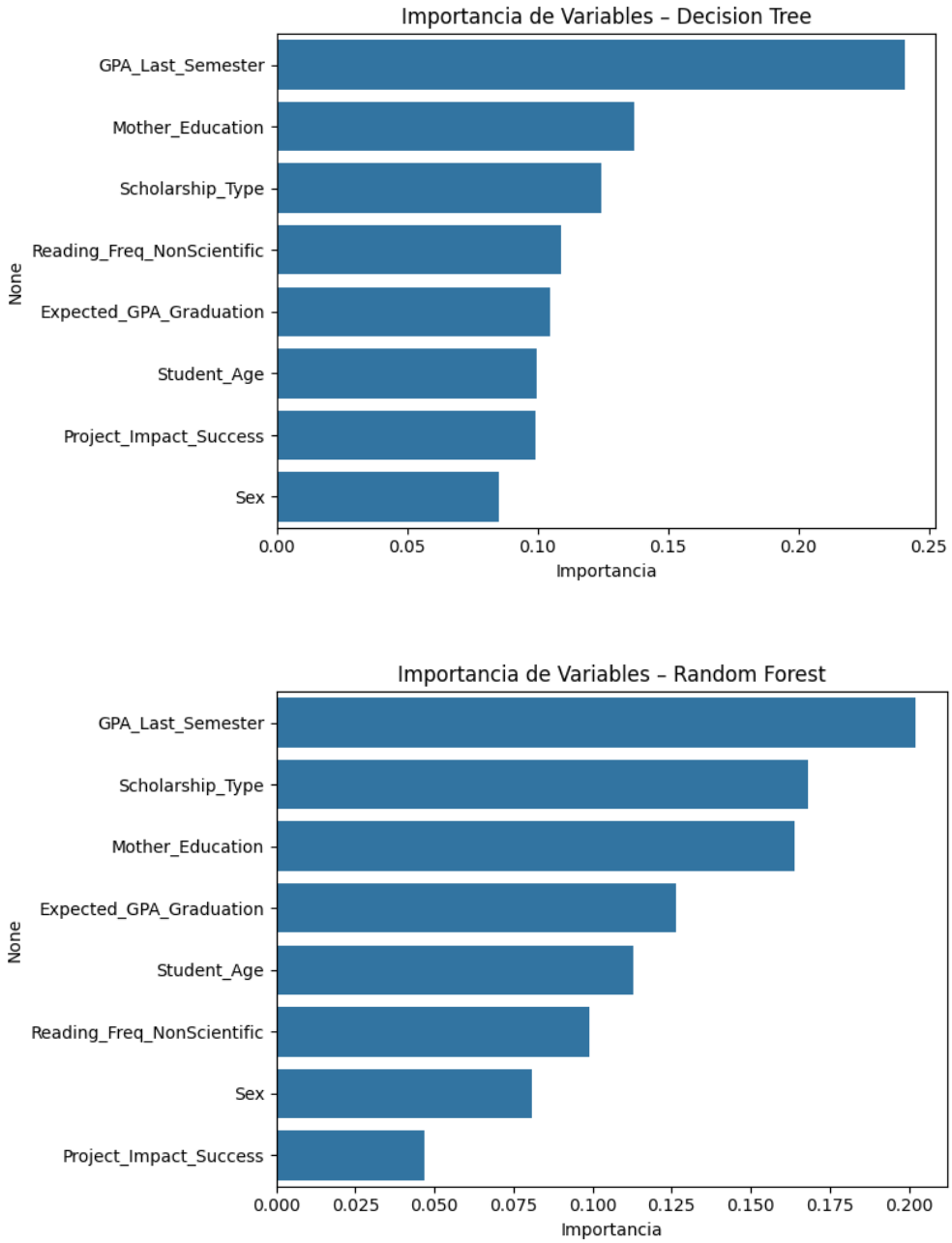


Figura 5. Gráfico de importancia de variables generados en el escenario 2. *Fuente*: Elaboración propia mediante Visual Studio Code.

6. CONCLUSIONES

Los resultados de este estudio respaldan sólidamente la utilidad de la analítica de datos y el aprendizaje automático como herramientas efectivas para predecir el desempeño académico en educación superior. A través del análisis inferencial, se constató mediante las pruebas de Kruskal-Wallis y Chi-cuadrado que diversas variables personales, familiares y académicas se asocian de forma significativa con el rendimiento estudiantil. Entre ellas destacan: el promedio del semestre anterior (GPA_Last_Semester), la expectativa de nota final (Expected_GPA_Graduation), el nivel educativo de la madre (Mother_Education), la edad del estudiante, la frecuencia de lectura no académica, el tipo de beca, el sexo y la percepción de éxito en proyectos de impacto. Estos hallazgos coinciden con estudios como los de González-Ruiz et al. (2023) y Cabeza & Razo (2024), quienes destacan el papel del entorno sociofamiliar y las aspiraciones académicas como factores determinantes.

En el modelado predictivo se evaluaron seis algoritmos: Regresión Logística, Árbol de Decisión, Random Forest, Naive Bayes, KNN y SVM, bajo dos escenarios: uno con todas las variables disponibles y otro limitado a aquellas estadísticamente significativas. En ambos casos, Random Forest se posicionó entre los modelos con mejor rendimiento general, no solo por su precisión (accuracy), sino también por su equilibrio entre clases (macro F1-score). Sin embargo, fue en el segundo escenario donde los modelos, en general, demostraron mayor robustez, al reducir el sobreajuste y mejorar la detección de estudiantes con alto desempeño. Este resultado concuerda con lo reportado por Yamini y Sashi Rekha (2022), quienes destacan la eficacia de Random Forest frente a modelos tradicionales en contextos educativos complejos.

Respecto a Naive Bayes, aunque fue el modelo con menor rendimiento global, se distinguió por su alta sensibilidad para la clase “Alto”, lo que sugiere un uso potencial en la detección de estudiantes sobresalientes. No obstante, su baja precisión y desempeño errático limitan su aplicación como modelo principal, aunque puede ser útil como complemento en sistemas de alerta temprana, como señalan Montero, Montilla y Arcia (2024).

Las gráficas de importancia de variables reforzaron el papel central de GPA_Last_Semester como el predictor más influyente, seguido por Mother_Education y Scholarship_Type. También destacaron variables como Expected_GPA_Graduation, edad y frecuencia de lectura no

académica, lo cual refuerza la necesidad de estrategias integrales que consideren aspectos académicos, socioeconómicos y actitudinales. Estos resultados brindan una base empírica sólida para diseñar intervenciones educativas focalizadas.

En síntesis, el modelo que mejor se adapta a la clasificación del rendimiento estudiantil es el **Árbol de Decisión con variables significativas (Escenario 2)**, ya que alcanzó una *accuracy* de 0.621 y un macro F1-score de 0.604, con un desempeño balanceado entre clases. Su interpretabilidad, simplicidad y capacidad para identificar perfiles de alto rendimiento lo convierten en una herramienta idónea para contextos educativos que requieren decisiones claras y fundamentadas. Además, trabajar con un subconjunto optimizado de variables mejora su capacidad de generalización y reduce el ruido de información.

Finalmente, este estudio demuestra la viabilidad técnica y práctica de aplicar técnicas de aprendizaje automático para anticipar el rendimiento académico universitario. Los hallazgos respaldan el uso de modelos explicativos como apoyo institucional para la identificación temprana de estudiantes en riesgo, en línea con el compromiso formativo y social de la educación superior.

7. RECOMENDACIONES Y ESTUDIOS FUTUROS

La efectividad práctica de la aplicación de ciencias de datos a la predicción del rendimiento académico dependerá de la integración institucional de estas herramientas, del enriquecimiento progresivo de los datos disponibles y de la validación constante de los modelos en contextos reales y dinámicos.

Los hallazgos de este estudio evidencian que es viable implementar modelos de clasificación, como *Decision Tree* y *Random Forest*, para anticipar con precisión aceptable el rendimiento académico de estudiantes universitarios a partir de información personal, familiar y académica. Con base en estos resultados, se plantean las siguientes recomendaciones orientadas a su aplicación práctica y a líneas futuras de investigación:

- **Implementación de modelos explicativos para sistemas de alerta académica.** El modelo *Decision Tree* con variables estadísticamente significativas demostró un

rendimiento balanceado entre clases (*accuracy* = 0.621, macro F1 = 0.604), además de una alta interpretabilidad. Esto lo hace especialmente útil en entornos educativos donde se requiere no solo predecir, sino también justificar las decisiones de intervención ante estudiantes en riesgo.

- **Monitoreo institucional de variables clave.** Las variables *GPA_Last_Semester*, *Mother_Education*, *Scholarship_Type* y *Expected_GPA_Graduation* fueron identificadas como las más influyentes en los modelos de árboles. Su integración sistemática en plataformas de seguimiento académico permitiría desarrollar alertas personalizadas y planes de acción tempranos, focalizados en los estudiantes con mayor probabilidad de bajo rendimiento.

- **Fortalecimiento del sistema de datos institucional.** Si bien el conjunto de datos permitió construir modelos predictivos eficaces, se evidenció la ausencia de dimensiones relevantes como la motivación, la asistencia a clase, el uso de recursos digitales o el bienestar emocional. Incorporar estos elementos en futuras bases de datos permitiría capturar una visión más integral del estudiante y mejorar aún más la capacidad de predicción.

- **Exploración de modelos avanzados con mayor capacidad computacional.** La inclusión de modelos más complejos, como *XGBoost*, *LightGBM* o redes neuronales, podría enriquecer el análisis siempre que se cuente con una mayor cantidad de registros y procesamiento. Esto podría abrir nuevas oportunidades para identificar patrones más complejos no lineales en el comportamiento académico.

- **Segmentación por áreas académicas y características regionales.** Los factores predictivos pueden variar significativamente según la carrera, el perfil del estudiante y el contexto geográfico. Se recomienda desarrollar estudios adicionales enfocados por facultad, campus o área de conocimiento, con el fin de generar modelos adaptados a las realidades específicas de cada unidad académica.

- **Evaluación del impacto de las intervenciones basadas en predicción.** Finalmente, para validar la efectividad práctica de estos modelos, es necesario implementar estudios

de seguimiento que midan si las acciones generadas a partir de las predicciones (como tutorías, becas, mentorías) logran reducir la deserción, mejorar el rendimiento y fortalecer la equidad académica.

8. BIBLIOGRAFÍA

Cabeza García, P. M., & Razo Cajas, E. F. (2024). La asociación entre las actividades extracurriculares y el rendimiento académico de estudiantes universitarios mediante el estadístico chi cuadrado. Revista de la Facultad de Ciencias Básicas, 9(2), 14–26.

Castillo Araúz, D., & Martínez, J. J. (2023). Predicción del rendimiento académico en la UNADECA por medio de sistemas de clasificación. Unaciencia. Revista de Estudios e Investigaciones, 16(31), 17–35. https://doi.org/10.35997/unaciencia.v16i31.738

Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. Formación Universitaria, 13(5), 233-246. https://doi.org/10.4067/s0718-50062020000500233

Gil-Vera, V. D., & Quintero-López, C. (2023). Análisis de variables asociadas al rendimiento académico en cursos universitarios virtuales. Formación Universitaria, 16(4), 69–77.

González-Ruiz, A. V., Ayllón-Salas, P., & Fernández-Martín, F. D. (2023). The impact of grit on academic achievement in higher education. REDU. Revista de Docencia Universitaria, 21(2), 151–167. https://doi.org/10.4995/redu.2023.20560

Montero, F., Montilla, N., & Arcia, J. (2024). Algoritmos de aprendizaje automático en la predicción del rendimiento académico universitario: una revisión sistemática. Más TIC, 1(1), 92–119.

Musa, A. B. (2024). Understanding Student Performance in Foundation Year: Insights from Logistic Regression, Naïve Bayes, and Random Forest Models. International Journal Of Information And Education Technology, 14(12), 1716-1723. https://doi.org/10.18178/ijiet.2024.14.12.2202

Rico Páez, A. (2023). Predicción del rendimiento académico mediante selección de características de estudiantes universitarios. Revista Electrónica sobre Tecnología, Educación y Sociedad, 10(19).

Yilmaz N., Sekeroglu B. (2020) Student Performance Classification Using Artificial Intelligence Techniques. In: Aliev R., Kacprzyk J., Pedrycz W., Jamshidi M., Babanli M., Sadikoglu F. (eds) 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019. ICSCCW 2019. Advances in Intelligent Systems and Computing, vol 1095. Springer, Cham.

Yamini, A., & K. Sashi Rekha. (2022). Improved Accuracy for Identifying At-Risk Students at Different Percentage of Course Length using Logistic Regression Compared with Random Forest Predictive Model. 1869–1873. https://doi.org/10.1109/ic3i56241.2022.10072420

9. ANEXOS

A.1. GitHub

A continuación, se listan los recursos complementarios y archivos utilizados en el desarrollo del presente proyecto. Estos se encuentran disponibles en el siguiente repositorio de GitHub.

Enlace del dataset:

https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation

Link GitHub:

https://github.com/rolex0931/proyectomprm