Proyecto: Predicción del Rendimiento en Matemáticas de Estudiantes mediante Aprendizaje de Máquina

John Alexander Acevedo Serna – jaacevedos@eafit.edu.co
José Manuel Camargo Hoyos – jmcamargoh@eafit.edu.co
Santiago Rodríguez Duque – srodrigu16@eafit.edu.co
EAFIT

Resumen

Este proyecto aplica un proceso completo de aprendizaje automático (Machine Learning) siguiendo la metodología CRISP-DM para abordar un problema educativo: predecir si un estudiante aprobará el examen de matemáticas a partir de características demográficas y académicas. El conjunto de datos utilizado (*Students Performance in Exams*) contiene información de 1000 estudiantes, incluyendo género, nivel educativo de los padres, tipo de almuerzo y preparación previa para exámenes, junto con sus calificaciones en matemáticas, lectura y escritura.

El trabajo inicia con la comprensión del negocio, identificando la necesidad de anticipar el riesgo académico para facilitar intervenciones tempranas que mejoren el rendimiento y reduzcan la deserción escolar. Posteriormente se realiza un análisis exploratorio de datos (EDA) para evaluar la calidad de la información, descubrir patrones y relaciones entre variables, y construir una "data card" que describe el conjunto de datos y posibles sesgos.

1. Introducción

En el ámbito educativo, anticipar el rendimiento de los estudiantes puede ayudar a diseñar estrategias de apoyo y mejorar los resultados académicos. El dataset "*Students Performance in Exams"* provee información demográfica y académica (género, nivel educativo de los padres, tipo de almuerzo, preparación previa, entre otros) junto con las calificaciones en matemáticas, lectura y escritura.

En este proyecto se busca desarrollar un modelo de aprendizaje automático capaz de predecir si un estudiante aprobará matemáticas, identificando factores asociados al éxito o fracaso. Este tipo de análisis permite tomar decisiones preventivas, como asignar tutorías o recursos adicionales a estudiantes en riesgo.

2. Objetivos

2.1. Objetivo General

Desarrollar un modelo de machine learning que permita predecir si un estudiante aprobará el

examen de matemáticas a partir de variables demográficas y académicas (género, nivel educativo de los padres, tipo de almuerzo, preparación previa, entre otras), con el fin de identificar tempranamente a estudiantes en riesgo académico y apoyar la toma de decisiones en intervenciones educativas.

2.2. Objetivos Específicos

- Analizar y caracterizar el conjunto de datos Students Performance in Exams, identificando la calidad de la información, variables relevantes y posibles sesgos.
- Definir y preparar la variable objeto pass_math, que indica si un estudiante aprueba o no matemáticas.
- Realizar un análisis exploratorio de datos (EDA) que permita entender patrones, correlaciones y factores asociados al desempeño.
- Construir un modelo baseline que sirva como punto de referencia inicial.
- Establecer métricas de evaluación adecuadas para clasificación binaria, considerando posibles desbalances de clases.

3. Trabajos Relacionados

Cortez & Silva (2008) - Using Data Mining to Predict Secondary School Student Performance

Este trabajo es uno de los primeros referentes académicos sobre predicción de rendimiento escolar. Los autores recopilaron datos de estudiantes de educación secundaria en Portugal, incluyendo factores como características demográficas, sociales y escolares (edad, sexo, consumo de alcohol, asistencia, entre otros). Aplicaron técnicas de minería de datos y aprendizaje automático como árboles de decisión, redes neuronales y regresión logística para predecir calificaciones finales. Un hallazgo clave fue que variables relacionadas con el entorno familiar y hábitos de estudio tienen fuerte impacto en el desempeño. Este artículo es relevante porque valida el enfoque de usar datos no académicos inmediatos para anticipar el rendimiento.

Dey et al. (2015) - Predicting Students' Performance using Advanced Learning Analytics

En este estudio se aplicaron técnicas de analítica de aprendizaje avanzada (Learning Analytics) para pronosticar el desempeño estudiantil. Los autores integraron datos demográficos, académicos y de participación en cursos en línea para entrenar modelos como máquinas de soporte vectorial (SVM) y árboles de decisión. Resaltan la importancia de seleccionar métricas apropiadas cuando existe desbalance de clases (ej. estudiantes que reprueban son minoría), recomendando métricas como F1-score y ROC-AUC en lugar de solo Accuracy.

Sahai et al. (2020) - Performance Prediction of Students Using Machine Learning Techniques

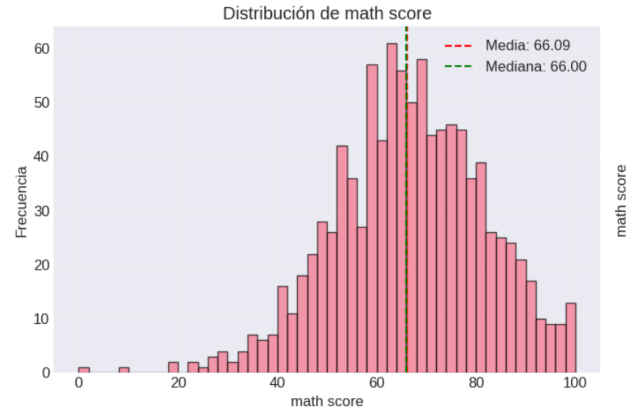
Este trabajo compara varios algoritmos de clasificación (KNN, Random Forest, XGBoost, SVM) para predecir el desempeño de estudiantes, mostrando que los métodos de ensamble logran mayor estabilidad y capacidad de generalización. Además, destacan la importancia de que las instituciones entiendan cómo el modelo llega a sus predicciones; para esto recomiendan herramientas de interpretabilidad, como el análisis de importancia de variables (qué factores pesan más en la predicción) y métodos como SHAP. Esto es relevante para nuestro proyecto porque no solo buscamos un modelo con buen desempeño, sino uno que pueda ser explicado a profesores o coordinadores para apoyar decisiones pedagógicas.

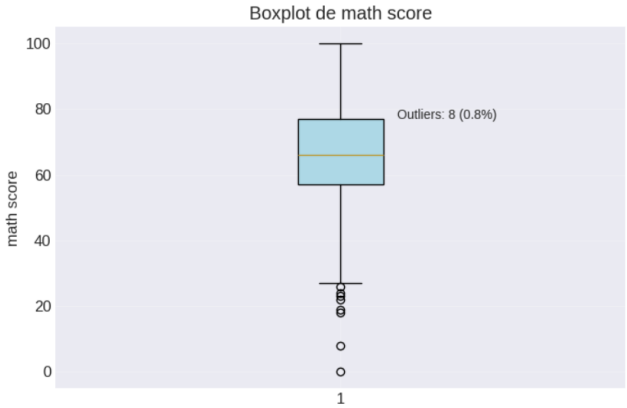
4. Datos y EDA

4.1. Descripción y Calidad de los Datos

El conjunto de datos contiene información de 1000 estudiantes con 8 variables, clasificadas principalmente como categóricas y tres variables numéricas de puntaje.

Distribución de la Variable Objetivo (math score): La distribución de la nota de matemáticas es casi simétrica, con la media y la mediana centradas en un valor similar (aproximadamente 67 puntos). Se identificó una presencia muy baja de outliers (0.8%), concentrados en el rango de notas más bajas.

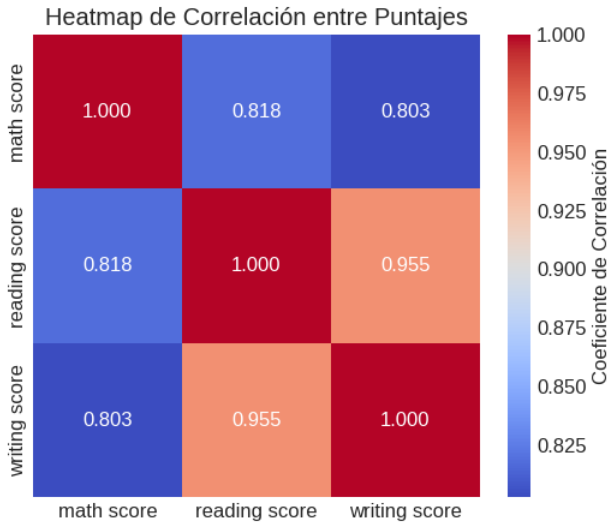




4.2. Hallazgos Clave del Análisis Bivariado

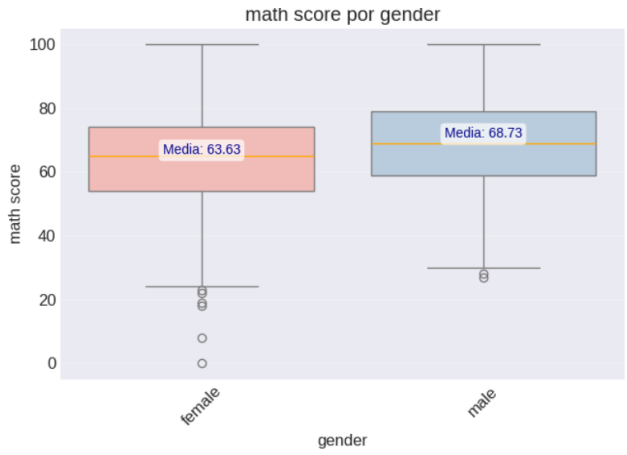
El análisis bivariado se centró en identificar qué variables categóricas introducen una variación significativa en la nota de matemáticas y qué variables numéricas son redundantes.

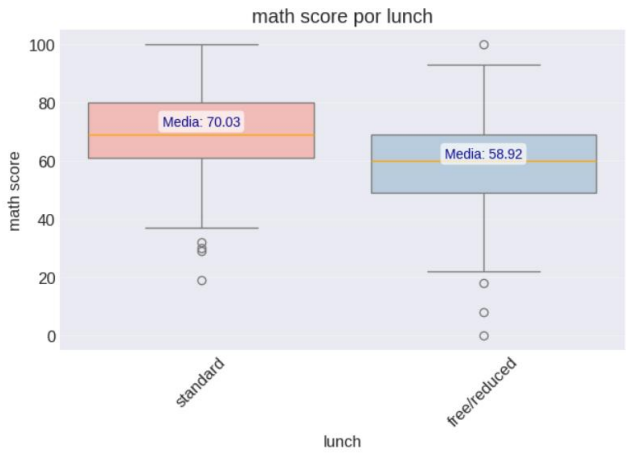
Al analizar la relación entre las variables de rendimiento, se encontró que el puntaje de matemáticas presenta una correlación extremadamente alta (por encima de 0.8) con los resultados de las otras competencias. Esta alta correlación es un indicio de multicolinealidad.

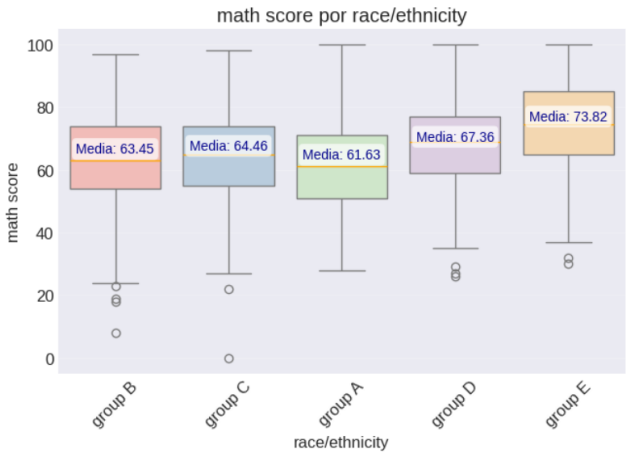


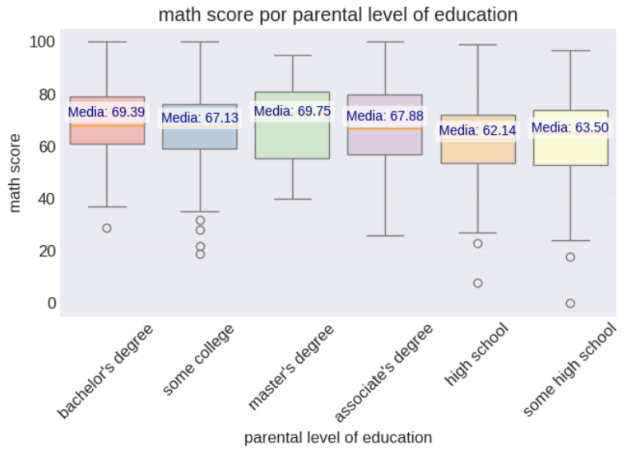
Para el objetivo de predecir el riesgo académico, esta relación no es representativa de la causa del problema, sino del efecto (habilidad académica general), volviendo a las notas unas variables redundantes. Por lo tanto, se justifica descartar estas variables en la fase de preparación de datos, enfocando el modelado en predictores demográficos y comportamentales que sí ofrezcan una relación real con la nota final.

A diferencia de los puntajes académicos redundantes, el análisis de las variables categóricas (demográficas y comportamentales) reveló diferencias estadísticamente significativas en la nota de matemáticas, lo que las convierte en predictores cruciales









5. Conclusiones Preliminares

- La calidad del dataset es alta para lo que se requiere, evidenciada por la distribución simétrica de la nota de matemáticas y la mínima presencia de outliers (0.8%).
- El análisis reveló una multicolinealidad extrema entre las notas académicas (correlación >0.8), lo que las hace redundantes para la predicción y justifica su descarte.
- La estrategia de modelado se centrará en variables categóricas, como gender y test preparation course, las cuales demuestran una influencia estadísticamente significativa y accionable sobre el rendimiento. Esto asegura que el modelo se enfoque en factores que explican la causa del riesgo y faciliten la intervención para alcanzar la meta de aprobación

6. Enlace al Repositorio

https://github.com/JoseCamargo10/machinelearning-project.git

7. Bibliografía

Cortez, P., & Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance*. https://www.researchgate.net/publication/22878

0408_Using_data_mining_to_predict_secondary_school_student_performance

Dey, et al. (2015). *Predicting Students' Performance using Advanced Learning Analytics*. https://www.researchgate.net/publication/315837527_Predicting_Student_Performance_using_Advanced_Learning_Analytics

Sahai, et al. (2020). *Performance Prediction of Students Using Machine Learning Techniques*. https://www.researchgate.net/publication/390154829_Academic_Performance_Prediction_Using_Machine_Learning_Approaches_A_Survey