# Taller de tesis I: Entrega 1

Edgardo Palazzo

21 de abril de 2023

Para el trabajo de especialización voy a analizar las bases de datos de estudiantes de la cátedra de física del CBC, que contienen un historial del desempeño de cada estudiante en las materias de la cátedra, y no contienen información personal. Los datos fueron provistos en distintas tablas de MS Access con diferentes estructuras que ya fueron unificadas en un archivo csv. Luego de un trabajo de limpieza y estandarización preliminar, cuento con alrededor de 230 mil registros, cada uno representando a un estudiante en cada cuatrimestre, para los años desde 2011 hasta 2019. La información en cada registro es la siguiente: año, cuatrimestre, DNI, nombre y apellido, comisión, horario, aula, sede, materia, carrera, notas de parciales y finales.

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 información externa a la institución[1], consideradas como variables endógenas de los estudiantes, especialmente las características socioeconómicas de los mismos[2]. Sin embargo, 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, analizando solo variables relacionadas con el entorno en el que cada estudiante cursa las materias de las que se dispone información, planteando la siguiente pregunta: **¿es posible determinar si el nivel de deserción o el rendimiento académico están influenciados por características de los cursos que se puedan deducir de los datos mencionados?**

Actualmente estoy finalizando la limpieza de los datos, y realizando un análisis exploratorio para decidir sobre posibles imputaciones y la generación de nuevas variables. En la ingeniería de variables que seguirá a este análisis se van a generar variables tales como: cantidad de alumnos en el curso, grupo de edad (estimada a partir del DNI), si es recursante, si es extranjero, entre otras, algunas a nivel individual y otras calculadas por grupos. También se van a generar las variables objetivo, que pueden ser a nivel estudiante como sería la categórica si abandonó antes de rendir los parciales, o numérica con la nota final. Pero también puede ser a nivel curso, calculando el porcentaje de deserción o el de aprobación, o la nota final promedio.

Para responder la pregunta planteada voy a explorar dos caminos. Por un lado realizar regresiones lineales múltiples o logísticas para determinar si existen variables que expliquen algunas de las características expresadas en las variables objetivo.

Por otro lado, utilizar XGBoost para realizar predicciones sobre deserción o alguna medida de desempeño académico. De ser posible entrenar un algoritmo con un rendimiento que sea mejor que el azar, indicaría que algunas de esas variables, identificadas por nivel de importancia según el algoritmo, están influenciando en el rendimiento de los estudiantes.

## Referencias


[1] 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.

[2] 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.

[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.