Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Modelos de Aprendizaje Automático para identificación de estudiantes en riesgo de abandono en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires

Tesis para la Licenciatura en Ciencias de Datos

Sol Anabella Calloni

Director: Mtr. Martín Pustilnik
Codirector: Dr. Guillermo Durán
Buenos Aires, 2025

RESUMEN

En promedio, 19.96 % de los inscriptos al Ciclo Básico Común (CBC) en la Facultad de Ciencias Exactas y Naturales (FCEN) de la Universidad de Buenos Aires (UBA) egresan, mientras que en el Sistema Universitario Argentino egresa el 23.06 %.

Entendemos que el abandono estudiantil es, tal vez, el factor individual más importante que explica estos porcentajes de egreso.

Con el fin de enfocarse en la emisión de alertas tempranas en lugar de identificar un "abandono definitivo", se definió un umbral de nivel de actividad de los estudiantes por semestre a partir del cual se los considera en "riesgo de abandono".

Entre las acciones para prevenir el abandono estudiantil, la FCEN realiza encuestas a ingresantes desde 2002, inicialmente en papel y, a partir de 2023, en formato digital con identificación nominal, lo que permite vincular las respuestas con datos del CBC y de las carreras. Estas encuestas sustentan un sistema de tutorías dentro del programa +Acompañamiento[1]. Asimismo, desde 2009, la FCEN cuenta con el "Programa Ingresantes CBC Exactas", que ofrece una Charla de Bienvenida, un Curso Previo de Matemática (CPM) y tutorías docentes para estudiantes del CBC de las carreras de la facultad.

En este trabajo implementamos modelos de Aprendizaje Automático basados en los datos del Sistema de Información Universitaria Guaraní (SIU-Guaraní) del Ciclo Básico Común (CBC) y del SIU-Guaraní del FCEN, con la perspectiva teórica de autores de referencia y la de otros actores de la misma universidad.

Una vez entrenados, son capaces de detectar estudiantes con alto riesgo de abandono, a la vez que permiten indagar en algunos de los motivos subyacentes.

Se realizó una investigación bibliográfica de los modelos empleados hasta la fecha, haciendo foco en aquellos que utilizaran Aprendizaje Automático. Luego, se desarrollaron modelos que proporcionan alertas tempranas de abandono en el contexto del FCEN, para poder intervenir y asistir a las personas antes de que abandonen. Se encontraron variables cuyo abandono condicional era significativamente distinto al abandono poblacional, por lo que se las podría utilizar para mejorar futuros modelos.

Se utilizaron métricas como Exactitud, área bajo la curva ROC (AUC ROC) y Exactitud Balanceada para medir el rendimiento de los modelos, alcanzando 0.845 de Exactitud Balanceada para el mejor de ellos.

Palabras clave: "Aprendizaje Automático", "Random Forest", "Riesgo de abandono", "Alerta temprana de riesgo de abandono"

[1] exactas.uba.ar/acompanamiento/

ABSTRACT

On average, 19.96 % of students who enroll in the Ciclo Básico Común (CBC) at the Facultad de Ciencias Exactas y Naturales (FCEN) of the Universidad de Buenos Aires (UBA) go on to graduate, compared with a 23.06 % graduation rate across the Argentine University System.

We understand that student dropout is, perhaps, the most significant individual factor explaining this phenomenon.

To focus on issuing early warnings rather than identifying "definitive dropout", a threshold for student activity level per semester was defined, from which a student is considered at "risk of dropping out".

Among the actions to prevent student dropout, FCEN has been conducting surveys for incoming students since 2002, initially on paper and, since 2023, in a digital format with nominal identification, which allows linking the responses with data from the CBC and degree programs. These surveys support a tutoring system within the +Acompañamiento program[2]. Likewise, since 2009, FCEN has had the "Programa de Ingresantes CBC Exactas", which offers a Welcome Talk, a Preliminary Mathematics Course (CPM), and a faculty tutoring system for CBC students of the faculty's degree programs.

In this work, we implement Machine Learning models based on data from the Sistema de Información Universitaria Guaraní (SIU-Guaraní) for the CBC and the SIU-Guaraní for FCEN, incorporating the theoretical perspective of reference authors and that of other actors from the same university.

Once trained, these models are capable of detecting students at high risk of dropping out, while also allowing for an investigation into some of the underlying reasons.

A bibliographic investigation of the models used to date was carried out, focusing on those that utilized Machine Learning. Subsequently, models were developed that provide early dropout warnings in the context of FCEN, in order to intervene and assist individuals before they drop out. Variables were found whose conditional dropout was significantly different from the population dropout, and thus could be used to improve future models.

Metrics such as Accuracy, area under the ROC curve (AUC ROC), and Balanced Accuracy were used to measure the performance of the models, achieving a Balanced Accuracy of 0.845 for the best of them.

Keywords: "Machine Learning", "Random Forest", "Dropout Risk", "Early Dropout Risk Warning"

[2] exactas.uba.ar/acompanamiento/

AGRADECIMIENTOS

En primer lugar, les quiero agradecer a mis directores, el Mtr. Martín Pustilnik y el Dr. Guillermo Durán. Su guía, paciencia y dedicación fueron fundamentales a lo largo de este trabajo. Les agradezco por su ayuda, por resolver cada una de mis dudas con claridad y por su rol crucial en la gestión de los datos que hicieron posible esta tesis.

También quiero agradecer muy especialmente a Claudia Zelzman (Directora de la Dirección de orientación Vocacional de Exactas y a cargo de las tutorías del CBC) y Luciana Fager (coordinadora del Programa +Acompañamiento). Desde que conocieron el proyecto, su disposición fue excelente. Aunque los datos de las encuestas no pudieron ser incorporados en este trabajo, valoro y admiro el enorme trabajo que hacen cada día para acompañar a los estudiantes.

Asimismo, agradezco a Felipe Vega Terra (Director del CBC), Marta Hughes (Secretaría de Planificación del CBC), Guido Rodriguez Migueres (Director de Alumnos de la FCEN) y Pablo Mislej. Su colaboración fue esencial para comprender la estructura de los datos y facilitar el acceso a los mismos. Una mención especial para Mariana Baleani y, muy particularmente, para Gonzalo Monasterio, por el meticuloso trabajo de descarga y armado de las bases de datos, realizado con una paciencia y atención al detalle que fueron determinantes para el éxito de este análisis.

Quiero agradecerle también a la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. Gracias a la universidad pública, gratuita y de calidad, tuve la oportunidad de formarme, y en la FCEN encontré un espacio donde siempre me sentí cómoda y motivada para aprender más. Es tanto lo que me ha dado que quise que esta tesis tuviera la posibilidad de retribuirle.

A mi familia, gracias por el apoyo incondicional, por el aliento en los momentos de frustración y por darme la fuerza necesaria para dedicarme a mis estudios durante todos estos años. Gracias también a Delta, cuya presencia fue clave para mantener la cordura y el foco a lo largo de la carrera y en el tramo final de esta tesis.

Por último, a mis amigos, quienes hicieron que la carrera valiera la pena de una forma completamente distinta. A mis compañeras de Álgebra y Análisis I, Lu y Valen, las "Chicas Superpoderosas", y a Marti, para formar juntos a "Las mariposas". A Agus, que con una invitación a almorzar nos sumó a su grupo "Ponerse al día". Y a Goico, que nos presentó a Pedro para terminar de formar a "Las mariposas al día". Les agradezco por cada llamada de estudio, por los mates, por el aliento en los momentos difíciles y por recordarme siempre que había vida más allá de la facultad. Haber transitado estos años con ustedes (Juani, Delfi, Joaco incluidos) no solo lo hizo posible, sino inolvidable.

Índice general

1. INTRODUCCIÓN

1.1. Motivación

En el Sistema Universitario Argentino, el promedio de egreso en tiempo teórico es del 23.06 % (Tabla 1.1), mientras que en la Universidad de Buenos Aires (UBA) este valor es del 23.17 % (Tabla 1.2). En la Facultad de Ciencias Exactas y Naturales (FCEN) de la UBA, el porcentaje de egresados es aún menor, alcanzando solo el 19.96 % de los estudiantes inscriptos en el Ciclo Básico Común (CBC) para alguna de las carreras de la facultad (Tabla 1.3). Entendemos que el abandono estudiantil es, tal vez, el factor individual más importante que explica estos porcentajes de egreso.

La UBA se fundó en 1821 y desde 1949 es pública y gratuita con gran integración con la comunidad y alto interés por la permanencia de sus estudiantes. Sin embargo, las estadísticas reflejan que el promedio de egreso es de 23.17 %, evidenciando la necesidad de abordar el abandono de los estudios universitarios.

Los estudiantes de la FCEN inscriptos en los años 2021 y 2022, en su mayoría (52.26 %) tienen entre 18 y 20 años al momento de inscribirse y el mayor porcentaje (63.10 %) se corresponde con el sexo masculino. A su vez, encontramos que la media del tiempo de viaje en transporte público desde su domicilio a la facultad es de una hora.

Entre las acciones para abordar la prevención del abandono, la FCEN ha implementado diversas iniciativas. Desde 2023, el programa +Acompañamiento [14] ofrece un sistema de tutorías basado en encuestas enviadas a los ingresantes, permitiendo identificar estudiantes en riesgo y brindarles apoyo personalizado. Asimismo, desde 2009, el "Programa Ingresantes CBC Exactas" proporciona una Charla de Bienvenida, un Curso Previo de Matemática (CPM) y tutorías docentes para estudiantes del CBC que cursan carreras de la facultad. Estas acciones reflejan el compromiso institucional con la retención estudiantil.

En la literatura, los modelos de Aprendizaje Automático han demostrado ser herramientas efectivas para identificar variables predictoras del abandono estudiantil.

1.1.1. Sistema Universitario Argentino

La información presentada a continuación proviene de la "Síntesis de Información Universitaria" [36] elaborada por la Subsecretaría de Políticas Universitarias de Argentina, que recoge y analiza los datos más recientes sobre el sistema universitario del país.

Hasta el año 2024, el Sistema Universitario Argentino está compuesto por ciento cuarenta y dos instituciones, de las cuales ciento veinte son universidades[1] y veintidós son institutos universitarios[2]. Dentro de las mismas sesenta y cuatro son instituciones de tipo estatal nacional, ocho estatal provincial, una internacional y el resto son privadas.

La educación superior en Argentina se caracteriza por una fuerte presencia del sector público. En el año 2023 el 78.4 % de los estudiantes de pregrado y grado asistían a universidades estatales, consolidando su papel central en la formación académica del país.

[1] Donde [36] define universidad como "institución universitaria que desarrolla su actividad en una variedad de áreas disciplinarias no afines, orgánicamente estructuradas en facultades, departamentos o unidades académicas equivalentes."

[2] Donde [36] define instituto universitario como "institución universitaria que circunscribe su oferta académica a una sola área disciplinaria."

Mientras que en todas las provincias hay al menos una institución universitaria estatal, las instituciones privadas se concentran en las zonas de mayor densidad poblacional, como la Ciudad Autónoma de Buenos Aires.

Para el año 2023, el Sistema Universitario Argentino registró 2,746,768 estudiantes, con 759,718 nuevos inscriptos y 160,117 egresados en los niveles de pregrado, grado y posgrado. A pesar del crecimiento sostenido en la cantidad de inscriptos en la última década, sólo el 23.3 % de los estudiantes logra graduarse en el tiempo teórico estipulado para su carrera[3].

En la Tabla 1.1 se presenta la cantidad de nuevos estudiantes[4] y egresados de pregrado y grado por año. El porcentaje presentado por año es una estimación de los egresados en tiempo teórico de cinco (5) años. No conocemos el valor exacto pero usamos esta variable como estimador y creemos que es fiel a la realidad porque en cada año se suman egresados de tiempos anteriores y se suman a años posteriores.

Año	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Promedio
Nuevos	423,920	425,415	445,763	458,565	489,701	516,305	547,661	596,446	641,929	710,699	719,699	713,467	557,464.17
Egresados	110,360	117,719	120,631	124,960	124,674	125,328	132,744	135,908	122,679	142,826	145,728	139,182	128,561.58
Porcentaje (5 años)	-	-	-	-	29.41 %	29.46 %	29.78 %	29.64 %	25.05 %	27.66 %	26.61 %	23.33 %	**23.06 %**

Tab. 1.1: Datos de los alumnos de las universidades argentinas: Nuevos (Nuevos Inscritos), Egresados y Porcentaje (Porcentaje estimado de egresados en tiempo teórico de 5 años) para las universidades Argentinas (2012-2023). Fuente: Elaboración propia a partir de estadísticas publicadas en [37, 36].

En la misma, se puede observar cómo en ninguno de los años se llega al 30 % de egresados en tiempo teórico. Desde 2016 hasta 2019 el porcentaje se mantiene cerca del 29 %, luego de la pandemia en 2020 ese numero disminuyó. *Si se hiciera una regresión lineal de la evolución de la cantidad de egresados, podríamos predecir que el porcentaje se va a mantener cercano a dichos valores.*

1.1.2. Universidad de Buenos Aires (UBA)

Una de las instituciones universitarias públicas más importantes de Argentina es la Universidad de Buenos Aires (UBA), fundada en 1821, se encuentra dividida en trece facultades [5]. Una de ellas es la Facultad de Ciencias Exactas y Naturales (FCEN), en donde se dictan las carreras: 'Ciencias Biológicas', 'Ciencias de la Atmósfera', 'Ciencias de la Computación', 'Ciencias Físicas', 'Ciencias Geológicas', 'Ciencias Matemáticas', 'Ciencias Químicas', 'Ciencia y Tecnología de Alimentos', 'Paleontología', 'Oceanografía', 'Ciencias de Datos' y 'Profesorados en Ciencias', lo cual incluye: 'Ciencias de la Atmósfera', 'Biología', 'Computación', 'Física', 'Geología', 'Matemática' y 'Química' [10].

Todas las carreras de la UBA tienen un primer año obligatorio llamado el Ciclo Básico Común (CBC). El mismo, busca brindar una formación básica integral, dura generalmente

[3] "Egreso en tiempo teórico: Porcentaje de egresadas/os que completan sus carreras de grado en el tiempo previsto por el plan de estudio. Para el cálculo de este indicador se tomó el promedio de la "duración teórica"de todas las ofertas académicas de grado vigentes (5 años). Este indicador permite una aproximación al conocimiento de la proporción de las /os estudiantes que finalizan sus estudios en los plazos establecidos."[36]

[4] "Nuevas/os Inscriptas/os: Las/os nuevas/os inscriptas/os son las/os estudiantes que ingresan por primera vez en una oferta académica. Componen esta población, las /os que por primera vez ingresan a una determinada oferta habiendo cumplido con los requisitos administrativos y académicos establecidos por cada institución; y las/os nuevas/os inscriptas/os por equivalencia, es decir, aquellas/os que se inscriben por primera vez en la oferta, pero con materias aprobadas "por equivalencia"de otra oferta (en la misma institución u otra institución)."[36]

un año (dos cuatrimestres), y consta de seis asignaturas: dos comunes a todas las carreras, dos de orientación según el área de estudio y dos específicas de la carrera elegida [35]. Se puede cursar hasta tres materias por cuatrimestre, presencialmente o a distancia mediante UBA XXI. Además, UBA XXI ofrece en verano e invierno cursos intensivos de algunas de las materias del CBC [39].

La UBA utiliza el Sistema de Información Universitaria - Guaraní (SIU-Guaraní), un sistema de gestión académica que posibilita el registro de inscripciones a materias y exámenes de los estudiantes. El mismo también contiene información personal, como lo es situación laboral o lugar de vivienda.

Se han realizado una serie de trabajos que estudian el abandono de los estudios en la UBA, como la tesis de Jenik "*Characterization of procrastination patterns of university students in academic subjects*" 2015 [16]. En su trabajo, Jenik estudia la "procrastinación" en las carreras de la FCEN, donde la procrastinación la define como la cantidad de cuatrimestres que el estudiante demora en rendir el examen final de cada materia. Además, "Predicción de abandono en ingresantes a las carreras de la Facultad de Ciencias Exactas y Naturales" [31] es otro trabajo en donde se busca entrenar un modelo usando resultados de encuestas realizadas a ingresantes de diversas carreras de la FCEN. Por último, en la tesis "Discontinuar los estudios en la Universidad" [30] se analizan una serie de entrevistas a personas que tomaron la decisión de abandonar alguna carrera de la UBA, entre ellos estudiantes de Ciencias Químicas.

Según Tinto (1982) [33] la congruencia normativa es un factor importante en el abandono. En este sentido, la FCEN tiene una organización diferente a otras facultadas, está basada en departamentos en lugar de cátedras, lo que influye en la dinámica académica y administrativa. Otro factor relevante es la política de la FCEN respecto a la validez de los trabajos prácticos aprobados, que tienen una vigencia de ocho cuatrimestres [11], lo que permite a los estudiantes hasta cuatro años para rendir sus exámenes finales. Esta política contrasta con otras facultades de la UBA, como Medicina, donde los estudiantes tienen dos ciclos lectivos para rendir [18]. Como se muestra en el trabajo de Jenik (2015) [16], que los estudiantes puedan rendir los exámenes finales hasta cuatro años después de haber cursado la materia, puede aumentar el abandono en algunos casos.

Además, como se menciona en la tesis "Discontinuar los estudios en la Universidad" [30], la FCEN se encuentra en Ciudad Universitaria, una ubicación geográfica distinta al resto dentro de la Ciudad Autónoma de Buenos Aires, lo que también puede afectar la experiencia estudiantil y, en consecuencia, al abandono.

A su vez, en la FCEN es muy común que los estudiantes completen todas las materias necesarias para recibirse, pero no lleven a cabo la tesis de licenciatura. Este fenómeno no es exclusivo de una carrera en particular, sino que es un problema generalizado en la facultad. En el caso de la Licenciatura en Ciencias de Datos, una carrera relativamente nueva, se tomó la decisión de incorporar la tesis como parte de una materia dentro de un cuatrimestre, con el objetivo explícito de combatir este problema y facilitar que los estudiantes finalicen sus estudios de manera más efectiva.

Estas diferencias organizativas y académicas hacen que la FCEN sea un caso particular dentro de la UBA, lo que justifica un análisis específico para entender y predecir el abandono en esta facultad.

No contamos con datos de la UBA anteriores a 2017, pero podemos ver que el promedio de egreso se asemeja al nacional (Tabla 1.2). No podemos saber si estos números son consecuencia de la pandemia o es una tendencia que se mantiene, igualmente es muy bajo

23.17 % de egresados en tiempo teórico.

Año	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Promedio
Nuevos	-	-	-	-	-	64,955	71,463	73,642	83,912	91,681	80,592	64,881	75,875.14
Egresados	-	-	-	-	-	18,176	16,476	17,297	7,718	18,903	25,929	18,571	17,581.43
Porcentaje (5 años)	-	-	-	-	-	-	-	-	-	29.10 %	36.28 %	25.22 %	**23.17 %**

Tab. 1.2: Datos de los alumnos de la UBA: Nuevos (Nuevos Inscritos al CBC), Egresados y Porcentaje (Porcentaje de egresados en tiempo teórico de 5 años) para la UBA. Fuente: Elaboración propia a partir de estadísticas publicadas en [34].

Si analizamos la tabla de los estudiantes del CBC para carreras de la FCEN (Tabla 1.3), se observa que los ingresantes representan en promedio un 3 % del total de ingresantes a la UBA. Como las carreras de esta facultad pueden durar más de cinco años si consideramos un año del CBC, se tomó la decisión de presentar el porcentaje de egresados pensando en el caso donde los estudiantes tardan cinco o seis años en concluir su carrera de grado. Además, en la Tabla 1.3 se puede observar que el porcentaje de egreso de la facultad en el año 2023 es mucho menor al que se tiene en la métrica de la UBA en general.

Año	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Promedio
Nuevos	1717	1595	1656	1679	1773	1608	1655	2374	2718	3425	3684	3204	2257.33
Egresados	-	-	-	-	-	507	412	547	146	763	421	358	450.57
Porcentaje (5 años)	-	-	-	-	-	31.79 %	24.88 %	32.58 %	8.23 %	47.45 %	25.44 %	15.08 %	**19.96 %**
Porcentaje (6 años)	-	-	-	-	-	29.53 %	25.83 %	33.03 %	8.70 %	43.03 %	26.18 %	21.63 %	**19.96 %**

Tab. 1.3: Datos de los alumnos inscriptos al CBC de la FCEN: Nuevos (Nuevos Inscritos al CBC de alguna carrera de la FCEN), Egresados, Porcentaje de egresados en tiempo teórico de 5 años y 6 años para la FCEN. Fuente: Elaboración propia a partir de estadísticas publicadas en [21, 34].

Año	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Promedio
Nuevos	994	976	841	828	798	890	831	935	672	1951	1230	1448	1032.8
Egresados	-	-	-	-	-	507	412	547	146	763	421	358	450.57
Porcentaje (5 años)	-	-	-	-	-	51.95 %	48.99 %	66.06 %	18.30 %	85.73 %	50.66 %	38.29 %	**43.63 %**

Tab. 1.4: Datos de los alumnos de la FCEN: Nuevos (Nuevos Inscritos), Egresados, Porcentaje de egresados en tiempo teórico de 5 años y 6 años para la FCEN. Fuente: Elaboración propia a partir de estadísticas publicadas en [21, 34].

Por otro lado, se desarrolló la Tabla 1.4 donde se presenta la cantidad de ingresantes a la FCEN, subconjunto de la Tabla 1.3. Esto implica que se consideran solamente a aquellas personas que se encuentran en condición de comenzar la carrera, para lo cual deben tener aprobadas todas las materias del CBC. En este caso, el promedio de egreso aumenta a 43,63 %. Sin embargo, al analizar el promedio de ingresantes de ambas tablas, queda a la vista que hay un alto porcentaje de personas, aproximadamente 55 %, que no logra superar las materias del CBC.

En las Tablas 1.3 y 1.4 se observa una caída en la cantidad de egresados en el 2020 con un posterior aumento en el 2021, lo cual podría atribuirse a la pandemia. En los años 2022 y 2023 cae el porcentaje de egresados y no podemos saber si es una tendencia que se mantendrá o es consecuencia de la pandemia.

A su vez, Zelzman (2022) en su tesis de maestría [40] plantea un índice de continuidad (IC) CBC carreras FCEN (Ecuación 1.1) con el objetivo de cuantificar la proporción de estudiantes que se inscribieron al CBC de alguna carrera de la FCEN y logran ingresar a la facultad.

$$\text{IC} = \frac{\sum_{i=2}^{4} \text{número de ingresantes a la FCEN}_i}{\sum_{i=1}^{3} \text{número de ingresantes al CBC}_i} \times 100 \tag{1.1}$$

Donde:
"el subíndice refiere al año considerado, donde para el CBC se toman tres años consecutivos, y para la FCEN lo mismo, corrido un año. Es decir, este cociente toma un lapso de 3 años, en tanto se estima como plazo máximo demandado para la cursada y aprobación del CBC"[40].

Si tomamos el complemento, estaríamos calculando la proporción de estudiantes que se inscribieron al CBC de alguna carrera de la FCEN pero no ingresaron a la facultad (Ecuación 1.2).

$$100 - \text{IC} = (1 - \frac{\sum_{i=2}^{4} \text{número de ingresantes a la FCEN}_i}{\sum_{i=1}^{3} \text{número de ingresantes al CBC}_i}) \times 100 \tag{1.2}$$

Donde:
IC se encuentra definida en la Ecuación 1.1.

Usando este índice obtenemos IC 2020-2022 = 47.10 %, es decir estaríamos estimando que 52.9 % de estudiantes del CBC no ingresaron a la FCEN.

1.2. Uruguay

Se puede considerar el caso de Uruguay que presenta un sistema universitario comparable al argentino, pues se encuentran dos universidades públicas de acceso gratuito, cinco universidades privadas y diez institutos universitarios privados. Sin embargo, tiene la particularidad de que hace unos años las universidades no se encontraban distribuidas a lo largo del territorio del país por lo que en los últimos tiempos se estuvieron implementando medidas para poder ampliar el territorio cubierto por las mismas. Se completó la misma estructura de tabla para los datos de estudiantes universitarios de Uruguay (Tabla 1.5), basándonos en los datos del Ministerio de Educación y Cultura del país.

Año	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Promedio
Nuevos	23,275	27,442	29,064	29,808	31,879	33,509	35,419	37,787	37,560	45,980	43,050	44,249	34,918.50
Egresados	7,826	8,034	7,205	7,804	8,703	8,825	7,969	7,721	7,602	8,775	8,812	8,809	8,359.00
Porcentaje (5 años)	-	-	-	-	37.39 %	32.16 %	27.42 %	25.90 %	23.85 %	26.19 %	24.88 %	23.31 %	**23.94 %**

Tab. 1.5: Datos de los alumnos de las universidades uruguayas: Nuevos (Nuevos Inscritos), Egresados, Porcentaje de egresados en tiempo teórico de 5 años para Uruguay. Fuente: Elaboración propia a partir de estadísticas publicadas en [12].

Al analizar la Tabla 1.5, en promedio egresan 23.94 %, valor muy similar al de la Argentina. Si se compara el porcentaje de egresados por año, notamos que en general en la Argentina, a partir de 2018, el porcentaje es superior.

1.3. ¿Qué hace actualmente la FCEN?

Entre las acciones para prevenir el abandono estudiantil, la FCEN implementa desde 2002 encuestas a ingresantes, inicialmente en formato papel, enfocadas en aspectos vocacionales y trayectorias educativas. A partir de 2018, estas encuestas pasaron a ser digitales,

incorporando preguntas sobre aspectos socioeconómicos. Desde 2023, las encuestas son nominales y forman parte del programa +Acompañamiento [14], que incluye un sistema de tutorías basado en las respuestas obtenidas. A cada una de las respuestas se les asigna un puntaje y rangos de riesgo de abandono, lo que llaman "semáforo". De esta manera, en caso de que el estudiante decidiera dejar sus datos, se le ofrece el servicio de tutoría considerando los resultados del semáforo. A partir de las encuestas a los ingresantes, López y Rosenfeld, Spognardi y Cerdeiro desarrollaron un trabajo [31], en donde entrenaron modelos con el objetivo de identificar a los estudiantes que abandonarían alguna de las materias donde se realizaron las encuestas.

A su vez, para combatir el abandono del CBC de la carrera, desde 2009 la FCEN cuenta con el "Programa Ingresantes CBC Exactas"donde ofrecen una Charla de Bienvenida, un Curso Previo de Matemática (CPM) y un sistema de tutorías docentes a estudiantes del CBC de alguna carrera de la facultad.

1.4. Organización del trabajo

El presente trabajo se realizó en un lapso de seis meses, en el contexto de la materia "Taller de Tesis"para la Licenciatura en Ciencias de Datos de la Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Buenos Aires. El mismo se centra en la predicción de riesgo de abandono en la misma facultad y ha sido estructurado para facilitar su comprensión y seguimiento.

En la Sección 2, "Objetivos de Tesis", se establecen los objetivos principales y específicos del trabajo, que incluyen el desarrollo de un modelo de Aprendizaje Automático para identificar estudiantes en riesgo de abandono que ayuden a la facultad en la implementación de medidas de retención.

La Sección 3, "Estado del Arte", presenta una revisión de la literatura sobre el abandono estudiantil, con énfasis en modelos previos y enfoques de Aprendizaje Automático. También se define el concepto de "riesgo de abandono"basado en un umbral de actividad académica, fundamentado en datos de la FCEN.

En la Sección 4, "Metodología", se describen las fuentes de datos utilizadas, provenientes del SIU-Guaraní del CBC y de la FCEN, junto con los algoritmos seleccionados y las métricas empleadas para evaluar el desempeño de los modelos.

La Sección 5, "Análisis de Datos", explora los datos recopilados para identificar patrones relacionados con el abandono, incluyendo un análisis condicional que examina factores como la edad, la situación laboral y el nivel educativo de los padres.

En la Sección 6, "Modelado", se detalla el preprocesamiento de datos y la ingeniería de atributos, como la creación de variables derivadas del tiempo de viaje, la edad de inscripción y los clusters de materias del CBC, para optimizar el entrenamiento de los modelos.

La Sección 7, "Experimentos", describe los experimentos realizados con modelos de *Random Forest*, comparando configuraciones con y sin bootstrap, y presenta los resultados obtenidos en términos de exactitud, exactitud balanceada y AUC ROC.

La Sección 8, "Conclusión", sintetiza los principales hallazgos del trabajo, destacando la efectividad de los modelos para predecir el riesgo de abandono y su potencial para apoyar las estrategias de retención de la FCEN.

Finalmente, la Sección 9, "Trabajos Futuros", propone líneas de investigación adicionales, como la integración de datos de encuestas, la implementación de reportes perso-

nalizados y la exploración de otros modelos de Aprendizaje Automático para mejorar la predicción del abandono.

2. OBJETIVOS DE TESIS

2.1. Objetivo general

El objetivo general del trabajo consiste en desarrollar un modelo de Aprendizaje Automático que permita identificar alumnos de la FCEN en riesgo de abandono. Adicionalmente suministrar la importancia de cada variable.

Esto le daría a la facultad la posibilidad de implementar medidas que fomenten la permanencia de los alumnos en los estudios, basados en su probabilidad de abandono. Además, al estudiar la importancia de variables para el modelo se podrían identificar medidas específicas de ayuda.

2.2. Objetivo específicos

1. Compilar una base de datos a partir de la información contenida en el SIU-Guaraní de estudiantes del CBC anotados para alguna carrera de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires y de estudiantes de dicha facultad.

2. Preprocesar los datos para poder utilizarlos en el entrenamiento de modelos de aprendizaje automático y obtener variables generadas a partir de las variables originales.

3. Entrenar y comparar los modelos para la predicción de alarma de abandono en el contexto de la UBA.

4. Identificar y recomendar un conjunto de variables críticas para modelar el fenómeno del abandono, extrayendo las más correlacionadas del conjunto actual de datos y sugiriendo nuevas variables relevantes, identificadas a partir de la literatura y estudios previos.

5. Implementar un sistema de reportes para identificar a las personas en riesgo de abandono, destacando específicamente las variables más relevantes que influyen en cada caso particular. Este reporte servirá como punto de partida para que el personal de la UBA tome la iniciativa de contactar a los estudiantes y brindarles el apoyo necesario.

3. ESTADO DEL ARTE

3.1. Alarma de abandono

En este trabajo, se utiliza el concepto de "semestre relativo" para analizar la trayectoria académica de los estudiantes, dado que cada uno puede comenzar sus estudios en la facultad en años distintos o incluso en momentos diferentes dentro de un mismo año (primer o segundo semestre). Así, el semestre 0 corresponde al primer semestre en el que se registra actividad académica, lo que permite comparar sus avances a lo largo del tiempo, independientemente de cuándo iniciaron. Por ejemplo, para un estudiante puede ser el primer semestre de 2021, mientras que para otro puede ser el segundo semestre de 2022.

En la Figura 3.1 se presenta el caso de las cohortes de 2021 a modo de ejemplo. Una explicación detallada del cálculo del semestre relativo se presenta en la Sección 6.1.1.

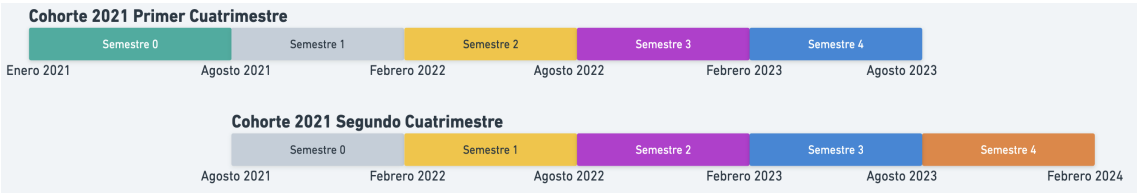


Fig. 3.1: Definición de semestre relativo de los estudiantes inscriptos en la FCEN en el año 2021. Fuente: Elaboración propia.

A partir de los datos de la FCEN elaboramos el Gráfico 3.2 que resume la actividad estudiantil de todas las cohortes para los semestres relativos 0 a 4 (Ecuación 3.1). Encontramos que la cantidad de alumnos con actividad = 0 aumenta a medida que aumenta el semestre relativo (decrece la actividad), y que la actividad para los alumnos con actividad positiva decrece en la misma dirección. Esto nos llevó a elaborar un umbral para encender la alarma de abandono antes de que la actividad llegue a 0.

La Ecuación 3.1 define la actividad de cada alumno j para el semestre número i.

$$TotalDeActividad(alumno_j, semestre_i) = \#inscripciones + \#TPsAprobados + \#finales \quad (3.1)$$

Donde:
#inscripciones = "cantidad de inscripciones a materias del $alumno_j$ en el $semestre_i$"
#TPsAprobados = "cantidad de materias aprobadas por el $alumno_j$ en el $semestre_i$"
#finales = "cantidad de inscripciones a exámenes finales del $alumno_j$ en el $semestre_i$"

Mas adelante utilizaremos estas y otras variables (actividad en semestres anteriores) para el modelo, a fin de predecir la actividad en el semestre siguiente.

Establecemos si un alumno se encuentra en "riesgo de abandono" a partir de su actividad total por semestre.

En el contexto de esta tesis, un alumno se encuentra en riesgo de abandono si no consigue sumar 3 en el semestre que queremos predecir, dada la Ecuación 3.1 (Ecuación 3.2). En trabajos futuro se podrían estudiar otros umbrales para la definición de riesgo de abandono.

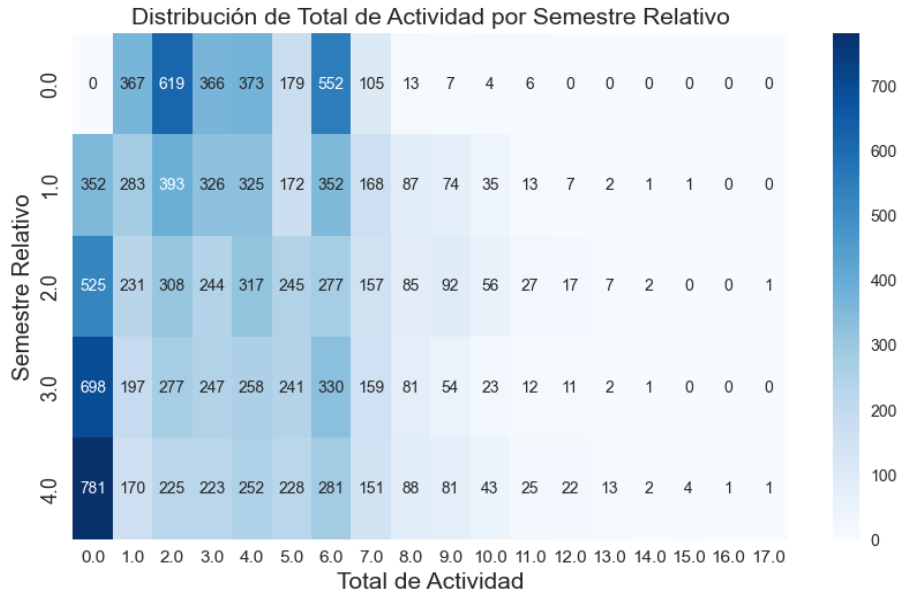


Fig. 3.2: Actividad estudiantil por semestre relativo. Fuente: Elaboración propia a partir de datos de la FCEN.

$$RiesgoDeAbandono(alumno_j, semestre_i) = \begin{cases} 1 & \text{si } TotalDeActividad(alumno_j, semestre_i) < 3 \\ 0 & \text{si } TotalDeActividad(alumno_j, semestre_i) \geq 3 \end{cases} \quad (3.2)$$

Donde:
$TotalDeActividad(alumno_j, semestre_i)$ se encuentra definida en la Ecuación 3.1

Un alumno puede tener un total de actividad equivalente a 3 mediante alguno de estos cuatro eventos:

- Inscribirse a dos materias y aprobar una
- Inscribirse a tres exámenes finales, independientemente de si los aprueba
- Inscribirse a tres materias pero no aprobar ninguna
- Inscribirse a una materia, aprobar los trabajos prácticos y el examen final

Las Tablas 3.1 y 3.2 resumen la información de la Figura 3.2. La Tabla 3.1 presenta por semestre relativo la cantidad de personas con $0 < actividad < 4$. Se puede observar como la cantidad decrece a medida que avanzan los semestres. La Tabla 3.2 contiene por semestre relativo la cantidad de personas con $actividad = 0$. Se puede observar como la cantidad crece a medida que avanzan los semestres.

Por cuestiones administrativas, la mayoría de los finales se contabilizan al finalizar el semestre (en el semestre siguiente). Considerando que es la misma materia, estos finales se compatibilizaran como si fueran del semestre actual.

El objetivo del presente trabajo es identificar aquellos alumnos que se encuentran en riesgo de abandono en su cuarto semestre.

Semestre	Cantidad de personas
0	1,352
1	1,002
2	783
3	721
4	618

Tab. 3.1: Cantidad de estudiantes inscriptos a la FCEN por semestre cuya actividad esta entre 1 y 3. En la misma se puede observar cómo dicho valor desciende al avanzar los semestres. Fuente: Elaboración propia a partir de datos de la FCEN.

Semestre	Cantidad de personas
0	0
1	352
2	525
3	698
4	781

Tab. 3.2: Cantidad de estudiantes inscriptos a la FCEN por semestre cuya actividad = 0. En la misma se puede observar cómo dicho valor aumenta constantemente. Fuente: Elaboración propia a partir de datos de la FCEN.

Por otro lado, es importante considerar que *el modelo busca hacer una predicción por* ***alumno (persona)****. Un alumno puede representar a más de un estudiante.* Si una persona se encuentra en más de una carrera, se podría corresponder con más de un estudiante.

3.2. Trabajos previos

Para poder comenzar el trabajo de entrenamiento de un modelo que identifique a aquellos alumnos que se encuentran en riesgo de abandono, se hizo un estudio de trabajos previos. En esta sección se presenta un resumen de los mismos, desde los primeros modelos explicativos hasta los desarrollos recientes con Aprendizaje Automático, con un enfoque particular en los trabajos relacionados con la predicción del abandono estudiantil en el contexto de la UBA y la FCEN.

3.2.1. Primeros estudios del abandono

Inicialmente, se realizaron múltiples trabajos en donde se analizaban los datos de estudiantes universitarios que abandonaron sus respectivas carreras y se los comparaba con datos de estudiantes que no lo hicieron, en búsqueda de características que los diferenciaran y establecer las razones por las cuales los alumnos abandonaban la universidad. Entre los trabajos más importantes, se encuentran los estudios realizados por Tinto [32, 33] y Pascarella [23].

Según el modelo teórico de Vincent Tinto, se pueden distinguir diferentes formas de abandono de la educación superior. Tinto enfatiza la importancia de no agrupar bajo una misma categoría conductas de abandono que son fundamentalmente diferentes. Principalmente, su modelo busca diferenciar entre:

- **Abandono por fracaso académico:** Este tipo de abandono ocurre cuando un estudiante es obligado a dejar la institución debido a un rendimiento académico

insuficiente, como malas notas o incumplimiento de las normas académicas.

- **Abandono voluntario:** Este se produce cuando un estudiante decide por propia voluntad cesar su actividad académica en la institución antes de finalizar sus estudios. Este tipo de abandono puede deberse a diversas razones que no están directamente relacionadas con el rendimiento académico, como falta de integración social o académica, incongruencia con el clima intelectual o social de la institución, cambios en los objetivos personales o profesionales, o la percepción de que existen alternativas más beneficiosas.

Además de esta distinción fundamental, Tinto también menciona que el abandono puede ser:

- **Temporal:** El estudiante puede pausar sus estudios con la intención de retomarlos en el futuro.
- **Permanente:** El estudiante no tiene la intención de regresar a la institución ni continuar su educación superior en otro lugar.
- **Transferencia a otra institución:** El estudiante abandona una institución para continuar sus estudios en otra. Tinto señala que no se debe equiparar este comportamiento con el abandono definitivo del sistema de educación superior.

Tinto argumenta que la falta de distinción entre estas formas de abandono en investigaciones previas ha llevado a hallazgos contradictorios y conclusiones engañosas. Su modelo se centra en explicar los procesos de interacción entre el individuo y la institución que conducen a estas diferentes formas de persistencia o abandono. En su trabajo posterior, también reconoce que su modelo inicial no profundiza lo suficiente en las diferencias en las trayectorias educativas de estudiantes de distintos géneros, razas y estatus socioeconómicos.

Para esta tesis es relevante entender el modelo conceptual de Tinto, Figura 3.3. El mismo plantea que varios factores pueden llevar a un estudiante a abandonar sus estudios superiores. Estos factores se interrelacionan y se influyen mutuamente a lo largo del tiempo.

Lo que se representa en la primera columna de la Figura 3.3 es que los estudiantes llegan a la universidad con una variedad de atributos, experiencias y antecedentes familiares que pueden influir en su persistencia. La interacción de esas características definen el nivel de compromiso para permanecer en la institución y finalizar la carrera universitaria, como se ve en la segunda columna. A su vez, ambas formas de compromiso se ven reflejadas en aspectos del sistema académico y social como calificaciones obtenidas, la percepción que tiene el estudiante sobre su desarrollo intelectual y el nivel de interacción con compañeros y profesores de la universidad. Estas últimas, definen la integración académica y social y generan nuevos niveles de compromiso con el objetivo final y con la institución que afectan a la decisión de abandonar.

Eventos fuera de la universidad pueden afectar la decisión de un estudiante de abandonar, aunque Tinto sugiere que estos impactos se reflejan en los cambios en el compromiso del estudiante con el objetivo de graduarse y con la institución.

Pascarella, junto con Terenzini, [23] realizaron un importante trabajo empírico para poner a prueba el modelo teórico de abandono de Tinto. Su investigación proporcionó evidencia que apoyaba la hipótesis de que la integración social y académica de los estudiantes en la institución juega un papel crucial en su decisión de persistir o abandonar.

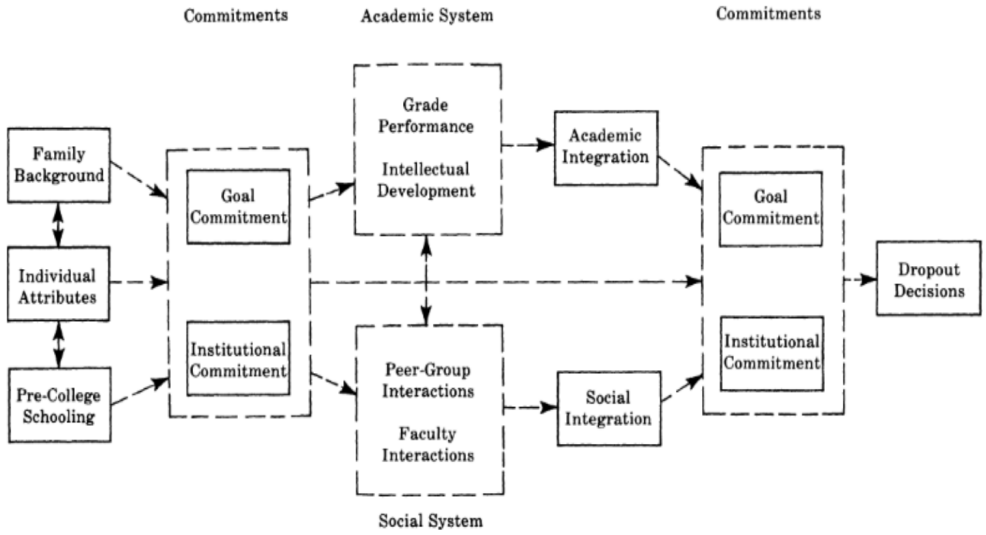


Fig. 3.3: Modelo conceptual propuesto por Tinto: el sentido de las flechas indican cómo un atributo tiene efecto sobre otro y los cuadrados agrupan los atributos que se utilizan para medir los conceptos más abstractos como compromiso, el sistema académico y social. Fuente: [32].

3.2.2. Modelo de ecuaciones estructurales

Como explican Ortiz y Fernandez-Pera en [22], un modelo de ecuaciones estructurales (SEM) es una técnica estadística multivariada que permite analizar de forma simultánea relaciones complejas entre variables observadas[1] y latentes[2]. Esto significa que, a través de diagramas "path", se pueden representar hipótesis teóricas donde variables medidas y constructos no observados se relacionan mediante efectos directos e indirectos, integrando técnicas como la regresión lineal y el análisis factorial. SEM permite, además, controlar el error de medición y realizar comparaciones entre grupos dentro de un mismo modelo. En la Figura 3.4 se puede observar el diagrama propuesto en el paper a modo de ejemplo.

Bean y Metzner adaptaron este tipo de modelos en su trabajo "*A Conceptual Model of Nontraditional Undergraduate Student Attrition*"[2] y presentan un modelo conceptual del abandono de estudiantes no tradicionales.

Se debe considerar que los SEM se suelen basar en encuestas, por lo tanto sería necesario encuestar al total de la población para obtener los mejores resultados. Como esto no siempre es posible, se utilizan técnicas estadísticas y modelos de Aprendizaje Automático.

3.2.3. Aprendizaje Automático y abandono de estudios en Argentina

El Aprendizaje Automática es un campo de la inteligencia artificial que se ocupa de diseñar y entrenar algoritmos capaces de aprender patrones directamente de los datos. Estos algoritmos pueden ser utilizados tanto para predecir resultados como para descubrir patrones ocultos.

En la actualidad, existen múltiples trabajos que desarrollan modelos de Aprendizaje Automático para la predicción de abandono en carreras universitarias.

[1] Variables que se pueden registrar de manera directa mediante instrumentos o pruebas.
[2] Variables que se infieren a partir de varios indicadores o variables observadas y no se pueden medir.

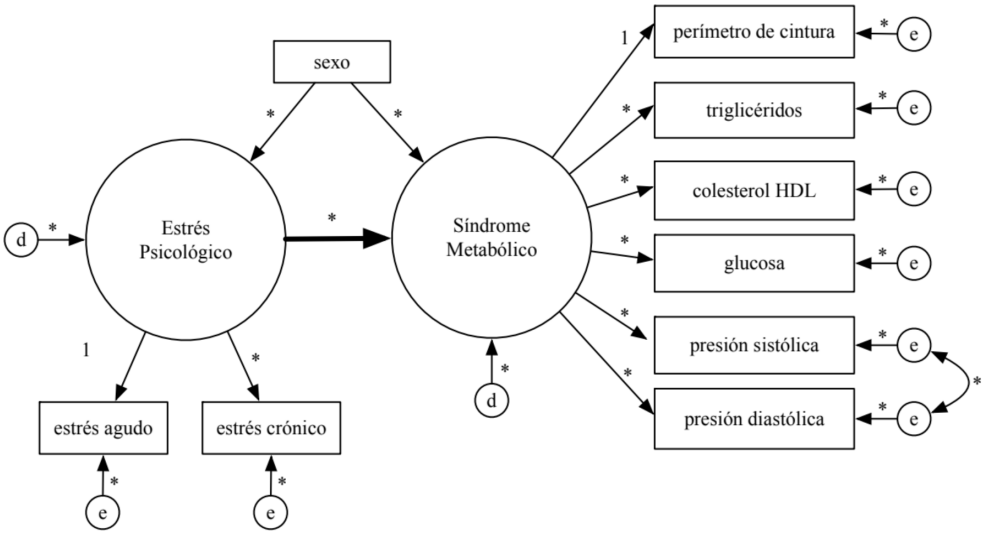


Fig. 3.4: Ejemplo de Diagrama "path" o SEM en el ámbito de las ciencias médicas y de la salud: Las variables observadas son representadas por rectángulos y comienzan en minúscula y los factores son representados por círculos y comienzan con mayúscula. Las relaciones entre variables son indicadas por flechas, donde una flecha indica una relación directa entre las dos variables, siendo aquella a la cual la flecha apunta la variable dependiente y una línea con doble flecha indica una covarianza entre las variables. Las esferas con las letras "e" y "d" se corresponden con el error de variables observadas y factores, respectivamente. Este ejemplo ilustra la estructura de un modelo SEM, aunque su contexto difiere del abandono estudiantil analizado en este trabajo. Fuente: [22].

En su trabajo García de Fanelli [15] hace un resumen de los distintos métodos de análisis del abandono en las universidades y define distintos tipos de factores que inciden sobre el éxito escolar, dando por resultado la Tabla 3.3. La misma será utilizada a lo largo de esta tesis para asignar los tipos de variables analizadas para el entrenamiento de los modelos.

Encontramos dos trabajos relativos a la UBA, ya mencionados en la introducción: "*Characterization of procrastination patterns of university students in academic subjects*" [16] y "Predicción de abandono en ingresantes a las carreras de la Facultad de Ciencias Exactas y Naturales" [31]. El primero estudia la "procrastinación" en las carreras de la facultad en estudiantes de los años 2002 a 2011, sin embargo contaba con información limitada, variables como el género se tuvieron que deducir de los nombres. Por otro lado, el segundo trabajo es más reciente pero cuanta sólo con datos de encuestas realizadas a estudiantes de los primeros años de la carrera, por lo que no considera variables como el historial académico.

En este sentido, un ejemplo relevante es el proyecto desarrollado en la Universidad Nacional de Hurlingham (UNAHUR)[3] por Pustilnik y Ndukanma [25] donde implementaron modelos de Aprendizaje Automático, como XGBoost, para predecir el abandono estudiantil a partir de datos del sistema SIU-Guaraní y censos opcionales completados por los estudiantes. Además, realizaron un análisis de datos para identificar factores de riesgo de abandono. La validación de su metodología se basó en métricas como el Área Bajo la Curva ROC (AUC ROC) y la Exactitud Balanceada, alcanzando un AUC ROC de 0,83

[3] Universidad pública y gratuita, ubicada en el conurbano bonaerense.

FACTORES INDIVIDUALES		
Demográficos	**Socioeconómicos**	**Académicos**
Sexo	Ingreso del hogar	Promedio escuela secundaria
Edad	Nivel educativo padres	Gestión pública - privada escuela secundaria
Nacionalidad- Raza	Nivel ocupacional padres	Título de la escuela media
Estado civil	Actividad económica	Horas y esfuerzo dedicados al estudio
Residencia	Cantidad de horas de trabajo	Aspiraciones y motivaciones al ingreso
Cantidad de hijos	Fuente financiamiento de los estudios	Rendimiento académico primer año
FACTORES ORGANIZACIONALES		
Políticas académicas	**Plan de estudio**	**Recursos**
Mecanismo de admisión	Duración del programa	Formación y habilidad de los docentes
Orientación vocacional	Flexibilidad de cursado	Relación docente - alumno
Comunicación institucional	Amplitud de oferta horaria	Servicios de bienestar estudiantil
Condición alumno regular	Cantidad de horas de cursado	Becas
Prácticas de enseñanza	Mecanismos de evaluación	Infraestructura y equipamiento
Seguimiento alumnos	Estrategias innovadoras primer año	Gasto por alumno
Tutorías	Dificultad materias primer año	Cultura organizacional

Tab. 3.3: Factores mencionados en la literatura. Fuente: [15] sobre la base de otros autores clásicos.

con XGBoost, lo que demuestra la robustez de estos enfoques en contextos universitarios públicos nacionales.

Se pueden encontrar trabajos con otros enfoque como predecir si un alumno abandonará una materia, un ejemplo de este enfoque es la tesis de Erausquin "Caracterización de trayectorias educativas a partir de producciones de código"[13].

A su vez, actualmente se encuentran disponible la compra de software, como Ed Machina [17], para la identificación de estudiantes en riesgo de abandono. Sin embargo, para poder hacer uso de ellos se debe pagar y dar acceso a la información de los estudiantes de la universidad a las empresas que ofrecen el software.

Dado nuestro conocimiento de la institución y los trabajos sobre cómo interactuan los alumnos de la FCEN, nos pareció interesante el caso de estudio. Al igual que en el trabajo de Pustilnik y Ndukanma se utilizarán datos de SIU-Guarani. Se pudo tener acceso a los datos del CBC y de la carrera de cada alumno y a las encuestas de la FCEN de 2023 y 2024 en donde, se espera conseguir información relativa a antecedentes familiares, situación laboral, entre otras. En la siguiente sección se puede encontrar más información sobre las fuentes de datos.

4. METODOLOGÍA

4.1. Fuentes de datos

Como se explicó en la Sección 1.1.2, todas las carreras de la UBA tienen un primer año obligatorio llamado Ciclo Básico Común (CBC), que consta de seis materias, algunas de las cuales se pueden cursar por la plataforma UBAXXI. Cada materia tiene dos evaluaciones parciales, calificadas como aprobadas o desaprobadas. Si el estudiante aprueba ambos parciales con un promedio de al menos 7 puntos y cumple con un mínimo del 75 % de asistencia, puede promocionar la materia directamente sin rendir examen final. Si aprueba los dos parciales con un promedio entre 4 y menos de 7 puntos, accede al examen final. Si aprueba uno y desaprueba el otro, puede rendir un examen complementario (recuperatorio); al aprobarlo, pasa al final. Si desaprueba ambos parciales o el recuperatorio, tiene la opción de recursar la materia. Los exámenes finales se califican de 0 a 10, aprobando con un mínimo de 4 [9, 6].

En el estudio del Estado del Arte, García [15] establece que los datos académicos, como promedio en la escuela secundaria, son relevantes para la predicción de abandono en la universidad. En el presente trabajo no contamos con acceso a información sobre los estudios secundarios de los alumnos, sin embargo los datos que se encuentran en SIU-Guaraní nos permiten generar un historial académico al estudiar las notas obtenidas y el tiempo que les tomó finalizar las materias.

A su vez los datos del SIU-Guaraní de la FCEN, no sólo contienen datos académicos, si no que también cuenta con datos demográficos. Como se explicó en la introducción de este trabajo, la FCEN dicta dieciocho carreras distintas, las mismas tienen materias en común y, exceptuando los profesorados, los planes de todas las carreras exigen un conjunto de materias optativas. Estas materias optativas pueden ser materias obligatorias de otras carreras.

Las encuestas realizadas a los ingresantes a la FCEN permitirían reforzar la información sobre factores socioeconómicos.

Finalmente, los datos solicitados provinieron de SIU-Guaraní del CBC, del SIU-Guaraní de la FCEN y de una encuesta realizada a los ingresantes a la FCEN como parte del programa +Acompañamiento. Los datos de SIU-Guaraní son de febrero de 2020 a febrero 2025, mientras que las encuestas se comenzaron a realizar en el segundo cuatrimestre de 2023. Para el entrenamiento del modelo se hizo uso de los datos provenientes del sistema SIU-Guaraní de estudiantes inscritos a la facultad en los años 2021 y 2022.

4.2. Adquisición de datos

Para unificar la información de las diversas fuentes, se requirió utilizar el Documento Nacional de Identidad (DNI) de los estudiantes. Al tratarse de información sensible, antes de comenzar el trabajo se recibió la autorización de la FCEN y del CBC para estudiarlos. Se definió que el modelo final podrá ser utilizado únicamente por miembros de la facultad y los datos fueron desechados una vez finalizado el entrenamiento del mismo.

A su vez, se trabajó con los datos almacenados de manera local, facilitando el control total sobre la seguridad y la integridad de la información.

4.3. Variables a estudiar

En la Tabla 4.1 se presentan los principales campos utilizados para el análisis de datos y experimentos. La clasificación de cada una de las variables se hizo siguiendo los tipos de factores definidos por García [15], exceptuando el caso de tiempo de viaje que surge del trabajo de Pustilnik y Ndukanma [25].

En la Sección 6.2 se detalla cómo se obtuvieron las variables calculadas.

Campo o Grupo	**Observación**	**Clasificación**
DNI	Documento Nacional de Identidad del estudiante, fue utilizado para poder unificar las tablas y desechado	
Tiempo de viaje	Tiempo de viaje del alumno desde su vivienda hacia la FCEN	Calculada (Numérica)
Nota de cluster de materia del CBC	Promedio de nota obtenido en las materias del cluster	Académico (Numérica)
Fecha de examen del cluster del CBC	Promedio de fecha asociada a las cursadas/exámenes del CBC o bien fecha del acta para UBAXXI.	Académico (Numérica)
Cantidad de veces que rendio las materias del cluster del CBC	Promedio de cantidad de veces en que el alumno rindió cada materia del cluster	Académico (Numérica)
UBA XXI	Indica si alguna materia del cluster se cursó por la plataforma UBA XXI	Políticas académicas y Plan de estudio (Categórica)
Edad	Edad del estudiante al inscribirse a la FCEN	Demográfico (Numérica)
Género	Femenino o Masculino	Demográfico (Categórica)
Carrera	La carrera donde el estudiante tiene más inscripciones a exámenes y cursadas de la FCEN	Plan de Estudio (Categórica)
Nivel Estudio Madre y Padre	Nivel máximo de estudios alcanzado por los padres según lo informado por el alumno.	Socioeconómico (Categórica)
Situación laboral	Indica si el alumno trabaja y/o busca trabajo según su declaración. Es una variable categórica.	Socioeconómico (Categórica)
Año inscripción facultad	Año en que el alumno se inscribió en la Facultad.	Plan de estudio (Numérica)
TP aprobado	Para cada una de las primeras diez materias inscriptas, se indica si aprobó o no	Académico (Categórica)
Fecha TP	Fecha en que se subió el acta de la materia.	Académico (Numérica)
Nota Final	Para cada una de los primeros diez finales inscriptos, se indica si aprobó o no	Académico (Categórica)
Fecha Final	Fecha en que se subió el acta del examen final.	Académico (Numérica)
Variables resumen	Cantidad de inscripciones, TPs aprobados y finales inscriptos por semestre	Académico (Numérica)

Tab. 4.1: Campos utilizados para el análisis, se agruparon variables para facilitar la lectura. Fuente: Elaboración propia.

4.4. Métodos utilizados

En este trabajo se emplearon dos familias de algoritmos basados en árboles de decisión: los Árboles de Decisión individuales y el ensamble *Random Forest*. A continuación se describen las características de ambos algoritmos, así como conceptos clave relacionados que serán importantes para la posterior comprensión de los experimentos.

4.4.1. Métricas de homogeneidad de una región

En los algoritmos de aprendizaje supervisado de tareas de clasificación, trabajamos con un conjunto de datos (instancias) donde cada uno tiene asignado una clase.

Las métricas de homogeneidad de una región nos permiten cuantificar cuán "pura" o "mezclada" está una colección de datos en términos de sus etiquetas de clase en un

problema de clasificación. A continuación presentamos dos de ellas:

Impureza de Gini (*Gini Impurity*)

La Impureza de Gini mide la probabilidad de que una instancia particular sea clasificada erróneamente si esta fuese etiquetada aleatoriamente siguiendo la distribución de clases dentro de la región. La formula para calcularla es:

$$Gini(S) = 1 - \sum_{k \in clases(S)} (p_S(k))^2$$

Donde:
S es el conjunto de instancias, *clases(S)* es el conjunto de clases posibles y $p_S(k)$ es la probabilidad de que una instancia del conjunto S sea de la clase k.

Entropía (*Entropy*)

Esta métrica está basada en la Entropía de Shannon, la cual mide el nivel promedio de "información", "sorpresa" o "incertidumbre" inherente a los posibles resultados de una variable aleatoria. Una entropía de 0 indica una región perfectamente homogénea (una sola clase), mientras que una entropía alta indica una mezcla uniforme de clases. A continuación se presenta la ecuación:

$$H(S) = - \sum_{k \in clases(S)} p_S(k) \log_2 p_S(k)$$

Donde:
S es el conjunto de instancias, *clases(S)* es el conjunto de clases posibles y $p_S(k)$ es la probabilidad de que una instancia del conjunto S sea de la clase k.

4.4.2. Reducción de impureza

A partir de la Impureza de Gini y la Entropía, se pueden definir métricas que cuantifican cuánto mejora una partición de un conjunto de datos en términos de homogeneidad de clases. Es decir, permiten medir si dividir un conjunto original en subconjuntos genera grupos más "puros" respecto a las clases presentes.

Supongamos que tenemos un conjunto de datos S y lo dividimos en dos subconjuntos $S_{\{\leq\}}$ y $S_{\{>\}}$ donde el primero tienen las instancias de S cuyo atributo a es menor o igual a c y el segundo contiene al resto. La reducción de impureza que genera esta partición puede expresarse como:

$$Gain(S, \langle a, c \rangle) = I(S) - \left(Prop_{\{\leq\}} \cdot I(S_{\{\leq\}}) + Prop_{\{>\}} \cdot I(S_{\{>\}})\right) \tag{4.1}$$

Donde:
I() representa la medida de impureza (ya sea Gini o Entropía), $Prop_{\{\leq\}}$ es la proporción de instancias del conjunto S que en el atributo a tienen un valor menor o igual a c, $Prop_{\{>\}}$ es 1- $Prop_{\{\leq\}}$ y $S_{\{\leq\}}$ y $S_{\{>\}}$ son los conjuntos definidos previamente.

4.4.3. Técnicas y algoritmos utilizados

Árboles de Decisión

Los Árboles de Decisión son modelos predictivos que representan una serie de decisiones estructuradas jerárquicamente en forma de árbol. Cada decisión corresponde a una condición sobre algún atributo de los datos, y conduce a diferentes ramas. Al seguir las ramas desde la raíz hasta una hoja, se obtiene una predicción, ya sea una clase (en clasificación) o un valor numérico (en regresión). En la Figura 7.3 se presenta un ejemplo de un Árboles de Decisión.

Este algoritmo de aprendizaje supervisado fue introducido por Quinlan J.R. en 1986 [26], y puede utilizarse tanto para tareas de clasificación como de regresión. Una ventaja significativa de los Árboles de Decisión es su robustez frente a datos faltantes, lo que los hace particularmente adecuados para conjuntos de datos como los provenientes de SIU-Guaraní del CBC y la FCEN. Como se detalla en la Sección 6.4, algunas variables, como las notas y fechas de materias, presentan registros incompletos. Los Árboles de Decisión pueden manejar estas ausencias sin requerir imputación de valores, analizando la reducción de impureza mediante estrategias que asignan las instancias con datos faltantes sin necesidad de imputación. Esta característica, junto con su interpretabilidad y flexibilidad, influyó en su elección como base para los modelos desarrollados en este trabajo.

En este trabajo se utilizó la librería scikit-learn donde los Árboles de Decisión para tareas de clasificación se implementan mediante la clase *DecisionTreeClassifier* [28], que utiliza una versión optimizada del algoritmo CART (*Classification and Regression Trees*) [3].

CART construye árboles binarios seleccionando, en cada nodo, el par (atributo, umbral) que maximiza la reducción de impureza, medida comúnmente como se definió en la Ecuación 4.1. Este proceso se repite recursivamente hasta que no haya nodos con más de una clase y posteriormente se define un criterio de poda, lo cual implica eliminar hojas del árbol. En el Algoritmo 1 se muestra cómo sería el proceso iterativo de la construcción del Árbol de Decisión con el algoritmo CART.

Algorithm 1 Algoritmo CART

Sea S una muestra de instancias con atributos A. Para construir un árbol de decisión ejecutamos:

1: Elegir el par $a \in A, c \in \mathbb{R}$ entre los posibles pares que maximice la reducción de impureza de S para nodo actual.
2: Crear dos hijos del nodo actual.
3: Dividir las instancias de S en los nuevos nodos, según $\langle a, c \rangle$:
$S_{\leq c} \leftarrow \{x \mid x \in S \wedge x[a] \leq c\}$
$S_{>c} \leftarrow \{x \mid x \in S \wedge x[a] > c\}$
4: Repetir para cada hijo en el que haya instancias de más de una clase.

Posteriormente se poda el árbol según un criterio definido.

Algoritmo 1: Construcción de un Árbol de Decisión para clasificación utilizando el Algoritmo CART.

Se pueden definir ciertos criterios de parada, como una profundidad máxima del árbol o un número mínimo de muestras por hoja, lo cual implicaría que el árbol no necesariamente crece hasta que cada hoja tenga una única clase. En los experimentos de este trabajo, se

definirá una altura máxima de los árboles.

Una limitación importante de los Árboles de Decisión es su alta varianza. Pequeñas variaciones en los datos de entrenamiento pueden producir árboles muy distintos, lo que impacta negativamente en la generalización del modelo. Además, si no se aplican técnicas de regularización (limitar la profundidad del árbol o establecer un número mínimo de muestras por hoja), los árboles tienden al sobreajuste (*overfitting*), es decir, el modelo se ajusta excesivamente a las particularidades del conjunto de datos de entrenamiento en lugar de aprender patrones generales [19].

Random Forest

Para mitigar la varianza inherente a los Árboles de Decisión individuales, se puede utilizar el algoritmo *Random Forest* [4]. Este es un método de ensamble, una técnica que consiste en combinar las predicciones de múltiples modelos (en este caso, Árboles de Decisión) para construir un modelo final más preciso y robusto. Se encuentra implementado en scikit-learn mediante la clase *RandomForestClassifier* [29]. Para definir la clase que se le asigna a cada instancia "la implementación de scikit-learn combina clasificadores promediando su predicción probabilística"(traducido del inglés de [27]), esto quiere decir que cada árbol devuelve la probabilidad de que la instancia pertenezca a cada una de las clases, se calcula el promedio y, según el umbral que se defina, se determina la clase a la que pertenece. En el caso de tener dos clases, seleccionar un umbral igual a 0.5 significa que a cada instancia se le asigna la clase para la cual la probabilidad sea mayor a 50 %.

La diversidad de los árboles dentro del ensamble se logra mediante dos formas de aleatoriedad:

- Usando *Bootstrap*: Consiste en que para cada árbol armamos un *dataset* del mismo tamaño que el original, tomando instancias con reposición.
- Seleccionar *features* al azar por nodo: Para cada nodo, elegimos al azar m atributos para considerar en la selección del "atributo que mejor separe".

Estas fuentes de aleatoriedad permiten que cada árbol explore distintas particiones del espacio de decisiones, distintas maneras de clasificar los datos y tomar decisiones. Como resultado, se espera que el modelo en su conjunto sea menos sensible a errores derivados de haber entrenado un solo árbol que pudo haber quedado atrapado en un mínimo local o haber sido influenciado por un conjunto de entrenamiento poco representativo.

4.4.4. Importancia de atributos

La importancia de atributos se refiere a medidas que indican cuánto contribuye un atributo particular a las predicciones de un modelo determinado. A cada atributo se le asigna una puntuación que permite clasificarlos según el impacto que tienen en las predicciones. Es importante notar que esta técnica no refleja el valor predictivo intrínseco de un atributo por sí solo, sino su importancia para un modelo específico.

Dentro del ámbito de los modelos basados en árboles, como los Árboles de Decisión y *Random Forests*, un método para determinar la importancia de los atributos es la Disminución Media de la Impureza (MDI) o Importancia Gini.

En este caso, la importancia de un atributo dentro de un único Árbol de Decisión se calcula sumando todas las reducciones de impureza (Ecuación 4.1) a las que contribuyó

en cada división donde fue empleado, por lo que la manera de medir la impureza depende del criterio utilizado para construir el árbol. Si un atributo no se usa nunca en un árbol, su importancia para ese árbol es cero.

Para el caso del *Random Forests*, se calculan las importancias para cada atributo en cada árbol individual del bosque y estas puntuaciones se promedian a lo largo de todos los árboles que lo componen. El atributo *feature_importances_* en scikit-learn proporciona este valor promedio. Además, las puntuaciones finales se normalizan típicamente para que sumen 1.

Finalmente, es importante considerar que MDI puede sobreestimar la importancia de características que tienen un gran número de valores únicos. Además, en presencia de características altamente correlacionadas, el modelo puede elegir una característica de un grupo correlacionado para una división y, al haber reducido ya la impureza, es menos probable que utilice las otras características correlacionadas en divisiones posteriores, por lo que estas últimas obtendrían un valor menor al calcular las reducciones de impureza en las que contribuyeron.

4.4.5. Evaluación de modelos

Al entrenar un modelo de Aprendizaje Automático, queremos evaluar su capacidad para generalizar su rendimiento a conjuntos de datos nuevos y nunca antes vistos. Para ello, se puede dividir al conjunto de datos en dos partes: un conjunto de entrenamiento, que se utiliza para ajustar el modelo, y un conjunto de testeo, que se utiliza para evaluarlo.

Sin embargo, este enfoque de una sola partición puede ser poco confiable. Existe el riesgo de que el conjunto de testeo no sea suficientemente representativo de los datos reales, lo que puede dar lugar a una estimación poco precisa o sesgada del rendimiento del modelo.

La validación cruzada (*cross-validation*) es una técnica estadística que busca producir una evaluación más estable y exhaustiva que una sola partición. Consiste en dividir el conjunto de datos en múltiples particiones de entrenamiento y testeo, evaluando el modelo en cada una y promediando los resultados. Existen diversas variantes de validación cruzada, siendo una de ellas la que utilizamos en este trabajo: *Monte Carlo Cross-Validation* [38].

Monte Carlo Cross-Validation consiste en repetir varias veces el siguiente procedimiento:

1. Dividir aleatoriamente el conjunto de datos en dos subconjuntos: uno para entrenamiento y otro para testeo.
2. Entrenar el modelo con los datos de entrenamiento.
3. Evaluar el modelo con los datos de testeo.

Este proceso se repite varias veces, utilizando diferentes particiones aleatorias en cada caso. Finalmente, se promedian los resultados obtenidos para obtener una estimación más robusta del rendimiento del modelo.

4.4.6. Métricas

En el presente trabajo, para definir que tan bueno es un modelo dado un conjunto de datos, se hace uso de tres métricas principales, Exactitud (*Accuracy*), Exactitud Balan-

ceada (*Balanced Accuracy*) y área bajo la curva ROC (AUC ROC) [1, 20].

Exactitud (*Accuracy*)

La exactitud se define como la proporción de instancias correctamente clasificadas sobre el total de instancias evaluadas, se presenta la ecuación:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Donde:
TP (*True Positives*) es el número de estudiantes en riesgo (clase 1) predichos correctamente como clase 1, TN (*True Negatives*) es el número de estudiantes que no se encuentra en riesgo de abandono (clase 0) predichos correctamente como clase 0, FP (*False Positives*) es el número de no riesgo etiquetados erróneamente como riesgo, y FN (*False Negatives*) es el número de riesgo etiquetados erróneamente como no riesgo. Por lo tanto, $TP + TN$ es equivalente a la cantidad de instancias clasificadas correctamente por el modelo y $TP + FP + FN + TN$ es equivalente al total de instancias dentro del dataset.

Esta métrica tiene la ventaja de ser sencilla de interpretar. Sin embargo, es una métrica que nos dice cuántos errores tuvo, pero no nos dice nada de cómo fue ese error. Además, en un conjunto de datos con clases desbalanceadas, un modelo podría obtener una alta exactitud simplemente prediciendo siempre la clase mayoritaria. Por ejemplo: Si tenemos 100 instancias, 10 de la clase 1 y 90 de la clase 2, si siempre predecimos la clase 2, vamos a tener un 0.9 de exactitud cuando en realidad nunca se predijo correctamente la clase 1.

En el presente trabajo, la exactitud se calculó tomando el umbral 0.5. Es decir, a cada instancia se le asignó la clase que obtuviera una probabilidad mayor a 0.5.

Exactitud Balanceada (*Balanced Accuracy*)

Para contrarrestar el sesgo hacia la clase mayoritaria, la exactitud balanceada promedia la sensibilidad y la especificidad:

$$BalancedAccuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Donde:
$\frac{TP}{TP+FN}$ es la sensibilidad (*recall o True Positive Rate*) de la clase 1 (riesgo de abandono), es decir, la proporción de estudiantes en riesgo correctamente identificados. Mientras que $\frac{TN}{TN+FP}$ es la especificidad (o *True Negative Rate*) de la clase 0 (no riesgo de abandono), o la proporción de estudiantes sin riesgo clasificados correctamente.

En el presente trabajo, la exactitud balanceada se calculó tomando el umbral 0.5. Es decir, a cada instancia se le asignó la clase que obtuviera una probabilidad mayor a 0.5.

AUC ROC (Área Bajo la Curva ROC)

Por otro lado, la curva ROC traza, a distintos umbrales de decisión, la relación entre el *recall* de la clase positiva o tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR = $\frac{FP}{FP+TN}$). El área bajo esta curva (AUC ROC) cuantifica la capacidad del modelo para ordenar correctamente instancias de forma independiente al umbral elegido. En la

Figura 4.1 se puede observar un ejemplo de la curva ROC obtenida luego de entrenar un modelo SVM, un algoritmo de Aprendizaje Automático cuyo objetivo es encontrar la mejor "frontera"posible para separar las instancias de distintas clases.

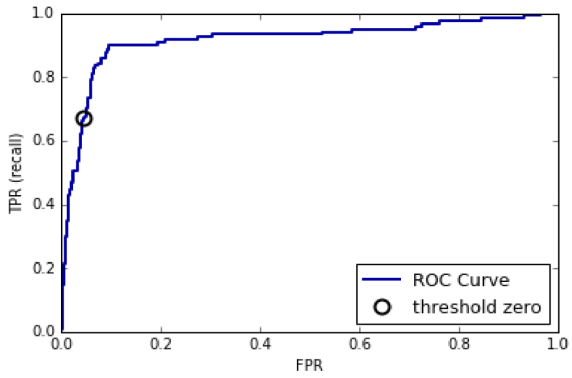


Fig. 4.1: Curva ROC para un modelo SVM, se encuentra marcado con un circulo el TPR y FPR obtenido al definir un umbral equivalente a 0. Fuente: [20].

La curva ideal debe estar cerca del margen superior izquierdo, ya que queremos un clasificador que tenga alto *recall* pero que tenga un FPR bajo. En este caso, el AUC ROC tiene un valor de 1.

5. ANÁLISIS DE DATOS

A continuación se presenta el análisis de los datos utilizados para los posteriores experimentos. Se detalla el preprocesamiento de los datos y se estudia la distribución de los mismos. A su vez, se presentan los principales resultados del análisis de abandono condicional.

5.1. Preprocesamiento

En este trabajo se utilizaron datos de alumnos inscriptos en la FCEN durante los años 2021 y 2022, provenientes del SIU-Guaraní del CBC y de la FCEN. Las encuestas realizadas a ingresantes, al estar disponibles a partir de 2023, quedaron fuera del rango temporal considerado, por lo que no se pudo incorporar información proveniente de las mismas.

Antes de llevar a cabo los experimentos, fue necesario realizar una etapa de preprocesamiento para adaptar los datos al formato requerido por los algoritmos de Aprendizaje Automático, particularmente los Árboles de Decisión.

En primer lugar, múltiples columnas contenían fechas asociadas a materias cursadas y exámenes rendidos. Dado que los Árboles de Decisión no pueden operar directamente con variables de tipo fecha, se optó por transformar estos valores a un formato numérico. Para ello, se tomó como punto de referencia el día 1 de enero de 2020, y se calcularon los días transcurridos desde esa fecha hasta cada una de las fechas presentes en el conjunto de datos.

En segundo lugar, dado que los registros recibidos del CBC incluyen una gran cantidad de materias, para reducir la cantidad de columnas con las cuales trabajar y, a su vez, reducir la cantidad de datos faltantes que implicaría crear una columna por materia, se las agruparon en siete grupos (*clusters*). Aunque existen numerosas materias disponibles, todas las carreras exigen que el estudiante apruebe seis asignaturas obligatorias. Con este criterio, se construyeron grupos de materias agrupando aquellas con contenidos similares dentro del plan académico y considerando los planes de estudio de las carreras. Esta agrupación también permitió abordar el caso de estudiantes que cursaron materias ajenas a su orientación específica, evitando que estas excepciones impactaran negativamente en el análisis. Para cada cluster se definieron cuatro columnas para la nota obtenida, fecha del examen, cantidad de veces que el alumno rindió y si cursó por UBAXXI. Al hacer un análisis de datos faltantes, se observó que las columnas asociadas con el cluster de materias que no forman parte del plan de estudio de la FCEN tenían más del 90 % de datos faltantes, por lo que no fue utilizada para el entrenamiento del modelo. Esto se encuentra resumido en la Ecuación 5.1.

$$\begin{aligned} NotaCluster_{j_a} &= promedio_{i=1\ldots n_a}(nota_i) \\ \#vecesQueRindioCluster_{j_a} &= promedio_{i=1\ldots n_a}(\#vecesQueRindio_i) \\ FechaCluster_{j_a} &= promedio_{i=9\ldots n_a}(fechas_i) \\ UBAXXI_{j_a} &= \begin{cases} 1 & \text{si} \quad \exists i \in [1, n_a] : UBAXXI_i = 1 \\ 0 & \text{si no} \end{cases} \end{aligned} \tag{5.1}$$

Donde:
$NotaCluster_i$ es el promedio de notas del cluster, $\#vecesQueRindioCluster_i$ indica la cantidad de veces que el alumno rindió las materias del cluster i, $FechaCluster_i$ el promedio de la fecha en que rindió y $UBAXXI_i$ indica si el alumno cursó alguna de las materias del cluster por UBAXXI. El parámetro n depende de cada alumno (a) y se corresponde con el total de materias por cluster.

Para construir una distribución empírica sobre el comportamiento académico de los estudiantes de la FCEN, se procesaron los registros de actas correspondientes a materias cursadas y exámenes rendidos. *Con el objetivo de capturar el orden y la frecuencia con que los estudiantes rinden, se generaron cuatro columnas por cada una de las materias cursadas, respetando el orden en que fueron aprobadas.* A continuación se muestra un ejemplo de dos materias para un alumno:

$$\begin{aligned}(FechaInscripcion_0, TPAprobado_0, FechaFinal_0, NotaFinal_0);\\ (FechaInscripcion_1, TPAprobado_1, FechaFinal_1, NotaFinal_1); ...\end{aligned} \tag{5.2}$$

Donde:
$FechaInscripcion_i$ es la fecha de inscripción en formato numérico, $TPAprobado_i$ indica si el estudiante aprobó o no los trabajos prácticos de la materia, $FechaFinal_i$ es la fecha en que rindió el examen final en formato numérico y $NotaFinal_i$ es la clasificación obtenida en el examen final.

Se consideraron las primeras nueve materias rendidas por cada estudiante incluyendo tanto trabajos prácticos como finales. Es importante aclarar que, en los casos donde una materia fuera cursada en más de una ocasión, cada instancia fue considerada por separado. Por otro lado, para aquellos estudiantes que presentaban más de diez registros entre trabajos prácticos y finales, se generaron cuatro columnas adicionales que contenían los promedios de las materias restantes y si aprobó los trabajos prácticos de alguna materia, permitiendo así conservar la información. Esto se encuentra resumido en la Ecuación 5.3.

$$\begin{aligned} FechaInscripcion_{9_a} &= promedio_{i=9...n_a}(fechas_i)\\ TPAprobado_{9_a} &= \begin{cases} 1 \quad \text{si} \quad \exists i \in [9, n] : TPAprobado_i = 1\\ 0 \text{ si no} \end{cases}\\ FechaFinal_{9_a} &= promedio_{i=9...n_a}(fechasFinal_i)\\ NotaFinal_{9_a} &= promedio_{i=9...n_a}(notasFinal_i) \end{aligned} \tag{5.3}$$

Donde:
Las fechas y notas de finales son el promedio acumulado en la décima columna, y se indica si aprobó los trabajos prácticos de alguna materia. El parámetro n depende de cada alumno (a) y se corresponde con el total de materias.

En este contexto, los datos faltantes en las columnas asociadas a las materias de la FCEN reflejan la ausencia de actividad académica por parte del estudiante, como la no inscripción o no rendición de exámenes en dichas materias. Esta interpretación permitió conservar los valores nulos como una representación válida del comportamiento académico, en lugar de imputarlos, lo que resultó coherente con la capacidad de los Árboles de Decisión para manejar datos faltantes, como se detalla en la Sección 4.4.3.

Además, se definió la carrera de los estudiantes en función del registro de las actas. A su vez, en función del año de inscripción y primer registro de acta de cada alumno, se definió el semestre en que el alumno comenzó sus estudios en la facultad, su **cohorte**.

También, a partir de los datos del CBC y el año de inscripción a la FCEN, se calculó la cantidad de años transcurridos desde el inicio del CBC.

Se generó una nueva variable que representa el tiempo estimado de viaje de los estudiantes desde su domicilio hasta la facultad, definiendose cero y más de cuatro horas (14,400 segundos) como atípico. Este atributo se incorporó con el objetivo de analizar cómo la distancia geográfica puede influir en el rendimiento académico, inspirado en el trabajo de Pustilnik y Ndukanma [25].

Por último, se realizó un análisis de datos atípicos. Durante esta revisión, se detectó un caso donde la edad registrada de un estudiante superaba los cien años, un valor que se consideró invalido. En consecuencia, este dato fue tratado como valor faltante. Dado que los Árboles de Decisión pueden trabajar de manera directa con valores nulos, sin necesidad de imputación previa, se optó por dejar estos campos como tal.

El método utilizado para calcular cada una de las variables se detalla más adelante, en el Capítulo 6.

5.2. Estudio de distribución de datos

5.2.1. Cohorte

Para el entrenamiento del modelo se utilizaron datos de estudiantes inscriptos en la FCEN en los años 2021 y 2022. Como se observa en la Figura 5.1, la distribución de estudiantes por cohorte muestra que la mayor proporción corresponde a la cohorte del primer cuatrimestre de 2021 (2021C1), con 34.39 % del total y le sigue la cohorte 2022C1, con 29.10 %. Esto podría estar indicando que en general los estudiantes comienzan sus estudios en la FCEN en el primer semestre de cada año.

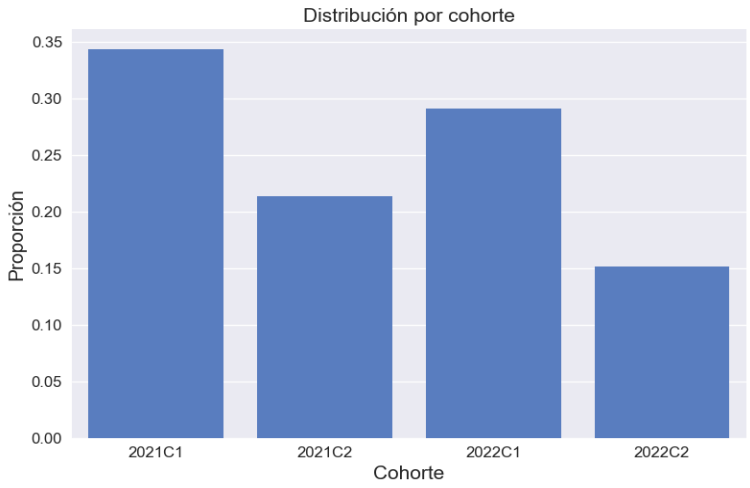


Fig. 5.1: Distribución de las cohortes de los años 2021 y 2022. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.2. Sexo

La distribución según el sexo (Figura 5.2) indica una mayor proporción de personas identificadas con el sexo masculino, que representan el 63.10 % del total, en tanto que las personas identificadas con el sexo femenino constituyen el 36.90 %.

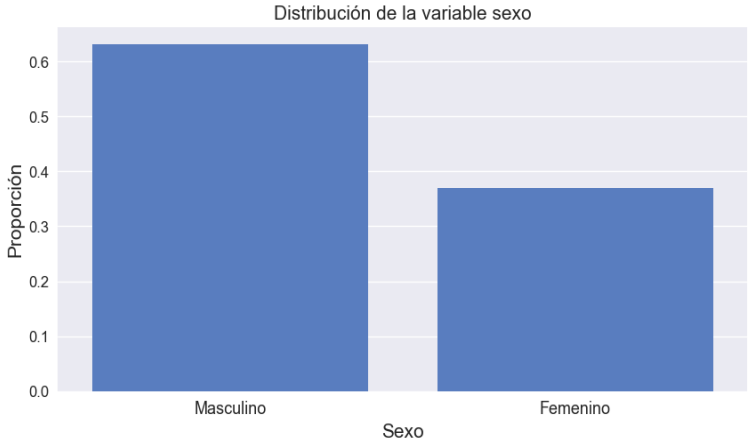


Fig. 5.2: Distribución de la variable sexo de las personas de la FCEN inscriptas en los años 2021 y 2022. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.3. Edad

La Figura 6.3 presenta la distribución etaria de las personas inscriptas en la FCEN durante los años 2021 y 2022, luego de haber eliminado los casos identificados como *outliers*, en el Capítulo 6 y la Sección 5.3 se detalla cómo se calculó la misma. El 52.26 % de los estudiantes se encuentra en el rango de 18 a 20 años, seguido por quienes tienen entre 20 y 30 años con 27.27 %. En menor medida, también se registran inscripciones en los grupos de 16 a 18 años (15.30 %) y mayores de 30 años (5.17 %), aunque estos últimos representan una fracción considerablemente más baja del total.

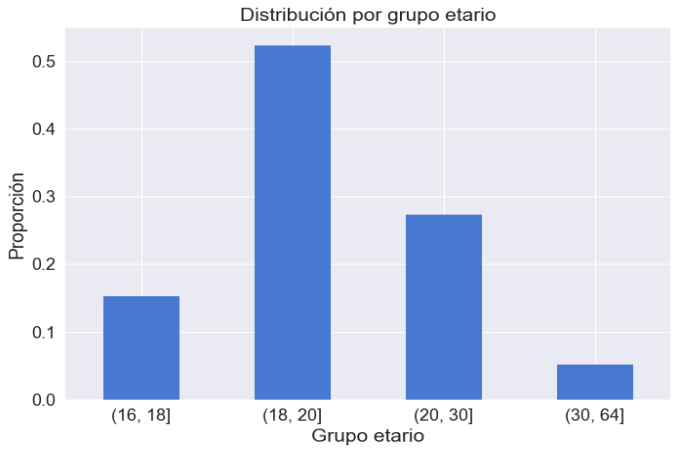


Fig. 5.3: Distribución de la edad de las personas de la FCEN inscriptas en los años 2021 y 2022. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.4. Nivel de estudio de los padres

La distribución del nivel educativo de los padres y madres de las personas inscriptas en la FCEN durante los años 2021 y 2022 revela que, el orden de las proporciones es muy similar. En ambos casos, la categoría más frecuente corresponde a estudios universitarios completos, superando el 20 % para ambos padres (21.84 % padres y 25.43 % madres). Sin embargo, la categoría que le sigue es la de datos faltantes, donde alcanza el 20.11 % del registro de los padres y 17.17 % de las madres. Se destaca que, en el caso de los padres, existe una mayor proporción de estudios universitarios incompletos (12.31 %), mientras que para las madres, una proporción comparable corresponde a estudios superiores completos (13 %).

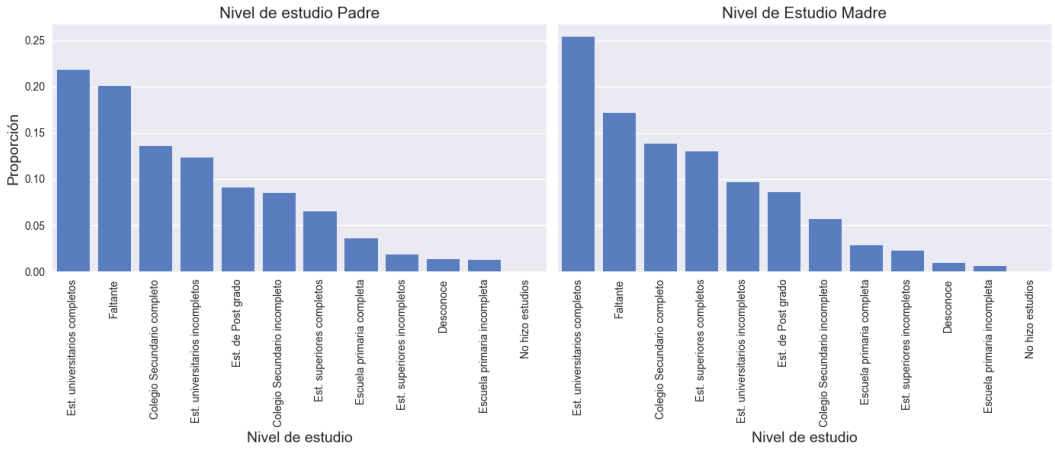


Fig. 5.4: Distribución del nivel de estudio de los padres de las personas de la FCEN inscriptas en los años 2021 y 2022. En la figura de la izquierda se presenta la distribución del nivel de estudio del padre, mientras que en la derecha se presenta la distribución para el caso de la madre. "Faltante"se corresponde con aquellos casos donde no se tenía una categoría seleccionada por el alumno. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.5. Situación laboral

La distribución de la situación laboral de las personas inscriptas en la FCEN durante los años 2021 y 2022 (Figura 5.5) muestra que la categoría más frecuente corresponde a quienes no trabajan ni se encuentran en búsqueda activa, abarcando el 42.42 % del total. Le siguen, en proporciones similares, las personas que declararon trabajar al menos una hora (28.91 %) y aquellas que, si bien no trabajan, se encuentran buscando empleo (27.71 %). Los casos sin respuesta, categorizados como "Faltante", representan un porcentaje muy bajo (0.96 %).

Este patrón del perfil de estudiante universitario de la FCEN, podría estar relacionado con que gran parte de las materias tienen una carga horaria que supera las 10hs semanales. La elevada proporción de estudiantes que no trabajan ni buscan empleo podría estar vinculada al nivel de dedicación requerido por las carreras de la FCEN, lo cual limita la posibilidad de compatibilizar el estudio con una actividad laboral. A su vez, el hecho de que un porcentaje significativo trabaje al menos una hora o se encuentre en búsqueda activa indica que existe una fracción considerable del estudiantado que intenta insertarse en el mercado laboral durante sus estudios.

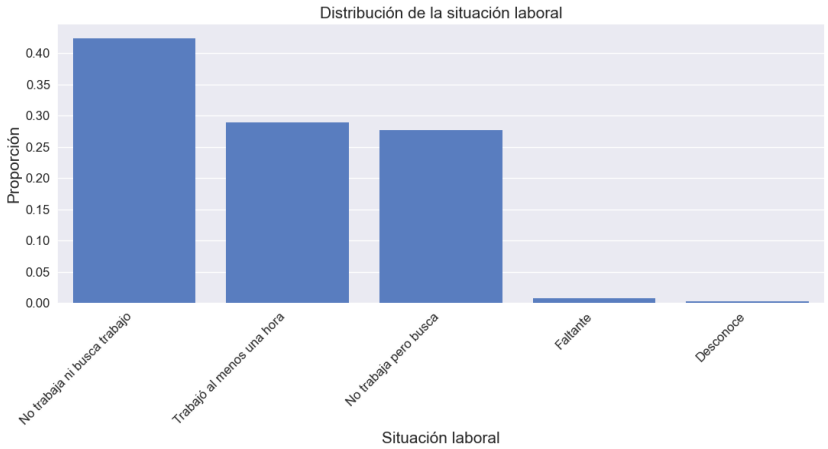


Fig. 5.5: Distribución de la situación laboral de las personas de la FCEN inscriptas en los años 2021 y 2022. "Faltante"se corresponde con aquellos casos donde no se tenía una categoría seleccionada por el alumno. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.6. Carrera

La Figura 5.6 muestra la distribución de la carrera principal de las personas inscriptas en la FCEN. La carrera principal fue definida como aquella en la que cada persona registró la mayor cantidad de inscripciones, ya sea a materias o a finales, durante sus primeros cinco semestres.

Se observa que el 25.01 % de las personas tiene como carrera principal Licenciatura en Ciencias de la Computación, lo que indica que una de cada cuatro personas se anota mayoritariamente a materias asociadas a dicha carrera. Le siguen en proporción las carreras de Biología (20.76 %), Física (17.10 %) y Ciencias de Datos (14.51 %), cada una con entre el 14 % y 21 % del total. Luego, la carrera Química tiene 5.56 % y el resto de las carreras representa menos del 5 % cada una. Para simplificar la visualización, se agruparon bajo la categoría "Otras carreras" aquellas opciones cuya proporción fue inferior al 2 % del total. Este comportamiento no necesariamente refleja la inscripción original al CBC, pues gran parte de las carreras de la FCEN tienen las mismas materias para poder inscribirse en la facultad.

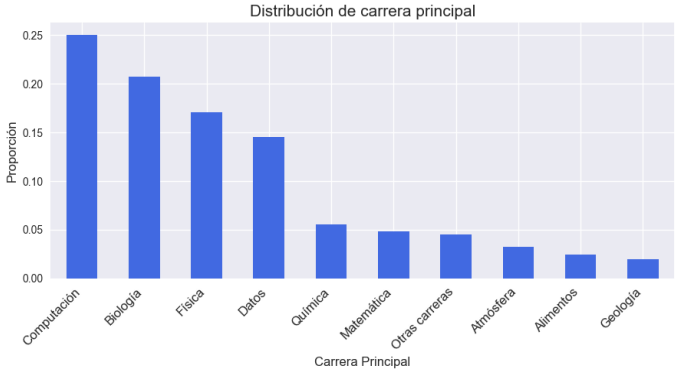


Fig. 5.6: Distribución de la carrera principal de las personas de la FCEN inscriptas en los años 2021 y 2022. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.7. Tiempo de viaje

En la Figura 5.7 se presenta la distribución de la variable calculada tiempo de viaje. La media del tiempo de viaje en transporte público durante la mañana es de 1.06 horas, mientras que la mediana es de 56.92 minutos. Además, el mínimo calculado fue de 1361 segundos (22.6 minutos) y el máximo 13882 segundos (3.85 horas). Debe considerarse que previamente se desecharon los registros con tiempo de viaje superior a cuatro horas, dado que se los definió como atípicos.

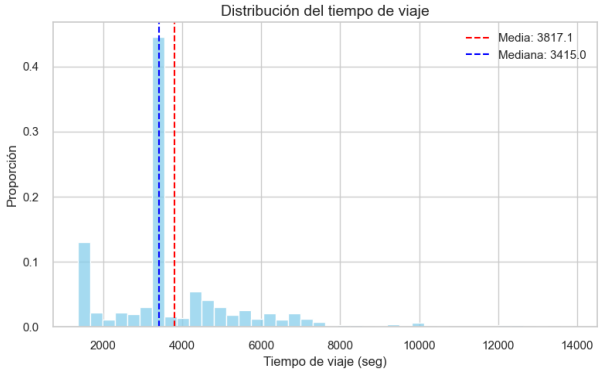


Fig. 5.7: Distribución del tiempo de viaje de los alumnos de la FCEN inscriptos en los años 2021 y 2022. En rojo se marca la media del tiempo de viaje y en azul la mediana. Fuente: Elaboración propia a partir de datos de la FCEN.

5.2.8. Riesgo de abandono

En la Figura 5.8 se observa la distribución de la variable *target*. El 45.39 % del conjunto de datos corresponde a personas en riesgo, según el umbral de actividad académica por semestre definido en la Sección 3.1. Esta distribución muestra que la variable a predecir no se encuentra fuertemente desbalanceada. Sin embargo, se optó por utilizar Exactitud Balanceada (*Balanced Accuracy*) como una de las métricas principales de evaluación (ver Sección 4.4.6). Esta métrica permite evaluar el desempeño del modelo considerando por separado la sensibilidad (capacidad para detectar correctamente casos en riesgo) y la especificidad (capacidad para identificar correctamente casos sin riesgo), contrarrestando el sesgo hacia la clase mayoritaria.

5.3. Análisis de datos faltantes y atípicos

Durante el análisis exploratorio se identificaron distintos casos de valores faltantes y atípicos en el conjunto de datos. A continuación se detalla cómo fueron tratados según el tipo de variable.

Por un lado, se optó por no imputar los valores faltantes, ya que los algoritmos basados en Árboles de Decisión, como los utilizados en esta tesis (Árboles de Decisión y *Random Forest*), son capaces de manejar valores faltantes de forma nativa sin necesidad de realizar una imputación previa.

Respecto a los datos atípicos, uno de los casos más relevantes fue el del tiempo de viaje desde el domicilio del estudiante hasta la Facultad. Se consideraron como extremos aquellos valores iguales a cero, que no resultan verosímiles, y aquellos mayores a cuatro

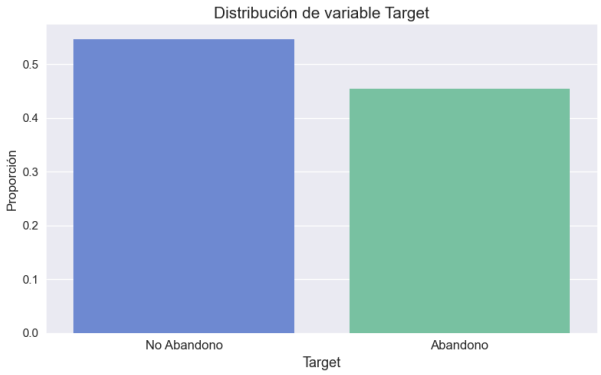


Fig. 5.8: Distribución de la variable a predecir (*target*). La variable toma valor 1 cuando el estudiante se encuentra en riesgo de abandono. Fuente: Elaboración propia a partir de datos de la FCEN.

horas. Estos registros fueron tratados como datos faltantes, quedando así un total de 196 registros sin tiempo de viaje.

De manera similar, se identificó a una persona cuya edad registrada era superior a los cien años, por lo que se lo tomó como dato atípico y se marcó como dato faltante.

Otro aspecto relevante del análisis de datos faltantes se presentó al trabajar con información proveniente de dos fuentes distintas, el SIU-Guaraní del CBC y el de la FCEN. Al intentar cruzar ambos conjuntos de datos, se observó que solamente 1,372 personas, es decir, el 52.95 % del total de personas registradas en la FCEN, contaban con registros en el CBC. Esto implicó que para el 47.05 % restante, toda la información proveniente de la base del CBC se presentaba como faltante. En consecuencia, para las variables vinculadas a las actas de materias del CBC, se conservaron los datos faltantes tal como estaban en los registros originales, sin aplicar imputaciones. Sin embargo, se descartaron los registros correspondientes a materias de otras carreras, dado que representaban menos del 9 % de los registros.

En lo que respecta a las materias registradas en la FCEN, si bien algunas personas tenían registros de más de diez materias distintas, solo 1,015 personas alcanzaban a tener información correspondiente a al menos nueve materias (39.17 % del total). Por este motivo, se definió un corte en nueve columnas individuales para representar las primeras materias de cada estudiante. Para aquellos casos en que se contaba con más de nueve materias, se construyeron columnas adicionales con el promedio de las variables asociadas a esa décima materia, como se detalla en la Ecuación 5.3. Cuando las personas no llegaban a completar las nueve materias, los valores correspondientes se conservaron como datos faltantes, sin aplicar técnicas de imputación. En este contexto, la ausencia de valores reflejan que el estudiante no tuvo actividad académica.

5.4. Análisis de abandono condicional

5.4.1. Grupo etario

Al analizar la variable target condicional a los distintos grupos etarios, se observa una relación positiva entre la edad y el riesgo de abandono. La Figura 5.9a muestra la cantidad de personas clasificadas como en riesgo de abandono (target = 1) y no en riesgo (target

= 0) en cada grupo etario. La mayoría de los estudiantes se concentra entre los 18 y 20 años. A medida que aumenta la edad, crece la proporción de personas clasificadas como en riesgo de abandono. Esta tendencia se visualiza con mayor claridad en la Figura 5.9b, donde se presenta la proporción de riesgo de abandono por grupo etario. Se utilizó un modelo de regresión lineal para mostrar la relación con los bins que no son proporcionales con la edad. No obstante, el abandono aumenta con la edad (no linealmente).

(a) Cantidad de personas por grupo etario y valor de variable target.

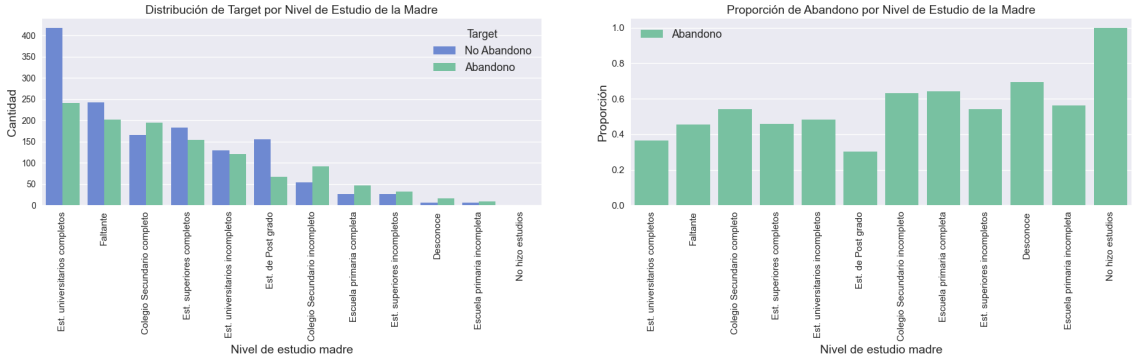
(b) Proporción de riesgo de abandono por grupo etario con recta de regresión para visualizar la tendencia creciente del target=1 a medida que aumenta la edad (MSE=0.0006).

Fig. 5.9: Distribución de la variable target por grupo etario. Fuente: Elaboración propia a partir de datos de la FCEN.

5.4.2. Nivel de estudio de los padres

En relación con el nivel educativo de los padres, se observa una variación significativa en la proporción de estudiantes en riesgo de abandono (ver Figuras 5.10 y 5.11). Cuando la madre o el padre posee estudios universitarios completos, la proporción de abandono es significativamente menor que la media (36.57 % y 39.05 % respectivamente), valor inferior al promedio general observado en la población. En contraste, cuando el nivel educativo de la madre corresponde a estudios secundarios completos, se registra un aumento en la proporción de abandono (54.04 %), que se intensifica aún más en los casos en que los estudios secundarios están incompletos (63.01 %). Esta misma relación se verifica también en el caso del padre, a menor nivel educativo, mayor proporción de estudiantes clasificados en riesgo. Es particularmente notorio el incremento observado cuando el padre ha finalizado la escuela primaria como nivel máximo alcanzado, donde la proporción de abandono es 71.74 %. Por otro lado, se destaca que en los casos en los que el padre posee estudios de posgrado, la proporción de estudiantes en riesgo cae significativamente por debajo del 40 %.

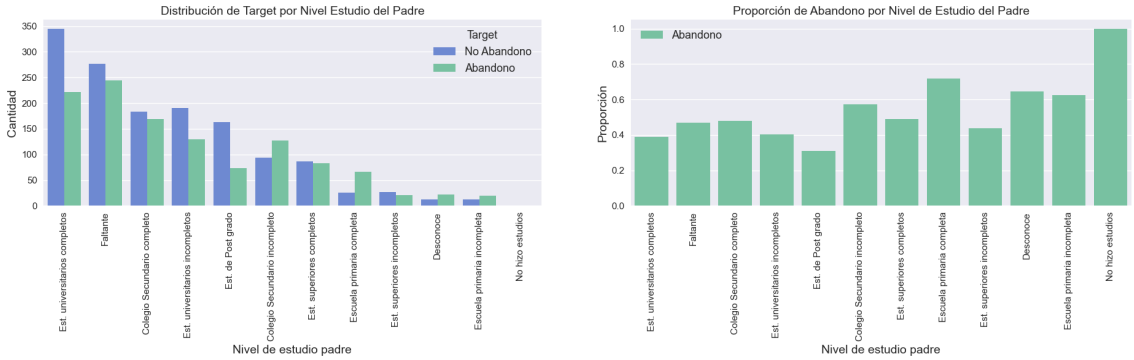
En resumen, los datos muestran a mayor nivel de estudios alcanzado por los padres la probabilidad de abandono de los estudiantes disminuye.



(a) Distribución de la variable target por nivel de estudio de la madre.

(b) Proporción de riesgo de abandono por nivel de estudio de la madre.

Fig. 5.10: Distribución de la variable target por nivel de estudio de la madre. Fuente: Elaboración propia a partir de datos de la FCEN.



(a) Distribución de la variable target por nivel de estudio del padre.

(b) Proporción de riesgo de abandono por nivel de estudio del padre.

Fig. 5.11: Distribución de la variable target por nivel de estudio del padre. Fuente: Elaboración propia a partir de datos de la FCEN.

5.4.3. Situación laboral

En la Figura 5.12 se analiza la relación entre la situación laboral de los estudiantes y la proporción de abandono. Al observar el gráfico de proporciones (Figura 5.12b), se identifica una tendencia creciente en el riesgo de abandono a medida que aumenta el grado de inserción laboral. Específicamente, quienes declararon que no trabajan ni buscan trabajo presentan una proporción menor de abandono (32.94 %), mientras que esta proporción se incrementa en el grupo que no trabaja pero busca empleo (49.72 %), y alcanza su valor más alto entre quienes trabajan al menos una hora semanal (59.81 %).

5.4.4. Carrera

En la Figura 5.13 se presenta la distribución de abandono según la carrera principal de cada estudiante. Al observar las carreras con mayor cantidad de registros, tales como Computación (46.14 %), Biología (41.45 %), Física (44.50 %) y Datos (44.41 %), se advierte que la proporción de abandono se mantiene cercana a la media poblacional, en torno al

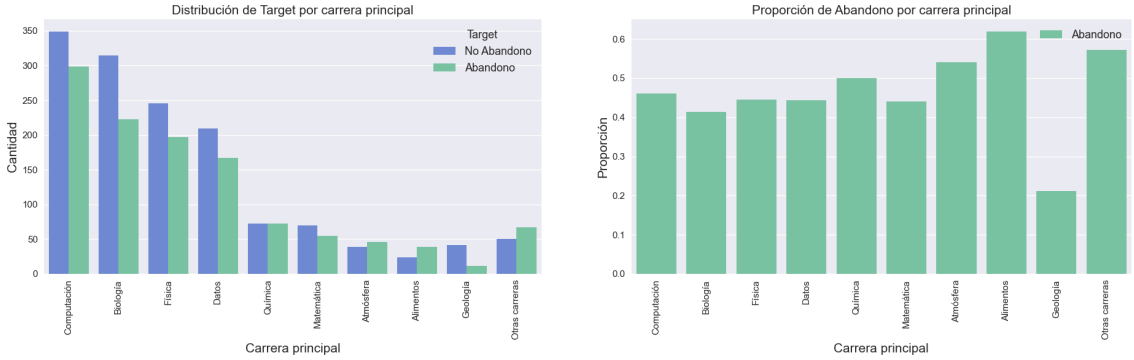
(a) Distribución de la variable target por situación laboral.

(b) Proporción de riesgo de abandono por situación laboral.

Fig. 5.12: Distribución de la variable target por situación laboral. Fuente: Elaboración propia a partir de datos de la FCEN.

45 %, por lo que no se observan grandes diferencias entre ellas en este aspecto.

Sin embargo, hay algunas carreras donde la proporción de estudiantes en riesgo de abandono es mayor. En particular, Ciencias de la Atmósfera (54.12 %), Ciencia y Tecnología de Alimentos (61.90 %) y la categoría "Otras carreras" (57.26 %), que agrupa trayectorias con menor representación, muestran porcentajes de abandono que superan el 50 %. Esto sugiere que en ciertos casos puede haber condiciones específicas de esas carreras que estén relacionadas con un mayor riesgo de abandono.



(a) Distribución de la variable target por carrera principal.

(b) Proporción de riesgo de abandono por carrera principal.

Fig. 5.13: Distribución de la variable target por carrera principal. Los nombres de las carreras se encuentran abreviados. En "Otras carreras"se agruparon el resto de las carreras del conjunto de datos para facilitar la lectura. Fuente: Elaboración propia a partir de datos de la FCEN.

5.4.5. Total de actividad

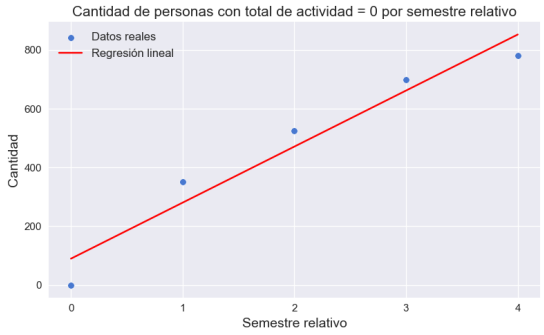
A partir de las Tablas 3.1 y 3.2, se construyeron los gráficos de la Figura 5.14 para analizar la evolución de la actividad académica al avanzar los semestres.

En la Figura 5.14a se observa una tendencia creciente en la cantidad de personas sin actividad académica a medida que transcurren los semestres relativos. Esta tendencia es capturada por la recta de regresión lineal.

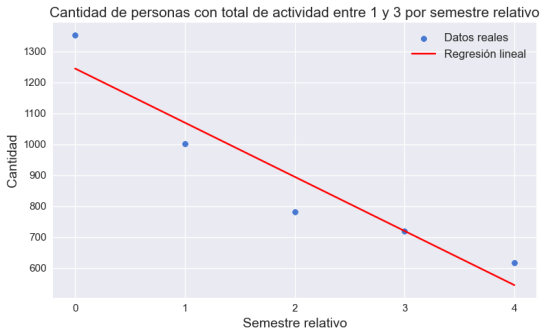
En cambio, la Figura 5.14b muestra cómo disminuye el número de personas cuya actividad académica por semestre se encuentra entre 1 y 3. Esta caída progresiva también es representada mediante una regresión lineal con pendiente negativa.

Considerando la Ecuación 3.1, en donde se define el total de actividad, la disminución en la cantidad de gente que logra acumular 3, nos habla de dos tipos de procrastinación, la de trabajos prácticos y de exámenes finales [16].

Este comportamiento es consistente con lo discutido en la Sección 3.1, donde se plantea el abandono como un proceso progresivo. En ese sentido, es esperable que, con el paso de los semestres, la cantidad de personas con una actividad baja (total de actividad entre 1 y 3) disminuya porque empiezan a registrar actividad nula. Es decir, muchas de esas trayectorias que inicialmente muestran cierta participación, aunque limitada, terminan derivando en inactividad total, lo que se refleja en el aumento de personas con actividad cero.



(a) Cantidad de personas con total de actividad 0 por semestre relativo con recta de regresión para visualizar la tendencia creciente a medida que pasan los semestres (MSE=4500.08).



(b) Cantidad de personas con total de actividad entre 1 y 3 por semestre relativo con recta de regresión para visualizar la tendencia decreciente a medida que pasan los semestres (MSE=6789.34).

Fig. 5.14: Distribución de la variable target por situación laboral. Fuente: Elaboración propia a partir de datos de la FCEN.

6. MODELADO

Con el objetivo de estudiar el abandono en el cuarto semestre de la carrera, se utilizaron datos del año 2020 en adelante. Se encontraron alumnos con año de inscripción 2021 con registro de materias cursadas en 2020, se supuso que esta inconsistencia fue consecuencia de la pandemia y fueron desechados. Las encuestas realizadas a ingresantes, al estar disponibles a partir de 2023, quedaron fuera del rango temporal considerado. Por lo tanto, el conjunto de datos final contiene 2591 personas.

La estructura final de los datos por persona, se resume visualmente en la Figura 6.1. En las siguientes subsecciones se detalla la ingeniería de atributos.

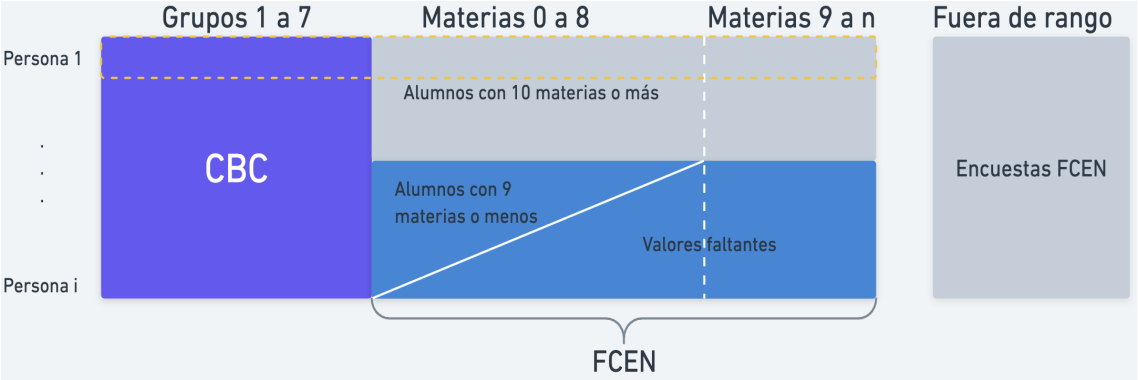


Fig. 6.1: Estructura del conjunto de datos por persona. Se integran los datos del CBC (agrupando las materias en 7 clústeres) y de la FCEN (separando las primeras 9 materias y resumiendo la información de las restantes). La información de las encuestas quedó fuera del alcance temporal de este trabajo. Fuente: Elaboración propia.

6.1. Ingeniería de atributos

En esta sección haremos un recorrido por las variables calculadas a partir de las distintas fuentes de datos.

6.1.1. Semestre relativo y cohorte

Para poder definir si un estudiante se encuentra en riesgo de abandono en su cuarto semestre de carrera, debemos saber en qué momento comenzó sus estudios en la universidad. Por lo tanto, no es suficiente conocer el año de inscripción, debemos identificar cuál fue el primer semestre en el que se registró su actividad académica, lo cual podemos definir a partir de los datos disponibles en SIU-Guaraní de la FCEN.

Las personas pueden comenzar sus estudios en la facultad en el primer semestre del año o en el segundo semestre y, debemos considerar el momento del año en que se suben los registros de inscripciones al sistema. En la Figura 6.2 se presenta la distribución de los registros disponibles en SIU-Guaraní, donde se observa que los registros suelen subirse en los meses de julio, agosto y diciembre. Es por esto que se hizo un corrimiento de

los semestres, de manera que siempre se consideran julio y agosto dentro de un mismo semestre.

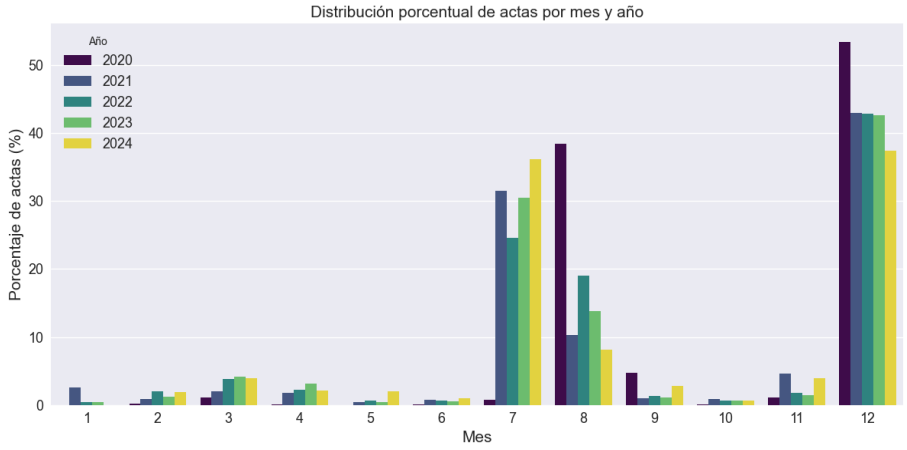


Fig. 6.2: Distribución porcentual de actas registradas entre los años 2020 y 2024 por mes y año. Fuente: Elaboración propia en función de los datos de la FCEN.

A cada alumno se le asignó una cohorte en función del año de inscripción a la FCEN y el semestre en donde tuviera su primer registro. En la Figura 3.1 se muestra de manera ilustrativa cómo se define el número de semestre para los inscriptos en el año 2021.

En el Anexo .1.1 se presenta el listado de semestres relativos por cohorte.

6.1.2. Tiempo de viaje

Como parte de las variables socioeconómicas, inspirados en los resultados obtenidos por Pustilnik y Ndukanma [25], se calculó el tiempo de viaje de los estudiantes utilizando la librería de Google Distance Matrix [24].

Se definieron 0 y más de 4 horas (14,400 segundos) como atípico, dado que se encontraron casos en donde los tiempos de viaje superaban las 24 horas. El tiempo de viaje se puede calcular usando la información otorgada por los estudiantes durante su incripción al CBC o a la FCEN. En el primero se calculó en función de la dirección y localidad, mientras que en el segundo se contaba con el código postal, localidad, departamento, provincia y país. A su vez, la librería utilizada necesita que se especifique el horario de llegada y permite indicar el medio de transporte, por lo tanto, el tiempo de viaje presentado se corresponde con el tiempo que tardan los estudiantes en llegar un día de semana a las 9 de la mañana al Pabellón 1 de la FCEN.

En aquellos casos donde se contaba con información de ambas fuentes, se priorizó la registrada por la FCEN, con el justificativo de que es la fuente de información más reciente.

6.1.3. Clusters de Materias del CBC

Considerando la bibliografía estudiada, como el trabajo de García [15], se definieron variables para tener un historial académico de los alumnos usando los datos del CBC.

Se tomó la decisión de agrupar las materias del CBC considerando la cantidad de información disponible y para limitar la cantidad de columnas con las que se iba a trabajar. Se pueden observar los clusters definidos en la Tabla 6.1. En general, se agruparon aquellas

materias con temarios similares y las equivalentes. Sin embargo, el cluster 1 junta las materias Biología, Biología e Introducción a la Biología Celular, Álgebra y Álgebra A, dado que las primeras dos se corresponden solo con dos carreras de la FCEN y, considerarlo de manera separada implicaba generar un gran desbalance en los datos. A su vez, en dichas carreras no se considera Álgebra como materia obligatoria del CBC, por lo que dependiendo de la carrera las personas tendrán registro solamente para una de las materias del cluster.

Finalmente, las materias de los clusters 1 a 6 se corresponden con materias de carreras de la FCEN o equivalentes. Mientras tanto, el cluster 7 se corresponde con todas las materias que pertenecen al CBC de otras carreras y no presentan equivalencias para el CBC de la FCEN.

Cluster	**Materias**
1	Biología, Biología e Introducción a la Biología Celular, Álgebra y Álgebra A
2	Física
3	Introducción al Conocimiento de la Sociedad y el Estado
4	Introducción al Pensamiento Científico e Introducción al Pensamiento Computacional
5	Matemática 9h y Análisis Matemático A
6	Química
7	Sociología, Principios de DD HH y Derecho Constitucional, Ciencia Política, Principios Generales de Derecho Privado, Semiología, Introducción al Conocimiento Proyectual I, Taller De Dibujo, Economía, Filosofía, Psicología, Historia Económica Social Gral., Introducción al Conocimiento Proyectual II, Antropología, Historia Económica y Social General, Metodología De Las Ciencias Sociales, Trabajo Y Sociedad, Principios Generales del Derecho Latinoamericano, Física e Introducción a la Biofísica, Matemática, Análisis Matemático I y Análisis Matemático

Tab. 6.1: Cluster de materias del CBC basado en similitud de temas. Fuente: Elaboración propia.

Las materias de los clusters 3 y 4 fueron obligatorias para todas las carreras de la FCEN hasta el año 2024, a partir del cual hubo un cambio en el plan de estudios de la Licenciatura en Ciencias de la Computación.

El cluster 5 está compuesto por materias relacionadas con Análisis Matemático y Álgebra. En particular, Álgebra A no forma parte del temario de la FCEN pero se la considera equivalente a Álgebra[1]. Mientras tanto, la materia Matemática 9h pertenece a la Licenciatura y Profesorado en Biología [8].

Una vez definidos los clusters de materias para cada uno de ellos se definieron las siguientes columnas, el calculo de cada una se encuentra detallado en la Ecuación 5.1:

- $NotaCluster_j$: Promedio de nota que sacó, pues cada alumno puede tener más de una materia por cluster. Las filas en donde se encontraban registros con Ausente (A) o No Regularizó (NR) fueron reemplazados por la nota numérica 0. Los aprobados por resolución (AP) fueron reemplazados por el promedio de la persona en el resto de las materias del CBC.

- $\#vecesQueRindioCluster_j$: Cantidad de veces que el alumno rindió la materia del cluster j. Si cursó más de una materia del cluster, se coloca el promedio.

- $FechaCluster_j$: Se registró la fecha promedio, considerando que un alumno pudo cursar más de una materia del cluster.

[1] Resoluciones 78.281/2014 y 23.327/2017

- $UBAXXI_j$: Indica si el alumno cursó alguna de las materias del cluster por la plataforma UBAXXI.

Los datos faltantes en las columnas asociadas a los clusters de materias del CBC reflejan que las materias correspondientes fueron cursadas o rendidas antes del 1 de marzo de 2020, fecha a partir de la cual tenemos registros. Esto indica que el estudiante demoró más tiempo en completar el CBC, lo que constituye una característica relevante del comportamiento académico y concuerda con lo que plantea Zelzman quien estima que el plazo máximo de cursada del CBC es de tres años [40]. Por ello, se optó por conservar los valores nulos en lugar de imputarlos, aprovechando la capacidad de los Árboles de Decisión para manejar datos faltantes, como se detalla en la Sección 4.4.3.

Como se detalla en la Sección 5.3, no se utilizaron las columnas asociadas al cluster 7 para el entrenamiento del modelo.

6.1.4. Años de cursada del CBC

Con la información de las materias rendidas por estudiantes incriptos al CBC de alguna de las carreras de la FCEN, se decidió estimar la cantidad de años que transcurrieron desde que comenzaron el CBC hasta que iniciaron sus estudios en la facultad.

Por como se definieron los clusters, cada estudiante inscripto en la facultad entre los años 2021 y 2022 debe tener al menos una materia cursada por cluster, exceptuando el septimo cluster. En el caso de que se cumpla esa condición y sabiendo que tenemos los registros del CBC de marzo de 2020 en adelante, podemos suponer que el estudiante inscripto en 2021 demoró un año en hacer el CBC y, en caso contrario, comenzó el CBC por lo menos dos años antes. Para los estudiantes inscriptos en la facultad en el año 2022, en caso de no tener un registro para todos los clusters, podemos estimar que comenzó dicho etapa por lo menos tres años antes. En caso contrario, en función del primer registro del CBC del estudiante se puede estimar el tiempo que se demoró en cursar dicha etapa al saber que en 2022 comenzó la facultad.

6.1.5. Edad de inscripción

Al tener acceso a la edad del estudiante al día 7 de mayo de 2025, se estimó la edad de los estudiantes en su año de inscripción a la facultad. Si se inscribieron en 2021, entonces se le restaron cuatro años a la edad registrada en 2025, pero si su año de inscripción fue 2022, se le restaron solamente tres años.

Se encontró un caso en donde la edad registrada era mayor a cien, por lo que se lo tomó como dato atípico y fue marcado como dato faltante.

6.1.6. Materias de la FCEN

Considerando la bibliografía estudiada, como el trabajo de García [15], se definieron variables para tener información académica de los alumnos en los primeros años de la FCEN.

Por cada semestre relativo i (i tomando valores 0 a 3 inclusive) se definieron las siguientes columnas, todas se corresponden con las definidas para el calculo de total de actividad por semestre relativo (Ecuación 3.1):

- $\#inscripciones_i$: Cantidad de inscripciones en el $semestre_i$.

- $\#TPsAprobados_i$: Para los trabajos prácticos (excepto casos específicos) no tenemos nota numérica, solamente si aprobó, reprobó o dejó la materia, por lo que tomamos la cantidad de materias con todos los trabajos aprobados.
- $\#finales_i$: Cantidad de exámenes finales inscriptos en el $semestre_i$.

Al igual que con el CBC, se agregaron las materias y finales rendidos como columnas. Sin embargo, en el CBC se definieron clusters mientras que en este caso se tomó el orden en que los alumnos rindieron. Para los primeras 9 materias de cada estudiante se definieron las siguientes columnas con el objetivo de tener una distribución empírica de qué rinden los estudiantes y cada cuánto:

- $FechaInscripcion_x$: Es un número que indica los días desde 1/1/2020. Se corresponde con la fecha de inscripción a la materia x.
- $TPAprobado_x$: 1 si aprobó los trabajos prácticos de la materia, 0 si no.
- $FechaFinal_x$: Es un número que indica los días desde 1/1/2020. Se corresponde con la fecha del exámen final de la materia x.
- $NotaFinal_x$: Nota obtenida en el exámen final de la materia x.

En caso de que el estudiante se inscribiera a la materia más de una vez, se definieron columnas distintas informando sobre las distintas para cada uno de los registros.

Se añadió un conjunto de cuatro columnas más que se completan con el promedio del resto de las materias que tenga la persona, como se presenta de manera ilustrativa en la Figura 6.3 y se detalla en la Ecuación 5.3.

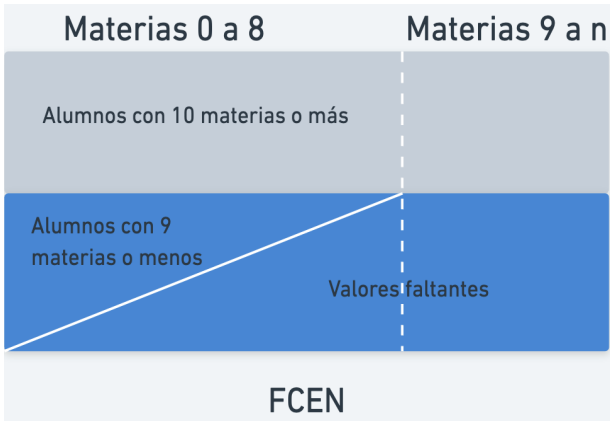


Fig. 6.3: Definición de columnas con información de las materias por alumno. Los rectángulos representan el universo de alumnos divididos en dos grupos disjuntos, aquellos que tienen diez materias o más y aquellos que tienen nueve materias o menos. Los estudiantes con diez materias o más tendrán un promedio de sus notas para poder resumirla en una sola columna. Fuente: Elaboración propia.

En este contexto, los datos faltantes en las columnas asociadas a las materias de la FCEN, incluyendo las primeras nueve materias y las columnas de promedios, reflejan la ausencia de actividad académica del estudiante, como la no inscripción en materias o la no rendición de trabajos prácticos o exámenes finales. Esta interpretación permitió conservar

los valores nulos como una representación válida del comportamiento académico, en lugar de imputarlos, lo que se alinea con la robustez de los Árboles de Decisión frente a datos faltantes, como se describe en la Sección 4.4.3.

6.1.7. Carrera Principal

Cada alumno de la FCEN puede inscribirse a múltiples carreras, por lo que se tomó la decisión de definir una "carrera principal"por persona.

Al inscribirse a una materia o exámen final por medio del sistema SIU-Guaraní, el estudiante debe indicar la carrera. En consecuencia, se estableció la cantidad de registros que tenía cada persona por carrera y aquella con la mayor cantidad se definió como la "carrera principal".

6.1.8. Variables categóricas

Para el trabajo con los modelos seleccionados y el posterior análisis de importancia de atributos, se hizo una tarea de label encoding de las variables categoricas, como sexo, cohorte, carrera, nivel de estudio de los padres y situación laboral.

Al utilizar label enconding cada categoría se indexa con valores numéricos, esto puede engañar a los árboles de decisión, pues toman decisiones binarias como $variable \leq x$ o $variable \geq x$, y al tratarse de variables categóricas estaría realizando una partición artificial. Una forma de evitar esto es utilizar la técnica de one-hot-encoding, sin embargo genera una variable nueva por cada categoría y puede dificultar el estudio de importancia de atributos del modelo.

En el caso de las carreras, se agruparon en una misma categoría aquellas que cuya proporción fue inferior al 2 % del total.

6.2. Importancia de atributos

Es fundamental destacar que, al realizar el análisis de importancia de atributos en los modelos de Aprendizaje Automático, las variables identificadas como relevantes para las predicciones del modelo no necesariamente reflejan las causas reales del abandono estudiantil. El estudio de las verdaderas razones del abandono requiere investigaciones complementarias que vayan más allá del marco predictivo del modelo.

Esta limitación no deriva de las técnicas empleadas para evaluar la importancia de los atributos, como las métricas basadas en la contribución de cada variable a la reducción de la impureza en los Árboles de Decisión, sino de la complejidad del problema con el que se está trabajando.

6.3. Conjunto de datos final

Se resumen los datos utilizados para el entrenamiento de los modelos en la Tabla 6.2

Campo o Grupo	**Descripción**
Tiempo de viaje	Tiempo de viaje en segundos del alumno desde su vivienda hacia la FCEN
Edad	Edad del estudiante al inscribirse a la FCEN
Género	Femenino o Masculino
Carrera	La carrera donde el estudiante tiene más inscripciones a exámenes y cursadas de la FCEN
Nivel Estudio Madre y Padre	Nivel máximo de estudios alcanzado por los padres según lo informado por el alumno
Situación laboral	Indica si el alumno trabaja y/o busca trabajo según su declaración. Es una variable categórica
Datos cluster j del CBC	Se corresponde con las cuatro columnas explicadas en la Ecuación 5.1
Año inscripción facultad	Año en que el alumno se inscribió en la Facultad
Datos materias i de la FCEN	Se corresponde con las cuatro columnas explicadas en las Ecuaciones 5.2 y 5.3. Hay 10 de estos grupos.
Variables resumen	Cantidad de inscripciones, trabajos practicos aprobados y finales inscriptos por semestre relativo en la FCEN. Explicadas en la Ecuación 3.1

Tab. 6.2: Campos utilizados para el entrenamiento de modelos, se agruparon variables para facilitar la lectura. Fuente: Elaboración propia.

7. EXPERIMENTOS

7.1. Definición de hiperparámetros

Al trabajar con Árboles de Desición, debemos definir los hiperparámetros con los que se va a trabajar. Es por ello que antes de comenzar los experimentos, graficamos la curva de complejidad de un Árbol de Decisión en función de su altura máxima y usando el criterio de división Gini, el resto de hiperparámetros se dejaron por default:

```
modelo = DecisionTreeClassifier(
            criterion='gini',  # Gini
            splitter='best',   # Selecciona la mejor división
            max_depth=max_depth,   # Límite de profundidad del árbol
            min_samples_split=2,   # Mínimo de muestras para dividir un nodo
            min_samples_leaf=1,    # Mínimo de muestras en una hoja
            max_features=None, # Considera todas las características
            random_state=42    # Para reproducibilidad
            )
```

Estos parámetros y las métricas que resultaron de los experimientos se encuentran definidos el siguiente código fuente [7], al estar trabajando con datos personales de los alumnos, los mismos no se encuentran publicados.

Para cada altura del árbol se usó la técnica de validación cruzada (*cross validation*) con cinco carpetas (*folds*), respetando la distribución de la variable a predecir. Luego, se hizo un promedio de la métrica AUCROC entre los folds de *train* y, por otro lado, los de *test*, obteniendo así la Figura 7.1 donde se puede observar que con altura máxima 3 se alcanza el mejor AUCROC en *test*.

Por lo tanto, para el resto de los experimentos se definió que la altura máxima de los árboles sea 3.

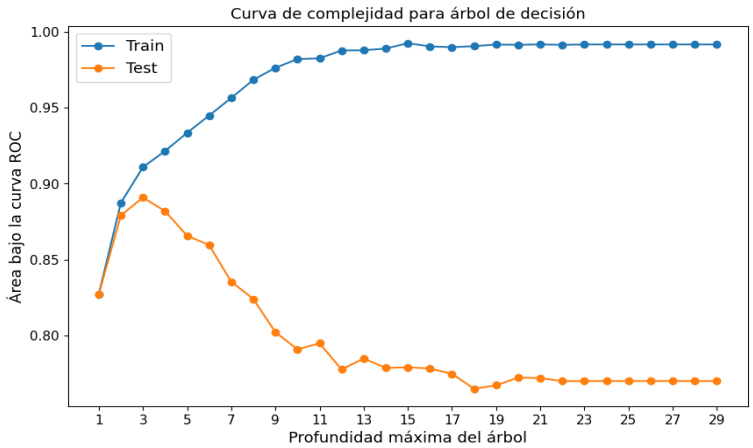


Fig. 7.1: Curva de complejidad para Árbol de Decisión con distintos niveles de profundidad máxima. Fuente: Elaboración propia.

7.2. Primer Experimento: Ensamble con subespacio aleatorio

El primer experimento tiene como objetivo construir una línea base de evaluación, a partir de la cual se buscará mejorar el rendimiento en los experimentos posteriores. Para ello, se dividió al conjunto de datos en dos, 80 % para entrenamiento (*train*) y 20 % para evaluación (*test*). La partición se realizó de forma estratificada para mantener la proporción de la variable a predecir, pero se hizo una única vez, sin repetir el proceso con diferentes semillas o particiones.

Con los datos de entrenamiento se entrenaron ocho modelos de ensamble de Árboles de Decisión con subespacio aleatorio. En todos los casos se utilizaron árboles con altura máxima fija de 3, según lo definido previamente a partir del análisis de la curva de complejidad (ver Figura 7.1). Este método se caracteriza porque cada árbol del ensamble se entrena con el conjunto de entrenamiento completo (sin utilizar bootstrap), mientras que la aleatoriedad se introduce al considerar en cada nodo la raíz cuadrada de n variables al azar (n siendo la cantidad total de columnas) y se selecciona la mejor usando el criterio de Gini o Entropía según la configuración.

Las combinaciones evaluadas surgen del cruce entre dos funciones de impureza (Gini y Entropía) y cuatro cantidades distintas de árboles: 5, 10, 15 y 20 árboles. De esta manera, se busca observar el impacto de estas dos variables sobre el rendimiento del modelo, manteniendo constante la profundidad.

Una vez entrenados, los modelos fueron evaluados sobre el conjunto de *test* utilizando tres métricas: Exactitud, Exactitud Balanceada y AUC ROC. Estas métricas permiten tener una mirada amplia sobre el rendimiento del modelo, considerando tanto su capacidad de clasificación general como su comportamiento frente a una variable *target* moderadamente desbalanceada, como se explicó en el Capítulo 4.

Los resultados se presentan en la Tabla 7.1, en donde todos los modelos dieron métricas superiores a 0.80. La configuración de 10 árboles y criterio de corte dado por Entropía, obtuvo los mejores resultados en Exactitud (0.842) y Exactitud Balanceada (0.835). Mientras que el mejor AUC (0,925) se obtuvo con el modelo de 15 árboles y usando el criterio de corte Gini.

Cant. de árboles	**Gini**			**Entropía**		
	Exact.	**Exact. Bal.**	**AUCROC**	**Exact.**	**Exact. Bal.**	**AUCROC**
5	0.832	0.826	0.910	0.836	0.830	0.913
10	0.838	0.832	0.918	**0.842**	**0.835**	0.918
15	0.829	0.822	**0.925**	0.829	0.822	0.919
20	0.827	0.821	0.917	0.834	0.828	0.921

Tab. 7.1: Resultados Experimento 1: Exactitud, Exactitud Balanceada, y AUCROC para criterio Gini y Entropía. Se repitió para 5, 10, 15 y 20 árboles. Todas las métricas fueron calculadas con el conjunto de *test*. Fuente: Elaboración propia.

A su vez, en la Tabla 7.2 se presenta la media y varianza de las métricas calculadas en los Árboles de Decisión que conforman los modelos entrenados. En la misma se observa una baja varianza entre los árboles y si lo comparamos con la Tabla 7.1 notamos que la media de los árboles es inferior a los resultados obtenidos en el ensamble de modelos.

Si bien este experimento permite obtener una primera aproximación al desempeño del modelo, presenta dos limitaciones importantes. Por un lado, al utilizar una única partición en *train* y *test*, existe la posibilidad de que el conjunto de evaluación no sea representativo, lo que puede afectar la confiabilidad de las métricas. Por otro lado, al no utilizar la técnica

Cant. de árboles	Gini						Entropía					
	Exact.		Exact. Bal.		AUCROC		Exact.		Exact. Bal.		AUCROC	
	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.
5	0.820	0.015	0.818	0.011	0.871	0.010	0.820	0.015	0.817	0.010	0.877	0.007
10	0.814	0.028	0.811	0.026	0.870	0.022	0.815	0.028	0.811	0.026	0.873	0.022
15	0.810	0.024	0.807	0.023	0.861	0.024	0.812	0.024	0.808	0.022	0.872	0.022
20	0.814	0.022	0.810	0.021	0.867	0.024	0.816	0.022	0.812	0.020	0.876	0.021

Tab. 7.2: Media y varianza de los árboles de decisión que conforman los modelos entrenados cuyos resultados se presentan en la Tabla 7.1. Las métricas presentadas son Exactitud (media y varianza), Exactitud Balanceada (media y varianza) y AUC ROC (media y varianza). Todas las métricas fueron calculadas con el conjunto de *test*. Fuente: Elaboración propia.

de *bootstrap*, todos los árboles del ensamble se entrenan sobre el mismo conjunto de datos, lo que limita la diversidad entre ellos y, por ende, la capacidad del modelo para reducir la varianza y evitar mínimos locales. Estas limitaciones son abordadas en el segundo experimento.

7.2.1. Importancia de atributos

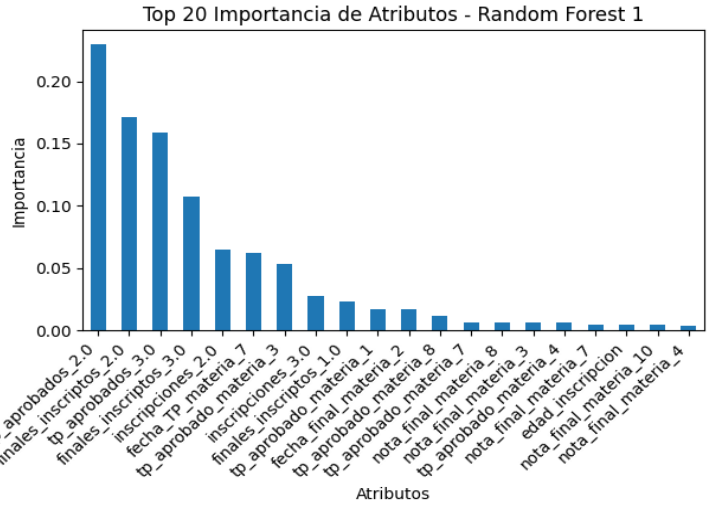


Fig. 7.2: Importancia de atributos del modelo *Random Forest* con 10 árboles y criterio de medición de impureza Gini. Fuente: Elaboración propia.

Al analizar la importancia de atributos de los modelos entrenados, notamos que en todos las variables con mayor importancia estaban relacionadas con las que se utilizan para el cálculo de total de actividad (ver Ecuación 3.1). Esto puede ser consecuencia de que los Árboles de Decisión pueden inferir la suma, representada en el árbol, cuando se divide el nodo en dos hojas y, en el análisis de datos, quedó en evidencia la correlación lineal entre el total de actividad por semestre relativo y la variable a predecir.

Se presenta modo de ejemplo la Figura 7.2, en donde se pude observar la importancia de atributos de las variables utilizadas para el entrenamiento del modelo. Entre las variables con mayor importancia se encuentran las que contabilizaban las materias con trabajos prácticos y finales aprobados en los semestres relativos 2 y 3. Le siguen la variable con la cantidad de inscripciones de los semestres relativos 2 y 3. El resto de las variables se encuentran relacionadas con fechas o las notas obtenidas en exámenes finales y trabajos

prácticos.

A su vez, en la Figura 7.3 se presenta uno de los árboles que conforman al modelo con 10 árboles y criterio de medición de impureza Gini. Se puede observar cómo la suma del total de actividad (Ecuación 3.1) en el semestre relativo 3, queda representada en la rama de la izquierda, cuyos nodos se encuentran en azul. Al estar trabajando con varias con valores enteros, vemos que en dicha rama se separan a aquellos alumnos cuyo total de actividad en el semestre 3 es nula.

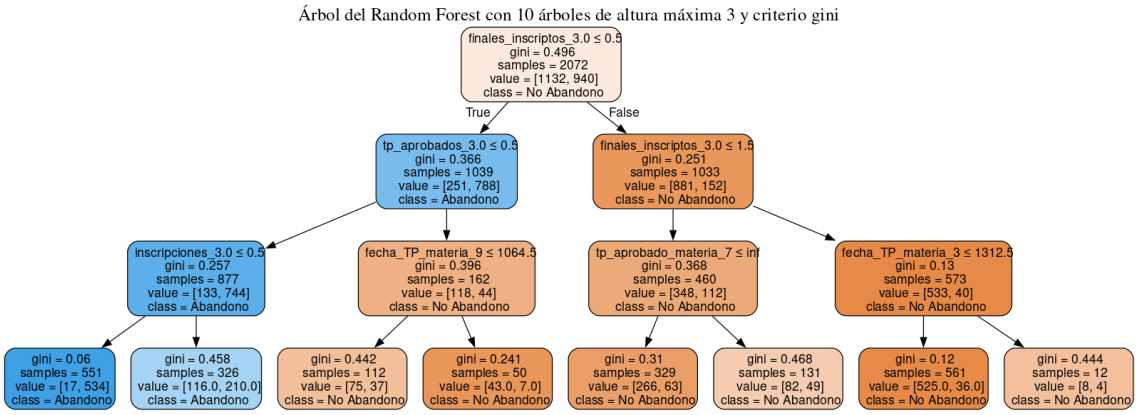


Fig. 7.3: Instancia de árbol de *Random Forest* y criterio de impureza Gini. Se puede observar que la rama celeste calcula la variable suma de total de actividad (Ecuación 3.1). Fuente: Elaboración propia.

7.3. Segundo Experimento: Random Forest

Para superar las limitaciones del primer experimento, en esta segunda fase se entrenaron modelos de *Random Forest*, incorporando ahora el muestreo con reemplazo (*bootstrap*) en la construcción de los árboles del ensamble, y utilizando la técnica de *Monte Carlo Cross-validation* para una evaluación más robusta. En este caso, se realizaron cinco iteraciones independientes para cada configuración, cada una con una partición aleatoria del conjunto de datos, reservando un 20 % para testeo en cada caso. El promedio de las métricas obtenidas en estas iteraciones se utilizó como estimador del desempeño del modelo, con el objetivo de mitigar la posibilidad de que los resultados estén influenciados por una partición poco representativa, tal como se advirtió en el experimento anterior. Esto se resume en el Algoritmo 2.

Los resultados obtenidos (Tabla 7.3) muestran mejoras en algunas de las métricas con respecto al primer experimento. En términos de Exactitud, los mejores desempeños promedio (0.850) se alcanzaron con modelos entrenados con 5 y 10 árboles. Para el caso de 5 árboles, tanto el criterio de impureza de Gini como el de Entropía ofrecieron resultados similares, mientras que con 10 árboles el mejor resultado se obtuvo empleando el criterio de Gini.

Respecto a la Exactitud Balanceada, el mejor valor promedio (0.845) fue obtenido con 5 árboles y el criterio Gini. En cuanto al AUC ROC, el mejor resultado (0.921) se observó con 20 árboles utilizando indistintamente Gini o Entropía.

Algorithm 2 Segundo Experimento

for cada configuración del modelo **do**
  Inicializar lista de métricas para la configuración actual
  for $i = 1$ **to** 5 **do**
    $D_{test\_i} \leftarrow$ Seleccionar 20 % de los datos de forma aleatoria
    $D_{train\_i} \leftarrow$ El 80 % de datos restantes
    $modelo\_i \leftarrow$ Entrenar el modelo con D_{train_i}
    $metricas\_i \leftarrow$ Calcular métricas de rendimiento de $modelo\_i$ usando D_{test_i}
    Agregar $metricas\_i$ a la lista de métricas
  Calcular el promedio de las métricas obtenidas para la configuración actual

Algoritmo 2: Experimento 2.

Comparando estos resultados con los del primer experimento, se observa una leve mejora general en las métricas. La Exactitud pasó de un máximo de 0.842 en el primer experimento a 0.850. La Exactitud Balanceada mejoró de 0.835 a 0.845. En el caso del AUC ROC, si bien el primer experimento había alcanzado un valor ligeramente superior de 0.925, el segundo experimento obtuvo un valor muy cercano de 0.921, lo cual sigue indicando un buen desempeño del modelo.

Cant. de árboles	Gini						Entropía					
	Exact.		Exact. Bal.		AUCROC		Exact.		Exact. Bal.		AUCROC	
	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.
5	**0.850**	**0.005**	**0.845**	**0.006**	0.914	0.005	**0.850**	**0.011**	0.844	0.012	0.916	0.004
10	**0.850**	**0.011**	0.844	0.012	0.918	0.004	0.849	0.013	0.842	0.014	0.920	0.003
15	0.845	0.011	0.839	0.012	0.918	0.004	0.845	0.013	0.840	0.014	0.920	0.004
20	0.848	0.013	0.842	0.014	**0.921**	**0.004**	0.849	0.012	0.843	0.013	**0.921**	**0.003**

Tab. 7.3: Resultados Experimento 2: Exactitud, Exactitud Balanceada, y AUCROC para criterio Gini y Entropía. Se repitió para 5, 10, 15 y 20 árboles. En cada celda se muestra media y varianza del experimento. Todas las métricas fueron calculadas con el conjunto de *test*. Fuente: Elaboración propia.

7.3.1. Importancia de atributos

Dado que en este segundo experimento se realizaron cinco iteraciones por cada configuración del modelo, se calculó la importancia promedio de cada atributo considerando su valor en cada una de las ejecuciones.

Los resultados obtenidos son similares a los del primer experimento, los atributos más importantes continúan siendo aquellos vinculados al cálculo del total de actividad por semestre (Ecuación 3.1). En particular, destacan los valores asociados a los semestres relativos 2 y 3, así como también las fechas y las notas de los exámenes rendidos. En la Figura 7.4 se presentan los veinte atributos con mayor importancia promedio del modelo *Random Forest* con 20 árboles y criterio de medición de impureza Entropía.

Si bien la distancia de viaje, u otros datos personales de los alumnos como la edad y el nivel de estudios de los padres resultaron relevantes en otros trabajos [15, 25], en esta tesis resultaron ser menos relevantes que las fechas. Es decir, dentro del marco de este modelo predictivo, aportaron más información para predecirlo. Es importante aclarar que las variables que resultan importantes para el modelo pueden no ser las causas reales del abandono, pudiendo existir factores subyacentes que aún deben estudiarse.

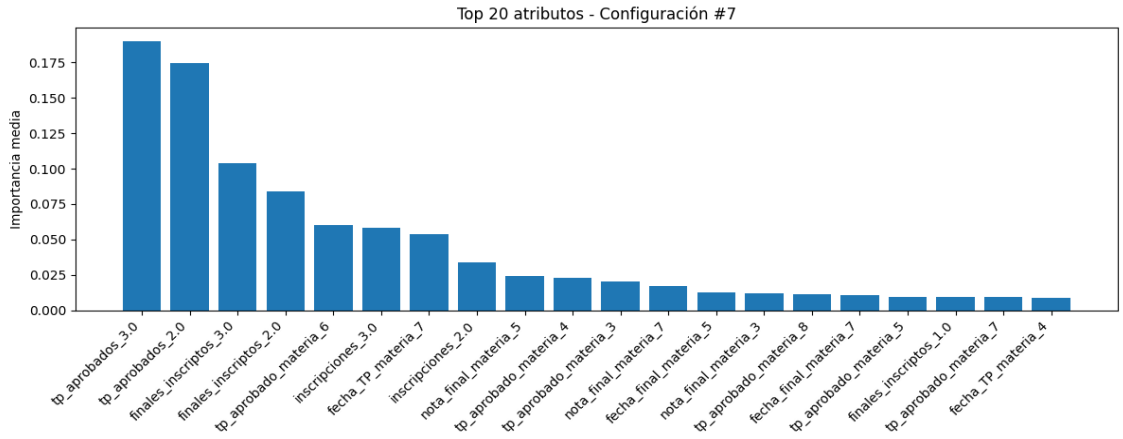


Fig. 7.4: Los veinte atributos con mayor importancia promedio del modelo *Random Forest* con 20 árboles y criterio de medición de impureza Entropía. Fuente: Elaboración propia.

8. CONCLUSIÓN

El presente trabajo surge motivado por una problemática del Sistema Universitario Argentino: las bajas tasas de egreso en tiempo teórico. A nivel nacional, el promedio de egresos es del 23.06 %, un valor que se refleja en la FCEN (UBA), donde apenas alcanza el 19.96 % para quienes inician el CBC. Aunque los números mejoran al considerar únicamente a quienes ingresan a la facultad, el porcentaje de egresados sigue siendo inferior al 50 %, lo cual refuerza la necesidad de herramientas efectivas para identificar y acompañar a estudiantes en riesgo de abandono.

En este contexto, la FCEN ya ha implementado diversas iniciativas, como tutorías y encuestas a ingresantes, para mitigar el abandono. Nuestro trabajo busca colaborar con esos esfuerzos mediante el desarrollo de modelos de Aprendizaje Automático que permitan generar alertas tempranas de abandono. Para ello, definimos el "riesgo de abandono" basada en un umbral de actividad académica menor a 3, que permite identificar estudiantes en riesgo antes de que efectivamente cesen su actividad académica (Ecuación 3.2). En el Gráfico 3.2 se evidencia que la cantidad de actividad disminuye a medida que avanzan los semestres, lo que motivó la elección del umbral 3, pudiéndose buscar otros umbrales en trabajos futuros.

Realizamos un estudio exhaustivo de trabajos previos, en donde se ve que inicialmente las investigaciones buscaban explicar el abandono de los estudiantes, basándose en la comparación entre los estudiantes que abandonaban con quienes no lo hacían.

Con el objetivo de predecir el abandono, se comienzan a usar los modelos de ecuaciones estructurales principalmente basados en encuestas, donde la explicabilidad estaba dada por la encuesta en sí misma. Sin embargo, esto requería encuestar al 100 % de la población para implementar el modelo. En la práctica, en general no se puede encuestar a toda la población, en ese sentido se comenzaron a usar algoritmos de Aprendizaje Automático para poder inferir las variables importantes del abandono.

Para poder hacer uso de estas técnicas, nos apoyándonos en referencias como García de Fanelli [15], el trabajo de Pustilnik y Ndukanma [25] e investigaciones realizadas dentro de la propia facultad, entre otros.

Trabajamos con datos provistos por el SIU-Guaraní del CBC y de la FCEN correspondientes a los años 2021 y 2022. No se utilizaron los datos de encuestas por estar disponibles recién a partir de 2023. En el análisis de datos observamos que la variable a predecir (riesgo de abandono) se encontraba levemente desbalanceada, por lo cual decidimos evaluar nuestros modelos utilizando métricas como la Exactitud, la Exactitud Balanceada y el área bajo la curva ROC (AUC ROC).

A partir del análisis exploratorio de los datos surgieron hallazgos relevantes, se encontró una relación entre el nivel educativo de los padres y el riesgo de abandono, a mayor nivel educativo, menor riesgo. Asimismo al comenzar la carrera: los estudiantes con mayor edad, aquellos que trabajaban o buscaban empleo, presentaban un mayor riesgo de abandono que el abandono poblacional medio. No se hallaron diferencias significativas según la carrera elegida.

A continuación, se entrenaron y compararon dos tipos de ensambles basados en árboles de decisión: un ensamble con subespacio aleatorio y un modelo de *Random Forest*. Se definió una altura máxima de los árboles igual a 3, para evitar el sobreajuste. Los experi-

mentos mostraron que la importancia relativa de los atributos se mantiene estable entre ambas configuraciones.

En todos los modelos entrenados se encontró que las variables que contabilizaban las materias con trabajos prácticos y finales aprobados en los semestres relativos 2 y 3 resultaron ser importantes. Sin embargo, es importante aclarar que las variables que resultan importantes para el modelo pueden no ser las causas reales del abandono, pudiendo existir factores subyacentes que aún deben estudiarse.

Comparando los resultados de ambos experimentos, se observa una leve mejora general en las métricas del segundo experimento. El mejor resultado de Exactitud en el primer experimento fue de 0.842 mientras que en el segundo se obtuvo 0.850. Al comparar los mejores resultados de ambos experimentos, la Exactitud Balanceada del segundo experimento mejoró de 0.835 a 0.845. En el caso del AUC ROC, si bien el primer experimento había alcanzado un valor ligeramente superior de 0.925, el segundo experimento obtuvo un valor muy cercano de 0.921, lo cual sigue indicando un buen desempeño del modelo.

Estos resultados respaldan la hipótesis de que es posible construir modelos que anticipen el riesgo de abandono de manera efectiva. Esta herramienta puede servir como insumo valioso para la FCEN (¡e incluso otras facultades!), permitiendo focalizar esfuerzos de acompañamiento en quienes más lo necesitan.

Queda como línea futura de trabajo la integración de encuestas a ingresantes cuando estén dentro del rango de fechas trabajadas, la implementación de reportes personalizados y el seguimiento longitudinal de cohortes para mejorar continuamente los modelos y estrategias de retención estudiantil.

9. TRABAJOS FUTUROS

A partir del trabajo realizado, se identifican diversas líneas de acción que podrían enriquecer y extender el análisis efectuado en esta tesis, con el objetivo de perfeccionar los modelos predictivos de riesgo de abandono y potenciar las estrategias de acompañamiento institucional.

En primer lugar, se propone como línea futura de trabajo la integración de los datos provenientes de las encuestas a ingresantes al momento en que estos datos se encuentren disponibles dentro del rango temporal utilizado para el modelado. Esto permitirá complementar la información académica y demográfica con dimensiones socioeconómicas enriqueciendo así la representación del perfil estudiantil [15].

En segundo lugar, quedó pendiente la implementación de un sistema de reportes personalizados, que permita comunicar los resultados del modelo a los equipos de tutoría o personal administrativo. Estos reportes podrían incluir alertas por estudiante en riesgo junto con las variables que más influyeron en dicha predicción, facilitando acciones de intervención específicas y tempranas.

Una tercera dirección a desarrollar es el seguimiento longitudinal de cohortes, es decir, extender el horizonte temporal del análisis incorporando datos de años posteriores. Esto permitirá evaluar la evolución de los estudiantes a lo largo de su trayectoria, reentrenar modelos con nueva información.

También sería de interés explorar distintos valores de umbrales para definir el riesgo de abandono según la actividad total por semestre (Ecuación 3.1), como así también utilizar técnicas de validación cruzada específicas para identificar el umbral óptimo en función del trade-off entre sensibilidad y especificidad.

Otro aspecto a considerar en trabajos futuros es la incorporación y comparación de distintos modelos de Aprendizaje Automático, como XGBoost, LightGBM o redes neuronales, con el fin de mejorar la capacidad predictiva y robustez del sistema. Asimismo, se podría evaluar el uso de modelos explicativos (como SHAP) para aumentar la interpretabilidad de las predicciones.

Por otro lado, dado que el dataset utilizado contiene datos faltantes, sería valioso investigar métodos de imputación estadística o basada en modelos para completar esa información sin introducir sesgos.

Finalmente, durante el proceso de limpieza y selección de datos, se descartaron personas inscriptas en 2021 que tenían registros previos a esa fecha, lo cual podría estar vinculado a la pandemia.

Asimismo, se encontró que en las actas de la facultad correspondientes a los años 2021 y 2022 figuraban estudiantes para los cuales no se encontraron registros del CBC entre los años 2020 y 2025. Esta situación sugiere que podría ser de utilidad ampliar la base de datos del CBC y obtener una visión más completa de las trayectorias estudiantiles en la FCEN.

BIBLIOGRAFíA

[1] Josephine Akosa. «Predictive accuracy: A misleading performance measure for highly imbalanced data». En: *SAS Institute Inc.* 12 (2017), págs. 1-4.

[2] John P. Bean y Barbara S. Metzner. «A Conceptual Model of Nontraditional Undergraduate Student Attrition». En: *Review of Educational Research* 55.4 (1985), págs. 485-540. ISSN: 00346543, 19351046. URL: jstor.org/stable/1170245 (visitado 30-03-2025).

[3] L. Breiman et al. *Classification and Regression Trees*. Chapman y Hall/CRC, 1984. DOI: https://doi.org/10.1201/9781315139470.

[4] Leo Breiman. «Random forests». En: *Machine learning* 45 (2001), págs. 5-32. DOI: doi.org/10.1023/A:1010933404324.

[5] .UBA Universidad de Buenos Aires. *"Más datos UBA"*. URL: uba.ar/datosuba (visitado 31-05-2025).

[6] Universidad de Buenos Aires. *Normas Generales de la Universidad de Buenos Aires, Libro I, Título 16, Capítulo C, Artículo 213: Régimen para el Ciclo Básico Común de Inscripción, Cursado, Evaluación y Promoción de Asignaturas*. 2023. URL: codigo.rec.uba.ar/codigo_uba/libro-i-normas-generales-de-la-universidad-de-buenos-aires-1/titulo-16-ciclo-basico-comun/capitulo-c-regimen-para-el-ciclo-basico-comun-de-inscripcion-cursado-evaluacion-y-promocion-de-asignaturas/ (visitado 21-06-2025).

[7] Sol Calloni, Martin Pustilnik y Guillermo Durán. *"Tesis de Licenciatura de Sol Calloni"*. URL: github.com/solcalloni/modelos-para-prediccion-de-abandono (visitado 09-06-2025).

[8] UBA CBC. *"Carreras"*. URL: cbc.uba.ar/carreras (visitado 31-05-2025).

[9] UBA CBC. *"Toda la información que necesitás para ingresar al CBC"*. URL: cbc.uba.ar/inscripciones (visitado 31-05-2025).

[10] .UBA Exactas Facultad de Ciencias Exactas y Naturales. *"Carreras de Grado"*. URL: exactas.uba.ar/ (visitado 31-05-2025).

[11] Facultad de Ciencias Exactas y Naturales. «Resolución (CD) N 1482/98 de la Facultad de Ciencias Exactas y Naturales». En: *Expte. Nro 437.320/85* (1998).

[12] Ministerio de Educación y Cultura de Uruguay. *"Anuario Estadístico de Educación"*. URL: gub.uy/ministerio-educacion-cultura/datos-y-estadisticas/datos?page=0 (visitado 31-05-2025).

[13] Carla de Erausquin. «Caracterización de trayectorias educativas a partir de producciones de código». En: *Tesis para la Licenciatura en Ciencias de Datos, UBA* (2024).

[14] .UBA Exactas. *"+Acompañamiento"*. URL: exactas.uba.ar/acompanamiento/ (visitado 07-06-2025).

[15] Ana María 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* (2014). ISSN: 1852-8171. URL: ri.conicet.gov.ar/handle/11336/35674.

[16] Michael Jenik. «Characterization of procrastination patterns of university students in academic subjects». En: *Tesis para la Licenciatura en Ciencias de la Computación, UBA* (2015).

[17] Ed Machina. *"Inteligencia Artificial para la retención y el éxito estudiantil en tiempo real"*. URL: https://edmachina.com/ (visitado 31-05-2025).

[18] .UBA Ciencias Médicas. *"Preguntas frecuentes de alumnos de Licenciaturas / Tecnicaturas: ¿Cuánto tiempo de vigencia tiene la regularidad?"* URL: fmed.uba.ar/preguntas-frecuentes/preguntas-frecuentes-de-alumnos-de-licenciaturas-tecnicaturas (visitado 31-05-2025).

[19] A. C. Müller y S. Guido. *Introduction to machine learning with Python: A guide for data scientists (cap. 2: Supervised Learning)*. O'Reilly Media, Inc., 2016.

[20] A. C. Müller y S. Guido. *Introduction to machine learning with Python: A guide for data scientists (cap. 5)*. O'Reilly Media, Inc., 2016.

[21] Dirección de Orientación Vocacional de la FCEN (DOV). *2.c) Cuadro evolución matrícula FCEN 2004-2024*. URL: exactas.uba.ar/extension/ov/#menu-7 (visitado 31-05-2025).

[22] Manuel S. Ortiz y Montserrat Fernández-Pera. «Modelo de Ecuaciones Estructurales: Una guía para ciencias médicas y ciencias de la salud». es. En: *Terapia psicológica* 36 (abr. de 2018), págs. 51-57. ISSN: 0718-4808. URL: scielo.cl/scielo.php?script=sci_arttext&pid=S0718-48082018000100051&nrm=iso.

[23] Ernest T. Pascarella y Patrick T. Terenzini. «Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model». En: *The Journal of Higher Education* 51.1 (1980), págs. 60-75. DOI: 10.1080/00221546.1980.11780030.

[24] Google Maps Platform. *"Descripción general de la API de Distance Matrix"*. URL: developers.google.com/maps/documentation/distance-matrix/overview (visitado 31-05-2025).

[25] Martin Pustilnik y Gianluca Ndukanma. «Modelos para la predicción del abandono en la Universidad Nacional de Hurlingham». En: *Biblioteca de la Universidad Nacional de la Plata* (jun. de 2023). URL: sedici.unlp.edu.ar/handle/10915/155839.

[26] J. Ross Quinlan. «Induction of decision trees». En: *Machine learning* 1 (1986), págs. 81-106. DOI: doi.org/10.1007/BF00116251.

[27] scikit-learn. *"1.11.2.1. Random Forests"*. URL: https://scikit-learn.org/stable/modules/ensemble.html#random-forests (visitado 31-05-2025).

[28] scikit-learn. *"DecisionTreeClassifier"*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier (visitado 31-05-2025).

[29] scikit-learn. *"RandomForestClassifier"*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier (visitado 31-05-2025).

[30] Andrés Ignacio Santos Sharpe. «Discontinuar los estudios en la universidad». En: *Tesis Doctoral en Ciencias Sociales* (2018). URL: teseopress.com/discontinuidadestudiosuniversitarios.

[31] Tomás Spognardi, Manuela Cerdeiro y Matías López y Rosenfeld. «Predicción de abandono en ingresantes a las carreras de la Facultad de Ciencias Exactas y Naturales». En: (2024).

[32] Vincent Tinto. «Dropout from Higher Education: A Theoretical Synthesis of Recent Research». En: *Review of Educational Research* 45.1 (1975), págs. 89-125. ISSN: 00346543, 19351046. DOI: 10.2307/1170024.

[33] Vincent Tinto. «Limits of Theory and Practice in Student Attrition». En: *The Journal of Higher Education* 53.6 (1982), págs. 687-700. ISSN: 00221546, 15384640. URL: jstor.org/stable/1981525 (visitado 17-03-2025).

[34] UBA. *"Informacion estadistica uba"*. URL: informacionestadisticauba.rec.uba.ar/facultades/ (visitado 31-05-2025).

[35] UBA. *"Ingreso UBA"*. URL: uba.ar/ingresouba (visitado 31-05-2025).

[36] Subsecretaria de políticas universitarias. *"Síntesis de Información Estadísticas Universitarias"*. URL: argentina.gob.ar/sites/default/files/sintesis_anuario_2023-2024.pdf (visitado 31-05-2025).

[37] Subsecretaria de políticas universitarias. *"Síntesis de Información Estadísticas Universitarias"*. URL: argentina.gob.ar/sites/default/files/sintesis_2022-_2023.pdf (visitado 31-05-2025).

[38] Qing-Song Xu y Yi-Zeng Liang. «Monte Carlo cross validation». En: *Chemometrics and Intelligent Laboratory Systems* 56.1 (2001), págs. 1-11.

[39] UBA XXI. *"¿Qué son los cursos intensivos?"* URL: ubaxxi.uba.ar/faq-items/que-son-los-cursos-intensivos-de-verano-e-invierno/ (visitado 31-05-2025).

[40] Claudia Marcela Zelzman. «Acceso y continuidad en carreras de la Facultad de Ciencias Exactas y Naturales-UBA. Un estudio de perfiles de ingresantes tras la implementación de políticas institucionales.» En: *Tesis para maestria en Ciancias Sociales con Orientación en Eduación. FLACSO* (2022). URL: hdl.handle.net/10469/18655.

Apéndice

.1. Anexos

.1.1. Semestre relativo

A continuación se presentan las cohortes y semestres relativos definidos para cada una:

- **Cohortes 2021C1**

  Semestre 0: enero a agosto 2021

  Semestre 1: septiembre 2021 a febrero 2022

  Semestre 2: marzo a agosto 2022

  Semestre 3: septiembre 2022 a febrero 2023

  Semestre 4: marzo a agosto 2023

- **Cohortes 2021C2**

  Semestre 0: septiembre 2021 a febrero 2022

  Semestre 1: marzo a agosto 2022

  Semestre 2: septiembre 2022 a febrero 2023

  Semestre 3: marzo a agosto 2023

  Semestre 4: septiembre 2023 a febrero 2024

- **Cohortes 2022C1**

  Semestre 0: enero a agosto 2022

  Semestre 1: septiembre 2022 a febrero 2023

  Semestre 2: marzo a agosto 2023

  Semestre 3: septiembre 2023 a febrero 2024

  Semestre 4: marzo a agosto 2024

- **Cohortes 2022C2**

  Semestre 0: septiembre 2022 a febrero 2023

  Semestre 1: marzo a agosto 2023

  Semestre 2: septiembre 2023 a febrero 2024

  Semestre 3: marzo a agosto 2024

  Semestre 4: el resto de 2024 y 2025 (septiembre a diciembre 2024 y todos los registros de 2025)