TRABAJO FINAL

Clasificación del estado académico estudiantil

Andrés Quintero Garcia, Jackson Monk Medina, Juan José Ospina Muriel.

Universidad EIA

PROGRAMACIÓN 2025-1

Profesor: Andrés Quintero Zea

Envigado antioquia

Mayo 2025

Introducción:

El presente proyecto tiene como propósito desarrollar un sistema predictivo basado en técnicas de ciencia de datos para clasificar el estado académico final de estudiantes universitarios. La variable objetivo considera tres posibles resultados: deserción (Dropout), inscripción activa (Enrolled) o graduación (Graduate). Esta clasificación se realiza a partir de un conjunto de variables explicativas que abarcan aspectos personales, académicos, familiares y socioeconómicos del estudiante.

Actualmente en la educación superior, comprender y anticipar el desempeño académico de los estudiantes es muy importante para poder implementar estrategias para poder prevenir la deserción y mejorar la retención estudiantil. Las instituciones educativas podrían beneficiarse mucho de modelos predictivos usados para identificar patrones de riesgo para intervenir de forma temprana.

El enfoque adoptado en este proyecto sigue el flujo completo de un proyecto de ciencia de datos: desde la carga y exploración inicial del conjunto de datos, pasando por el preprocesamiento y limpieza, hasta la construcción, evaluación y comparación de modelos de aprendizaje supervisado. Para ello, se implementaron dos algoritmos de clasificación populares: Random Forest y XGboost, y se aplicaron técnicas como la imputación de valores faltantes, la codificación de variables categóricas, la estandarización de variables numéricas y la optimización de hiper parámetros.

El producto final de este proceso es un modelo predictivo robusto que puede ser integrado en sistemas de apoyo institucional para la toma de decisiones, ofreciendo una herramienta útil para ayudar a mejorar la trayectoria académica de los estudiantes.

Definición del problema:

El problema tratado en este proyecto corresponde a una tarea de clasificación supervisada multiclase, en la cual el objetivo es poder predecir el estado académico final en el que se encuenetra un estudiante universitario. Específicamente, se busca clasificar a cada estudiante en una de las siguientes categorías:

- **Dropout:** estudiante que abandona sus estudios.
- **Enrolled:** estudiante que continúa inscrito en la institución.
- **Graduate:** estudiante que completa exitosamente su programa académico.

La predicción se basa en un conjunto de variables que incluyen información personal (edad, estado civil, género), académica (calificaciones, materias cursadas, evaluaciones), familiar (educación y ocupación de los padres), y contexto económico (tasa de desempleo, inflación, PIB).

Este tipo de clasificación es bastante relevante para instituciones educativas, ya que les permite identificar patrones de riesgo y detectar estudiantes que podrían beneficiarse de una intervención temprana. Así, se pueden tomar decisiones informadas para poder reducir la deserción, mejorar la eficiencia terminal y optimizar los recursos institucionales.

Al formular este problema como un modelo predictivo, se facilita su integración en plataformas de análisis educativo y sistemas de alerta temprana, contribuyendo directamente a mejorar las tasas de permanencia y graduación.

Descripción del DataSet:

El dataset utilizado para este proyecto contiene un total de 4424 registros y 37 columnas, cada una representando distintas características asociadas al perfil del estudiante. Estas variables cubren un amplio espectro de información, que incluye:

- **Datos personales del estudiante**:

estado civil, edad, género y nacionalidad.

- **Historial académico:**

calificaciones de admisión, número de materias inscritas y aprobadas, evaluaciones, tasas de aprobación y desempeño en diferentes semestres.

- **Contexto familiar:**

nivel educativo y ocupación de los padres o tutores.

- **Condiciones económicas y macroeconómicas:**

tasa de desempleo, inflación y PIB en el momento de ingreso.

La variable objetivo del análisis es “Target”, el cual indica el resultado final del estudiante en su trayectoria académica y toma uno de tres posibles valores: Dropout, Enrolled o Graduate.

También es importante recalcar que el conjunto de datos original no contenía valores faltantes, pero para seguir los lineamientos del proyecto, se simuló aleatoriamente un 5% de datos nulos distribuidos entre diferentes celdas del dataset. Esta simulación tuvo como objetivo asegurar el cumplimiento de los requisitos metodológicos y permitir la aplicación de técnicas reales de imputación y limpieza de datos.

Análisis exploratorio:

Durante esta fase se realizó un análisis exploratorio de los datos con el fin de comprender mejor la estructura, distribución y relaciones entre las variables del dataset. Se comenzaron explorando las estadísticas descriptivas de las variables numéricas y categóricas, incluyendo medidas como mediana, media, desviación estándar, valores mínimos y máximos.

Primero, se evaluó la distribución de las variables numéricas usando histogramas, Esto permitió identificar asimetrías, valores extremos y tendencias generales en los datos. Usando una matriz de correlación visualizada con un mapa de calor, se detectaron relaciones fuertes entre variables académicas, como por ejemplo, entre el número de unidades curriculares aprobadas y las calificaciones obtenidas por los estudiantes. Estas correlaciones son indicativas de la coherencia interna de los datos y relevantes para el diseño de modelos predictivos.

Además, se construyó un mapa de calor de los valores faltantes, que permitió verificar visualmente la proporción y distribución de los datos nulos simulados. Esta visualización fue muy importante para planificar las estrategias de imputación implementadas en la siguiente etapa.

El análisis exploratorio proporcionó una visión integral del dataset y permitió formular hipótesis sobre la importancia relativa de ciertas variables en la predicción del estado académico de un estudiante.

Comparación de modelos:

Para determinar cuál de los dos modelos desarrollados ofrecía un mejor desempeño en la tarea de clasificación del estado académico final de los estudiantes, se realizó una comparación utilizando diversas métricas de evaluación.

Los dos modelos Random Forest y XGBoost fueron evaluados sobre el mismo conjunto de datos de prueba. Se analizaron las métricas estándar de clasificación multiclase, como accuracy, precision, recall y f1-score, tanto individualmente por clase como en sus promedios ponderados y macro.

Adicionalmente, se utilizó también el índice de Kappa de Cohen, una métrica que se usa para evaluar la concordancia entre las predicciones del modelo y las etiquetas reales, ajustando por el acuerdo esperado por azar. Esta medida es muy útil en contextos de clasificación multiclase como el presente.

Los resultados mostraron que el modelo XGBoost obtuvo un índice de Kappa superior al de Random Forest, lo que indica una mejor consistencia y fiabilidad en sus predicciones. Además, se construyeron y visualizaron las matrices de confusión para ambos modelos, lo que permitió identificar los patrones de aciertos y errores más más comunes por clase, y

evidenciar que XGBoost tuvo una mejor capacidad para distinguir entre las tres clases del problema (Dropout, Enrolled, Graduate).

Esta evidencia respalda la elección del modelo XGBoost como la mejor solución para el problema planteado.

Conclusión:

A lo largo de este proyecto se aplicaron muchas técnicas completas de ciencia de datos para abordar un problema de clasificación multiclase relacionado con el desempeño académico de estudiantes universitarios. Se desarrollaron dos modelos predictivos —Random Forest y XGBoost— y se implementaron procesos de limpieza, imputación de datos, codificación, estandarización y optimización de hiper parámetros.

Los resultados obtenidos demuestran que el modelo basado en XGBoost fue el más efectivo para predecir el estado académico final de los estudiantes, superando a Random Forest en algunas métricas clave como la precisión general y el índice de concordancia de Kappa. Gracias a su capacidad para manejar relaciones no lineales y su eficiencia computacional, este modelo representa una solución confiable para tareas de clasificación en entoros educativos.

Este modelo puede ser integrado como parte de sistemas de alerta temprana, permitiendo a las instituciones identificar cuáles estudiantes pueden estar en riesgo de deserción y que puedan tomar ciertas medidas preventivas antes de que esto pase. De esta manera, se puede contribuir directamente a mejorar las tasas de permanencia y de graduación de los estudiantes.

En conjunto, este trabajo demuestra el potencial del análisis predictivo para respaldar la toma de decisiones en el sector educativo.