Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Análisis de la influencia de factores exógenos en la deserción de estudiantes en el Ciclo Básico Común de la Universidad de Buenos Aires

Edgardo Palazzo
epalazzo@cbc.uba.ar

Buenos Aires, Argentina
Julio 2023

Índice

Resumen

En este trabajo se analizan las bases de datos de estudiantes de la cátedra de física del Ciclo Básico Común de la Universidad de Buenos Aires [1], que contienen un historial de calificaciones de cada estudiante en las materias de la cátedra, desde el año 2011 hasta el año 2019. El objetivo es determinar si el desgranamiento de los cursos está influenciado por las variables contenidas en los datos existentes, que en gran parte son consideradas como variables exógenas respecto al estudiante.

Calculando precisiones medias (AP) y analizando importancias de variables de modelos clasificadores entrenados con XGBoost, que predicen si un estudiante abandona o no después de rendir el primer examen parcial, se detectó una leve influencia de las variables exógenas en el desgranamiento de cursos.

1. Introducción

El enfoque general que se observa en los trabajos sobre deserción, o sobre rendimiento estudiantil, donde se aplican técnicas de minería de datos, es el análisis del problema incluyendo toda la información externa a la institución que se pueda conseguir, consideradas como variables endógenas respecto de los estudiantes, especialmente las características socioeconómicas de los mismos [1]. Los factores más comunes en este tipo de trabajos están relacionados a las notas previas y desempeño en clase del estudiante, a sus actividades de estudio en línea, a variables demográficas del estudiante, y a su información social [2]. Sin embargo, como propone Vincent Tinto, las instituciones educativas deben reconocer que la deserción, o un avance lento, también puede tener raíces en el entorno en el cual se les pide a los estudiantes que aprendan [3]. Este trabajo apunta precisamente en esa dirección, realizando un análisis que incluya variables relacionadas con el entorno en el que cada estudiante cursa las materias, para determinar si hay influencia, y en qué medida, de estas variables sobre el nivel de deserción.

Para responder la pregunta planteada se propone entrenar modelos clasificadores para realizar predicciones sobre desgranamiento, en particular, predecir cuáles estudiantes que rindieron el primer examen parcial luego abandonaron el curso antes de rendir el segundo y último examen parcial. Y en el caso de lograr entrenar modelos con rendimientos superiores al azar, analizar las importancias de las variables para evaluar sus influencias en la deserción de estudiantes.

La elección sobre predecir desgranamiento y no incluir además predicciones a la deserción de estudiantes al comenzar el curso no es arbitraria. Es normal que en muchos de los cursos del C.B.C. no se controle asistencia de los estudiantes, y en particular se sabe que no se controló la asistencia en los cursos correspondientes con los datos de este trabajo. Por lo tanto, entre los registros de estudiantes que no poseen calificaciones en el primer examen parcial, no es posible discriminar si se trata de estudiantes que abandonaron antes de la instancia del primer parcial, si se presentaron a alguna clase o si directamente desistieron de presentarse al curso antes de comenzar el cuatrimestre. En cambio, el análisis de los estudiantes que abandonan entre primer y segundo parcial es más sólido teniendo en cuenta los datos a disposición. Se trata observaciones que contienen una calificación en el primer examen parcial (en lugar de ser un dato faltante), y corresponden a estudiantes que hicieron el esfuerzo por continuar en la materia y cumplieron con la condición de rendir el primer examen parcial, pero no se presentaron a rendir el segundo examen parcial, representados con datos faltantes en la variable asociada.

Lo primero que sigue a continuación es un marco teórico que presenta el tipo de algoritmo utilizado para entrenar los modelos, la métrica seleccionada para evaluar sus rendimientos y las técnicas para analizar las importancias de las variables. Siguiente al marco teórico se presentan los datos y un análisis exploratorio de los mismos, que en parte fue retroalimentado por algunos resultados preliminares. Al finalizar la exploración se pasa a los métodos experimentales, donde se explica como se entrenaron distintos modelos clasificadores, para luego encontrar una discusión sobre sus resultados y la conclusión.

[1] https://www.cbc.uba.ar/

2. Marco teórico

La utilización de ciencia de datos aplicada a educación ha proliferado en tal medida que hasta se ha generalizado en inglés la denominación *"Educational Data Mining"* [4], técnica que también es muy aplicada en universidades argentinas (ver como ejemplo el trabajo de D. Kuna y otros [5]). Entre los métodos más usados para estudiar rendimiento de estudiantes sobresalen las redes neuronales y los árboles de decisión, siendo los que alcanzan mejores rendimientos predictivos [6]. Si bien las redes neuronales entregan las mejores predicciones, los árboles de decisión además permiten comprender los resultados de una forma más simple.

2.1. Algoritmo: XGBoost

Siendo que el objetivo del trabajo no es buscar los mejores algoritmos para predecir la deserción de estudiantes, sino entrenar modelos que permitan detectar qué factores influyen en dicha deserción, se optó por el algoritmo XGBoost [7], que incluye muchas de las bondades de los árboles de decisión pero que además ha demostrado ser muy eficiente y rápido para analizar conjuntos de datos tabulares muy grandes, características a tener en cuenta si se desea extender este trabajo a un conjunto de datos aumentado. Si bien no se buscará el modelo con mayor poder de predicción, para tener más certezas en el estudio de las importancias de las variables es necesario alcanzar buenos rendimientos. XGBoost permite mejorar los rendimientos con entrenamientos rápidos, considerando la gran cantidad de datos con la que se trabajó, y al mismo tiempo ofrece la posibilidad de interpretar los resultados.

2.2. Métricas: AUCPR y AP

Cuando los datos son desequilibrados, una de las métricas generalmente recomendadas para medir el rendimiento de modelos de clasificación binaria es el área bajo la curva *precision-recall* , o AUCPR (*area under the curve precision-recall*) [8]. Esta métrica resulta conveniente para este trabajo porque resume el rendimiento del modelo para todos los valores de corte, evitando tener que buscar un corte óptimo para las predicciones, el cual puede cambiar arbitrariamente por un futuro usuario según sus acciones a seguir con dichas predicciones. La línea de base de la métrica AUCPR es [9]:

$$y = \frac{P}{P+N} \tag{1}$$

siendo y la altura en una gráfica *precision-recall*, y donde P representa la cantidad de observaciones positivas y N las negativas. De esta forma el área bajo la curva *precision-recall* obtenida por un clasificador al azar resulta igual a la proporción de observaciones positivas sobre el total de observaciones. Sin embargo, pueden surgir malas interpretaciones al comparar rendimientos de diferentes modelos únicamente teniendo en cuenta dicha línea de base, debido a que en el espacio *precision-recall* existe una región inalcanzable que depende del desequilibrio del conjunto de datos [10]. La curva *precision-recall* siempre se encuentra por encima de esta región inalcanzable que define un mínimo para el AUCPR, y, si representamos al desequilibrio como $\pi = \mathrm{P}/(\mathrm{P}+\mathrm{N})$, este mínimo se puede calcular de la siguiente forma:

$$\mathrm{AUCPR}_{\min} = 1 + \frac{(1-\pi)\ln(1-\pi)}{\pi} \tag{2}$$

Para obtener el valor de AUCPR es necesario contar con la curva *precision* vs. *recall* completa, y esto solo puede ser aproximado por los experimentos. Un método para aproximar el AUCPR es calculando la precisión media, o AP (*average precision*), de la siguiente forma:

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n \tag{3}$$

donde P_n y R_n son los valores de *precision* y de *recall* alcanzados en el corte n-ésimo.

3. Datos

3.1. Descripción de los datos utilizados

Los datos fueron provistos en distintas tablas de MS Access con diferentes estructuras que fueron unificadas en una tabla dentro de un archivo `csv`. Cada registro de esta tabla de datos corresponde a la información de un estudiante en un curso. La tabla 1 muestra la cantidad de registros, de estudiantes y de sedes, y podemos notar que el número de estudiantes es muy inferior a la cantidad total de observaciones porque un gran número de estudiantes se inscribe a una de las materias de esta cátedra en más de una oportunidad.

Tabla 1: Cantidad de registros, estudiantes y sedes en los datos originales.

Ítem	Cantidad
observaciones	233615
estudiantes	120364
sedes	21

En la tabla 2 se describen brevemente las variables que integran esos datos. Aunque es sencillo comprender qué representa cada variable listada en dicha tabla, vale hacer las siguientes aclaraciones. Las notas de los exámenes de un estudiante pueden estar vacías (alguna o todas), y eso representará que el estudiante no rindió ese examen. Además, cuando el estudiante alcanza la condición que lo habilita a rendir un examen final, tiene tres oportunidades consecutivas para hacerlo, y las calificaciones de esas oportunidades se encuentran en las variables `Final`, `rem1` y `rem2`.

Tabla 2: Variables originales. En la columna de valores se muestra el contenido que se debería encontrar en cada variable. (`nan`: *not a number*)

Variable	Descripción	Valores
`anio`	El año en que el estudiante cursó en un determinado curso.	2011, 2012, . . . , 2019
`cuat`	El cuatrimestre en que el estudiante cursó.	1 o 2
`dni`	Documento Nacional de Identidad.	Ejemplo: 42000251
`COMISION`	Código utilizado para los cursos.	Ejemplo: 45301
`HORARIO`	Código utilizado para los diferentes turnos.	Ejemplo: 658
`AULA`	Número de aula donde se desarrolla el curso.	Ejemplos: 1, 13, 214
`SEDE`	Código de sede donde se desarrolla el curso.	Ejemplos: 1, 4, 28
`MATERIA`	Código de la materia que cursa el estudiante.	3 o 53
`pa1`	Nota del primer parcial.	0 a 10, o `nan`
`pa2`	Nota del segundo parcial.	0 a 10, o `nan`
`Final`	Nota del examen final.	0 a 10, o `nan`
`codCarrera`	Código que identifica la carrera del estudiante.	Ejemplos: 9, 45
`facultad`	Nombre de la facultad para la carrera del estudiante.	Ejemplo: MEDICINA
`rem1`	Nota del examen final en 2da oportunidad.	0 a 10, o `nan`
`rem2`	Nota del examen final en 3ra oportunidad.	0 a 10, o `nan`

3.2. Limpieza y preparación de los datos

Los datos fueron generados por múltiples usuarios con diversidad de criterios y en múltiples locaciones, por lo cual era esperable encontrarse con muchos datos erróneos o indeterminados, además de diferentes nomenclaturas. El trabajo de estandarización y limpieza fue bastante extenso, y a continuación solo se muestra un resumen.

Acciones relacionadas a nombres de facultades, carreras y códigos de carrera:

- Estandarización de los nombres de Facultades. Los casos indefinidos se reemplazaron por un código que los diferencie del resto.
- Los nombres de las carreras contenían diferentes denominaciones para una misma carrera y caracteres extraños, que fueron estandarizados según la información que se encuentra en la página del Ciclo Básico Común [11].
- Los registros con información faltante sobre facultad, carrera o código de carrera, se completaron utilizando los códigos o nombres en otros registros completos cuando fue posible encontrar alguna relación entre las observaciones.
- Se eliminaron los registros con códigos de carreras inexistentes o sin código ni información sobre carrera o facultad. (Cerca de 80 observaciones)

Luego de esta estandarización, alrededor de un 12 % de las observaciones contienen un código de carrera (99 o 999) que no está asociado a ninguna carrera ni facultad, y no contienen información adicional como el nombre de la carrera o la facultad, en ninguno de los registros de esos estudiantes. Luego de un análisis exploratorio se decidirá si imputar o no esos valores y cómo hacerlo.

En cuanto a las variables relacionadas con calificaciones, se encontraron 231 observaciones con valores no esperados, como por ejemplo 25 o 98. En los casos en que fue posible, se imputaron valores según la información de las otras notas. Al tratarse de muy pocos registros, cuando no había información concluyente simplemente se reemplazaron por valores posibles, sin dedicar demasiado tiempo a una imputación más inteligente. De ser necesaria una corrección a este método (luego de los análisis correspondientes), una posibilidad es reemplazar por las notas más probables o que respeten alguna distribución en el curso, sede o turno.

Para finalizar se puede mencionar que se imputaron valores faltantes en COMISION y AULA en 120 observaciones, utilizando valores posibles según la sede y el horario de cada registro.

3.3. Ingeniería de variables

En la tabla 3 se resume una descripción de las variables creadas en esta etapa del trabajo. A partir de ahora, cada observación tendrá la información sobre un estudiante en un curso y además información sobre el curso y los demás estudiantes del curso. El objetivo de la creación de estas variables es incluir factores exógenos que intuitivamente se relacionan con desempeño académico o deserción, como el número de estudiantes o la composición de los cursos según alguna característica, que están bajo el control de la universidad, y de esta forma poder analizar si las decisiones de la institución en estos aspectos tienen una influencia medible.

Los códigos de `COMISION` se repiten en cada cuatrimestre y las numeraciones de `AULA` tienen repeticiones en diferentes sedes. Para posibilitar análisis más específicos respecto de estas variables se generaron identificadores únicos de `curso` y de `sala`, contemplando que la sala sí puede repetirse en distintos horarios y cuatrimestres para una misma sede.

Tabla 3: Variables creadas.

Variable	Descripción
extranjero	0 o 1. Es extranjero si dni $>$ 90 millones.
curso	Identificación única de curso.
turno	A: muy temprano, B: media mañana, C: tarde, D: noche.
n_alum	Cantidad de estudiantes inscriptos en el curso.
p_ext	Porcentaje de extranjeros en el curso.
recurso	Cantidad de veces que se inscribió anteriormente.
p_recursa	Porcentaje de recursantes en el curso.
sala	Identificación única de aula.
pa1_prom	Promedio de calificaciones de parcial 1 en el curso.
pa2_prom	Promedio de calificaciones de parcial 2 en el curso.
final_prom	Promedio de calificaciones de final en el curso.
edad	Categoría estimada con dni.
prom_edad	Promedio de la variable edad en cada curso.
condición	Abandona1, Abandona2, Insuficiente, Examen, Promociona.
abandona1_p	Porcentaje en condición Abandona1 en el curso.
abandona2_p	Porcentaje en condición Abandona2 en el curso, sobre los que rindieron parcial 1.

El código HORARIO indica los días y horarios en que se cursa la materia. Según estos códigos se asignó la categoría turno a cada observación según el siguiente criterio: los cursos que comienzan al principio del día (7AM y 8AM), los que comienzan a media mañana (de 9AM a 11AM), los que comienzan a la tarde (de 1PM a 6PM) y los cursos de la noche (desde 7PM en adelante).

La variable edad es una categoría estimada a partir del dni. Para cada cuatrimestre se construye un histograma de los valores de dni formado con 10 intervalos regulares, y se extraen los límites de dichos intervalos. Luego a cada observación se le asigna la categoría de edad según a qué intervalo pertenece su dni en ese cuatrimestre. Un estudiante que recurse puede tener diferentes categorías de edad en los distintos cuatrimestres.

Los extranjeros poseen un código de dni que se diferencia de los nacionales (son valores mayores a 90 millones, cuando los números de dni nacionales no llegan a 60 millones), y por lo tanto no se puede determinar su categoría de edad. Para completar la variable edad en todas las observaciones, con los extranjeros se decidió imputarles un valor de dni extraído aleatoriamente del conjunto de dni sin extranjeros de cada cuatrimestre, manteniendo de esta forma la distribución de las edades, a costa de agregar errores en la edad de una porción de las observaciones. Si en el futuro se observa que esta categoría puede ser relevante, se propone repetir los experimentos para el conjunto de datos sin incluir extranjeros.

La categoría condición se determina según las siguientes reglas:

- Abandona1: no tiene notas en ningún examen, abandonó antes de rendir el primer parcial.
- Abandona2: tiene nota en el primer parcial pero no tiene nota de segundo parcial.
- Insuficiente: la suma de ambos parciales es menor a 8.
- Examen: la suma de ambos parciales es mayor o igual a 8 y menor a 13. Son estudiantes que deben rendir un examen final para aprobar la materia.
- Promociona: la suma de ambos parciales es mayor o igual a 13.

Para finalizar la modificación de los datos, la variable `dni` fue sustituida por una identificación única de estudiante diferente para anonimizar su posible aparición en códigos o resultados que se deseen distribuir, y las variables `COMISION`, `AULA` y `HORARIO` fueron eliminadas.

Se ha considerado generar más variables, como por ejemplo la composición del curso según facultades o carreras, o descripciones sobre las distribuciones de notas, pero la generación de variables como las mencionadas demandan tiempo para su generación y su verificación, y se decidió postergar la creación de más variables para una posible continuación de este trabajo, luego de analizar los resultados. Y por supuesto, siempre es interesante incluir la variable género, en este caso el de los estudiantes, pero lamentablemente no se cuenta con esa información en forma directa. Sí se tiene acceso a los nombres de los estudiantes, y sería posible extraer el género con diferentes probabilidades a partir de esa información, quedando este análisis también pendiente para una posible segunda etapa del trabajo.

4. Análisis exploratorio de datos

Finalizada la preparación de los datos y la creación de nuevas variables, se generaron reportes preliminares de análisis exploratorios automatizados, demasiados extensos para ser incluidos en este informe pero que se puede consultar en el repositorio del material de este trabajo (consultar el apéndice sobre la organización de dicho repositorio). Dicho informe es exhaustivo por demás y no toda la información que contiene es relevante, pero sirve como referencia del estado de los datos en este punto del trabajo.

Es importante en este punto observar que durante el análisis exploratorio se detectaron cursos completos sin calificaciones ingresadas. Estos cursos aportan información al análisis exploratorio pero no son válidos para un análisis de deserción. El trabajo que a continuación se relacione con calificaciones se realiza sobre una porción de los datos que se considera válida, es decir, que todos los cursos tienen notas cargadas. Esta porción de los datos contiene 106987 observaciones consideradas válidas para el estudio sobre desgranamiento, con 1763 cursos.

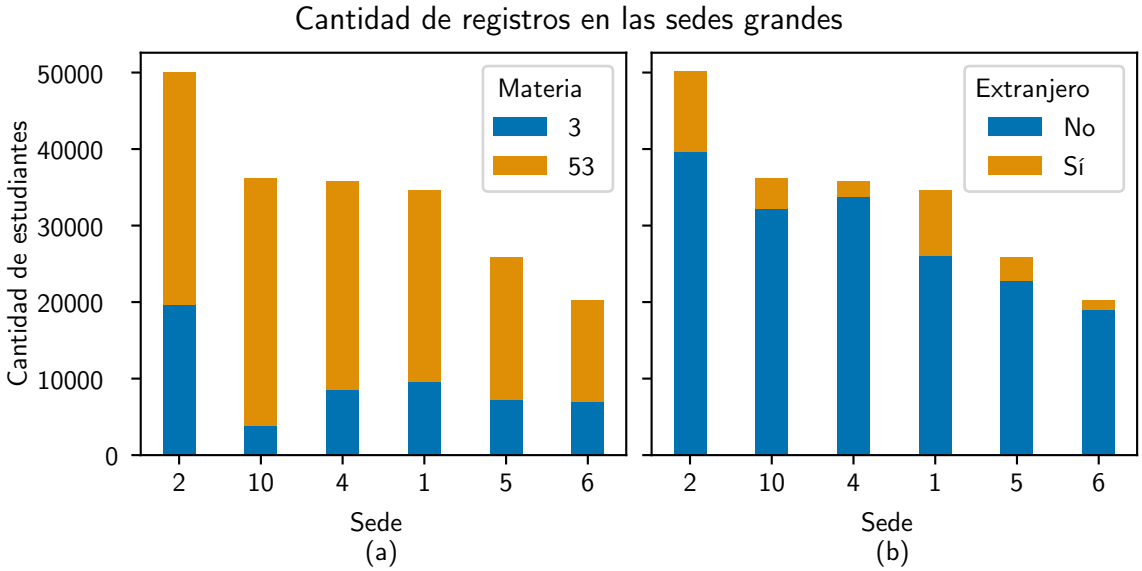

Figura 1: Cantidad de estudiantes que se inscribieron en las sedes más grandes en cada materia y la cantidad de extranjeros.

Como primer paso se realizó un estudio del balance de los datos en diferentes categorías, del cual se desprende que las 21 sedes se pueden caracterizar según el número de estudiantes como sedes grandes o

pequeñas, siendo que el 86 % de las observaciones corresponden a solo 6 sedes más grandes. Esta es una nueva variable que se puede crear respecto de la sede donde cursa cada estudiante.

Respecto a la cantidad de observaciones según otras categorías, a modo de ejemplo en la figura 1 (a) se muestran proporciones similares entre estudiantes de cada materia en las distintas sedes grandes, con la excepción de la sede 10 donde hay déficit de estudiantes de la materia 3, y en 1 (b) se observa que las proporciones de extranjeros también son similares en estas sedes, con una diferencia notable en la sede 1.

Por otro lado, en la figura 2 se pueden comparar las cantidades de inscriptos por turnos, donde hay una gran diferencia en el número de inscriptos al turno noche (D), y no se ven grandes diferencias entre sedes.

Estos desequilibrios en las categorías sede, materia, extranjero o turno deberán ser tenidos en cuenta al momento de interpretar resultados.

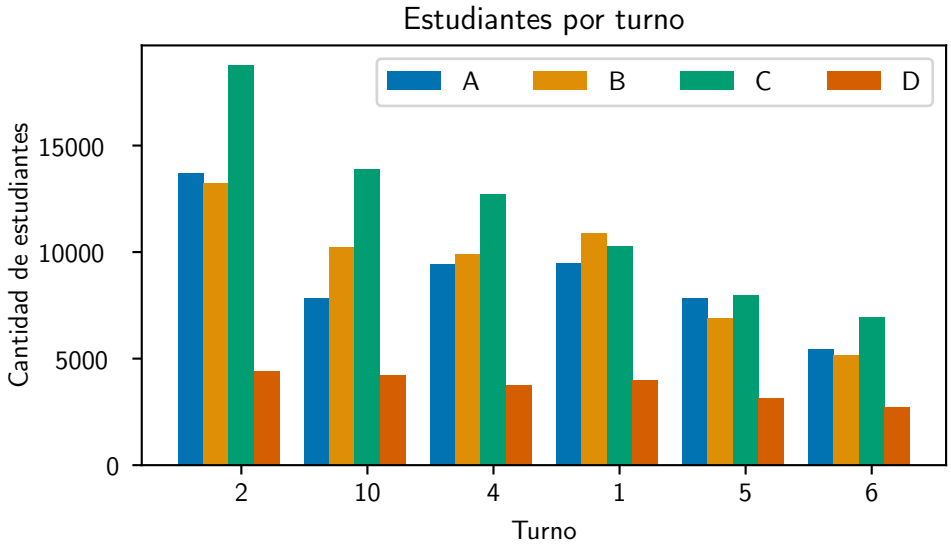

Figura 2: Cantidad de estudiantes que se inscribieron en las sedes más grandes en cada turno.

Para indagar sobre deserción, la variable objetivo es `Abandona1` o `Abandona2` o alguna combinación de ellas. En las figuras 3 y 4 se muestran gráficos de caja de porcentajes de estudiantes que abandonan, para las sedes más grandes (las primeras 6) y algunas de las sedes pequeñas. Los porcentajes antes del primer parcial representan la cantidad de estudiantes que no rindieron el parcial 1 sobre el total de estudiantes inscriptos en el curso, y los porcentajes después del parcial 1 representan la cantidad que rindió el segundo parcial sobre los que rindieron el primero. Es notable la mayor retención entre los estudiantes que rindieron el primer parcial, y como era de esperar, hay mayor variabilidad cuando se calculan por curso respecto al porcentaje global de cada sede en cada cuatrimestre.

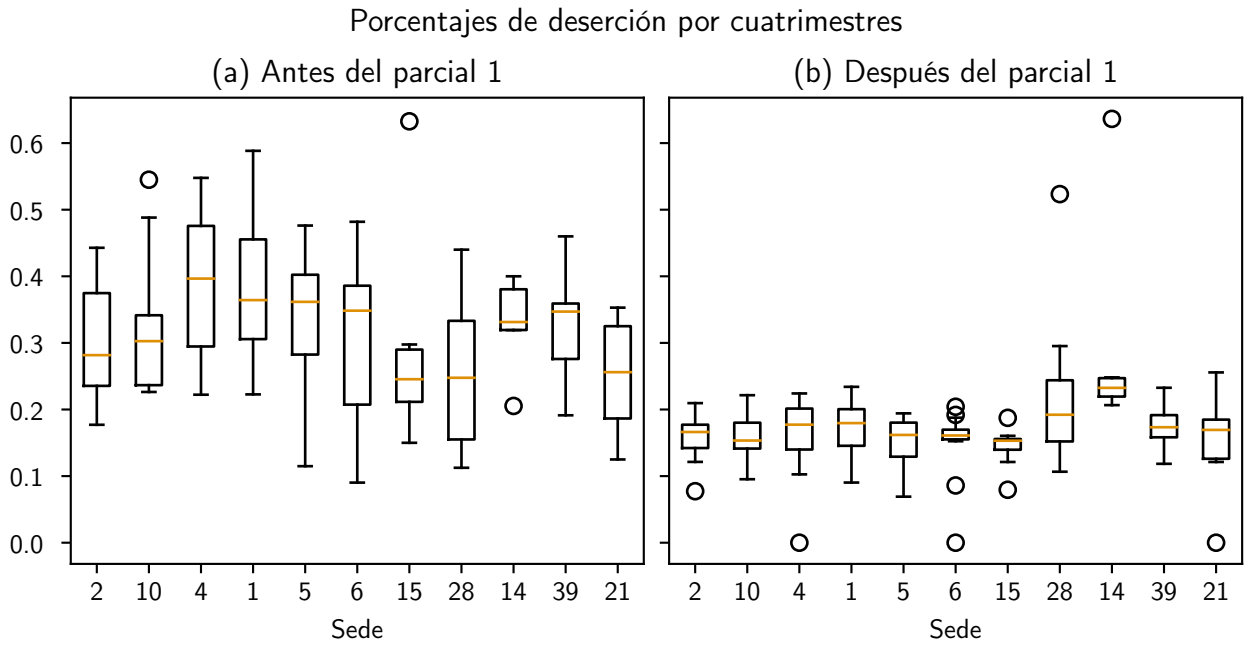

Figura 3: Dispersión de los porcentajes de estudiantes que abandonan antes o después del parcial 1 en cada cuatrimestre, por sedes.

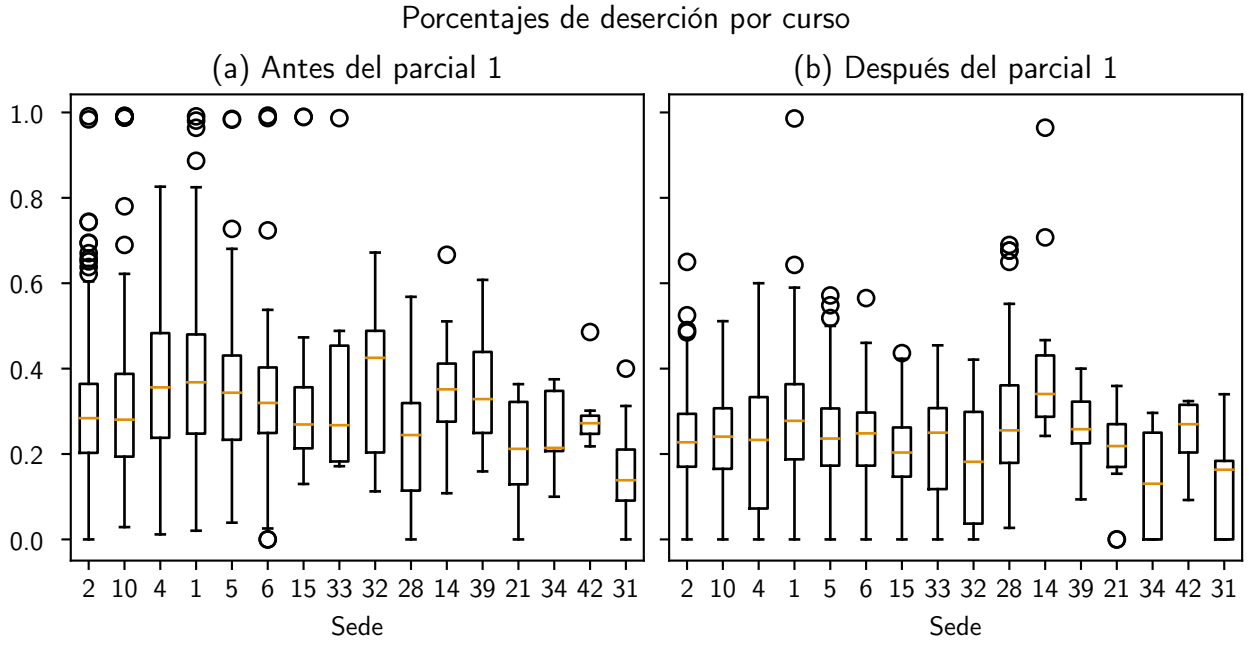

Figura 4: Dispersión de los porcentajes de estudiantes que abandonan antes o después del parcial 1 en cada curso, por sedes.

Un análisis posible sobre desempeño académico es a través de los promedios de calificaciones. La figura 5 muestra histogramas de las calificaciones obtenidas por cada estudiante en el examen parcial 1, en cada sede. Se observan distribuciones similares entre las sedes grandes y entre las sedes pequeñas, pero los histogramas de las sedes pequeñas están desplazados hacia las notas más bajas respecto de las sedes grandes. En cambio, si se analizan los histogramas de los promedios de notas de ambos parciales, no se encuentran diferencias apreciables entre las sedes grandes y pequeñas.

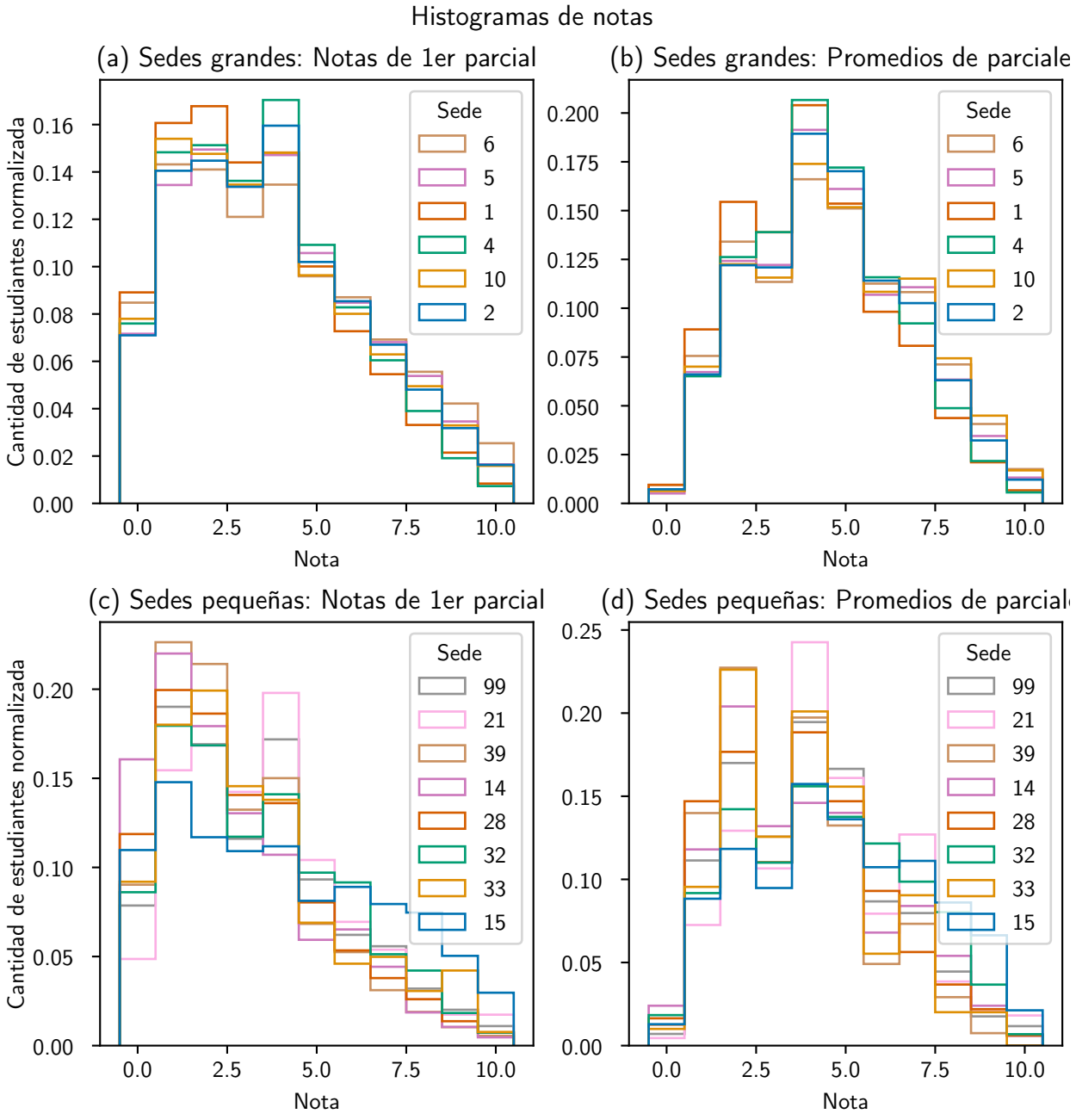

Figura 5: Histogramas de calificaciones discriminados por sede. Cada una de las sedes grandes agrupa más de 20 mil observaciones, mientras que cada sede pequeña agrupa menos de 10 mil. El código 99 agrupa las observaciones de todas las sedes con menos de mil registros cada una.

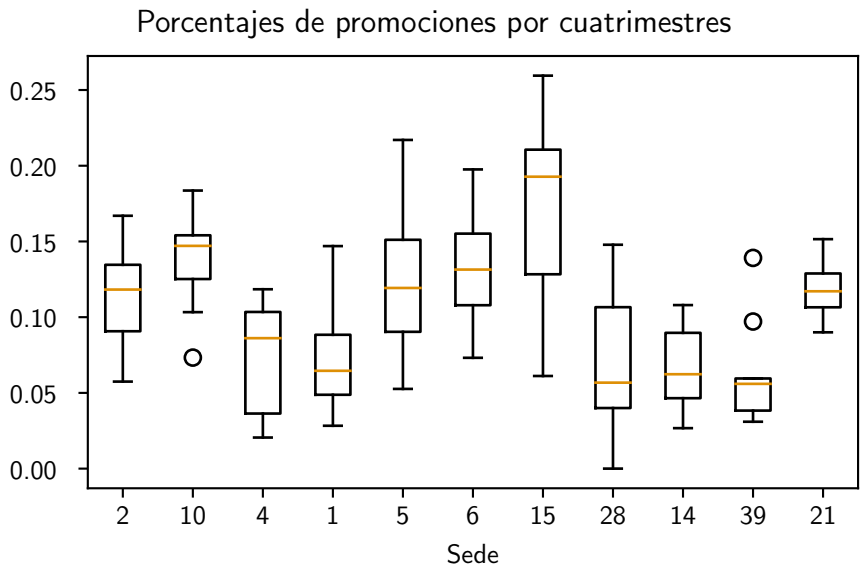

Figura 6: Dispersión de los porcentajes de estudiantes que promocionan cada cuatrimestre, por sedes.

Otra forma de medir desempeño académico es mediante la condición que alcanza cada estudiante al finalizar el curso, siendo `condicion` la variable objetivo en ese caso. Como ejemplo, la figura 6 contiene diagramas de caja de los porcentajes de estudiantes que alcanzan la condición de promoción (las notas más elevadas) en cada cuatrimestre, discriminado por sedes. Viendo que los resultados son dispares, una parte de los estudios futuros será dedicada a intentar explicar estas diferencias.

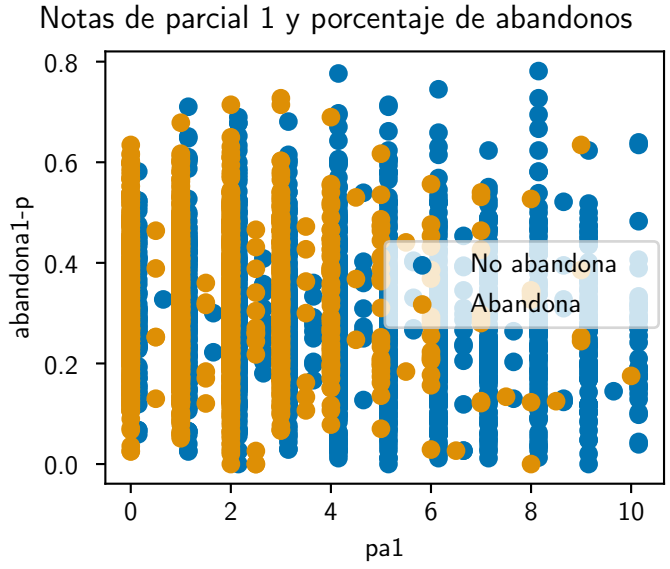

Figura 7: Organización de las observaciones según la nota del parcial 1 y el porcentaje de abandonos antes del primer parcial, discriminados por estudiantes que abandonan después del parcial 1 y los que rinden el segundo parcial.

Ya enfocados en los modelos clasificadores que nos interesan, entre los múltiples gráficos de a pares que se pueden obtener (consultar apéndice sobre dónde acceder a gráficos de a pares realizados con estos

datos), el de la figura 7 demuestra que la gran mayoría de los estudiantes que abandondan luego de rendir el primer examen parcial son los que obtuvieron calificaciones bajas en dicho examen. Y al mismo tiempo, el porcentaje de estudiantes que abandona antes del primer examen parcial en cada curso parece uniforme respecto de las notas de los estudiantes, con un leve descenso en las notas 9 y 10.

También considerando los factores que pueden ser utilizados para predecir abandono después del primer parcial, la figura 8 nos muestra que hay baja correlación entre las variables numéricas seleccionadas para entrenar modelos de predicción de estudiantes que abandonan entre primer y segundo parcial.

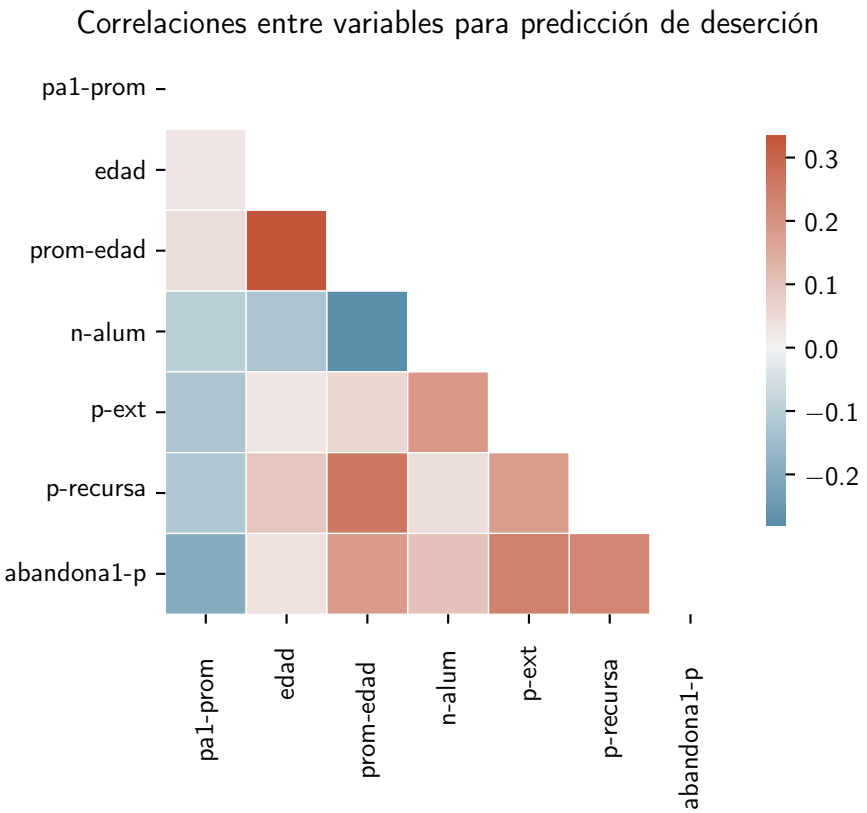

Figura 8: Mapa de correlaciones entre las variables numéricas que se consideran consistentes con un modelo de predicción de deserción entre primer y segundo parcial.

5. Métodos

5.1. Entrenamiento y evaluación

Como ya se mencionó, en todos los experimentos se entrenaron modelos de clasificación binaria con el algoritmo XGBoost. Teniendo en cuenta que el objetivo último no era la predicción, sino determinar si las predicciones se alejan del azar debido a la inclusión de variables exógenes, las rutinas de entrenamiento no fueron exhaustivas. De todas formas, para descubrir las variables o interacciones de variables más importantes es necesario incluir un ajuste de hiperparámetros en los entrenamientos, porque con

predictores muy débiles se pueden esperar muchas variaciones en el orden de importancia de las variables en diferentes experimentos, y perder consistencia.

El camino elegido fue la validación cruzada entre 5 subconjuntos, realizando una búsqueda aleatoria en grilla de hiperparámetros. En todos los casos los hiperparámetros y sus valores que se seleccionaron en cada entrenamiento son los presentados en la tabla 4, ya los parámetros de la validación cruzada son los mostrados en la tabla 5.

Tabla 4: Grilla para búsqueda de hiperparámetros en XGBoost.

hiperparámetro	**Valores o rango**
`objective`	binary:logistic
`eval_metric`	aucpr
`eta`	(0,005, 0,6)
`gamma`	(0, 20)
`max_depth`	(2, 20)
`lambda`	(0, 20)
`alpha`	0

Tabla 5: Parámetros para validación cruzada.

parámetro	**Valor**
`num_boost_round`	1000
`nfold`	5
`early_stopping_rounds`	10
`metric`	aucpr

Se separaron los datos en un 70 % para entrenamiento y un 30 % para evaluación, y los rendimientos de los modelos fueron cuantificados calculando el AP utilizando la librería *scikit-learn* [12], siempre en el mismo conjunto de evaluación.

5.2. Variables

Puede considerarse que la línea de base fue la conseguida con el modelo que incluye todas las variables consistentes con la predicción de abandono entre parcial 1 y parcial 2, listadas en la tabla 6. En los casos `codCarrera` o `facultad`, y `SEDE` o `sala`, se debió optar por una en ambos pares, ya que `facultad` es un agrupamiento de códigos de carreras, y `SEDE` un agrupamiento de salas.

Los múltiples experimentos se realizaron eliminando una o varias de las variables de la tabla 6, para cuantificar la importancia de la variable eliminada, o la interacción entre variables eliminadas, determinando cuánto descendía el valor de AP, y también para analizar cómo variaban las importancias en el nuevo modelo.

5.3. Importancia de variables y test de permutaciones

Con los modelos que ofrezcan algún poder de predicción se pueden realizar análisis de importancias de variables. XGBoost tiene sus propios métodos para informar la importancia de las variables en cada modelo, y en este trabajo se registraron las importancias según sus aportes al cálculo de *gain* total (*total gain*), siendo *gain* la forma en que XGBoost mide el aumento en rendimiento en cada división del árbol.

Tabla 6: Variables consistentes con la predicción de abandono entre parcial 1 y 2.

variables	
cuat	turno
extranjero	p_ext
n_alum	abandona1_p
pa1	pa1_prom
recurso	p_recursa
edad	prom_edad
codCarrera o facultad	
SEDE o sala	
MATERIA	

Sin embargo, los resultados que entrega XGBoost pueden derivar en interpretaciones erróneas, debido a la complejidad en el cálculo de dichas ganancias.

Un método alternativo y consistente (cada vez que se cambia un modelo para depender más en una variable, la importancia de esa variable no puede descender) es el ya mencionado: entrenar modelos eliminando las variables de a una (o de a pares para investigar interacciones o correlaciones) y cuantificar sus importancias según el descenso en el valor de AP. El inconveniente con este método es que demanda mucho tiempo computacional para entrenar todos los modelos.

Otro método consistente y muy fácil de aplicar es el test de permutaciones para importancias [13]. Con el modelo de base entrenado y su AP calculado con el conjunto de evaluación, lo que consideramos línea de base, se permutan los valores de una variable en el conjunto de evaluación y se calcula el AP de las predicciones en este conjunto permutado. La importancia de esa variable es la diferencia entre la línea de base y del AP en las predicciones sobre el conjunto con la columna permutada. Este método no requiere reentrenar modelos.

5.4. Repeticiones

Se realizaron algunos experimentos reentrenando el mismo modelo (las mismas variables) múltiples veces, cambiando las observaciones de los conjuntos de entrenamiento y de evaluación, con el objetivo de observar la variabilidad de los resultados de AP y las importancias de las variables según las observaciones disponibles en cada entrenamiento y evaluación.

6. Resultados y discusión

Los resultados de la métrica AP que se muestran en la tabla 7 pueden ser interpretados de la siguiente forma. Con un clasificador al azar [14] se obtuvo un AP que es igual a la proporción de casos positivos respecto del total (0,24), en acuerdo con lo indicado por la ecuación 1. El rendimiento sobre el conjunto de evaluación del modelo base, que incluye todas las variables consistentes para este experimento, está alejado del clasificador aleatorio, indicando que existen relaciones entre los factores incluidos y la variable objetivo. Sin embargo, si se observan las importancias de las variables en la tabla 8, encontramos que la variable `pa1` es la más importante y su importancia está muy lejos de las otras, lo cual pone en duda la interpretación del resultado de AP del modelo base. Este resultado es el usual en la mayoría de este tipo de trabajos de minería de datos en educación, donde el promedio de las notas del estudiante es determinante para decidir si completa sus estudios o no [6].

Tabla 7: Precision media obtenida sobre el conjunto de evaluación por los modelos: base que incluye todas las variables, modelo sin la variable pa1, y un clasificador aleatorio.

modelo	**AP**
base	0,587
sin `pa1`	0,358
aleatorio	0,242

Para intentar entender si el modelo solo depende de la nota del parcial 1 y las demás variables son inservibles, podemos recurrir al resultado obtenido con el modelo entrenado sin incluir la variable `pa1`. El valor AP de dicho modelo es menor al valor base, como es de esperar, pero sigue siendo significativamente superior al clasificador aleatorio. Por lo cual podemos afirmar que si bien la nota del primer parcial es una gran influencia en la variable objetivo, las demás variables en conjunto también demuestran un cierto poder de predicción sobre los abandonos. Releyendo la tabla de importancias luego de este análisis, podemos interpretar que las bajas importancias se deben a interacciones entre variables y/o correlaciones que no fueron detectadas anteriormente [13].

Tabla 8: Importancias informadas por XGBoost (*total gain* relativo a la menor), obtenidas por el método de permutaciones, y las obtenidas por el método de eliminación de variables.

XGBoost		**permutaciones**		**eliminación**	
variable	**importancia**	**variable**	**importancia**	**variable**	**importancia**
`pa1`	168,75	`pa1`	0,314	`pa1`	0,229
`abandona1_p`	12,62	`abandona1_p`	0,020	`abandona1_p`	≈ 0
`pa1_prom`	10,18	`edad`	0,016	`edad`	≈ 0
`prom_edad`	9,60	`p_recursa`	0,007	`p_recursa`	≈ 0
`p_ext`	8,76	`pa1_prom`	0,006	`pa1_prom`	≈ 0
`p_recursa`	8,50	`turno`	0,005	`turno`	≈ 0
`n_alum`	7,60	`facultad`	0,004	`facultad`	≈ 0
`facultad`	3,20	`prom_edad`	0,004	`prom_edad`	≈ 0
`turno`	2,91	`extranjero`	0,003	`extranjero`	≈ 0
`edad`	2,89	`SEDE`	0,002	`SEDE`	≈ 0
`SEDE`	2,77	`MATERIA`	0,002	`MATERIA`	≈ 0
`recurso`	2,38	`p_ext`	0,002	`p_ext`	≈ 0
`extranjero`	1,47	`cuat`	0,001	`cuat`	≈ 0
`cuat`	1,10	`n_alum`	0,001	`n_alum`	≈ 0
`MATERIA`	1,00	`recurso`	0,000	`recurso`	≈ 0

El experimento de entrenamientos múltiples utilizando distintas observaciones para los conjuntos de entrenamiento y de evaluación entrega resultados estables en las importancias, como se puede ver en el orden de las 5 más importantes variables en cada repetición en la tabla 9, y también en los valores de AP, con un valor medio de 0,588 y una desviación estándar igual a 0,005.

Recordando que en el análisis exploratorio se observó que una proporción muy grande de los casos positivos (abandona después del primer parcial) se corresponden a notas bajas en el primer parcial, se decidió entrenar el clasificador utilizando solo las observaciones con calificaciones en el primer parcial mayores a 3 ($\texttt{pa1} > 3$). En este caso, la proporción de casos positivos respecto del total es igual 4,2 %, un desequilibrio muy pronunciado. El AP de este modelo resultó igual a 0,126, es decir que sigue siendo

Tabla 9: Orden de importancias de las 5 variables más importantes en entrenamientos con diferentes observaciones en los conjuntos de entrenamiento y de evaluación.

variable	1	2	3	4	5	6	7	8	9	10
abandona1_p	2	2	2	2	2	2	2	2	2	2
edad					5					
p_ext									5	
p_recursa	5	5	5	3	3	5	5	5		4
pa1	1	1	1	1	1	1	1	1	1	1
pa1_prom	3	4	4	4		4	4	4	3	3
prom_edad	4	3	3	5		3	3	3	4	5
turno					4					

significativamente mayor a un clasificador al azar. Pero con estas condiciones la contribución de la nota del parcial 1 ya no es tan importante, como se puede observar en la taba 10.

Tabla 10: Importancias de las 5 variables más importantes informadas por XGBoost en el clasificador entrenado con la condición pa1 > 3.

variable	importancia
pa1	8,73
abandona1_p	7,56
pa1_prom	6,46
prom_edad	6,21
p_recursa	5,52

7. Conclusión

Se demostró que el método propuesto es adecuado para estudiar si ciertas variables exógenas influyen en los niveles de deserción, mediante la evaluación de clasificadores. Con el conjunto de datos utilizado en este trabajo, de calidad regular y limitado en variables, se logró determinar que las variables más importantes en la clasificación de abandonos con XGBoost son todas variables que expresan características de los cursos, exceptuando a la nota del primer parcial obtenida por el estudiante, cuando se la incluye.

Referencias

[1] Ana García de Fanelli. «Rendimiento académico y abandono universitario: Modelos, resultados y alcances de la producción académica en la Argentina». En: *Revista Argentina de Educación Superior* 8 (2014), págs. 9-38.

[2] A. Abu Saa, M. Al-Emran y K Shaalan. «Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques». En: *Tech Know Learn* 24.6 (2019), págs. 567-598. DOI: 10.1007/s10758-019-09408-7.

[3] Vincent Tinto. «Taking Student Retention Seriously: Rethinking the First Year of College». En: *NACADA Journal* 19 (sep. de 1999). DOI: 10.12930/0271-9517-19.2.5.

[4] Cristóbal Romero y Sebastián Ventura. «Educational Data Mining: A Review of the State of the Art». En: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.6 (2010), págs. 601-618. DOI: 10.1109/TSMCC.2010.2053532.

[5] Horacio Daniel Kuna, Ramón García Martínez y Fransisco Villatoro. «Identificación de causales de abandono de estudios universitarios: Uso de procesos de explotación de información». En: IV Congreso de Tecnología en Educación y Educación en Tecnología. 2009, págs. 172-177. URL: http://sedici.unlp.edu.ar/handle/10915/18991.

[6] Amirah Mohamed Shahiri, Wahidah Husain y Nur'aini Abdul Rashid. «A Review on Predicting Student's Performance Using Data Mining Techniques». En: *Procedia Computer Science* 72 (2015). The Third Information Systems International Conference 2015, págs. 414-422. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.12.157.

[7] Tianqi Chen y Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, págs. 785-794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[8] Jesse Davis y Mark Goadrich. «The Relationship between Precision-Recall and ROC Curves». En: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, págs. 233-240. ISBN: 1595933832. DOI: 10.1145/1143844.1143874. URL: https://doi.org/10.1145/1143844.1143874.

[9] Takaya Saito y Marc Rehmsmeier. «The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets». En: *PLOS ONE* 10.3 (mar. de 2015), págs. 1-21. DOI: 10.1371/journal.pone.0118432. URL: https://doi.org/10.1371/journal.pone.0118432.

[10] Kendrick Boyd y col. *Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation*. 2012. arXiv: 1206.4667 [cs.LG].

[11] CBC - Universidad de Buenos Aires. *CARRERAS - Según Facultades*. 2023. URL: https://www.cbc.uba.ar/carreras.

[12] F. Pedregosa y col. «Scikit-learn: Machine Learning in Python». En: *Journal of Machine Learning Research* 12 (2011), págs. 2825-2830.

[13] Silke Janitza, Carolin Strobl y Anne-Laure Boulesteix. «An AUC-based permutation variable importance measure for random forests». En: *BMC bioinformatics* 14 (abr. de 2013), pág. 119. DOI: 10.1186/1471-2105-14-119.

[14] Markus Ojala y Gemma C. Garriga. «Permutation Tests for Studying Classifier Performance». En: *Journal of Machine Learning Research* 11.62 (2010), págs. 1833-1863. URL: http://jmlr.org/papers/v11/ojala10a.html.

A. Repositorio de código y datos

El código utilizado en este trabajo se puede descargar del repositorio en GitHub:

https://github.com/epplugins/TT1

En dicho repositorio también se encuentran los datos, pero solo en su versión procesada y anonimizada, debido a que es un repositorio público. Debido a esto, las partes del código que se encargan del procesamiento y limpieza de los datos originales no pueden ser ejecutadas.