Detección de riesgo académico en estudiantes universitarios

Ana Cortazzo
Maestría en ciencia de datos y aprendizaje automático
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
ana.cortazzo@fing.edu.uy
Luciana Olazábal
Maestría en Ingieniería Ambiental
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
lolazabal@fing.edu.uy

Resumen

El objetivo de este proyecto es contribuir a la reducción del abandono y el fracaso académico en la educación universitaria mediante el uso de técnicas de aprendizaje automático para identificar estudiantes en riesgo de abandono en una etapa temprana de su trayectoria académica. Para lograr esto, se utilizará un conjunto de datos que contiene información recopilada al momento de la inscripción de los estudiantes, incluyendo su trayectoria académica, características demográficas y factores socioeconómicos. El problema se plantea como una tarea de clasificación de tres categorías: abandono, matriculado y graduado

I. Descripción del conjunto de datos y análisis exploratorio

Se seleccionó un conjunto de datos de *UCI Machine Learning Repository* sobre predicción del suceso académico de los estudiantes, los datos se puede encontrar en este enlace. Los datos cuentan con 4424 instancias y 37 atributos con información demográfica, socio-económica, sobre trayectoria de los estudiantes, datos sobre los cursos del primer y segundo semestre, y la columna `Target` que indica si el estudiante se graduó, abandonó o sigue matriculado al finalizar la duración normal del curso.

Se realizó un análisis exploratorio con el objetivo de determinar posibles problemas de calidad del conjunto de datos:

- Detección de valores nulos o faltantes (NA).
- Detección de un posible desequilibrio de clases en la variable `Target`, con una proporción desigual entre las categorías de abandono, matriculado y graduado.

Los datos no presentan valores NA, hay 7 atributos de tipo continuo (`float64`), 29 de tipo discreto (`int64`) y uno de tipo categórico (`object`), en el Anexo se presenta la Table I con el detalle de las variables y tipo de cada una. Se observa un desbalance de clase en la variable `Target`, que presenta 2209 Graduate, 1421 Dropout y 794 Enrolled. Se realizaron algunas visualizaciones exploratorias para entender la distribución de los datos, se muestran en la Figure 1.

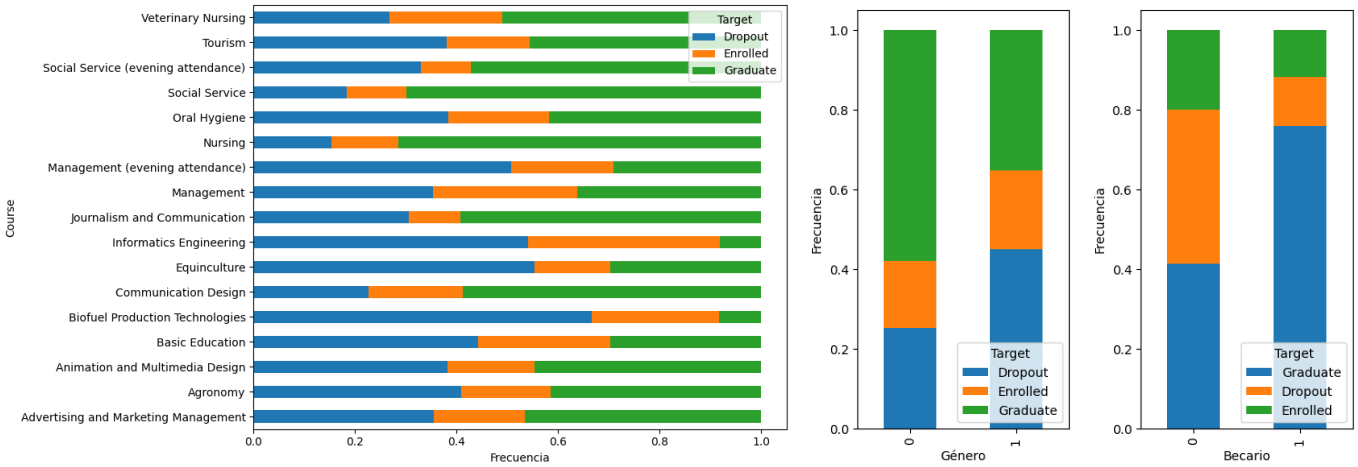
II. Problema

¿Es posible utilizar técnicas de aprendizaje automático para identificar de manera precisa y temprana a los estudiantes en riesgo de abandono académico en la educación universitaria conociendo su situación socio-económica, demográfica y académica al finalizar el segundo semestre? Es decir, predecir si el estudiante culmina la carrera (se gradúa) o abandona. En caso que la predicción detecte que el estudiante es propenso a abandonar, se pueden implementar estrategias dentro de la Universidad para brindarle acompañamiento al estudiante durante su trayectoria académica, con el fin que consiga graduarse.

III. Metodología

Para preprocesar el conjunto de datos, se proponen varios pasos para asegurar que los datos se encuentren en un formato adecuado para el análisis. Estos pasos incluyen:

1. **Limpieza de datos:** El conjunto de datos fue cuidadosamente inspeccionado en busca de valores faltantes, inconsistencias y errores. El análisis exploratorio mostró que no hay registros faltantes en los datos.
2. **Selección de características:** Para mejorar la eficiencia y efectividad de los modelos predictivos, se pueden aplicar técnicas de selección de características para identificar las características más relevantes para predecir la deserción estudiantil. Este paso ayuda a reducir la dimensionalidad y eliminar atributos irrelevantes o redundantes. Una técnica a ser aplicada puede ser a partir de la matriz de correlación entre las variables y seleccionar aquellas variables que tengan



(a) Carreras (b) Género (0-Mujer/1-Varón) (c) Becario (0-No/1-Sí)

Figura 1: Distribución de la variable `Target` en relación a carreras, género y becas

una correlación significativa, determinada mediante el coeficiente de correlación de Pearson, la Figure 2 muestra la matriz de correlación para estos datos (ver Anexo).

3. **Transformación de características:** Es posible que algunas características requieran transformación para hacerlas adecuadas para los modelos de ML. Esto podría implicar normalización o estandarización de variables numéricas, también se puede transformar la variable `Target` en numérica, asignando un número a cada categoría.
4. **División de datos:** El conjunto de datos se debe dividir en conjuntos de entrenamiento y prueba (`train/test`). El conjunto de entrenamiento se utiliza para entrenar los modelos de ML, mientras que el conjunto de prueba se utiliza para evaluar su rendimiento y capacidad de generalización. Se propone una división 80 % para entrenamiento y 20 % para prueba.
5. **Equilibrio de la distribución de clases:** Como se mencionó anteriormente, la distribución de clases de estudiantes que abandonan y no abandonan está desequilibrada, lo que puede afectar negativamente el rendimiento de algunos modelos. Se pueden aplicar técnicas como sobremuestreo o submuestreo para abordar este problema y lograr una distribución de clases más equilibrada.

Al realizar estos pasos de pre-procesamiento, el conjunto de datos se transforma en un formato adecuado para entrenar y evaluar los modelos de aprendizaje automático para predecir la deserción estudiantil. Es necesario seleccionar un modelo de aprendizaje automático apropiado para abordar la tarea de clasificación. Se propone la evaluación de varios modelos disponibles en la biblioteca `scikit-learn`:

- **Random Forest**: Ensemble de árboles de decisión que pueden manejar desbalance de clases. Pueden capturar relaciones no lineales entre las variables y la variable objetivo.
- **Máquinas de Vectores de Soporte (SVM)**: Las SVM son modelos que buscan encontrar un hiperplano que separe las clases de manera óptima en un espacio de alta dimensión. Pueden ser efectivas para problemas de clasificación desbalanceada.
- **Regresión Logística Multinomial**: Es un modelo lineal que se utiliza comúnmente en problemas de clasificación con más de dos clases. Puede manejar el desbalance de clases ajustando los pesos de las muestras o utilizando técnicas como la regularización.

Se propone la utilización de la técnica de validación cruzada para estimar su rendimiento. Se realizará una búsqueda de hiperparámetros para encontrar la configuración óptima del modelo. La evaluación del rendimiento del modelo se realizará utilizando métricas de evaluación apropiadas para problemas de clasificación multiclase, como accuracy, precisión, recall, y F1-score. También se analizará la matriz de confusión para entender el funcionamiento del modelo. Se podrían utilizar gráficos de barras, diagramas de dispersión, entre otros, para comparar el rendimiento de los modelos y resaltar las características más relevantes de cada uno.

ANEXOS

Tabla I: Descripción de las variables

	Nombre de la variable	Tipo
Datos demográficos	Marital status	discreta
	Nationality	discreta
	Displaced	binaria
	Gender	binaria
	Age at enrollment	discreta
	International	binaria
Datos socioeconómicos	Mother's qualification	discreta
	Father's qualification	discreta
	Mother's occupation	discreta
	Father's occupation	discreta
	Educational special needs	binaria
	Debtor	binaria
	Tuition fees up to date	binaria
	Scholarship holder	binaria
Datos macroeconómicos	Unemployment rate	continua
	Inflation rate	continua
	GDP	continua
Datos al momento de la inscripción	Application mode	discreta
	Application order	discreta
	Course	discreta
	Daytime/evening attendance	binaria
	Previous qualification	discreta
Datos al finalizar el 1° semestre	Curricular units 1st sem (credited)	discreta
	Curricular units 1st sem (enrolled)	discreta
	Curricular units 1st sem (evaluations)	discreta
	Curricular units 1st sem (approved)	discreta
	Curricular units 1st sem (grade)	continua
	Curricular units 1st sem (without evaluations)	discreta
Datos al finalizar el 2° semestre	Curricular units 2nd sem (credited)	discreta
	Curricular units 2nd sem (enrolled)	discreta
	Curricular units 2nd sem (evaluations)	discreta
	Curricular units 2nd sem (approved)	discreta
	Curricular units 2nd sem (grade)	continua
	Curricular units 2nd sem (without evaluations)	discreta
Variable objetivo	Target	categórica

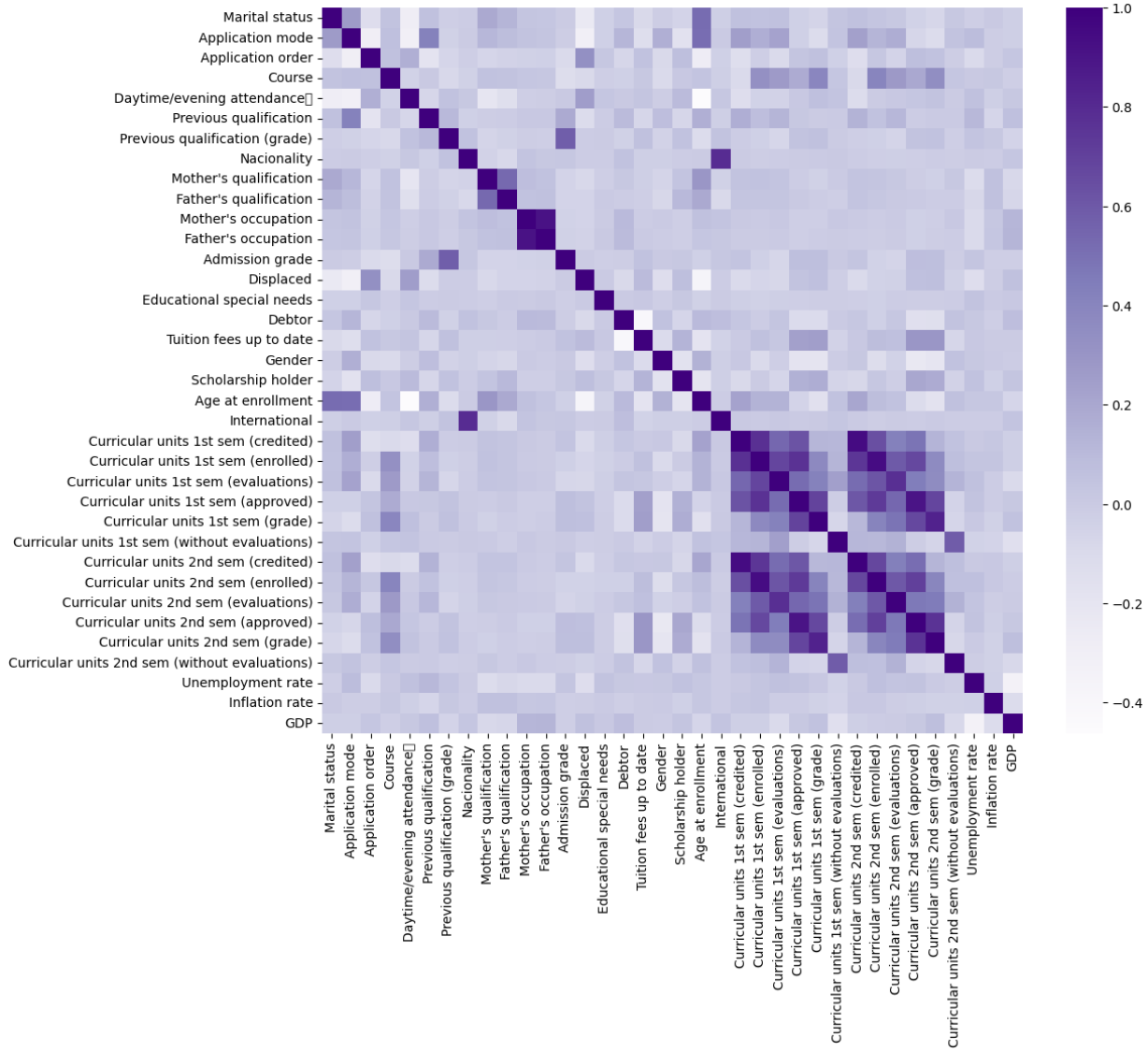


Figura 2: Matriz de correlación de las variables numéricas (coeficiente de Pearson). El coeficiente de correlación de Pearson, que se encuentra en el rango de -1 a 1, cuantifica la fuerza y la dirección de la relación lineal entre dos variables. Un valor de -1 indica una correlación negativa perfecta, donde las variables se mueven en direcciones opuestas de manera perfectamente predecible. Un valor de 0 indica una ausencia de correlación, lo que significa que no hay una relación lineal aparente entre las variables. Un valor de 1 indica una correlación positiva perfecta, donde las variables se mueven en la misma dirección de manera perfectamente predecible.