Taller de Tesis I
Entrega 2: Análisis Exploratorio de Datos

Edgardo Palazzo
epalazzo@cbc.uba.ar

31 de mayo de 2023

Resumen

Para el trabajo de especialización voy a analizar las bases de datos de estudiantes de la cátedra de física del Ciclo Básico Común de la Universidad de Buenos Aires, que contienen un historial del desempeño de cada estudiante en las materias de la cátedra, desde el año 2011 hasta el año 2019. El objetivo es estudiar la posibilidad de determinar si el nivel de deserción o el de rendimiento académico están influenciados por características de los cursos contenidas en los datos existentes, que en su mayoría son consideradas como variables exógenas respecto al estudiante.

En este trabajo se presentan los datos y la creación de nuevas variables, y se muestran algunos análisis exploratorios que puedan servir como referencia para las decisiones futuras al momento de entrenar modelos o aplicar técnicas que expliquen los datos.

1. Presentación de los datos

Los datos fueron provistos en distintas tablas de MS Access con diferentes estructuras que ya se unificaron en un único archivo `csv`. Cada registro de la tabla de datos originalmente 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 cursa una materia más de una vez.

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 representa 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 este estudiante cursó.	2011, 2012, . . . , 2019
`cuat`	El cuatrimestre en que este 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 cursa este estudiante.	Ejemplos: 1, 13, 214
`SEDE`	Código de sede donde cursa este estudiante.	Ejemplos: 1, 4, 28
`MATERIA`	Código de la materia que cursa este estudiante.	3 o 53
`pa1`	Nota del primer parcial.	Entre 0 y 10 o `nan`
`pa2`	Nota del segundo parcial.	Entre 0 y 10 o `nan`
`Final`	Nota del examen final.	Entre 0 y 10 o `nan`
`codCarrera`	Código que identifica la carrera.	Ejemplos: 9, 45
`facultad`	Nombre de la facultad correspondiente.	Ejemplo: MEDICINA
`rem1`	Nota del examen final en 2da oportunidad.	Entre 0 y 10 o `nan`
`rem2`	Nota del examen final en 3ra oportunidad.	Entre 0 y 10 o `nan`

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 más extenso que el resumen mostrado a continuación.

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

- Estandarización de los nombres de Facultades. Los casos indefinidos se reemplazaron por `nan`.
- 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 [1].
- En los registros con información faltante sobre facultad, carrera o código de carrera, se completó la información utilizando los códigos o nombres relacionados en otros registros completos.
- 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. 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.[2]

Tabla 3: Variables creadas.

Variable	Descripción
`extranjero`	0 o 1. Es extranjero si dni $>$ 90 millones.
`edad`	Categoría estimada con dni.
`curso`	Identificación única de curso.
`turno`	A: muy temprano, B: media mañana, C: media tarde, D: noche.
`n_alum`	Cantidad de estudiantes 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.
`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.

La variable `edad` es una categoría estimada a partir del `dni` de la siguiente forma. 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.

En el caso de los extranjeros 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 extraídos aleatoriamente de el conjunto de dni sin extranjeros de cada cuatrimestre. Si en el futuro se observa que esta categoría puede ser relevante, será necesario hacer análisis por separado sin incluir extranjeros.

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 cuatrimestres para una misma sede.

El código de `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 media tarde (de 1PM a 6PM) y los cursos de la noche (desde 7PM en adelante).

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 luego de obtener algunos resultados.

4. Análisis exploratorio

Finalizada la preparación de los datos y la creación de nuevas variables, se generó un reporte con un análisis exploratorio automatizado demasiado extenso para ser incluido en este informe pero que se puede consultar en el siguiente link: http://users.df.uba.ar/edmundo/eda_reporte_02-preliminar.html. 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 algún tipo de información pero no son válidos para un análisis de desempeño académico. 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 159120 observaciones válidas, con 1763 cursos.

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 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. En la misma figura 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.

Estas distribuciones de estudiantes según sede, materia, extranjero o turno deberán ser tenidas en cuenta al momento de interpretar resultados.

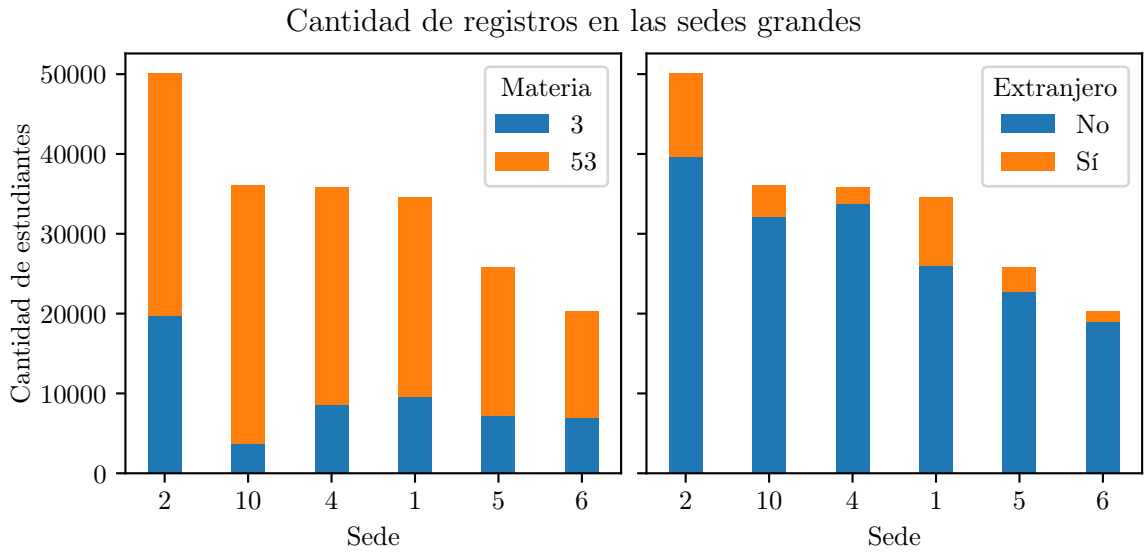

Figura 1: Cantidad de estudiantes que se inscribieron en las sedes más grandes en cada materia y la cantidad de extranjeros.

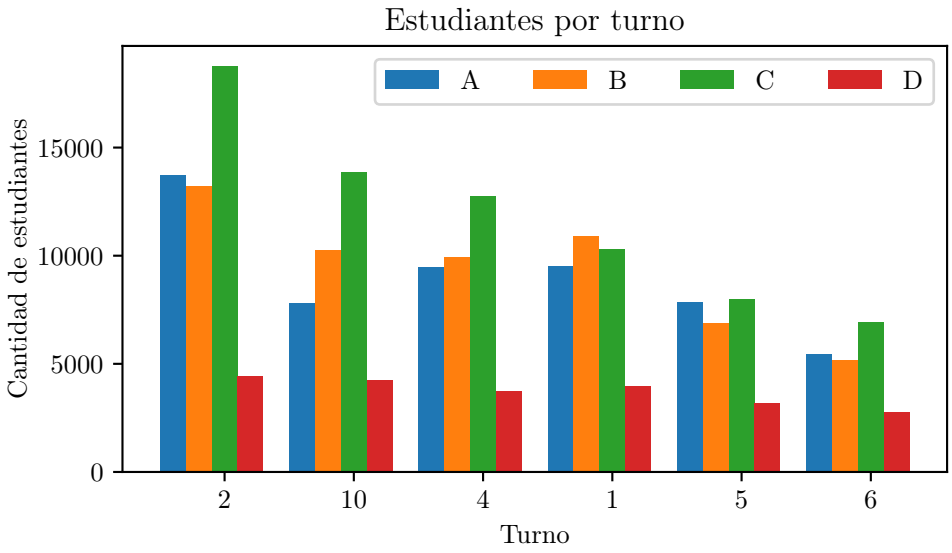

Figura 2: Cantidad de estudiantes que se inscribieron en las sedes más grandes en cada turno.

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, 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.

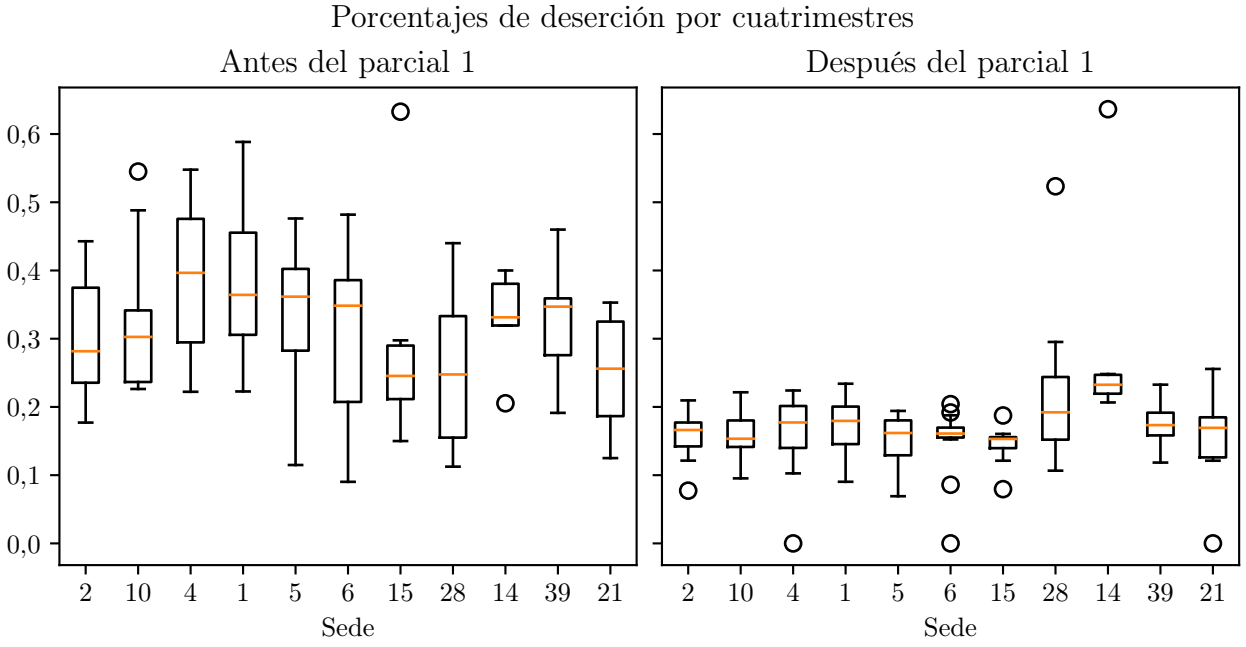

Figura 3: Dispersión de los porcentajes de estudiantes que abandonan antes o después del parcial 1 en cada cuatrimestre, por sedes.

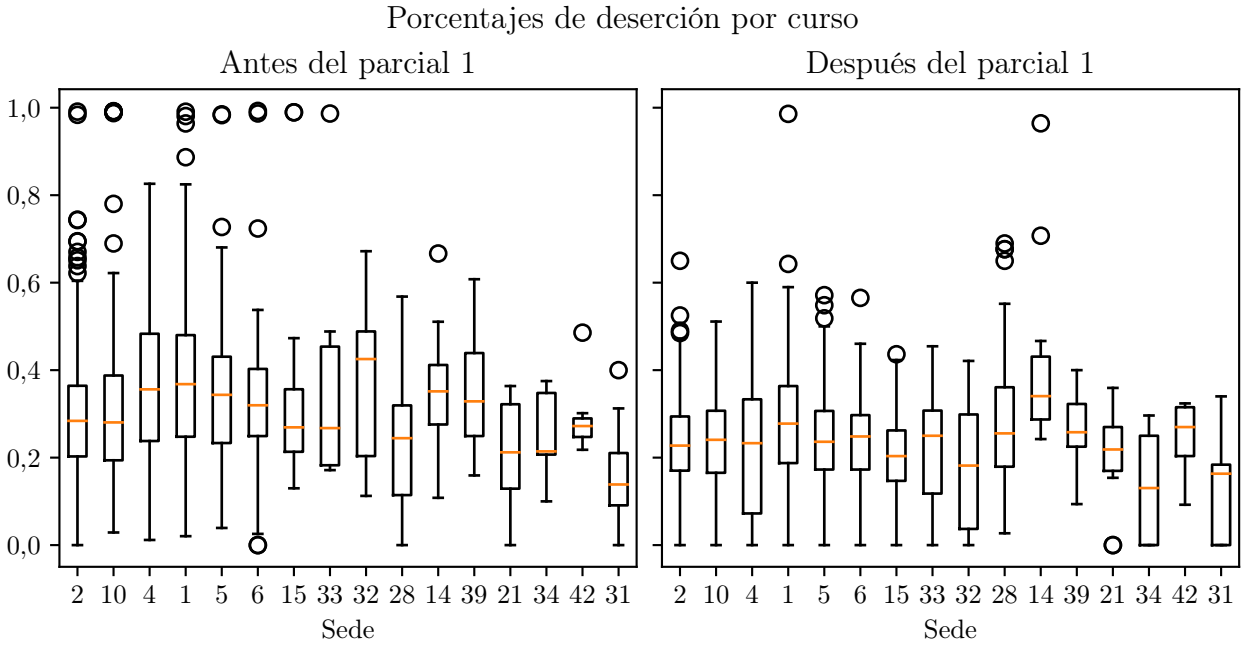

Figura 4: Dispersión de los porcentajes de estudiantes que abandonan antes o después del parcial 1 en cada curso, por sedes.

Una 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 5 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.

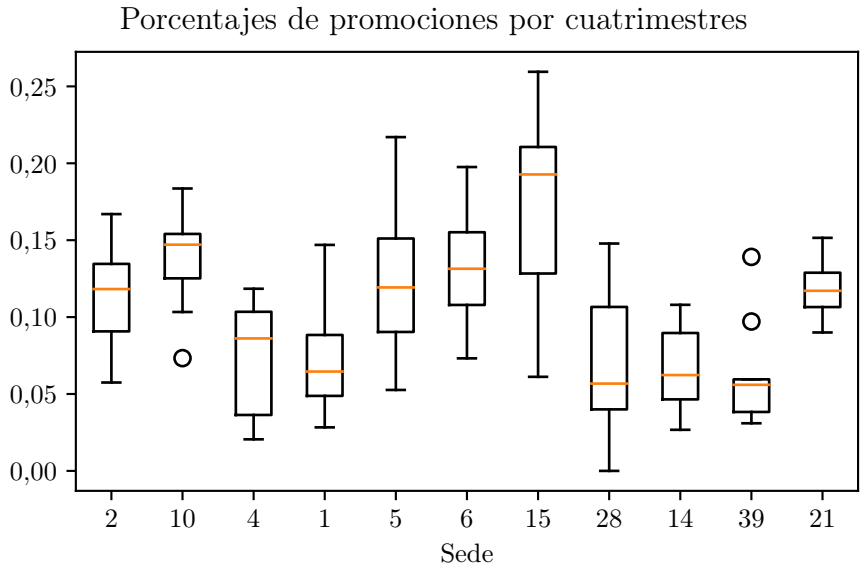

Figura 5: Dispersión de los porcentajes de estudiantes que promocionan cada cuatrimestre, por sedes.

Otro análisis posible sobre desempeño académico es a través de los promedios de calificaciones. La figura 6 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, mostrados en la figura 7, no hay diferencias apreciables entre las sedes grandes y pequeñas.

Hay muchos otros análisis posibles que se podrían hacer, como estudiar distribuciones dentro de una misma sede discriminadas por aula, por turno, por edades, o por materia. Pero primero se aplicarán las técnicas propuestas en la sección siguiente para descubrir variables importantes y luego se agregarán análisis para la comprensión de esos resultados particulares.

A modo de último comentario sobre exploración, siempre es interesante incluir la variable género, en este caso el de los estudiantes. No se cuenta con esa información en forma directa, pero sí se tienen los nombres de los estudiantes, y sería posible extraer el género con diferentes probabilidades a partir de esa información. Este análisis queda pendiente para una posible segunda etapa del trabajo.

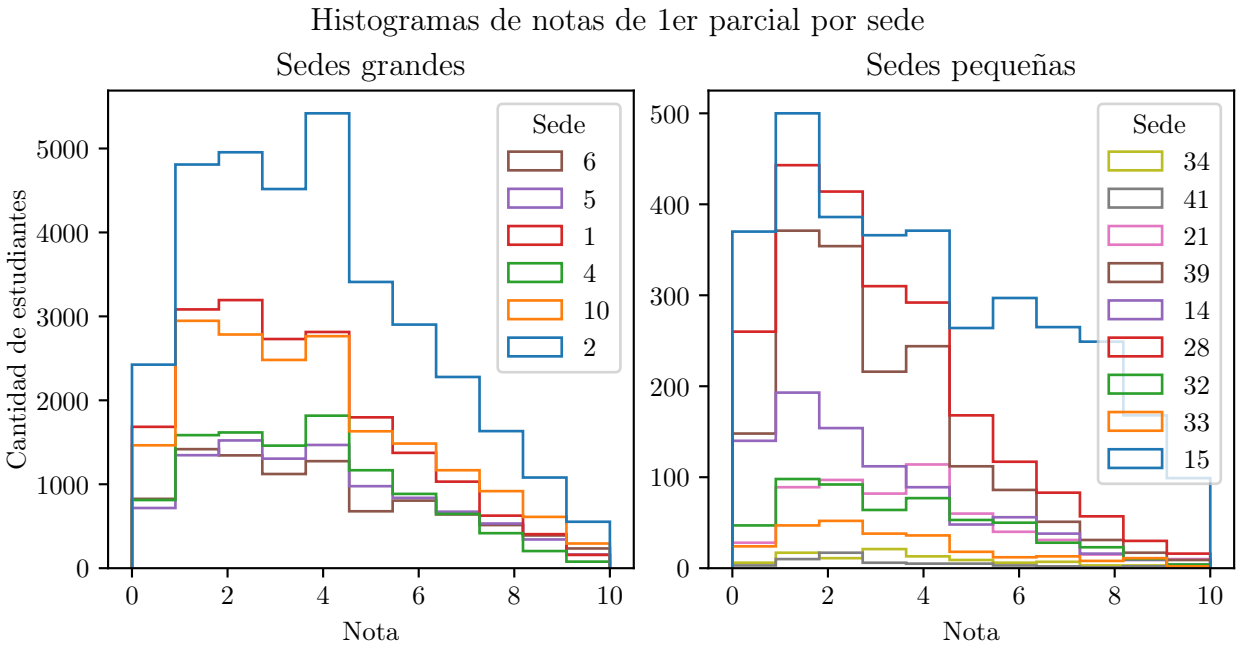

Figura 6: Histogramas de notas del parcial 1 por sede.

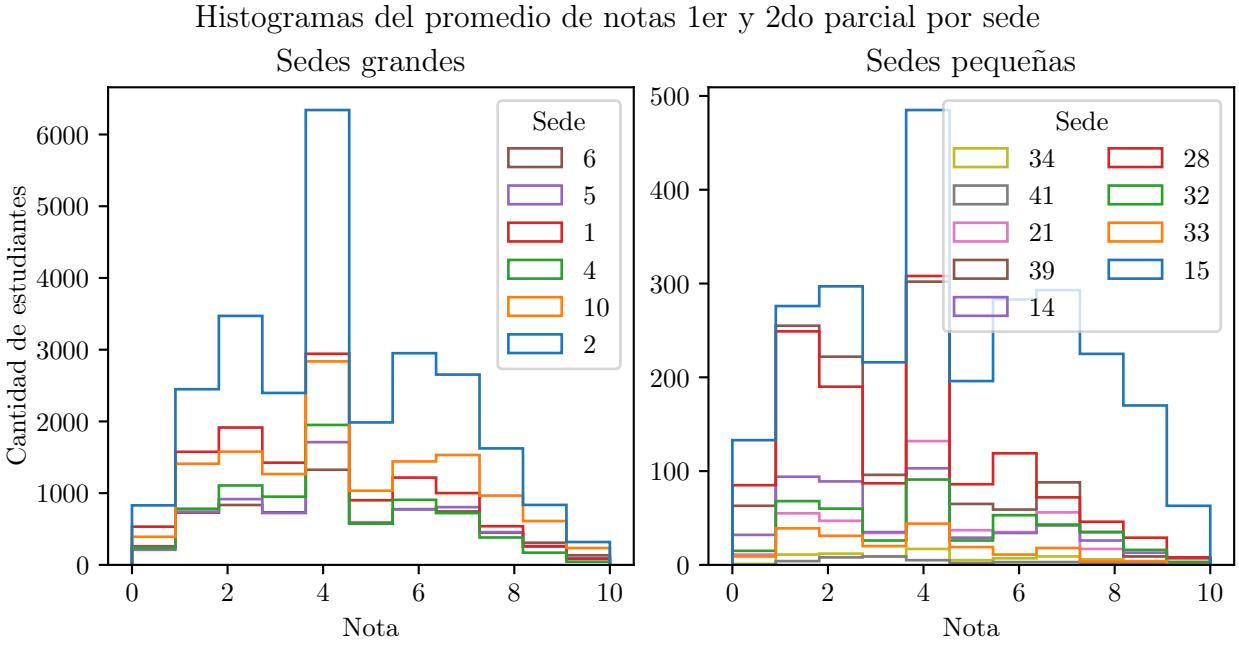

Figura 7: Histogramas del promedio de notas de ambos parciales por sede.

5. Metodologías propuestas

Se trabajará con dos técnicas diferentes: entrenar modelos predictivos y aplicar técnicas de regresión que expliquen alguna variable.

El primer enfoque consistirá en entrenar modelos predictivos utilizando la biblioteca XGBoost [3]. Se buscará predecir alguna de las variables objetivo que se analizaron durante la exploración: si abandona en el 1er o 2do parcial o la condición final de cada estudiante. Eso referido a predicciones sobre estudiantes individuales, pero también se buscará incluir predicciones respecto a cursos, considerando como variables objetivo a los porcentajes de deserción.

El objetivo último no es la predicción, sino medir cuanto se alejen las predicciones del azar, buscando indicios de que existen factores en estos datos que están influenciando esos resultados. Además estos modelos ayudarán a descubrir las variables o interacciones de variables más importantes.

En un enfoque alternativo se utilizarán modelos de regresión lineal múltiple, cambiando las variables y haciendo transformaciones fundamentadas en los diagnósticos anteriores, para buscar en qué medida las variables explican el porcentaje de deserción o de aprobación en cada curso.

6. Conclusiones

Según los análisis realizados hasta el momento, se puede decir que se posee una cantidad adecuada de observaciones, ya sea de estudiantes individuales o de cursos, y con una variabilidad apreciable, para poder utilizar algoritmos predictivos o técnicas explicativas, si los datos se agrupan adecuadamente.

Referencias

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

[2] 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.

[3] 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.