Entrega 2 - Proyecto

Estudiantes:

Angel Eduardo Morales

Andrés David Pérez Cely

Profesor: Ferney Maldonado López

Introducción

Este documento presenta un análisis del proyecto de datos del programa Talento Tech Bogotá, centrado en tres componentes esenciales del proceso de ciencia de datos: el preprocesamiento, el análisis exploratorio y la metodología de machine learning. El objetivo principal es dejar la información lista para la construcción de un modelo predictivo que permita identificar factores que influyen en el éxito académico de los estudiantes, generando un soporte para tomar mejores decisiones educativas. El análisis completo con el paso a paso desarrollado se encuentra en el siguiente repositorio de github: https://github.com/Davidp1905/Data-Analytics-Project.

5. Preprocesamiento de Datos

El preprocesamiento permitió convertir una base de datos cruda de más de 3.000 estudiantes en un conjunto de datos limpio, estandarizado y útil para análisis posteriores. Las tareas realizadas fueron:

5.1 Limpieza de Datos

- **Valores faltantes:** Se detectaron principalmente en las variables *género* y *tiempo_segundos*. Se imputaron valores numéricos con la mediana para evitar sesgos y en el caso del género se mantuvo la categoría vacía para no eliminar información potencialmente relevante.
- **Duplicados:** Se identificaron registros repetidos y se consideró su eliminación para evitar sobreajuste en modelos futuros.
- **Outliers:** Se utilizó el método IQR, sin hallar valores extremos significativos que afectaran el análisis.

5.2 Feature Engineering

Se crearon variables nuevas con significado estadístico y educativo:

- **promedio_lineas:** resume los conocimientos iniciales del estudiante en áreas técnicas.
- **promedio_areas:** captura las competencias digitales generales.
- **ratio_asistencia:** mide compromiso y participación (asistencia efectiva / horas programadas).

- **exito_academico:** variable objetivo binaria utilizada para clasificación (1: formado, 0: no formado).

Estas características ayudan a reducir ruido, mejorar interpretabilidad y aumentar el poder predictivo.

5.3 Reducción de Dimensionalidad

- Se usó **PCA** para reducir de 49 a 18 dimensiones manteniendo más del 90 % de la varianza total.
- Se eliminaron redundancias y multicolinealidad entre variables.
- Se mejoró el rendimiento potencial de los modelos y la velocidad de entrenamiento.

5.4 Transformación y Codificación

- Aplicación de **estandarización** (media 0, desviación 1) para asegurar que todas las variables numéricas tengan el mismo peso en algoritmos sensibles a escala.
- Verificación de codificación numérica en variables categóricas (por ejemplo, género y nivel educativo).

Gracias a este preprocesamiento, el dataset quedó listo para análisis exploratorio y modelado, asegurando calidad analítica y reproducibilidad.

6. Análisis Exploratorio de Datos (EDA)

El EDA permitió conocer la estructura del dataset, identificar patrones fundamentales y formular hipótesis para su posterior modelado.

6.1 Estadística Descriptiva y Visualización

- Se analizaron distribuciones en variables demográficas como género, estrato y nivel educativo.
- Se observó una **tasa de deserción alta** (70 %) y predominio masculino (65 %).
- La mayoría de estudiantes pertenecen a **estratos bajos** (2 y 3), indicando impacto social del programa.
- Se estudiaron asistencia y puntuaciones iniciales mediante histogramas y boxplots, lo que permitió detectar diferencias importantes entre estudiantes exitosos y no exitosos.

6.2 Correlaciones y Patrones

Se utilizaron diferentes métricas según el tipo de variables:

- **Pearson:** para correlaciones lineales entre variables continuas.
- **Spearman:** para relaciones monotónicas con variables ordinales.
- **Cramer's V y Phi:** para medir asociaciones entre variables categóricas.

Hallazgos importantes:

- La **asistencia** y el **conocimiento inicial** muestran relación positiva con el éxito académico.
- El **género** y el **estrato socioeconómico** no fueron determinantes directos en el desempeño, pero sí en la participación.

6.3 Hipótesis Formuladas

- Mayor nivel educativo previo implica mayor probabilidad de éxito.
- Mejores puntuaciones iniciales predicen mejor desempeño final.
- La asistencia es el principal indicador de permanencia.

Estas hipótesis son la base para el desarrollo del modelo predictivo del punto 7.

7. Metodología de Machine Learning

Se diseñó una metodología para entrenar un modelo que prediga el éxito académico y apoye intervenciones tempranas.

7.1 Selección de Modelos

Se compararán dos algoritmos:

- **Regresión Logística:** interpretable y útil para explicar qué variables influyen en el éxito.
- **Random Forest:** maneja relaciones complejas y reduce riesgo de sobreajuste.

7.2 Tratamiento de Desbalance

Debido al bajo porcentaje de éxito:

- Se utilizarán **class weights** y la técnica **SMOTE** para equilibrar las clases durante el entrenamiento.

7.3 Validación y Métricas

- División de 80 % entrenamiento y 20 % prueba, con **validación cruzada estratificada** para evaluar estabilidad.
- Métricas de desempeño: **F1-Score**, **Precision**, **Recall** y **AUC-ROC**.

Estas métricas permiten medir la capacidad real del modelo para identificar estudiantes en riesgo sin ignorar la clase minoritaria.

Conclusión

Las etapas de preprocesamiento, análisis exploratorio y definición metodológica concluyen con un dataset de alta calidad, patrones relevantes identificados y un plan claro para un modelo predictivo. Como resultado, se cuenta con una base sólida que permitirá en siguientes entregas construir una herramienta capaz de anticipar la deserción y apoyar estrategias de acompañamiento académico basadas en datos.