Tarea 2
IN5244-1 Ciencia de datos

Cristopher Urbina Luis Montero Kilian Ehrenreich

Abstract

Este informe se centra en la predicción del re-enrolamiento de estudiantes que se desvincularon del sistema escolar, en el contexto de la política de reportes del MINEDUC implementada el 28 de junio de 2022. La variable de interés, *ListedInDropoutReport*, señala si un estudiante fue incluido en el reporte enviado a las escuelas, y se estudia su relación con la probabilidad de reingreso antes de finalizar el año. Se emplean modelos de clasificación supervisada para predecir la reincorporación de los estudiantes, utilizando un enfoque que incluye técnicas de análisis de datos y algoritmos de machine learning.

1. Definición inicial del problema

El problema de ciencia de datos a abordar consistirá en el desarrollo de un modelo para predecir la reincorporación de estudiantes "desertores" a algún establecimiento del sistema de educación. El tipo de problema a resolver es de clasificación binaria, debido a que el algoritmo deberá predecir un "output" o salida de 2 alternativas: Reincorporado (1) o No Reincorporado (0). Por otra parte, el tipo de aprendizaje del algoritmo es supervisado, pues se cuenta con un vector con las etiquetas que informa sobre los resultados reales observados en el conjunto de datos, lo que además, permitirá evaluar la capacidad predictiva del modelo implementado.

2. Objetivos del proyecto

2.1. Objetivo general

Desarrollar modelos de clasificación binaria para predecir la probabilidad de reincorporación de estudiantes en las escuelas, utilizando datos administrativos escolares relacionados a su desvinculación y reingresos.

2.2. Objetivos específicos

- Determinar las variables clave que afectan la probabilidad de que los estudiantes se reincorporen al sistema educativo.
- Crear y validar modelos predictivos para estimar la reincorporación de los estudiantes al 31 de agosto del año 2022.
- Comparar el desempeño predictivo de los diferentes modelos entrenados y validados para determinar el que alcanza la mejor performance.
- Identificar aquellos atributos que tienen la mayor preponderancia en el desempeño del mejor modelo, de manera de dar interpretabilidad a las predicciones realizadas.
- Ofrecer recomendaciones basadas en los hallazgos del modelo para optimizar las estrategias de revinculación implementadas por las escuelas.

3. Metodología

La metodología para este trabajo se estructurará en varias etapas clave que guiarán el proceso de análisis de datos y el desarrollo de modelos de clasificación binaria. En primer lugar, se realizará una comprensión del problema acompañada de un análisis exploratorio de datos (EDA), donde se identificarán patrones, relaciones y distribuciones de las variables, utilizando visualizaciones, estadísticas descriptivas y análisis de correlación. Además, se procederá a la detección y tratamiento de valores atípicos y datos faltantes para asegurar la calidad del conjunto de datos.

Posteriormente, en la etapa de preparación de los datos, se transformarán las variables categóricas y numéricas para adecuarlas al modelado. Esto incluye la codificación de variables categóricas y la normalización o estandarización de las variables numéricas. Además, se dividirá el conjunto de datos en conjuntos de entrenamiento y prueba para poder validar el rendimiento de los modelos.

En cuanto a la selección de variables y la generación de características, se identificarán las variables más relevantes que podrían influir en el re-enrolamiento de los estudiantes. También se crearán nuevas características derivadas que mejoren el rendimiento predictivo del modelo, lo que se conoce como feature engineering.

A continuación, en la etapa de desarrollo de modelos de

clasificación binaria, se establecerá un modelo cuyo desempeño sirva como línea base para optimizar los modelos posteriores. Se entrenarán diferentes algoritmos de clasificación binaria, tales como random forest, algoritmos de boosting y redes neuronales MLP. Se ajustarán hiperparámetros para mejorar el rendimiento del modelo y se aplicarán técnicas de muestreo para manejar el desbalance de clases en el vector objetivo, tales como el sobremuestreo de la clase minoritaria o el submuestreo de la clase mayoritaria.

Para la evaluación del modelo, se medirán el rendimiento de los modelos utilizando métricas como la precisión, el recall, el F1-score y el área bajo la curva ROC (AUC-ROC). A partir de estas métricas, se compararán los modelos para seleccionar el que tenga el mejor desempeño.

Finalmente, se llevará a cabo la interpretación y validación de los resultados, donde se analizará la importancia de las características para entender qué factores clave influyen en el re-enrolamiento. Se validará el modelo seleccionado en el conjunto de prueba para asegurar que generaliza correctamente a nuevos datos.

En la última etapa, de despliegue y recomendaciones, se implementará el modelo final para realizar predicciones en nuevos datos de estudiantes desvinculados. Además, se elaborarán recomendaciones para las escuelas y el MINEDUC basadas en los resultados del análisis y el rendimiento del modelo, con el fin de mejorar las estrategias de revinculación de los estudiantes.

4. Análisis exploratorio de datos

El análisis exploratorio mostró que faltan 100 valores en la columna asociada a la variable *GPAIn2021*. Para tratar adecuadamente estos valores faltantes, primero queremos investigar si son MCAR (missing completely at random) o no. Para ello, realizamos la prueba MCAR de Little mencionado en la cátedra. El resultado (p=0.0) de esta prueba indica que los valores faltantes no son MCAR, es decir, no faltan totalmente al azar. Esto implica que los valores faltantes son MAR (missing at random) or NMAR (not missing at random). Dado que faltan menos del 2% de los valores, se puede recurrir a la supresión de casos en lugar de imputaciones complicadas. No existen duplicados en este conjunto de datos.

En relación con la distribución de las características, el conjunto de datos estudiado incluye niños y niñas menores de 18 años, cuyas edades se concentran mayoritariamente entre los 12 y 16 años. La asistencia escolar es, en su mayoría, cercana al 100%, con un pico de frecuencia alrededor del 80%. Según la distribución de los días de la semana en que los estudiantes abandonaron sus estudios, el lunes presenta la mayor frecuencia de deserciones, mientras que hay una disminución progresiva hacia el viernes, que es el día con la menor frecuencia de abandono.

La variable objetivo presenta un desbalance, donde la proporción de estudiantes revinculados al 30 de agosto de 2022 en comparación con los que no se revincularon es de aproximadamente 3:1. En cuanto al género, la distribución es uniforme. El promedio de las notas es bastante alto, lo que sugiere que la mayoría de los estudiantes no abandonan la escuela debido a un bajo rendimiento académico. Este hecho se relaciona con los estudiantes que se graduaron en 2021, quienes son mayoría en comparación con los que no lo hicieron.

Por otro lado, las personas listadas en el reporte de deserción son una minoría en comparación con las que no lo están, lo que a primera vista sugiere que estar en el reporte no influye significativamente en la revinculación al 30 de agosto. Los estudiantes migrantes representan una proporción muy baja en relación con el total. En cuanto a las escuelas, la cantidad de estudiantes es similar entre las escuelas municipales y no municipales. Los estudiantes prioritarios que abandonaron los estudios son casi el doble en comparación con los no prioritarios. Finalmente, parece haber más estudiantes que desertaron después del 30 de mayo en comparación con los que lo hicieron antes de esa fecha.

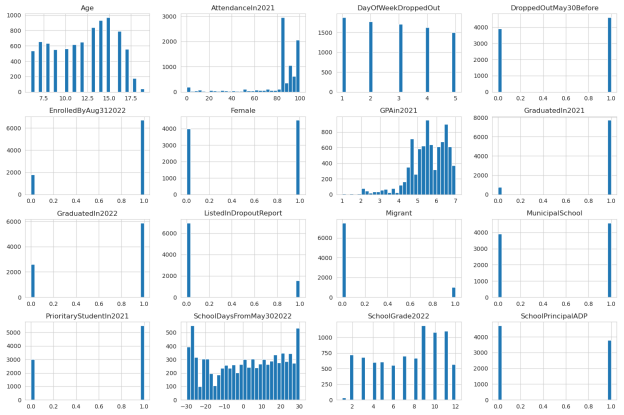

Figure 1. Histograma de las variables númericas

De los estudiantes listados en el reporte, existe una distribución desequilibrada entre aquellos que se revinculan y los que no. En cambio, los estudiantes prioritarios muestran una mayor proporción de revinculación en comparación con los no prioritarios.

La matriz de correlación revela que la variable *EnrolledByAug312022* está fuertemente correlacionada con *GraduatedIn2022* (0.65), lo que sugiere que los estudiantes graduados en 2022 tienen mayor probabilidad de revinculación. También se observa una correlación positiva, aunque más débil, con *GPAin2021* (0.21) y *GraduatedIn2021* (0.15).

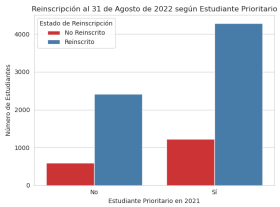

Figure 2. Estudiantes prioritarios que se revincularon o no al 31 de agosto de 2022

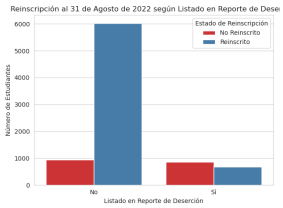

Figure 3. Estudiantes listados en el reporte que se revincularon al 31 de agosto de 2022

Por otro lado, *ListedInDropoutReport* muestra una correlación negativa de -0.40, indicando que los estudiantes listados en el reporte de deserción tienen menor probabilidad de revinculación. Las variables *Age* (-0.23) y *SchoolGrade2022* (-0.19) presentan correlaciones negativas moderadas, mientras que *PrioritaryStudentIn2021* muestra una correlación casi nula (0.05), indicando poca relación con la revinculación.

Estas correlaciones son clave para el desarrollo de un modelo de machine learning. Las variables *GraduatedIn2022* y *ListedInDropoutReport* deben considerarse como predictores importantes en el proceso de modelado. La fuerte correlación positiva de *GraduatedIn2022* sugiere que este es un factor clave en la predicción, mientras que la correlación negativa de *ListedInDropoutReport* indica que estar en el reporte de deserción reduce significativamente la probabilidad de re-enrolamiento.

Al seleccionar variables para entrenar el modelo, es importante centrarse en aquellas que tienen una correlación relevante con la variable objetivo, evitando incluir variables con poca relación, como *PrioritaryStudentIn2021*. Además, dado el desbalance en la variable objetivo, será crucial aplicar técnicas de muestreo o ponderación de clases para garantizar que el modelo no favorezca la clase mayoritaria, y se debe ajustar cuidadosamente la importancia de las variables dentro del modelo para mejorar tanto su precisión como su interpretabilidad.

Es importante señalar que no se incluirá la variable *GraduatedIn2022* en los modelos predictivos debido al riesgo de **data leakage**. Esta variable está relacionada con un evento que ocurre después del periodo que se busca predecir, es decir, la revinculación de los estudiantes al 31 de agosto de 2022. Incluir dicha variable introduciría información que no estaría disponible en el momento de realizar predicciones reales, lo que podría distorsionar los resultados y llevar a un modelo con rendimiento inflado. Por lo tanto, *GraduatedIn2022* será excluida del proceso de modelado para evitar este problema y asegurar que el modelo generalice correctamente a nuevos datos.

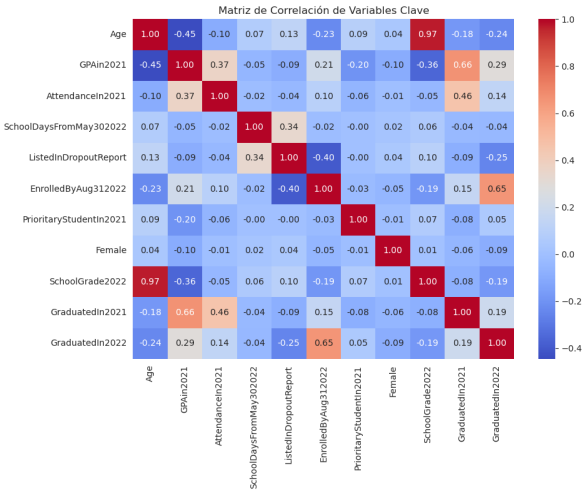

Figure 4. Matriz de correlaciones

4.1. Análisis de componentes principales

Como una forma de explorar posibles patrones de asociación entre las variables del dataset, se realizó un análisis de componentes principales (PCA). Una primera aproximación que maximiza la varianza explicada por el procedimiento, indica la extracción de 11 componentes.

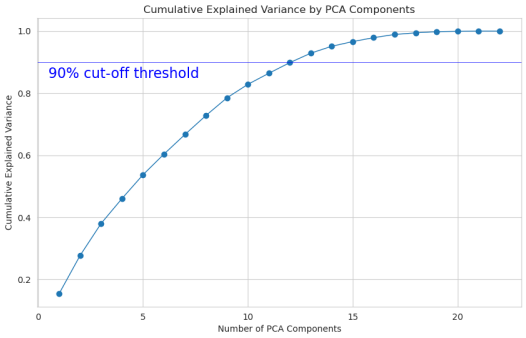

Figure 5. Varianza acumulada explicada

Por otra parte, una reducción que prioriza en número de potenciales componentes la contribución marginal de cada componente extraído, sugiere una extracción de entre 6 y 7 componentes.

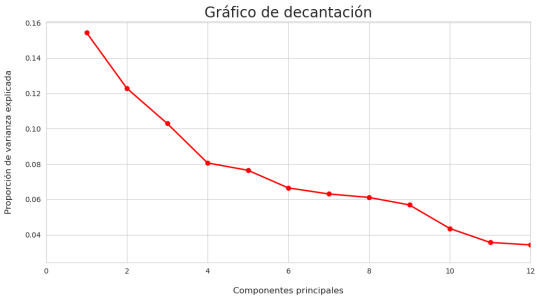

Figure 6. Proporción de varianza explicada en función de N° de componentes

Por otra parte, un análisis de las cargas de los 6 primeros componentes muestra las siguientes relaciones: El primer componente tiene fuertes cargas asociadas a las variables *MunicipalSchool* (0.68) y *SchoolPrincipalADP* (0.67). El segundo componente tiene altas cargas con las variables *DroppedOutMay30Before* (0.74), *SchoolDaysFromMay302022* (0.43) y *ListedInDropoutReport* (0.38). El tercer componente presenta altas cargas con las variables *Female* (0.55), *Migrant* (0.26) y *PrioritaryStudentIn2021* (-0.45). A su vez, el cuarto componente tiene altas cargas con *PrioritaryStudentIn2021* (0.74), *Female* (0.46) y *Migrant* (-0.27). El quinto componente presenta altas cargas con *Female* (0.65), *GraduatedIn2022* (0.49) y *DayOfWeekDroppedOut* –lunes (0.29). Por último, el sexto componente presenta las cargas más altas asociadas con *DayOfWeekDroppedOut* –martes (0.67), *GraduatedIn2022* (0.22) y *DayOfWeekDroppedOut* –lunes (-0.61).

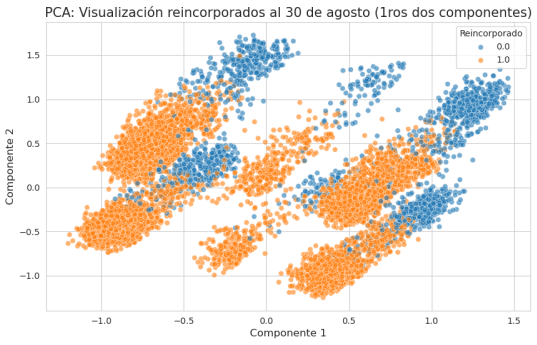

Figure 7. Visualización de componentes principales de los reincorporados al 31 de agosto de 2022

5. Propuesta inicial: Modelos de clasificación binaria

En relación con los tipos de atributos que predominan en el dataset (categóricos-binarios por sobre los numéricos), se estima en un primer momento, utilizar algoritmos basados en árboles de decisión (como RandomForest, GradientBoost y similares). También se contempla el diseño y entrenamiento de un modelo MLP para comparar su desempeño con el de los algoritmos basados en árboles. De esta manera se puede evaluar la calidad de las características obtenidas mediante feature engineering.

6. Conclusión preliminar

El análisis exploratorio y el uso de técnicas como PCA han permitido identificar relaciones clave entre las variables del conjunto de datos y su capacidad predictiva respecto al re-enrolamiento de estudiantes. Las variables como *MunicipalSchool*, *PrioritaryStudentIn2021* y *Female* han demostrado tener una influencia significativa en la predicción, mientras que variables como *GraduatedIn2022* y *ListedInDropoutReport* muestran una fuerte correlación, aunque *GraduatedIn2022* será excluida por riesgo de **data leakage**.

Además, el análisis de componentes principales sugiere que se pueden reducir las dimensiones a 6 o 7 componentes sin perder gran parte de la varianza explicada, lo que simplifica el modelo sin comprometer el rendimiento. Las cargas de los primeros componentes indican que variables relacionadas con el tipo de escuela, el género y el estatus de estudiante prioritario son los principales impulsores en la variabilidad de los datos.

Finalmente, el desbalance de clases observado en la variable objetivo (EnrolledByAug312022) resalta la importancia de aplicar técnicas de muestreo o ponderación de clases para asegurar un modelo equilibrado. Con base en los hallazgos preliminares, se continuará con el desarrollo y la validación de modelos predictivos que puedan ofrecer un enfoque robusto para predecir la reincorporación de estudiantes, con un enfoque especial en la interpretación de los resultados y en la generación de recomendaciones útiles para las escuelas y el MINEDUC.