Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Modelos de predicción del abandono en la Universidad Nacional de Hurlingham

Tesis Presentada para optar al Título de Magíster en Explotación de Datos y Descubrimiento de Conocimiento

Esp. Martin Ariel Pustilnik
Director: Dr. Emmanuel Iarussi

Lugar de trabajo: Laboratorio de Investigación y Desarrollo Experimental en Computación (LIDEC), Universidad Nacional de Hurlingham.

Fecha de defensa: 24 de Junio de 2025

Resumen

Modelos de predicción del abandono en la Universidad Nacional de Hurlingham

Se estima que en el sistema universitario sólo el 27,66% de los estudiantes que ingresan se gradúa en un tiempo teórico de 5 años. Para las carreras de informática, este número es incluso más bajo: cercano al 20%. Entendemos que el abandono estudiantil es, tal vez, el factor individual más importante que explica este fenómeno. La medición del abandono en sí presenta diversas complejidades. Con el fin de enfocarse en la emisión de alertas tempranas en lugar de identificar un "abandono definitivo", se considera en situación de abandono a aquel estudiante que tras haber iniciado sus estudios, no muestra actividad académica durante al menos un semestre. Esto puede deberse a una pausa en sus estudios, un cambio de universidad o de carrera, con la posibilidad de retomar su formación en un momento posterior.

La Universidad Nacional de Hurlingham (UNAHUR) es pública y gratuita, funciona desde 2016 con gran integración con la comunidad y alto interés por la permanencia de sus estudiantes, pero con alta tasa de abandono estudiantil. Su matrícula crece aceleradamente y presenta alta incidencia de estudiantes de primera generación de universitarios (77% para 2023) y de bajos ingresos económicos. Entre las acciones para abordar la prevención del abandono, con la participación del autor de esta tesis, la UNAHUR ha estado desarrollando modelos de predicción de abandono utilizando técnicas de Aprendizaje Automático para identificar estudiantes en riesgo, con el objetivo de prevenir el abandono estudiantil de manera temprana.

Estos modelos se basan en los datos del Sistema de Información Universitaria Guaraní (SIU-Guaraní[1]) y en datos generados a partir de la ingeniería de atributos, 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.

En este trabajo 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 de la UNAHUR, para poder intervenir y asistir a las personas antes de que abandonen. Se probaron hipótesis para identificar qué variables influyen en el abandono, y así mejorar futuros modelos. Se proporcionaron recomendaciones sobre variables no relevadas y que deberían ser censadas. Se generó un reporte de personas en riesgo, indicando además de la probabilidad de abandono, los factores más significativos para cada individuo, permitiendo así iniciar la comunicación y explorar los motivos subyacentes sin tener que censar a toda la población estudiantil. Se utilizaron métricas como curva ROC y exactitud balanceada para medir la performance de los modelos, alcanzando un Área bajo la curva ROC de 0,88 para el mejor de ellos.

Palabras clave: Abandono Universitario, Modelo Predicción, Aprendizaje Automático

[1] https://www.siu.edu.ar/

Abstract

Dropout prediction models at the Universidad Nacional de Hurlingham

It is estimated that only 27.66% of students entering the university system graduate within the theoretical time of 5 years. For computer science degrees, this number is even lower: around 20%. We understand that student dropout is perhaps the most important individual factor explaining this phenomenon. Measuring dropout itself presents various complexities. In order to focus on issuing early alerts rather than identifying a "definitive dropout", a student is considered in a dropout situation if, after starting their studies, they do not show academic activity for at least one semester. This may be due to a pause in their studies, a change of university or degree, with the possibility of resuming their education at a later time.

The Universidad Nacional de Hurlingham (UNAHUR) is a public and free institution that has been operating since 2016 with great integration with the community and high interest in student permanence, but with a high dropout rate. Its enrollment is growing rapidly and has a high incidence of first-generation university students (77% by 2023) and low-income students. Among the actions to address dropout prevention, with the participation of the author of this thesis, UNAHUR has been developing dropout prediction models using Machine Learning techniques to identify students at risk, aiming to prevent student dropout early.

These models are based on data from the Guaraní University Information System (SIU-Guaraní) and data generated from feature engineering, with the theoretical perspective of reference authors and other actors from the same university. Once trained, they are able to detect students at high risk of dropping out, while allowing us to investigate some of the underlying reasons.

In this work, a bibliographical investigation of the models used to date was carried out, focusing on those that used Machine Learning. Then, models that provide early warnings of abandonment in the context of UNAHUR were developed to intervene and assist students before they drop out. Hypotheses were tested to identify which variables influence dropout to improve future models. Recommendations were provided on unrevealed variables that should be surveyed. A report of individuals at risk was generated, indicating in addition to the probability of dropping out, the most significant factors for each individual, thus allowing specialists to initiate communication and explore the underlying reasons without having to survey the entire student population. Metrics such as ROC curve and balanced accuracy were used to measure the performance of the models, achieving an Area under the ROC curve of 0.88 for the best model.

Keywords: University Dropout, Prediction Model, Machine Learning

Agradecimientos

En primer lugar quiero agradecer a mi familia, en especial a mis hijas, Matilda y Valentina por su amor y apoyo para terminar el trabajo. En segundo lugar quiero agradecer a Emmanuel Iarussi por haber aceptado dirigir la tesis, de quien aprendí la metodología para hacer investigación. También a mis compañeros de trabajo del LIDEC[2] de la Universidad Nacional de Hurlingham, por su ayuda para resolver algunos problemas que se fueron presentando a lo largo de la tesis. Por último quería agradecer a los profesores y autoridades de la maestría por el conocimiento brindado y por la buena predisposición ante cada inquietud de mi parte.

¡Muchas gracias a todos!

[2] https://unahur.edu.ar/laboratorio-de-investigacion-y-desarrollo-en-computacion-lidec/

Índice

1. Introducción

1.1. Motivación y Problema

Se estima que en el sistema universitario sólo el 27,66% de los estudiantes que ingresan se gradúa en un tiempo teórico de 5 años. La Tabla 1 resume esta información entre los años 2016 y 2021. Se muestran los Nuevos Inscriptos, Egresados y el Porcentaje de graduados en tiempo teórico.

Excepto durante la pandemia, en donde el porcentaje fue significativamente inferior (25,05% y 27,66%), el valor se mantuvo en torno al intervalo (29,41%; 29,78%). En el período 2016-2021 el porcentaje promedio de graduados en tiempo teórico[3] fue del 28,5% y del 31,1% si se considera un periodo de 7 años[4]. El tiempo teórico nos brinda una aproximación porque cada año puede recoger alumnos que ingresaron hace más de 5 años, y esto vale para cualquier año donde se tome la muestra. Según Marino et al. (2023), para las carreras de informática, las cuales son actualmente de interés estratégico para el desarrollo del país, este número es incluso más bajo: cercano al 20%. Según Istvan et al. (2022) en Chile se gradúa alrededor del 57%, pero al tener un sistema universitario arancelado las comparaciones son difíciles. Como se explica en la Sección 2.1 (Sobre la noción del abandono), la medición del abandono en sí presenta diversas complejidades. Por ese motivo utilizaremos una estimación basada en las estadísticas públicas del Departamento de Información Universitaria (2022). Entendemos que el abandono estudiantil es, tal vez, el factor individual más importante que explica estos porcentajes de egreso.

Año	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Promedio
Nuevos	423.920	425.415	445.763	458.565	489.701	**516.305**	547.661	596.446	641.929	710.699	525.664
Egresados	110.360	117.719	120.631	124.960	124.674	125.328	132.744	135.908	122.679	**142.826**	125.750
Porcentaje					29,41%	29,44%	29,78%	29,64%	25,05%	**27,66%**	**28,5%**

Tabla 1: Nuevos (Nuevos Inscriptos), Egresados y Porcentaje (Porcentaje de egresados en tiempo teórico de 5 años) para las universidades Argentinas (2012-2021). Fuente: Elaboración propia a partir de estadísticas publicadas en el Departamento de Información Universitaria (2022).

La Universidad Nacional de Hurlingham (UNAHUR) es pública y gratuita, funciona desde 2016 con gran integración con la comunidad y alto interés por la permanencia de sus estudiantes, pero con alta tasa de abandono estudiantil, su matrícula crece aceleradamente y presenta alta incidencia de estudiantes de primera generación de universitarios y de bajos ingresos económicos. Según el informe elaborado en 2023 por la Secretaría de Evaluación y Planeamiento, en el primer semestre de 2023 se inscribió el 71,5% (23.260 de 32.513) de los estudiantes de pregrado y grado 2022, el 3% (960) egresó, y el 25,5% (8.293) no se inscribió en el primer semestre. El 77% de estos estudiantes son primera generación universitaria en su familia.

Entre las acciones para abordar la prevención del abandono, la UNAHUR ha estado desarrollando modelos de predicción de abandono utilizando técnicas de Machine Learning como los mencionados en el Apéndice 8.1 (Antecedentes personales en el tema), para identificar estudiantes en riesgo, con el objetivo de prevenir el abandono estudiantil de manera temprana. Estos modelos se basan en los datos del Sistema de Información Universitaria Guaraní (SIU-Guaraní[5]) y en datos generados a partir de la ingeniería de atributos, 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.

[3] Se considera como tiempo teórico, el tiempo estipulado en el informe del Departamento de Información Universitaria (2022).
[4] Se considera el porcentaje de alumnos egresados en 2021 respecto de los ingresantes en 2017 (para 5 años) o 2015 (para 7 años). Como pueden egresar alumnos de cualquier año anterior es importante destacar que es un valor teórico.
[5] https://www.siu.edu.ar/

1.2. Objetivos

Los objetivos generales de este trabajo de Tesis de Maestría en Explotación de Datos y Descubrimiento del Conocimiento son:

- Desarrollar modelos que proporcionen alertas tempranas de abandono en el contexto de la UNAHUR, para poder intervenir y asistir a las personas antes de que abandonen.
- Probar hipótesis para identificar qué variables influyen en el abandono, y así mejorar el modelo.
- Proporcionar recomendaciones sobre variables no relevadas y que deberían ser censadas.
- Generar un reporte de personas en riesgo, indicando además de la probabilidad de abandono, los factores más significativos para cada individuo, permitiéndonos así iniciar la comunicación y explorar los motivos subyacentes sin tener que censar a toda la población estudiantil.

Objetivos específicos:

- Compilación de una base de datos a partir de la información contenida en el SIU-Guaraní de UNAHUR (Dataset).
    - El Dataset tiene más de 10.000 campos y 500 tablas que se deben relevar para ser mapeados a variables de los modelos, a los que no se puede acceder de manera directa, tanto por la integridad de la base en sí como por la la privacidad de los alumnos.
    - Para simplificar el proceso, se ejecutó una consulta parametrizada que obtiene información de cuatro semestres consecutivos (Sección 3.4.1., Consulta a la Base). Esa vista se elimina luego de generar las variables y entrenar el modelo.
- Preprocesamiento de los datos para los modelos.
    - Los datos de Guaraní no siempre están completos ni son correctos. Se realizó una limpieza sobre la vista para generar una versión utilizable de las variables.
    - Se derivaron variables generadas a partir de las originales, tales como la cantidad de evaluaciones en un semestre, siguiendo las recomendaciones y directrices establecidas tanto en los primeros modelos conceptuales como en los enfoques más contemporáneos y avanzados que aplican Aprendizaje Automático.
- Entrenar modelos de Aprendizaje Automático para la predicción del abandono en el contexto de la UNAHUR.
    - Se desarrollaron y entrenaron una serie de modelos utilizando los datos disponibles para predecir el riesgo de abandono. Se buscó iterar y mejorar progresivamente estos modelos, ajustando y optimizando sus parámetros y características, con el fin de incrementar su precisión y confiabilidad en la predicción del riesgo de abandono.
- Obtener una lista de variables importantes.
    - Se identificó 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. Esto implica superar desafíos asociados a la recolección de algunas variables aún no incluidas pero esenciales, y planificar su integración en futuras consultas y análisis.
- 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.

- Mediante un algoritmo, se cruzó la información de las variables más importantes con los datos de las personas en riesgo y sus detalles personales, generando así un reporte que resalta la variable más significativa que afecta a cada individuo en riesgo. Este reporte servirá como punto de partida para que el personal de la UNAHUR tome la iniciativa de contactar a los estudiantes y brindarles el apoyo necesario, adaptando las intervenciones a las necesidades específicas de cada uno, ya sea que el riesgo esté asociado al tiempo de viaje a la universidad, a la baja cantidad de exámenes rendidos en el último semestre, u a otros factores críticos.

1.3. Organización del trabajo

El presente documento, que se centra en la predicción del abandono en la Universidad Nacional de Hurlingham, ha sido estructurado para facilitar su comprensión y seguimiento. En la Sección 2 se hace una inmersión profunda en el estado del arte, donde no sólo se presenta una revisión actualizada de investigaciones y desarrollos relevantes en la materia, sino también un acercamiento histórico que contextualiza el problema del abandono educativo en un marco temporal. Con esa base se proponen una serie de factores que son plausibles de ser analizados en la UNAHUR. La Sección 3 está dedicada a la revisión y exploración de la base de datos del SIU-Guaraní, así como de otros conjuntos de datos que han sido fundamentales para el desarrollo de esta investigación. En este apartado, se profundizará en la naturaleza, características y relevancia de estos datos para el estudio del abandono. Continuando con la Sección 4, se describen en detalle los métodos y técnicas que han sido empleados en las fases de exploración y predicción del abandono estudiantil, ofreciendo un panorama claro de la metodología implementada. La Sección 5 se centra en presentar los modelos de predicción desarrollados y los experimentos realizados. Esta sección servirá para entender la base teórica y práctica de las herramientas predictivas utilizadas. En la Sección 6 se presentan los resultados obtenidos de los distintos experimentos, junto con una explicación detallada de las métricas empleadas para evaluar y comparar dichos resultados. Finalmente, la Sección 7 ofrece una discusión sobre los hallazgos, conclusiones y una reflexión sobre posibles caminos a seguir a partir del trabajo presentado en esta Tesis.

2. Estado del arte

2.1. Sobre la noción de abandono

En la bibliografía, el abandono se define como el cese de la actividad académica del alumno, antes de finalizar sus estudios. Típicamente se considera abandono a cuatro o más semestres sin actividad como se menciona en Losio y Macri (2015). Tanto Pascarella y Terenzini (1980) como Tinto (1982) suelen utilizar los términos abandono/deserción (drop-out en inglés) indistintamente. En español deserción tiene una connotación marcial y da a entender que el fenómeno es responsabilidad principalmente del estudiante. Como mostraremos más adelante, el abandono es un fenómeno multivariado, cuyas variables no son exclusivas del estudiante. Por esos motivos utilizaremos el término "Abandono" en lugar de "Deserción". La medición del abandono presenta diversas complejidades. Parrino (2012) propone múltiples métodos para estimar este fenómeno:

-Interanual *(event rates*): la proporción de estudiantes de un grupo de edad que deja de estudiar cada año sin completar el nivel correspondiente.
-Tasa de abandono por edad (TAE): agrupa datos del abandono para un rango determinado de edad. En la Ecuación 1 se muestra la TAE para alumnos de 20 a 30 años. La cantidad de estudiantes que abandonan surgen de encuestas.

$$TAE(20,30) = abandonan(20,30) \,/\, (abandonan(20,30) + graduados(20,30)) \qquad (1)$$

-Abando por Cohorte (AC): aquí se mide el abandono para una cohorte en particular. En Parrino (2016) se propone calcular el abandono para una carrera *C* y un año *a*, a partir de los alumnos de la misma cohorte, restando los que egresan en duración teórica (dt) o hasta dos años después y finalmente, restando los alumnos de esa cohorte que continúan cursando en el año *a+dt+2*. Por ejemplo, si tomamos un tiempo teórico de 5 años (*dt = 5*) para la Tecnicatura en Sistemas (C = TS), el AC se puede calcular con la Ecuación 2.1, a partir de instanciar la Ecuación 2.1. En la Ecuación 2.3 se calcula el valor para esta instancia, en este caso, un 40% de los alumnos que comenzaron la TS en el año 2000 no se inscribió ni terminó la carrera en el 2007. Este porcentaje de abandono constituye un valor teórico ya que los alumnos pueden volver a inscribirse como explicaremos más adelante. En la ecuación 2.3 se asume que el 40% de los alumnos que comenzaron en el año 2000 abandonaron en 2007.

$$AC(c_a) = inscriptos(c_a) - alumnos(c_{a+dt+2}) - \sum_{i=0}^{2} egresados(c_{a+dt+i}) \qquad (2.1)$$

$$AC(TS_{2000}) = inscriptos(TS_{2000}) - alumnos(c_{2007}) - \sum_{i=0}^{2} egresados(TS_{2005+i}) \qquad (2.2)$$

$$AC(TS_{2000}) = 100 - 10 - 50 = 40 \qquad (2.3)$$

Landi y Giuliodori (2001), y García de Fanelli (2013) proponen algunas de estas métricas, pero existen muchas otras en la bibliografía. Con el fin de alinearse con el objetivo central de la tesis, la perspectiva sobre el abandono se ajustó para enfocarse en la emisión de alertas tempranas en lugar de identificar un "abandono definitivo". En este contexto, se considera en situación de abandono a aquel estudiante que, tras haber iniciado sus estudios, no muestra actividad académica durante al menos un semestre, como se indica en la Ecuación 3. Esto puede deberse a una pausa en sus estudios, un cambio de

universidad o de carrera, con la posibilidad de retomar su formación en un momento posterior. Esta conceptualización se sustenta en la definición de "primera deserción" proporcionada por Tinto en 1982, ajustándose así a los propósitos de esta investigación y permitiendo una intervención más oportuna y efectiva. En la Ecuación 3 se considera la cantidad de evaluaciones de una persona durante el último semestre para encender la alarma de posible abandono.

$$Abandono(persona) = \begin{cases} 0 & \text{si } cantEval(persona) > 0 \\ 1 & \text{si } cantEval(persona) = 0 \end{cases} \quad (3)$$

2.2. Modelos conceptuales.

El abandono se puede explicar como el resultado de distintas variables que afectan al estudiante. Braxton et al. (1997) propone agrupar los modelos explicativos en cinco categorías reconociendo diferentes enfoques: psicológico, sociológico, económico, organizacional y de interacciones. Posteriormente se ha incorporado a los modelos explicativos un modelo integrado que abarca más de uno de los enfoques mencionados anteriormente. Díaz Peralta (2008) resume las teorías que fundamentan los primeros modelos conceptuales.

2.2.1 Modelos Psicológicos

Desde el enfoque psicológico, Fishbein y Ajzen (1975) señalan que el abandono depende rasgos de la personalidad del estudiante. Explican el abandono mediante la Teoría de la Acción Razonada. Los autores plantearon que las actitudes están moldeadas por las características que los individuos asocian con un determinado objeto, basándose en sus creencias acerca de dicho objeto. En el modelo, se propone una fórmula que especifica cómo se integran las creencias importantes de los individuos para formar una actitud general. La Ecuación 4.1 ejemplifica la aplicación práctica de esta teoría (para un *individuo*$_i$ y una creencia *c*), presentando un método para calcular la actitud de un individuo hacia un objeto o meta específica.

$$ActitudObjeto(individuo_i) = \sum_{i=1}^{cant.creencias\ individuo\ i} probabilidad(c_{ij})\ valoración(c_{ij}) \quad (4.1)$$

Donde ActitudObjeto representa la actitud del *individuo*$_i$ hacia el objeto en cuestión, y se calcula sumando el producto de la *probabilidad*(c_{ij}) (cuán probable es que ocurra una creencia particular) y la *valoración*(c_{ij}) (la importancia asignada a una creencia) para cada creencia del individuo. Las valoraciones pueden variar entre -3 y +3, indicando una percepción negativa o positiva respectivamente. Para el caso del abandono podría pensarse el siguiente ejemplo: Si el individuo$_1$ cree que *"Estudiar 5hs por semana"*, *"Asistir a Clase"* y el *"Tiempo de viaje hasta la facultad"* son importantes para *"Finalizar los estudios"*, se le pregunta cuál es la probabilidad de que estos eventos ocurran y cuan importante son. De esta manera, la Ecuación 4.1 se instancia con los siguientes términos para el individuo$_1$ en la Ecuación 4.2. Donde c_{11}: Estudiar 5hs por semana, c_{12}: Asistir a clase, c_{13}: Tiempo de viaje (todos, para el individuo$_1$).

$$Finalizar\ los\ estudios(individuo_1) = p(c_{11})\ v(c_{11}) + p(c_{12})\ v(c_{12}) + p(c_{13})\ v(c_{13}) \quad (4.2)$$

Con este modelo se puede predecir si un individuo finalizará sus estudios viendo si se suman valores positivos y altos. Si para el individuo$_1$ es 20% probable estudiar 5hs, 50% probable asistir a clases y 100% probable tardar mucho en llegar a la facultad, y asigna una importancia de 2, 2 y -3 respectivamente a estas

creencias (considerando el tiempo de viaje como un factor negativo), el modelo computará -1,6 para estos valores. Esto se traduce en una mayor probabilidad de abandono que de finalización de los estudios, como se evidencia en la Ecuación 4.3 para el $individuo_1$:

$$Finalizar\ los\ estudios(individuo_1) = 0,2 * 2 + 0,5 * 2 - 1 * 3 = -1,6 \qquad (4.3)$$

Fishbein y Ajzen sostienen que la deserción es consecuencia de la disminución de las aspiraciones y motivaciones individuales iniciales, tema que es abordado y analizado en dos trabajos incluidos en la Tabla 1 anexo (Goldenhersh et al. 2011; Odetti et al. 2010). En 1990, Ethington testea el modelo de "Conductas de logros" de Eccles. Este modelo psicológico propone que la persistencia está correlacionada positivamente con las elecciones y comportamientos previos a la universidad. Ethington utilizó un software estadístico para calcular la influencia de las variables entre sí, y en particular en el abandono (*Persistence*) a partir de los datos de una encuesta realizada a 8.790 alumnos. La Tabla 2.1 muestra los valores (entre paréntesis) y errores calculados por el software estadístico utilizado. Las influencias en cero (0) no se grafican, indicando, por ejemplo, que las variables {1, 2 ... 12} no tienen influencia sobre las notas del secundario (Prior Achievement). La Tabla 2.2 resume la influencia directa (última columna de la Tabla 2.1) e indirecta (influencia agregada sobre otras variables), sumando así la "influencia total" sobre la persistencia en la última columna. Se puede apreciar que las notas del secundario (Prior Achievement), las variables socioeconómicas (Socioeconomic Status, SES), mayor título al que aspira (Degree Aspirations[6]) y la importancia asignada a la educación (value[7]) concluyendo que las "variables previas" son importantes, y a su vez dan otra mirada a la "Conducta del logro".

	Dependent Variables										
	3	4	5	6	7	8	9	10	11	12	13
1. SES	.066	.135*	.030	.153*	−.030	−.085	.003	−.048	−.098*	−.143*	.091**
	(.018)	(.065)	(.008)	(.076)	(−.026)	(−.079)	(.003)	(−.026)	(−.041)	(−.167)	(.028)
2. Prior Achievement	−.054	.480*	−.203*	.023	−.144*	−.245*	.045	−.007	.179*	.042	.145*
	(−.020)	(.321)	(−.073)	(.016)	(−.175)	(−.312)	(.050)	(−.006)	(.105)	(.068)	(.061)
3. Family Encouragement		.019	.049	−.046	.153*	.180*	.095	.123*	−.004	.123*	.026
		(.033)	(.047)	(−.086)	(.501)	(.616)	(.283)	(.251)	(−.006)	(.535)	(.029)
4. Academic Self-Concept				.235*	.219*	.086	−.042	.127**	.280*	−.034	.091
				(.241)	(.398)	(.164)	(−.070)	(.144)	(.244)	(−.082)	(.058)
5. Perception of Difficulty				−.018	−.013	−.037	.018	−.087	−.281*	−.028	.020
				(−.035)	(−.046)	(−.131)	(.057)	(−.183)	(−.457)	(−.127)	(.024)
6. Degree Aspirations									.180*	.135*	.162*
									(.152)	(.319)	(.100)
7. Political Goals									.045	.074	.017
									(.022)	(.098)	(.006)
8. Business Goals									−.025	.205*	−.010
									(−.011)	(.260)	(−.003)
9. Humanitarian Goals									.012	.118*	−.042
									(.006)	(.173)	(−.016)
10. Desire for Recognition									.071	.117**	.018
									(.055)	(.249)	(.010)
11. Expectations for Success											−.004
											(−.003)
12. Value											.155*
											(.041)
13. Persistence											
R^2	.006	.272	.043	.106	.062	.091	.011	.041	.395	.205	.138

Tabla 2.1: Error y coeficientes (entre paréntesis) del modelo calculados por el software estadístico. Efecto de las variables sobre la variable Persistence (No Abandono). No figuran los coeficientes en cero (0). *p < 0,01 **p < 0,05 . Fuente: Adaptado de Ethington (1990).

[6] Ethington discretiza 'Degree Aspirations' como: 1 = None, 2 = associate, 3 = bachelor, 4 = master, 5 = doctor. Indicando que a mayor la variable, mayor el grado al que aspira.

[7] Cuán valioso es recibirse. Si es importante para la sociedad. Si puedo obtener un mejor trabajo, entre otros.

	Direct	Indirect	Total
SES	.091**	.024	.115*
	(.028)	(.007)	(.035)
Prior Achievement	.145*	.062**	.207*
	(.061)	(.027)	(.088)
Family Encouragement	.026	.024	.050
	(.029)	(.028)	(.057)
Academic Self-Concept	.091	.050**	.141*
	(.058)	(.032)	(.090)
Perception of Difficulty	.020	−.011	.009
	(.024)	(−.014)	(.010)
Degree Aspirations	.162*	.020	.182*
	(.100)	(.012)	(.112)
Political Goals	.017	.011	.028
	(.006)	(.004)	(.010)
Business Goals	−.010	.032**	.022
	(−.003)	(.011)	(.008)
Humanitarian Goals	−.042	.018**	−.024
	(−.016)	(.007)	(−.009)
Desire for Recognition	.018	.018	.036
	(.010)	(.010)	(.020)
Expectations for Success	−.004		−.004
	(−.003)		(−.003)
Value	.155*		.155*
	(.041)		(.041)

Tabla 2.2: Error y coeficientes (entre paréntesis) del modelo calculados por el sistema. Efecto directo e indirecto en la Persistencia (No Abandono). *p < 0,01 **p < 0,05 . Fuente Ethington (1990).

Bean y Eaton en 2001 propusieron un modelo psicológico que explica el grado de persistencia. Creen que las variables como “Sexo”, “Religión” y “Grupo étnico” por si mismos no constituyen variables importantes. Pero las variables psicológicas como la Autopercepción/Self-assessments (i.e. ¿Tengo confianza en mi desempeño académico?), las creencias Normativas/Normative beliefs (i.e. ¿Mis personas cercanas me apoyan en la decisión de cursar la universidad?) y el Comportamiento previo/Past behavior (similar a lo planteado por Ethington), sí lo son. En la Figura 1 se muestra que cada estudiante ingresa con ciertas características psicológicas que influenciarán en su permanencia en la carrera. Durante su paso por la institución, el estudiante va interactuando con diferentes actores (desde burocráticos a docentes y otros alumnos). Esto va mejorando sus atributos psicológicos iniciales para, en un escenario ideal, permitir integrarse y alcanzar sus metas personales. En particular, se hace hincapié sobre cuánto control cree el alumno que tiene sobre lo que le sucede en la universidad y en su vida en general (*Locus of Control*). Un individuo con un “*Locus of Control*” interno cree que es responsable en sus propios éxitos o fracasos, mientras que una persona con “Locus of Control” externo cree que se deben al destino o a la oportunidad. Según los autores, la integración tiene lugar también por el mecanismo de adaptación a situaciones nuevas (*Coping Process*) y por cómo puede utilizar sus habilidades previas y experiencia para enfrentarse a experiencias nuevas (*Self-Efficacy Assessments*). Los autores concluyen que hay una serie de factores que influyen positivamente en la persistencia. A continuación se nombran algunos de ellos:

-Programas de acompañamiento/tutoría y seminarios de orientación para nuevos alumnos.

-Comunidades de estudio.

En UNAHUR hay varios ejemplos de estas prácticas, como el programa “1 estudiante - 1 compañero” o la utilización de la red social Discord® como grupo de estudio.

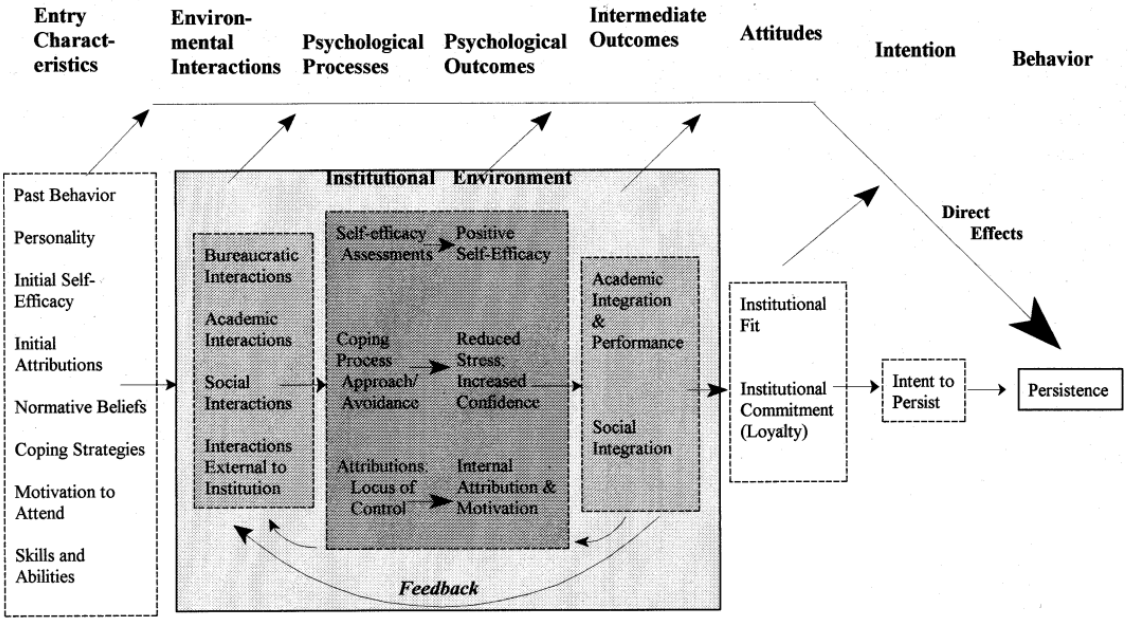


Figura 1: Modelo Psicológico de Bean y Eaton. Fuente: Bean y Eaton (2001).

2.2.2. Modelos sociológicos

Los modelos sociológicos destacan la importancia de factores externos al individuo en la incidencia de la deserción académica, complementando así las perspectivas psicológicas. Spady en 1970 realizó un estudio de las teorías del abandono. Estos estudios son principalmente censos a los que Spady denomina estudios "post-mortem", haciendo referencia a que tratan de explicar el abandono una vez que sucede. Un ejemplo de estos estudios es el trabajo de Iffert (1958), que realizó un censo nacional sobre una muestra de 12.667 alumnos de Estados Unidos. Como resultado, se determinó que sólo el 40% se recibió en el tiempo teórico de 4 años[8] y se estima que sólo se recibió el 60%. Otros estudios intentan establecer una correlación entre las variables y el abandono. Spady desaconseja este enfoque porque es difícil de determinar la causalidad basándose únicamente en la correlación. Los algoritmos de Aprendizaje Automático utilizados, incluidos los de esta tesis, hacen uso de esas "correlaciones". Como se mostrará más adelante, estas hipótesis están basadas en variables recomendadas en la literatura o en actores expertos del dominio para justificar la causalidad.

2.2.2.1 Modelo sociológico de Spady

Spady (1970) comienza diciendo: "El proceso de abandono se explica mejor mediante un enfoque interdisciplinario que explique la interacción entre los estudiantes y su entorno universitario. Sus atributos (disposiciones, intereses, actitudes y habilidades) están expuestos a influencias, expectativas y demandas de una variedad de fuentes (incluidos cursos, profesores, administradores, y compañeros)". El modelo representado en la Figura 2, representa la interacción del alumno en el sistema académico y social de la universidad. Las líneas representan causalidad (en dirección de la fecha) y la línea punteada en particular representa una retroalimentación, en el sentido que las variables no son estáticas y se van modificando durante el tiempo. En el "Sistema académico", las "Notas" (Grade Performance) y el "Desarrollo intelectual" (Intellectual Development), son las variables más influyentes. Para el "sistema social", la "Congruencia normativa" (Normative Congruence) representada por las características, la orientación de la institución y las "Relaciones entre pares" (Friendship Support), representan las variables más influyentes. Un ejemplo de

[8] El tiempo de teórico del College es de 4 años.

“Incongruencia normativa” es cuando un alumno quiere estudiar una orientación técnica pero se encuentra inscripto en un instituto con orientación humanística. La “Red familiar” (Family Background), es relevante como variable inicial del modelo, asignando importancia a los “Estudios de los padres”, su estado socioeconómico, y sus “notas previas”, entre otros factores.

Cuando alguno de esos dos sistemas falla, Spady afirma que falla la integración del alumno, favoreciendo la decisión de abandonar. Tomando como base el modelo de Durkheim (1951) sobre el suicidio, en donde el suicidio se desarrolla como consecuencia de una falta de “integración social”. Spady establece un paralelismo con el abandono académico, denominándolo “suicidio académico”.

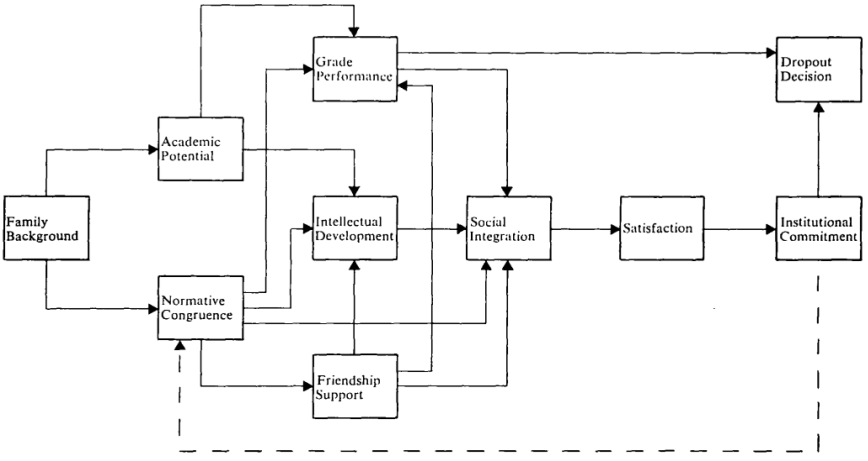


Figura 2: Modelo Sociológico explicativo de Spady, del proceso de abandono. Fuente: Spady (1970). Las líneas representan causalidad. La línea punteada representa la “retroalimentación” en el tiempo, generando nuevas interacciones.

2.2.3. Modelo de interacción de Tinto.

En muchos estudios se intenta explicar el abandono como una variable binaria en donde el alumno finaliza o abandona los estudios. Tinto (1975, 1989, 1993) diferencia el abandono voluntario (i.e. cambios de carrera o no estudiar en ese momento) del generado por la normativa institucional o el bajo desempeño académico de los estudiantes, entre otros motivos. Para ello, desarrolla un modelo longitudinal y de interacción (i.e. la probabilidad de abandono, va cambiando en el tiempo y según la interacción del alumno con la institución, docentes y sus pares). En la Figura 3 se muestran estas interacciones mediante líneas punteadas. Además, los compromisos (Commitments), cómo “ir a una institución más prestigiosa” o “dedicar más recursos financieros para la universidad” están enmarcados en rectángulos. Todos los modelos asumen que a mayor compromiso, menor es el grado de abandono.

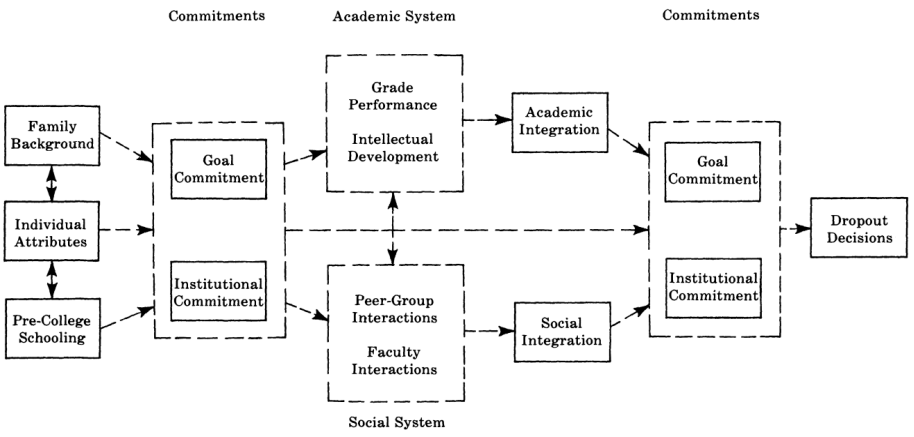


Figura 3: Esquema conceptual del modelo interactivo de Tinto. Fuente: Tinto (1975).

Al igual que en los modelos económicos (Sección 2.3.2) y sociológicos (Sección 2.2.2), Tinto también considera tanto las "variables previas", como la "trayectoria en el secundario" y los estudios de los padres como la teoría del intercambio de Nye (1976), en donde el abandono se explica en parte con "costo" de estudiar respecto del "beneficio" de finalizar los estudios.

Aunque hay un consenso sobre ciertas variables "importantes", no hay unanimidad sobre la magnitud del impacto de cada una sobre el abandono. Un ejemplo de esto, es el trabajo de Pascarella y Chapman (1983a, 1983b), en donde se llega a la conclusión de que la integración académica (i.e. cantidad de compañeros de estudio, cantidad de horas. en campus, participación en actividades de extensión, etc) es más fuerte que las metas institucionales (i.e. recibirse, tener cierto promedio, etc), pero esto no ocurre en todos los trabajos. Inclusive, las variables no son uniformemente importantes para todos los alumnos. Pascarella y Chapman afirman que la integración social no les resulta un factor importante al grupo de alumnos que abandona.

2.2.4. Modelos conceptuales integrados.

En la Figura 4 se integran las variables con los modelos conceptuales estudiados en Díaz Peralta (2008) mediante un esquema topológico como el que se explica en Collen y Gasparski (1995). El esquema muestra la asociación de las principales variables y autores de la literatura agrupadas en factores individuales, académicos, socioeconómicos e institucionales, como se sugiere en García (2014).

Hay dos tipos de relaciones. Una de ellas es la relación entre autores ( ), que se da cuando un autor toma como base o extiende un modelo previo. Por ejemplo vemos que Pascarella y Terenzini (1980) extienden el modelo clásico de Tinto (1975). La otra relación es entre variables ( ). Por ejemplo, vemos que la "Carga académica" está relacionada con el "Grado de satisfacción de la carrera". El esquema describe la categoría de cada variable (e.g. "Situación laboral" → categoría "Socioeconómica"); y sirve como punto de partida para explorar otras "variables derivadas". Un ejemplo de ello es como sugiere Parrino (2012) con "Baja selectividad". Esta variable se deriva de la variable "Normativa académica".

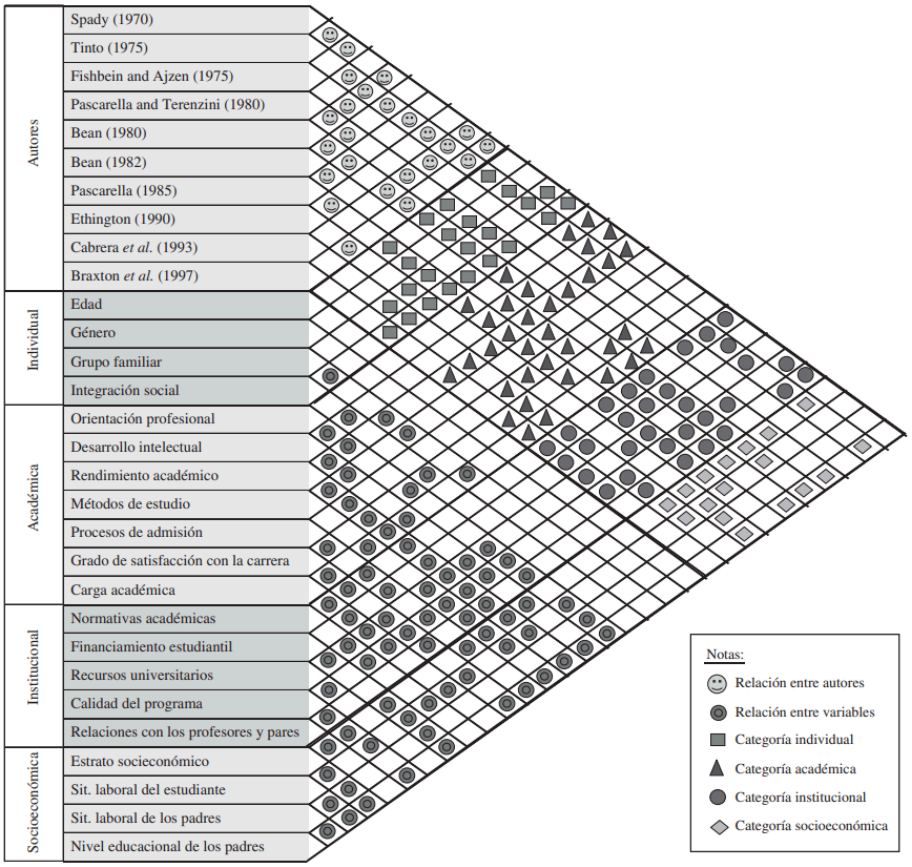


Figura 4: Esquema topológico de primeros modelos conceptuales. Relaciones entre variables y autores clásicos. Fuente: Díaz Peralta (2008).

2.2.5. Modelos conceptuales en Argentina

En la revisión de publicaciones de Argentina acerca del abandono realizada en García (2014), se aprecia el predominio de investigaciones sobre dos enfoques teóricos: económico y sociológico. Los economistas suelen utilizar estimaciones con métodos econométricos de vinculación entre variables. Los sociólogos, se basan en modelos como los presentados por Bourdieu y Passeron en Francia. Estos modelos consideran el papel tanto del capital cultural como del capital social del estudiante y su familia en el análisis de los factores que inciden sobre el éxito en la universidad. El Modelo Reproductivista de Bourdieu asume que en la educación se reproduce el modelo social, haciendo más difícil la adaptación a quienes ingresan con menor capital cultural (Bourdieu y Passeron, 1998). Se entiende por capital cultural, el conjunto de habilidades intelectuales, tanto producidas por el sistema escolar como por la familia. Este enfoque presta atención a los condicionantes culturales y socioeconómicos de los estudiantes y al choque cultural percibido por los estudiantes con la cultura institucional de la universidad. Este choque adquiere mayor relevancia para los estudiantes que son la primera generación en acceder a estudios de educación superior de su familia y su entorno.

2.3. Modelos predictivos.

2.3.1. Modelos estadísticos para predicción del abandono

Durante la década del ‘80 se desarrollaron los primeros modelos estadísticos para predecir el abandono, que utilizan las variables propuestas en modelos conceptuales (Sección 2.2.), como los propuestos por Bean y Metzner (1985). Estos modelos son adaptaciones del “Modelo de Ecuaciones Estructurales”

utilizados en ciencias médicas y sociales para cuantificar la importancia de un conjunto de variables como causalidad de otra, como se menciona en Ortiz y Fernández-Pera (2018). Para la predicción del abandono se utilizan variables recomendadas en la literatura, censadas a partir de encuestas, incluida la intención de abandono. Luego se procesan para obtener diferentes relaciones entre variables, por ejemplo, las correlaciones que tienen las variables entre sí. Basado en esas relaciones se entrena el modelo asignando un peso relativo para cada variable. Luego, se construye un modelo de regresión lineal con el que se intenta predecir el abandono para una cohorte basado en el entrenamiento del modelo con cohortes anteriores.

En Castillo Diaz (2022) se implementó un modelo estructural integrado a partir de una muestra de 653 estudiantes de primer ingreso de la Universidad Nacional Autónoma de Honduras (UNAH). En primer lugar, se realizó una encuesta mediante el Cuestionario de Vivencias Académicas, versión reducida (QVA-r). Elaborado por Almeida et al. (1999) y adaptado al castellano por Márquez et al. (2009). El QVA-r evalúa las vivencias académicas relacionadas con la adaptación universitaria estudiantil. La escala está compuesta por 60 ítems que se corresponden en escala Likert (Likert, 1932) e incluye 5 dimensiones: personal, interpersonal, carrera, hábitos de estudio e institucional. La colecta de datos se desarrolló en dos momentos: (1) al inicio del primer período académico del año 2022, en el cual los estudiantes completaron el cuestionario ad-hoc construido para indagar variables socioeconómicas y académicas generales y (2) al término del primer período académico, momento en que los estudiantes completaron el QVA-r y la subescala de intención de abandono. Luego se realizó un análisis bivariado de cada variable respecto de la intención de abandono para quedarse con variables estadísticamente significativas (p-valor < 0,05) que correlacionen con "Abandono". Considerando la distribución no normal de la variable intención de abandono se utilizó el test de Mann-Whitney para el contraste de variables dicotómicas (sólo dos valores), el test de Kruskal-Wallis para variables politómicas (tres o más valores) y el coeficiente de correlación de Spearman para las variables cuantitativas (Devore, 2016). En la segunda fase se realizó un análisis multivariado por medio de la evaluación de un modelo por ecuaciones estructurales. Este modelo incluyó únicamente las variables estadísticamente significativas de la primera fase. El ajuste del modelo se testeó con el índice de ajuste comparativo (CFI[9]) y el error de aproximación cuadrático medio (RMSEA), dando CFI > 0,90 y RMSEA < 0,08 respectivamente. Estos valores son adecuados para el test en términos de Schumacker y Lomax (2018). Los análisis fueron realizados por medio de los Softwares JASP y los paquetes semTools y lavaan de R[10].

Los resultados indicaron asociaciones estadísticamente significativas (p-valor < 0,01) entre todas las variables, evidenciando correlaciones negativas que oscilan entre -0,42 y -0,61 entre las vivencias académicas y la intención de abandono, es decir, a mayores puntajes de las cinco dimensiones de vivencias académicas (factores personales, interpersonales, carrera, hábitos de estudio e institucionales), menor serán los puntajes de la intención de abandono. Al analizar el modelo estructural predictivo integrado, los hallazgos indican un buen ajuste del modelo testado, con un CFI = 0,95, indicando que las variables no son independientes y el RMSEA = 0,07 indica que se explica la mayor parte de la varianza. La Figura 5 indica los betas (β_i) estandarizados respecto a las predicciones del modelo. Los hallazgos evidencian predicciones estadísticamente significativas ($p < 0,05$) sobre la intención de abandono por parte de las vivencias tempranas en la dimensión personal ($\beta_1 = - 0,21$), interpersonal ($\beta_2 = - 0,13$), carrera ($\beta_3 = - 0,51$), institucional ($\beta_4 = - 0,28$) y la opción de carrera en admisión ($\beta_5 = 0,17$). No obstante, en esta segunda fase de análisis, la dimensión de hábitos de estudio, el área de conocimiento y el desempleo en núcleo familiar pierden significancia predictiva ($p > 0,05$). Como se detalla en Ortiz y Fernández-Pera (2018), los óvalos son variables no censadas directamente, agrupadas/calculadas a partir de otras. Las flechas bidireccionales son la correlación entre dos variables mientras que las unidireccionales son los betas que entrenó el modelo.

[9] Indica un buen ajuste del modelo para valores próximos a 1.
[10] https://www.r-project.org/

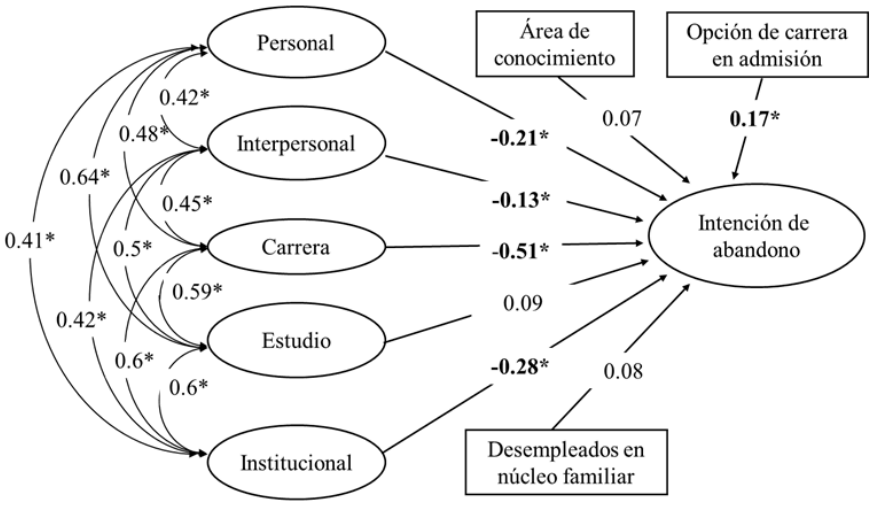


Figura 5: Modelo estructural integrado elaborado en Castillo Diaz. (2022) a partir de una encuesta, entrenando los betas para el modelo. Fuente: Castillo Diaz (2022).

Los modelos estadísticos se basan en hipótesis que "ya conocemos", ya sea porque están en la literatura o porque son relevantes en una universidad (i.e. provienen de encuestas). Como veremos en la siguiente sección, los modelos basados en Aprendizaje Automático ofrecen la importancia relativa de cada variable, lo que permite "testear" las hipótesis iniciales a posteriori, además de las ventajas ya mencionadas de los modelos estadísticos.

2.3.2. Modelos económicos

Entre los modelos que incluyen variables económicas, los modelos de Cabrera et al. (1992 y 1993) son unos de los más citados en la bibliografía. Cabrera extiende los modelos clásicos de Spady y Tinto para medir la influencia de variables económicas respecto del resto de las variables y en particular en la persistencia (no abandono). Agregó variables como la actitud financiera (Financial Attitudes), que mide el comportamiento y capacidad financiera del estudiante y su familia. Y la ayuda financiera (Financial Aid), que mide la ayuda financiera que otorga la institución, ya sea en becas o en el financiamiento de los estudios por préstamos u otros procedimientos. En la Figura 6 se ve el resultado del modelo económico basado en ecuaciones estructurales como los explicados en la sección previa. En cada línea se representa la correlación de cada variable sobre otra. A continuación se detallan algunos ejemplos del resultado del modelo:

Para la integración académica (Academic and Intellectual Development), la actitud financiera (alfa = 0,245), las notas del secundario (alfa = 0,164) y "otras variables significativas" (alfa = 0,147) tuvieron una correlación significativa. Sin embargo, la ayuda financiera no. Las variables "notas del secundario" (alfa = 0,321) y "otras variables significativas" (alfa = 0,058) tuvieron influencia en la "performance académica" (GPA: Grade Performance average). Como resultado general, Cabrera encontró que la ayuda financiera tiene un resultado significativo (positivo) en la persistencia.

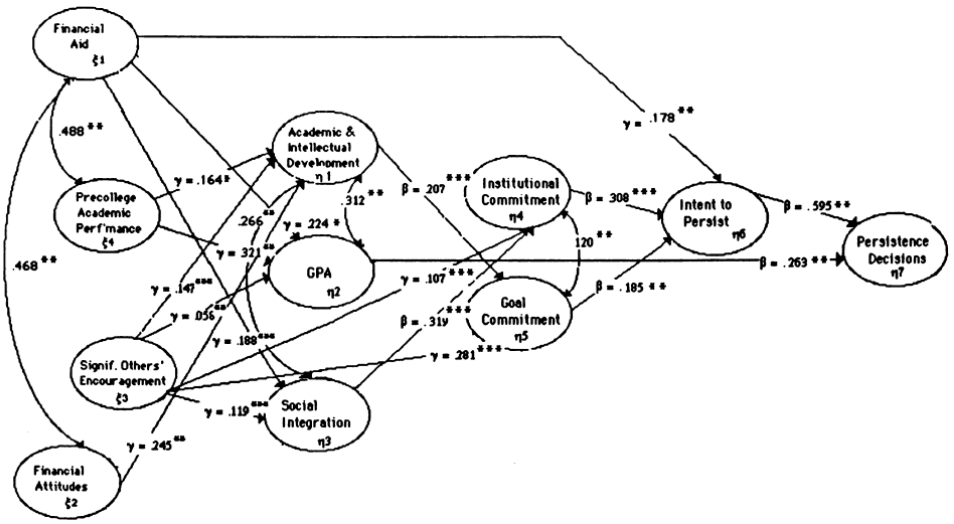


Figura 6: Modelo económico de Cabrera, basado en ecuaciones estructurales. *p-valor < 0,07; ** p-valor < 0,05; *** p-valor < 0,01. Fuente: Cabrera et al. (1992 y 1993).

Bernalet (2000) y John et al. (2000) clasifican estos modelos como de "costo-beneficio", ya que ponderan el "costo" de tener que estudiar durante un período de tiempo, trabajando menos horas o sin trabajar con el "beneficio" del título obtenido. Hay especialistas que consideran que focalizar las ayudas financieras y programas de acción de las instituciones educativas sobre grupos de estudiantes con dificultades puede influir en su permanencia. Cameron y Taber (2001) planearon modelos que miden los efectos sobre estas políticas.

2.3.3. Modelos organizacionales.

Autores como Berger y Milem (2000), Berger (2002) y Kuh (2002) sostienen que las variables organizacionales tienen mucha influencia en la permanencia. Berger y Braxton (1998) implementaron un modelo organizacional de ecuaciones estructurales, basado en datos recogidos en encuestas dirigidas a relevar las variables organizacionales recomendadas en la literatura, agregadas a las variables clásicas conocidas. La Tabla 3 es el diccionario de las variables utilizadas en el modelo.

Variable/ Tipo	En Inglés	Explicación
Background Characteristics		
SEXF; RACEW	Sex; Race	Género; Raza/ Etnia. En particular si es blanco.
INCOME	Income	Ingreso estimado de los padres.
HSGPA	High school grade point average	Notas promedio del secundario.
POLVIEW	Political View	Orientación política: Izquierda/Derecha.
INSTCOM1	Initial Institutional Commitment	Si fue su 1ra, 2da o 3ra elección.
Institucionales		
COMMUN	Institutional Communication	Cuán bien se comunican las reglas: académicas, sociales, de materias y de graduación.*
FAIR	Fairness in policy and rule enforcement	Cuán justas son las reglas: académicas, sociales, de graduación y becas. *
PARTIC	Participating in decision making	Cuánto se puede participar en las decisiones para el armado y aplicación de las reglas antes definidas.*
Integración social		
SIPEER	Peer Relations	Cuán de acuerdo está con las afirmaciones: las relaciones personales aumentan el crecimiento intelectual; he desarrollado buenas relaciones interpersonales; esas relaciones me dan un crecimiento personal positivo; tengo dificultad para hacer amigos y solo algunos me ayudan.
SIFAC	Faculty Relations	Cuánto influye la interacción con la facultad en mi crecimiento intelectual/personal.
Subsequent Institutional Commitment		
INSTCOM2	Subsequent Institutional Commitment	Cuán de acuerdo está con las afirmaciones: No es importante graduarme en esta universidad; ésta es la universidad adecuada.
Departure Decision		
IRENROLL	Intent to Re-enrol	Intención de inscribirse en el siguiente año.

Tabla 3: Diccionario de variables del modelo Berger y Braxton (1998). *Ítems derivados de Bean (1980). Fuente: Adaptado de Berger y Braxton (1998).

El modelo de Berger y Braxton tiene como hipótesis que el compromiso institucional inicial (IC1) y las variables previas (Background Characteristics) tienen tanto influencia al principio como al final de la carrera. Esto se manifiesta en la Tabla 4 con una fuerte correlación, y se muestra en la Figura 7 con una línea punteada. El estudio encontró que los estudiantes blancos eran los que mayor percepción tenían de que participaban de las decisiones (correlación (PARTIC,SEXF) = 0,49) y mayor predisposición a relacionarse con otros alumnos (0,11). Por otro lado, IC1 no tuvo ningún efecto estadísticamente significativo en las variables organizacionales. Estas variables sí tuvieron un efecto significativo en la integración social, por ejemplo la variables: “relaciones entre alumnos” (0,20), “Las reglas se aplican con justicia” (0,12) y “participación en las decisiones” (0,41).

Berger y Braxton concluyen que: comunicar las reglas de manera efectiva, hacerlas cumplir de manera justa y hacer participar a los estudiantes del armado de los materiales para las materias, es relevante para la retención durante primer año de la cursada. Esto se aplica también dentro de cada asignatura. Un ejemplo de esto último es la figura del “alumno asistente”, que UNAHUR práctica desde su creación.

Variable Name	Mean	Standard Deviation	1	2	3	4	5	6	7	8	9	10	11	12	13
1. INCOME	10.911	2.766	1.00												
2. HSGPA	7.122	0.988	−.12	1.00											
3. SEXF	1.494	0.500	.00	.13	1.00										
4. RACEW	1.838	0.368	.16	.05	−.02	1.00									
5. POLVIEW	2.773	0.828	−.09	−.01	.16	−.11	1.00								
6. INSTCOM1	3.280	0.949	−.02	.06	.09	.10	−.06	1.00							
7. COMMUN	14.624	2.905	−.02	.12	−.03	.05	−.13	.03	1.00						
8. FAIR	13.311	2.635	−.02	−.04	−.05	.09	−.02	.07	.56	1.00					
9. PARTIC	9.608	3.380	−.00	−.05	−.12	.49	−.12	.06	.14	.28	1.00				
10. SIPEER	17.771	3.255	.03	.09	.16	.12	−.08	.08	.27	.23	.05	1.00			
11. SIFAC	12.667	3.236	.01	.06	.01	.03	−.08	.06	.16	.22	−.35	.11	1.00		
12. INSTCOM2	9.307	2.285	−.04	.08	.04	.11	−.11	.11	.33	.32	.07	.58	.18	1.00	
13. IRENROL	12.850	2.935	−.02	.12	.10	.01	−.00	.09	.35	.27	.03	.48	.11	.62	1.00

Tabla 4: Promedio, Desviación estándar y Correlación entre variables del modelo. Fuente: Berger y Braxton (1998).

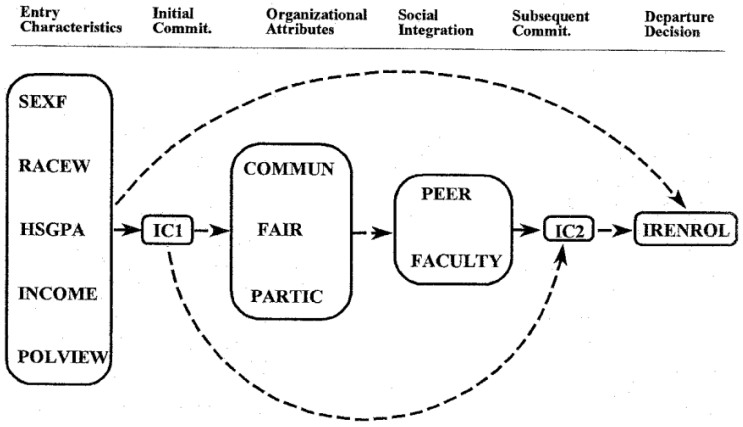


Figura 7: Modelo conceptual de Berger y Braxton. Las líneas punteadas representan influencia indirecta en otras variables. Fuente: Berger y Braxton (1998).

2.3.4. Modelos predictivos que utilizan técnicas de Aprendizaje Automático

Existen numerosos trabajos de modelos que utilizan técnicas de Aprendizaje Automático para predicción del abandono. En Mendoza Santamaria et al. (2022) se investigaron 57 publicaciones del 2017 al 2021 a partir de una búsqueda en Scopus y Web of Science con el siguiente conjunto de términos: {"*Student dropout*", "*Machine Learning*"} . A partir de ahí se consideraron publicaciones que:

- Estuvieran escritas en inglés.
- Están enfocadas a una población de educación superior con modalidad presencial.
- Corresponden a artículos de revista o artículos de conferencias.
- Incluyan técnicas de Aprendizaje Automático o de Minería de Datos que den respuesta a la pregunta de investigación: "¿Cómo predecir el riesgo de abandono, mediante técnicas de *Machine Learning*?

En la Tabla 5 se listan los 11 trabajos que cumplen con estos criterios, resumiendo los factores y técnicas más importantes utilizados en cada uno.

Nombre	Referencia	Factores de abandono	Algoritmos de Aprendizaje Automático
Predicting student dropout: A machine learning approach	Kemper et al. (2020)	Académicos Personales	Logistic Regression Decision Tree
Student dropout prediction	del Bonifro et al., (2020)	Académicos Personales	Support Vector Machine Linear Discriminant Analysis
Supervised learning in the context of educational data mining to avoid university students dropout	De Santos et al. (2019)	Académicos	Decision Tree Random Forest Support Vector Machine
A real-life machine learning experience for predicting university dropout at different stages using academic data	Fernandez-Garc ia et al. (2021)	Académicos	Support Vector Machine Ensemble model Random Forest
Educational data mining: Analysis of drop out of engineering majors at the UnB – Brazil	da Fonseca Silveira et al., (2019)	Académicos Socioeconómicos	Generalized Linear Model Random Forest Gradient Boosting
EvolveDTree: Analyzing Student Dropout in Universities	Santos et al., (2020)	Académicos Socioeconómicos	Decision Tree
Predictive modelling of student dropout using ensemble classifier method in higher education	Hutagaol y Suharjito, (2019)	Académicos Socioeconómicos Personales	Ensemble model Naïve Bayes K-Nearest Neighbor
Analysis of first-year university student dropout through machine learning models: A comparison between universities	Opazo et al., (2021)	Académicos Socioeconómicos Personales	Random Forest Gradient Boosting Decision Tree
Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study	Maksimova et al., (2021)	Académicos	Naïve Bayes Support Vector Machine Neural Network
Application of decision trees for detection of student dropout profiles	Timaran Pereira y Caicedo Zambrano, (2017)	Académicos Socioeconómicos Personales	Decision Tree
A machine learning approach to identifying students at risk of dropout: a case study	Lottering et al., (2020)	Académicos Socioeconómicos Personales	Support Vector Machine Logistic Regression Decision Tree

Tabla 5:Nombre (título), Referencia (autor/es), Factores de Abandono y Algoritmos de Aprendizaje Automático utilizados de los 11 estudios hallados en Mendoza Santamaria et al. (2022). Fuente: Adaptado a partir de Mendoza Santamaria et al. (2022).

Las variables utilizadas en los modelos presentados en esta tesis pertenecen principalmente a factores Académicos, Socioeconómicos y Personales, al igual que en el resumen de Mendoza Santamaria et al. (2022). Como veremos en la Sección 5.3 (Variables de la bibliografía utilizadas) existen Factores Organizacionales que influyen en el abandono, como “Qué becas y tutorías tiene asignada cada alumno”, mencionadas en García (2014). En los modelos propuestos se utilizaron técnicas propuestas en Mendoza Santamaria et al. (2022) Máquinas de Soporte Vectorial, Regresiones Logísticas, Árboles de Decisión *y Gradient Boosting*. Además se utilizaron técnicas más actualizadas con el estado del arte, como *XGBoost* (*Extreme Gradient Boosting*). En la sección 5.6 (Experimentos) se muestra como fueron entrenados los modelos.

2.3.5. Software de terceras partes

Actualmente existe software comercial que cubre parte de estos problemas. Por ejemplo, Ed Machina®[11], es una empresa que desde 2021 aplica Inteligencia Artificial al problema del abandono universitario mediante un sistema de gestión universitario propietario mencionado en Universidades Hoy (2021). No obstante el software no es gratuito ni de código abierto. Además, utilizarlo implicaría otorgar a la empresa el acceso a todos los datos personales de alumnos para poder entrenar sus modelos, previo acuerdo de confidencialidad. Estas restricciones hacen necesarias soluciones de código abierto en donde la universidad pueda garantizar la privacidad de los datos y donde las recomendaciones del sistema no estén sesgadas por beneficios comerciales para un tercero.

2.3.5. Marco conceptual para las variables utilizadas en el modelo

En García (2014) se analizan los factores de abandono en los estudios dentro del campo de las ciencias sociales. Existen dos grandes grupos: los factores centrados en el individuo y los centrados en la organización. En la Tabla 6 se resumen los factores mencionados en la literatura especializada. Entre los factores individuales, se incluyen variables demográficas, socioeconómicas y académicas. Entre los factores organizacionales, se mencionan variables dentro de la dimensión académica, el plan de estudios y los recursos (humanos, equipamientos e infraestructura, financieros y de gestión). Utilizaremos este marco conceptual para identificar de qué tipo es cada variable de nuestro modelo y a que factores afecta.

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

Tabla 6: Factores mencionados en la literatura. Fuente: García (2014) sobre la base de Berger (2000;2002); Berger y Braxton (1998); Cabrera y La Nasa (2001); Gansemer y Schuh (2006); Peltier et al. (1999); Terenzini et al. (2010); Tinto (1993) y **Tabla I Anexo**.

3. Variables y Dataset

3.1. Factores individuales: demográficos, socioeconómicos y académicos previos

La base de datos del SIU-Guaraní de UNAHUR (en adelante Dataset) es administrada por la Secretaría Académica desde que el estudiante es aspirante hasta que egresa. Al momento de la inscripción los estudiantes completan información que no está disponible en su totalidad en el SIU. Algunos de estos datos hacen referencia al capital cultural del estudiante y su entorno mencionados en Bourdieu y Passeron (1998). En cuanto a los datos académicos, al ingreso se solicita: el nombre de la institución para los niveles primario, el nombre del secundario, si tiene estudios superiores no universitarios o universitarios y la condición (finalizado, abandonado o en curso). Asimismo no se solicita el promedio de notas por nivel y se pide

[11] https://edmachina.com/

declarativamente el nivel de conocimiento sobre distintos idiomas. En cuanto al nivel socioeconómico e ingresos del hogar, no se realizan preguntas en forma directa, pero se indaga en el nivel educativo de los padres y su actividad laboral. También se pregunta dónde vive y con quién lo hace (amigos, familiares, pareja, hijos, etc.) y cómo financia sus estudios: aporte de familiares, planes sociales, trabajo, beca, entre otros. Además, se profundiza en los datos laborales: condición de actividad si trabaja o no trabaja, si busca o no busca trabajo, tipo de trabajo, cantidad de horas semanales de trabajo. También se pregunta la disponibilidad de computadora e Internet en el hogar, disponibilidad de cobertura de salud y si practica deporte, cuáles y dónde lo hace. Además, se preguntan cuestiones relacionadas al compromiso, aspiraciones y motivaciones al ingreso basadas en Tinto (1993), a través de las siguientes preguntas:
-Motivos por los que eligió esta Institución educativa (UNAHUR).
-Motivos de mayor peso en la elección de la Oferta (Carrera).

Según los informes de la Secretaría de Evaluación y Planeamiento, UNAHUR (2022a y 2022b), en 2020 el 84% de los estudiantes ingresantes son primera generación de estudiantes universitarios en las familias. El 64% de los padres y 51% de las madres de los estudiantes no han finalizado estudios secundarios. La mayor concentración de ingresantes que trabajan se da en el grupo etario de entre 25 y 34 años, alcanzando un 40,5%. La mitad de los ingresantes que no trabajan ni buscan hacerlo, pertenece al grupo etario de hasta 19 años. El 56,6% de los ingresantes trabaja, un 31,7% se encuentra desocupado ya que declaró no trabajar, pero estar buscando trabajo. Entre quienes trabajan, el 53% tiene hijos y el 77,3% cuenta con cobertura de salud. El 52,8% de los ingresantes en 2021 residen en el partido de Hurlingham. Al momento del ingreso se cuenta con datos laborales del estudiante, la disponibilidad de computadoras e Internet en el hogar, la educación de los padres, si el estudiante es primera generación de universitarios y la condición laboral de los padres. También se consulta sobre el compromiso inicial con la institución UNAHUR y en conjunto todo aporta al capital cultural del estudiante. Estos datos se completan en el momento de la inscripción pero no se encuentran actualizados en los alumnos avanzados.

3.2. Variables organizacionales: políticas académicas, plan de estudio, recursos

A través de la Dirección de Orientación y Acompañamiento se realizan distintas actividades para acompañar a los estudiantes en diferentes instancias. Se ofrecen talleres de “Orientación vocacional", actividades con escuelas secundarias de la zona, en las mismas escuelas y en la universidad. También se desarrollan varias actividades de acompañamiento durante la cursada como el programa “1 estudiante - 1 compañero” en donde estudiantes avanzados acompañan a los ingresantes en su primer año de cursada. Y talleres iniciales de matemática, y de lectura y escritura en la universidad. No todos estos datos se encuentran disponibles en el SIU. La estructura de su oferta académica se realiza en base a cuatro ejes reflejados a través de institutos: Instituto de Educación, Instituto de Salud Comunitaria, Instituto de Biotecnología e Instituto de Tecnología e Ingeniería. Las carreras afines cuentan con un coordinador que atiende las demandas de los docentes y de los estudiantes.

En Figura 8 se observa una matrícula de 18.454 estudiantes en 2020, con el predominio de estudiantes de los institutos de Salud Comunitaria y Educación, con un total de 18.454 estudiantes en 2020. El 60% de los estudiantes de grado y pregrado han aprobado y/o regularizado materias en 2020, de los cuales el 86% se inscribieron en 2021. Del 40% que no ha aprobado ni regularizado materias en 2020 sólo el 35% se inscribió en materias en 2021. En cuanto a los egresados en 2021 el 75% pertenecía al instituto de Educación y en segundo lugar al de Salud Comunitaria con el 21%, y el resto, muy por debajo, a los institutos más vinculados a lo tecnológico. La condición de Inscripto/Reinscripto, el año de ingreso y la cantidad de materias aprobadas en relación al total de las carreras en que se inscribió, son variables a agregar en futuras mejoras de los modelos.

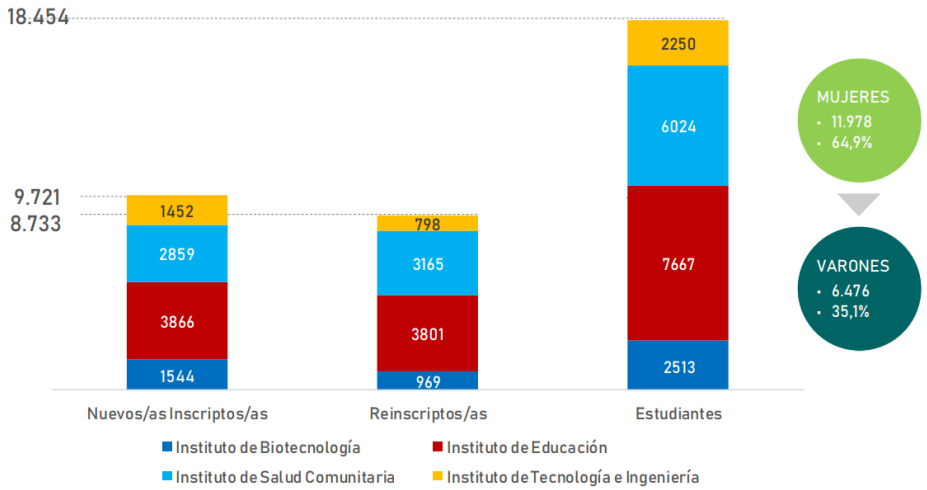


Figura 8: Matrícula 2020 según Nuevos Inscriptos (NI), Reinscriptos (RI) y Total de Estudiantes (TE), agrupado por instituto de pertenencia y género, donde TE = NI + RI. Fuente: Secretaría de Evaluación y Planeamiento, UNAHUR (2022a y 2022b).

3.3. SIU-Guaraní UNAHUR

En Dataset se almacenaron los datos personales y académicos de los alumnos. En la Figura 9 se ve la evolución de la matrícula en UNAHUR, desde su creación en 2016 hasta el año 2022. En 2022 se matricularon 25.861 personas, que representan a 37.394 alumnos. La discrepancia entre alumnos y personas se justifica por varios motivos: cada persona que ingresa a UNAHUR se inscribe al Curso de Preparación Universitaria (CPU), categorizado como carrera, aunque dura tres meses. Luego existe la posibilidad de anotarse o cambiar a más de una carrera y por último hay solapamientos técnicos, carreras incluidas dentro de otras (i.e Tecnicatura Universitaria en Programación está incluida en Licenciatura en Informática). La diferencia entre estudiante y persona refiere a que una persona puede estar inscripta en una o más ofertas, por lo que cuenta cómo un estudiante diferente en cada una de ellas. Por todo esto, las predicciones que realizaremos en esta Tesis serán sobre personas y su probabilidad de abandono de todas las ofertas a las que se inscribieron.

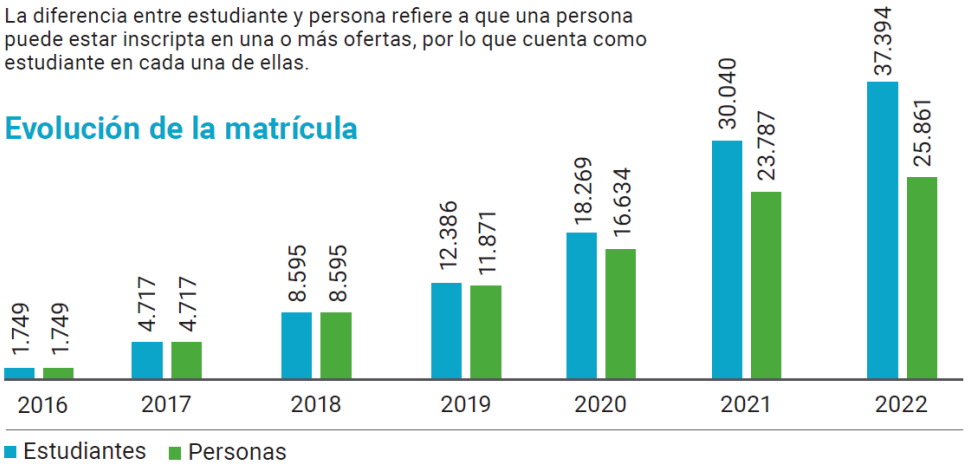


Figura 9: Evolución de la matrícula en UNAHUR, años 2016-2022. Fuente: elaboración propia a partir de Secretaría de Evaluación y Planeamiento (2023).

El Dataset tiene más de 10.000 campos y 500 tablas. Se relevó el subconjunto de datos necesarios para generar las variables de los sucesivos modelos. Al ser una base en producción no se puede acceder a ella de manera directa (i.e. entrenar los modelos a partir de la base). Para resguardar la integridad de la base y por

cuestiones de performance se accedió a ella a través de una consulta parametrizada que obtiene información de los alumnos activos en los semestres que se desee entrenar. Como se muestra en la Figura 10, los datos son extraídos a través de una consulta a SQL. El conjunto de features extraídas son guardadas en un archivo ".csv", donde cada fila le corresponde a una persona y cada columna a un feature específico. Como se explicará en la Sección 3.2.2 (Valores atípicos y Datos Faltantes), se eliminaron o corrigieron los datos para un mejor funcionamiento de los modelos. Por último, para garantizar la privacidad de los datos personales, se codificaron los valores alfanuméricos antes de subir a la nube y se eliminó la vista de la nube luego de entrenar el modelo. Se guardaron los modelos entrenados para poder realizar las predicciones y cálculo de métricas. Solo se preservó un identificador por persona para que los usuarios puedan identificar los alumnos en riesgo.

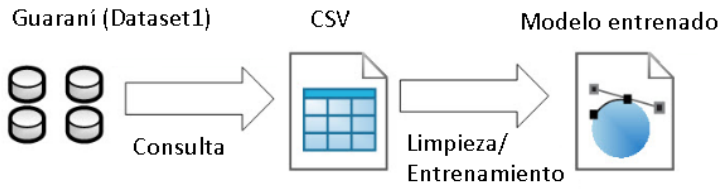


Figura 10: Proceso de extracción de datos. Fuente: Adaptado de Azevedo y Santos (2008); Schröer et al. (2021); Martínez-Plumed et al. (2021).

3.4. Pre-procesamiento

Se extrajeron los datos del SIU-Guaraní, se limpiaron datos atípicos y se imputan datos faltantes antes de entrenar los modelos. Se transforman las variables categóricas en numéricas para poder ejecutar modelos que no soportan ese tipo de variables.

3.4.1. Consulta a la base

Dataset se encuentra instalada en un servidor PostgreSQL® versión 12 y las consultas se realizan con el cliente DBeaver® versión 21. La metodología consiste en llenar una tabla como la que se muestra en la Figura 11, con las variables que se necesiten para cada persona (clave primaria de la tabla), para luego ser exportado en un archivo ".csv". La Figura 11 contiene la representación en SQL de algunas de las variables del modelo. Luego se realiza una consulta para poblar la tabla, que trae solo información de los últimos cuatro semestres de todos los alumnos regulares considerando todas las carreras donde se encuentren inscriptos. De esa manera se evita entrenar el modelo con alumnos que o bien ya egresaron o bien ya abandonaron, asumiendo que siempre pueden volver más adelante. Los cambios de universidad se consideran "Abandonos" desde el punto de vista de los modelos porque estos cambios no son censables por el momento. Es recomendable actualizar (reentrenar) el modelo en cada semestre para tener una caracterización del alumnado más cercana a la cohorte actual.

```
CREATE TABLE negocio.modelo1(
    persona int4 PRIMARY KEY,
    fechanacimiento date not null,
    Edad int2 not null,
    sexo bpchar(1) ,
    nacionalidad int2,
    email VARCHAR (100) ,
    cant_eval_m3 int2,
    cant_eval_m2 int2,
    cant_eval_m1 int2,
    cant_eval_m0 int2,
    ...
```

Figura 11: Tabla temporal para la consulta que genera Dataset. Fuente: Repositorio Git (2023a y 2023b).

El archivo se procesa en Python, en la plataforma Colab® de Google en donde también se entrena el modelo. Una vez entrenado, se procede a hacer las mediciones y métricas. Se eligió esa plataforma por la capacidad de procesamiento y almacenamiento en la nube y porque facilita el uso colaborativo de los modelos para los actores que tienen acceso. El IDE de Colab® es el navegador mismo y embebe el código, la consola y los gráficos (resultados) en una misma interfaz lo que lo hace muy versátil para prototipar e iterar el modelo. Como el servidor esta en Google®, también es muy potente para el procesamiento y uso de RAM, y al mismo tiempo permite cargar los datos directamente desde Google Drive®, lo cual facilita la carga, administración y uso colaborativo de archivos (i.e. ciertos usuarios pueden acceder a los archivos de mi Drive desde la nube). Como veremos en la Sección 5.5.1. (Estandarización y disociación de datos) los datos como la calle y el email se codifican a valores numéricos, tanto para que funcionen para todos los algoritmos como para disociar el dato de la persona.

3.4.2. Valores atípicos y Datos Faltantes

Realizamos un preprocesamiento para corregir los valores erróneos y/o atípicos (por ejemplo, entradas en la variable "Edad" con valor 0). Para variables como "Edad" y "Hace cuantos meses completó el censo" se eliminaron datos atípicos extremos según la siguiente definición: Un dato *d* es atípico extremo sii $d > Q_3 + 3 * DIQ$ *o* $d < Q_1 - 3 * DIC$, dónde Q_3 es el tercer cuartil y *DIQ* es la distancia entre Q_1 y Q_3, del test de Tukey (1977). La Figura 12 muestra los histogramas de frecuencias absolutas con y sin datos atípicos para la "Edad". Cuando fue posible se unificaron los valores, como la variable categórica "Número de Piso", en donde se unificaron "planta baja" y "piso 0". Se dejaron las variables con datos faltantes, y cuando la correlación entre registros lo permitió, se realizaron imputaciones mediante la técnica k-nearest neighbors (KNN) descrita en Mucherino et al. (2009). Por ejemplo, cuando no tenemos el barrio, se sustituye por el barrio más cercano (en latitud y longitud) al partido y calle informadas.

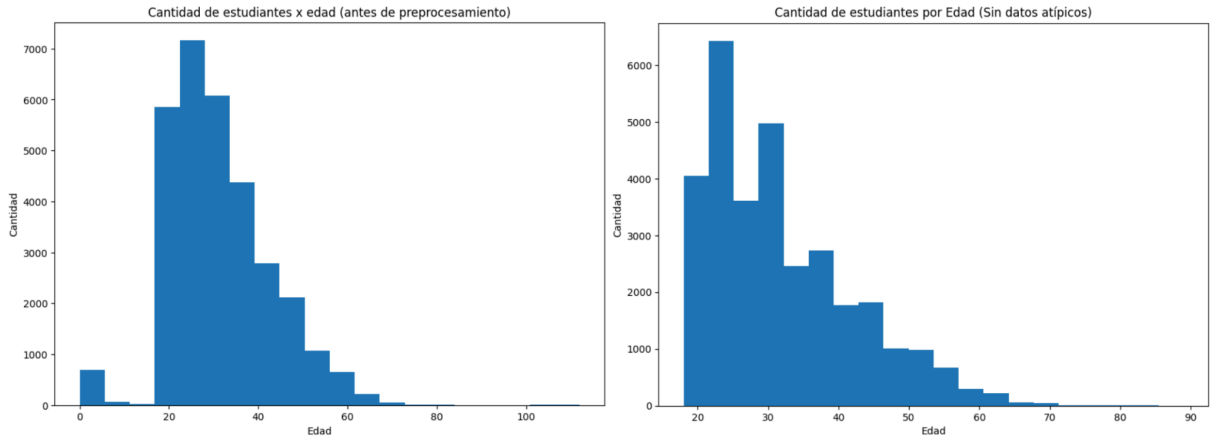


Figura 12: Frecuencia absoluta de la variable "Edad" antes y después de corregir los datos atípicos. Fuente: Elaboración propia.

3.4.3. Estandarización y disociación de datos

El primer paso fue estandarizar los datos para que los modelos que sólo trabajan con datos numéricos, como Máquina de Soporte Vectorial, pudieran funcionar. Primero se unificaron los datos alfanuméricos, pasando a mayúscula y unificando la codificación cuando fue posible. Por ejemplo, con errores de ortografía de las calles, mediante un diccionario de calles. Luego se transformaron en una representación numérica. Esa transformación protege los datos del alumno en términos de la ley actual de protección de datos[12]. Para los datos faltantes se asignó un valor especial. A los numéricos se les asignó un

[12] https://www.argentina.gob.ar/sites/default/files/antecedente_2018_25326.pdf

valor que no sesgara el modelo y para los datos categóricos se asignó un valor fuera del dominio. Para la cantidad de hijos se imputó la media para datos faltantes porque no se observó una diferencia significativa para los modelos asignando otro valor. Para que los modelos sean más explicativos y al mismo tiempo las heurísticas que particionan los nodos en los algoritmos basados en árboles funcionen mejor, se discretizan algunas variables continuas. “Cantidad de hijos” es técnicamente continua porque se imputó 0,7 hijos como dato faltante. Para salvar esa situación, se transformó la variable multiplicando su valor por un escalar (cantidad de hijos * 10), de manera de tener solamente 5 valores posibles, enteros. De esta manera 10 representa 1 hijo y 7 representa la imputación (el promedio de hijos, que es 0,7). Esta transformación, además, preserva el orden original. En la Tabla 7 se listan algunos ejemplos.

Variable	**Valor Asignado**	**Descripción**
Edad	media	Observamos que la media no sesgaba los modelos significativamente.
Meses Censo	255	Se asignó 255 para que no llenar el censo tenga un valor más alto que cualquiera que lo haya llenado por lo menos una vez.
#Hijos	7	Se imputo la media (0,7) y se multiplicó por 10 la variable (7) para generar una variable entera con valores: 0, 7, 10, 20 y30. De manera de generar un árbol con menos nodos que si tuviéramos una variable continua*.
Turno Preferido	-5	Mañana = 0; Tarde = 1; Noche = 2; Se hizo lo mismo con otras variables categóricas.
Calle	-6	Se corrigió la calle cuando fue posible.
Calle censo	-6	Se mantiene igual ID que en Calle adrede.
Número de Piso	-7	Se unificaron sinónimos como “0” y “Planta Baja”.

Tabla 7: Estandarización de datos. y Valor asignado en caso de tener datos faltantes. * “30” representa 3 hijos o mas, asi que se generan a lo sumo 5 nodos (0,7,10,20 y 30) para #Hijos y #Familiares. Fuente: Elaboración propia.

3.5. Análisis exploratorio de datos

3.5.1 Abandono condicional

De un total de 33.584 alumnos matriculados en UNAHUR hasta 2021, el 56,93% no trabajaban al comenzar los estudios. Observamos que para ciertas variables, la proporción de abandono no es la misma para todo su dominio. Variables como sexo, edad, cantidad de becas y cantidad de horas trabajadas tienen poca correlacion lineal con la probabilidad de abandono, pero mostraron tener diferente proporción de abandono según qué valores se tomen. Respecto de las variables socioeconómicas (i.e. “Cantidad de hijos”, “Sexo” , “Edad” y “Salud” (Cobertura) y “Tipo Vivienda”) relevantes para autores como Arias et al. (2015), Chavez (2020) y Marquez (2022), solo los alumnos agrupados por “Sexo” o por “Edad” registraron un porcentaje de abandono diferente al poblacional. Se utilizó el test chi cuadrado (χ^2) para variables categóricas sugerido en Solanas et al. (2005) para censar la independencia (o no) de todas las variables respecto del abandono. El test se explica en la Sección 4.3.2.4. (Test Chi Cuadrado de independencia para variables categóricas). Todas las variables que se listan en esta sección no pasaron el test $\chi 2$ con significancia 0,05. Luego se calculó la significancia Cramer’s V (Cramér H., 1946), para ponderar el grado real de dependencia de dichas variables.

3.5.1.1. Abandono condicional por Carrera

La carrera es una variable importante para discriminar el abandono. En la Figura 13 se muestra la proporción de abandono para carreras con más de 840 inscriptos, ordenadas de manera descendente comparadas con el abandono poblacional. En el eje *x* se muestra la carrera con la cantidad de inscriptos entre paréntesis. Las mayores proporciones (0,95 y 0,87) se explican por talleres o cursos como "Talleres deportivos, recreativos, culturales y de oficio (TDR)" y "Curso de manipulación de los alimentos (CMA)", que en realidad son actividades de extensión y en donde la mayoría de inscriptos no son alumnos de carreras regulares. Por lo tanto, es esperable que no se censen inscripciones luego de ese curso. No obstante, el modelo lo considera "abandono", queda en manos de los actores del sistema considerar esta información para tomar acciones o no con estos alumnos. El Curso de Preparación Universitaria (CPU) presenta la mayor proporción de abandono (0,7). Esto se explica porque el abandono durante el primer año es mucho mayor (tanto en cantidad como en porcentaje) al abandono de los años subsiguientes. El abandono censado una vez comenzada la carrera baja al 47%, como se indica en el trabajo de Giuliano y Pustilnik (2024). La mayoría de las carreras presentan un rango de proporciones que van de 0,37 a 0,46, pero existen excepciones. La "Licenciatura en Educación (LE)" presenta una proporción de 0,21 debido a que la mayoría de los alumnos se migró a un plan nuevo (2020) y aún no se estabilizó el abandono para ser censado correctamente. El resto de las excepciones corresponde a carreras que aún cuentan con pocos alumnos, dado que la universidad en sí comenzó a funcionar en 2016 y recién tuvo una inscripción mayor a 10.000 alumnos en 2019. A partir del test χ^2 y el valor de Cramer's V (0,33), la "Carrera" no resultó independiente del abandono.

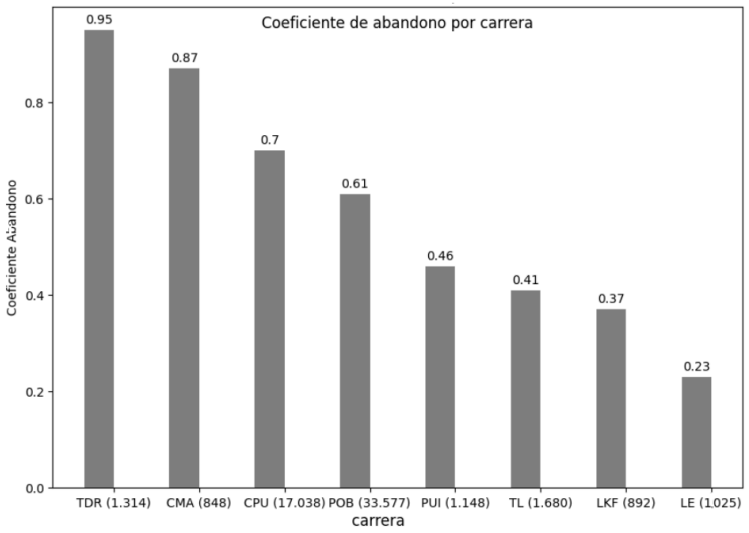


Figura 13: Coeficiente (proporción) de abandono por Carrera en orden descendente para carreras con más de 840 alumnos hasta 2021. Talleres deportivos, recreativos, culturales y de oficio (TDR); Curso de manipulación de los alimentos (CMA); Curso de Preparación Universitaria (CPU); Abandono poblacional (POB); Profesorado universitario de Inglés (PUI); Tecnicatura en Laboratorio (TL); Tecnicatura en Laboratorio (TL); Licenciatura en Educación (LE). Fuente: Elaboración propia.

3.5.1.2. Abandono condicional por Cantidad de Horas de Trabajo por Semana

En la Figura 14 se muestran los alumnos agrupados por cantidad de horas de trabajo por semana al comenzar su cursada. Hay 19.121 alumnos que no trabajan al comenzar la cursada. Como sugiere Fazio (2004), el mayor porcentaje de abandono no se alcanzó entre los alumnos que trabajaban, a menos que lo hagan por más de 35 horas por semana. Para los alumnos que trabajan entre 10 y 35 horas por semana el porcentaje de abandono está en rango del 56,42% al 56,66%.

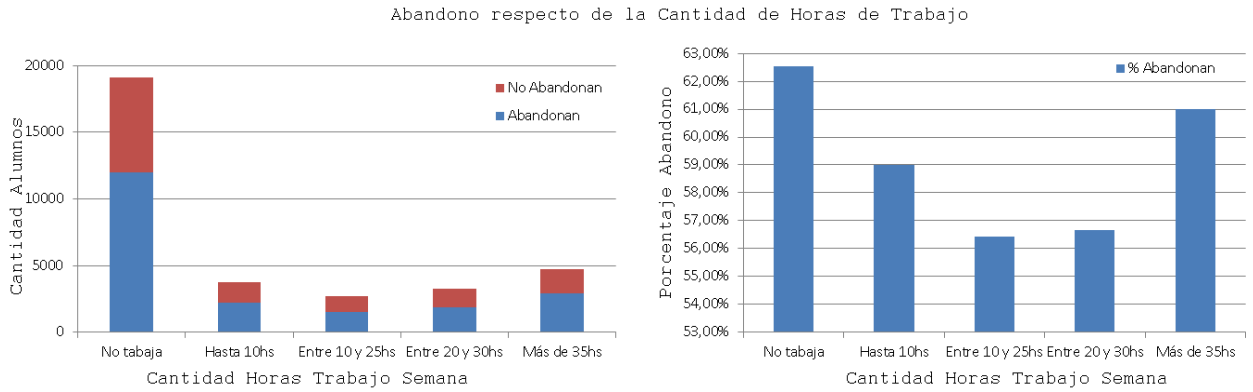


Figura 14: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Cantidad de horas de trabajo semanales" (izquierda). Porcentaje de alumnos que abandonan (derecha). Fuente: Adaptado de Pustilnik y Ndukanma (2023b).

3.5.1.3. Abandono condicional por Cantidad de Familiares/Hijos

Autores como Giovagnoli (2002) afirman que "los estudiantes que viven con sus familias tienen mayor riesgo de desertar que los alumnos que tienen que vivir en forma independiente". Sin embargo, medimos que los alumnos que conviven con uno o más familiares, mostraron tener porcentaje de abandono levemente menor (del 57,6% al 59,01%) respecto de los alumnos que viven solos. Adicionalmente, los que no completaron esa información, reportaron un porcentaje mayor (70,13%). Giovagnoli (2002) sugiere que los alumnos que vienen con familiares son "locales" a la universidad, mientras que los que viven solos son alumnos que vienen de más lejos, teniendo así costos de alquiler y de viaje subyacentes, que los hace más propensos a no terminar sus estudios. En las Figuras 15a y 15b se muestran las cantidades y porcentajes para alumnos que: o bien viven solos, o bien viven con al menos un familiar o no reportaron con quién viven.

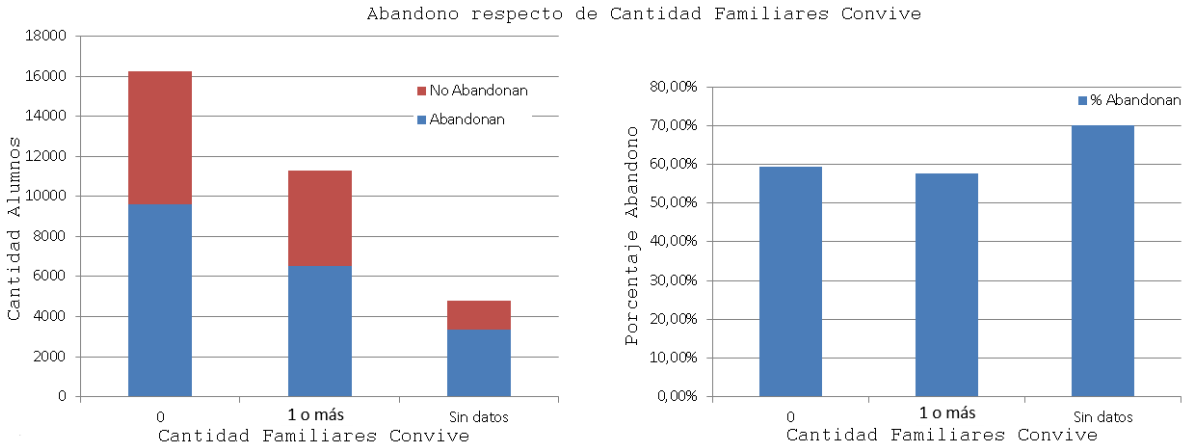


Figura 15a: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Cantidad Familiares Conviven" (izquierda). Porcentaje de alumnos que abandonan (derecha). 0 → vive solo; 1 o más → vive con familiares; Sin datos → dato faltante. Fuente:Fuente: Elaboración propia.

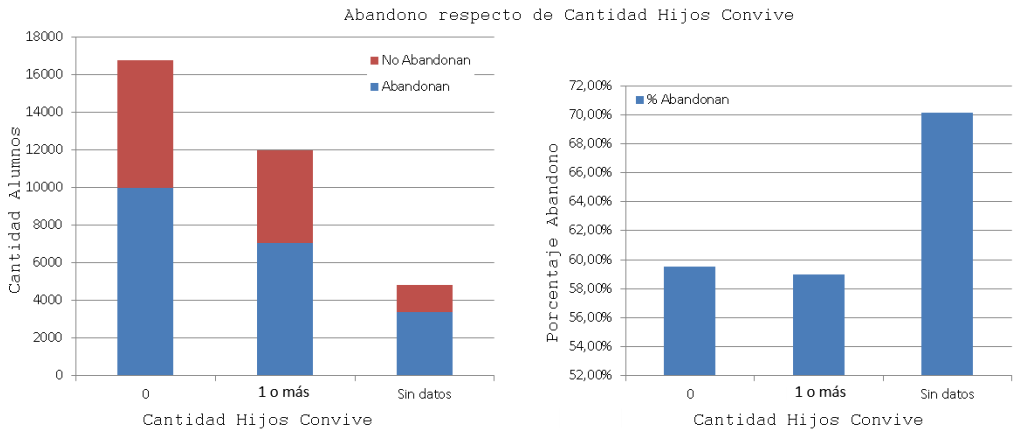


Figura 15b: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Cantidad Hijos Conviven" (izquierda). Porcentaje de alumnos que abandonan (derecha). 0 → vive solo; 1 o más→ vive con familiares; Sin datos → dato faltante. Fuente: Elaboración propia.

3.5.1.4. Abandono condicional por Tiempo de Viaje

En la Figura 16 se muestra el abandono por tiempo de viaje calculado en transporte público. Se registró un mayor abandono (64,51%) en los alumnos que tardaban entre 0,1 y 0,3hs (de 6 a 18 minutos) respecto de otros tiempos, inclusive más que los tardaban entre 0,3 y 2hs, cumpliendose parcialmente que a mayor distancia, mayor es el abandono, como mencionan en Giovagnoli (2002), Kuna et al. (2011) y Di Gresia et al. (2007). Recordemos que este tiempo de viaje fue estimado para transporte público, pudiendo pasar en realidad que los alumnos de distancias más lejanas se trasladen en vehículos personales, pudiendo tener en realidad un tiempo menor al calculado. Los tiempos de viaje mayores a 2 horas se consideraron datos atípicos. Estos datos, junto con los tiempos que no se pudieron calcular por contener direcciones inválidas, suman un 11,2% del total de alumnos y no fueron considerados en la estadística.

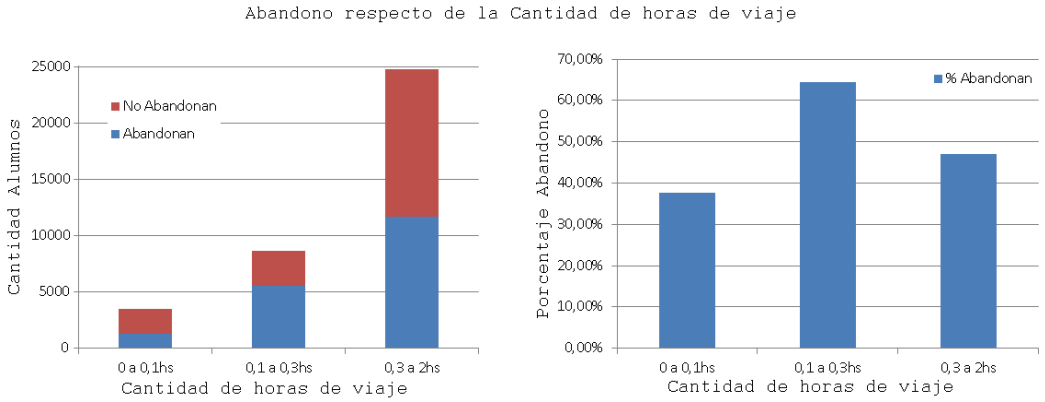


Figura 16: Cantidad de alumnos que abandonan. Porcentaje de alumnos que abandonan, según "Tiempo de viaje", si vinieran en transporte público hasta la universidad. Fuente: Elaboración propia.

3.5.1.5. Abandono condicional por Género

Autores como Arias et al. (2015) y Chavez (2020), afirman que las mujeres se ven más afectadas por las tareas domésticas. Marquez (2022) afirma que durante el COVID-19 se incrementaron estas desigualdades y sugiere que esto provoca más situaciones de abandono en las mujeres que en que el promedio. Sin embargo, en la Figura 17 se observa que durante ese periodo se matriculó una mayor cantidad de mujeres (22.425), y con un menor porcentaje de abandono (58,27%) respecto de los varones (66,06%). Este fenómeno podría explicarse por múltiples motivos. Al igual que sucede en las universidades privadas,

en donde el abandono promedio es menor que en las universidades públicas[13], hay alumnos que ni siquiera intentan o no pueden inscribirse debido a condiciones económicas, o por la sobrecarga en sus tareas domésticas, como es el caso de las mujeres. Este efecto hace que ese factor no sea censado, ya que esas "posibles alumnas" no figuran en el sistema. En la Figura 17 se muestran los resultados. En este análisis, no se consideraron otros géneros por no ser estadísticamente significativos ya que representan menos del 1% del total.

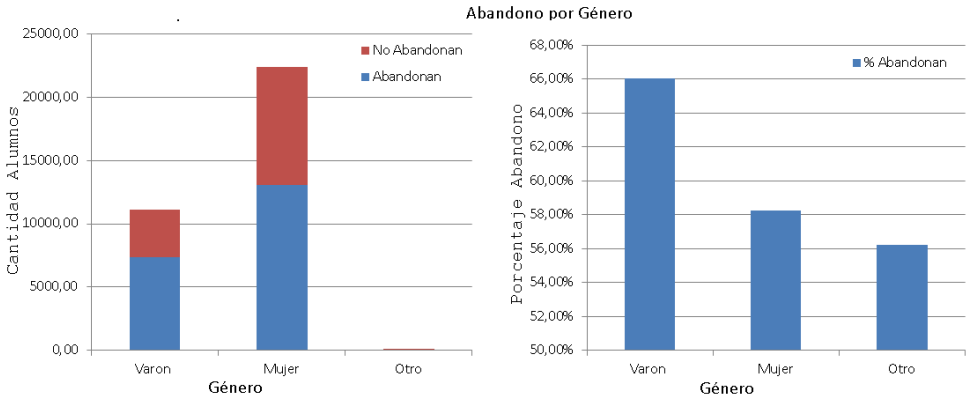


Figura 17: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Género" (izquierda). Porcentaje de alumnos que abandonan (derecha). Fuente: Elaboración propia.

3.5.1.6. Abandono condicional por Dominio de Email

Por recomendación de los directores de carrera censamos el abandono por "Dominio de Email". Hasta la actualidad no encontramos modelos que se beneficien de esta variable. Encontramos que los alumnos con dominio institucional (@unahur.edu.ar, @inta.gob.ar, etc) abandonan menos (32,08%) que los alumnos con otros dominios de email (63,58%). Los dominios no tradicionales (< 1% del total) no son estadísticamente significativos. Una hipótesis es que contar con un mail institucional denota una mayor adaptación a la vida académica (Tinto, 1982) que no tenerlo. La Figura 18 muestra el porcentaje de abandono según el dominio del email.

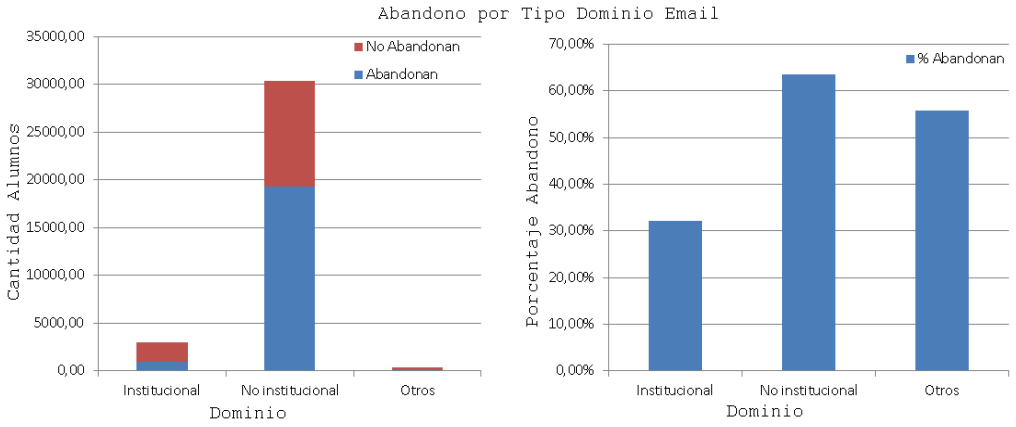


Figura 18: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Dominio de Email" (izquierda). Porcentaje de alumnos que abandonan (derecha). Fuente: Elaboración propia.

[13]El 81% de los alumnos de pregrado + grado cursan establecimientos estatales (Departamento de Información Universitaria, 2022).

3.5.1.7. Abandono condicional por Cantidad de Becas

Autores como Di Gresia et al. (2007), Cameron y Taber (2001) afirman que las becas, en especial si están dirigidas a grupos de riesgo, tienen un efecto positivo en la retención. En la Figura 19 se observa que los alumnos que perciben al menos una beca abandonan menos (entre el 50,16% y el 51,17%) que los alumnos que no perciben ninguna (61,15%).

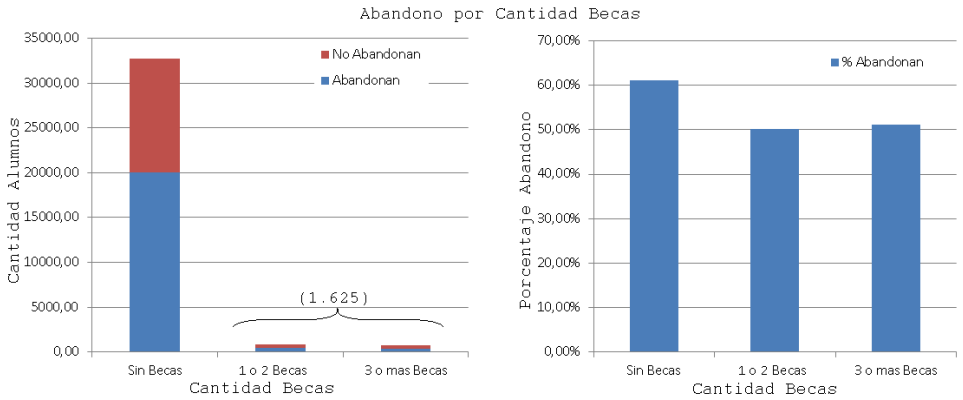


Figura 19: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Cantidad de Becas" (izquierda). Porcentaje de alumnos que abandonan (derecha). Fuente: Elaboración propia.

3.5.1.8. Abandono condicional por Edad

Hay autores como Antoni et al. (2007), Chudnovsky (2003) y Di Gresia (2009), que afirman que todas las edades tienen la misma probabilidad de abandono, o incluso que los de mayor edad abandonan menos. Sin embargo, en la Figura 20 se observa que los alumnos de 18 a 20 años abandonan menos (57,78%) que los alumnos mayores de 20 años (entre el 60,42% y el 62,67%).

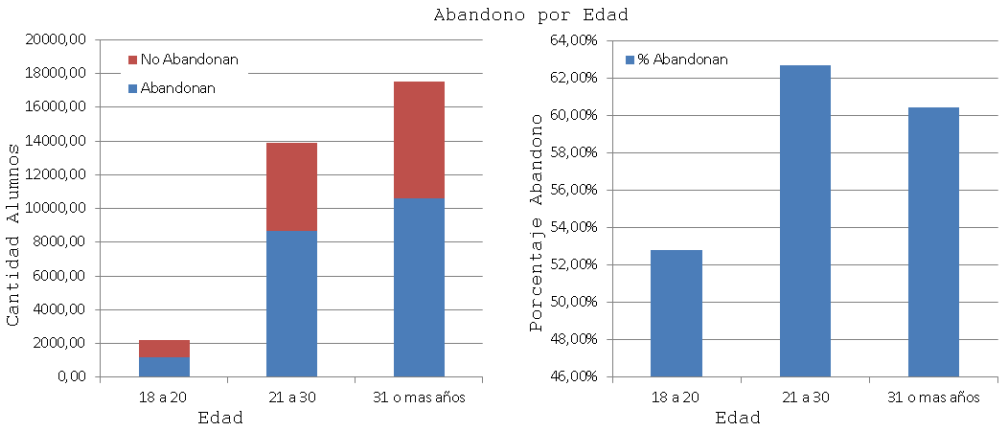


Figura 20: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Edad" (izquierda). Porcentaje de alumnos que abandonan (derecha). Fuente: Elaboración propia.

3.5.1.9. Abandono condicional por Turno

Autores como Berges et al. (2007) afirman que el turno no tiene una importancia significativa en el abandono. En la Figura 21 se muestra el abandono por turno ordenado de manera ascendente por abandono. El turno "Inactivo" representa a los alumnos turnos que actualmente no se utilizan. El "Turno Mañana" denominado "Turno mañana, martes, jueves y sábados" presentó el menor porcentaje de abandono (34,42%). No obstante, no se observa un menor abandono en turnos en el horario de la mañana en general, ya que

tenemos ejemplos similares para los turnos tarde y noche. Tanto para abandono por debajo como por encima de la media.

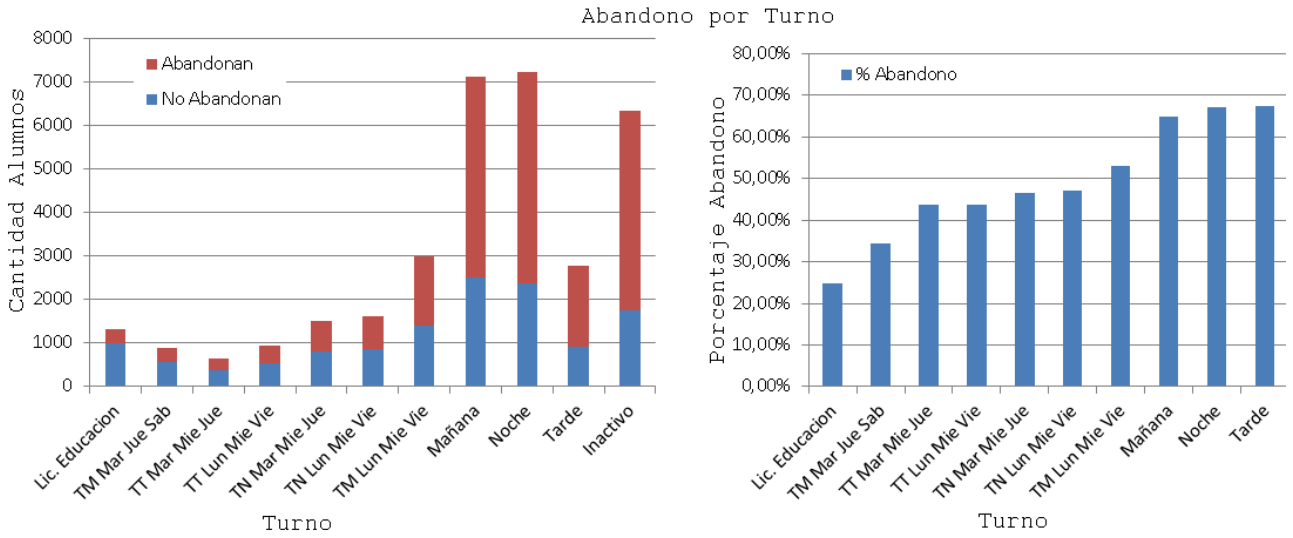


Figura 21: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Turno" (izquierda). Porcentaje de alumnos que abandonan (derecha). Ambos gráficos ordenados por porcentaje de abandono. Turno Mañana (TM), Turno Tarde (TT), Turno Noche (TT) . Fuente: Elaboración propia.

3.5.1.10. Abandono condicional por "Cantidad de Meses Censo"

"Cantidad de Meses Censo" representa: hace cuántos meses el alumno completó el censo. La explicabilidad del abandono por motivos académicos está mencionada por varios autores de la bibliografía. En la Figura 22 se visualiza un abandono promedio mayor (80,11%) para alumnos que no completaron el censo hace más de 50 meses respecto del grupo que lo completó hace menos de 50 meses (50,47%). Como veremos en la Sección Resultados, el porcentaje de abandono crece respecto de "hace cuantos meses se completó el censo".

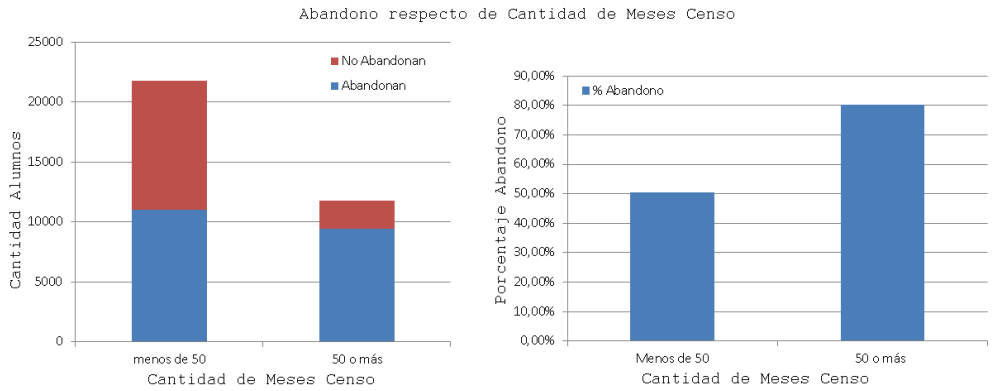


Figura 22: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Cantidad de Meses Censo" (izquierda). Porcentaje de alumnos que abandonan (derecha).

3.5.1.11. Abandono condicional por Cantidad de Evaluaciones Hace un Semestre.

En la Figura 23 se visualiza un abandono promedio menor (50,91%) para alumnos que rindieron al menos un examen hace un semestre (en el semestre anterior al que se quiere predecir) respecto del grupo que no rindió ninguno (79,04%). Como veremos en la Sección Resultados, el porcentaje de abandono decrece respecto de "Cuántos exámenes rindió hace un semestre". Esta relación también se presenta para 2, 3 y 4 semestres anteriores.

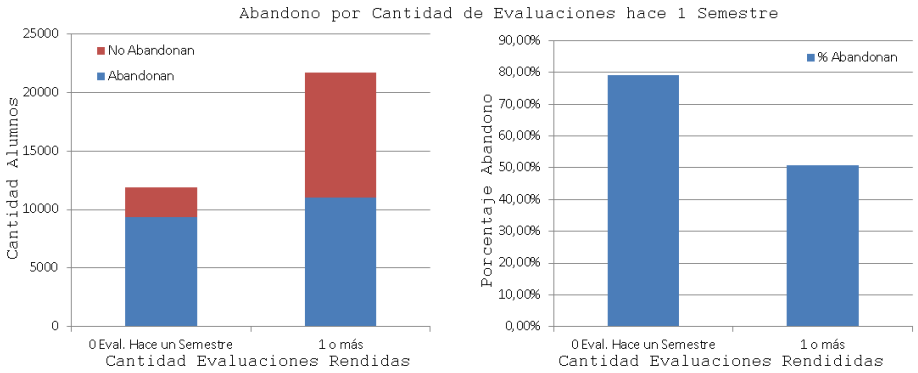


Figura 23: Cantidad de alumnos que abandonan/No abandonan, agrupados por "Cantidad de Evaluaciones Rendidas Hace un Semestre" (izquierda). Porcentaje de alumnos que abandonan (derecha).

3.6. Hipótesis a testear

Queremos testear hipótesis que surgen de la bibliografía y de las recomendaciones de actores de UNAHUR. En la Sección 6 (Resultados) testeamos todas las hipótesis midiendo el abandono condicional y su importancia para discriminar el abandono sobre algunas variables.

H1 Cantidad de Evaluaciones y Cantidad de Meses Censo: Queremos testear si la cantidad de evaluaciones rendidas de los alumnos a lo largo de su historia académica influyen en el abandono. Para reflejar el rendimiento en diferentes semestres, tendremos una variable diferente para cuantificar cuatro semestres consecutivos (i.e. cantidad de evaluaciones en 2020 primer semestre, cantidad de evaluaciones en 2020 2do semestre, etc). Además queremos censar cuándo fue la última vez que el alumno completó el censo opcional, que se realiza todos los años (Cuántos meses pasaron desde la última vez que lo completó). Estas variables surgen de García (2014). Varios autores de la *Tabla 1 anexo*: Berges et al. (2007); Cerro (2007); Ferreira (2007); Gertel et al. (2007); Giner et al. (2007); Giovagnoli (2007); Giuliodori et al. (2010); Goldenhersh et al. (2011); Kuna et al. (2011); López et al. (2012); Paz et al. (2011); Pron (2007); Ríos (2010); Sosa Escudero et al. (2009); Jaime (2004) y Di Gresia et al. (2007) y Di Gresia (2009) mencionan una o más variables relacionadas con el rendimiento académico, tanto antes (en el secundario) como durante la universidad. Completar el censo constituye una gestión adicional del estudiante, que puede ser interpretada como mayor integración a la vida universitaria.

H2 Tiempo de Viaje: Según la Secretaría de Evaluación y Planeamiento, UNAHUR (2022b) solo el 52,8% de los ingresantes en 2021 residen en el partido de Hurlingham. Para Kuna et al. (2011) y Di Gresia et al. (2007) el tiempo de viaje y el lugar de residencia son variables demográficas importantes. Queremos testear si un mayor tiempo de viaje produce un aumento en la tasa de abandono.

H3 Dominio de Email: Documentos como "Autodiagnóstico de las capacidades institucionales. Implementación de la educación remota en las universidades" (Dirección General de Educación Superior

Universitaria de Perú, 2021) sugieren que la comunicación a través del email institucional disminuye al abandono. Desde 2020 UNAHUR ofrece cursos mixtos (presencial-remoto). Queremos testear si los alumnos con dominio @unahur.edu.ar (o institucional) abandonan menos. Por otro lado, obtener y utilizar este mail constituye una gestión adicional del estudiante, que puede ser interpretada como mayor integración a la vida universitaria.

H4 Genero y Cantidad de Hijos: Como sugieren Arias et al. (2015), Chavez (2020) y Marquez (2022), la división no equitativa en las tareas de la casa (en particular durante la pandemia de COVID-19), impacta en el aumento de las desigualdades de género y en la sobrecarga de tiempo de trabajo de cuidados que pre existía en el conurbano bonaerense. Se desea testear cómo influyeron esas variables demográficas en la proporción del abandono. Hay estudios como Gertel et al. (2007) que sugieren que el sexo no influye de manera significativa mientras que hay otros, como Giner et al. (2007), que sugieren que sí.

H5 Cobertura de Salud: Según la Secretaría de Evaluación y Planeamiento, UNAHUR (2022b) solo el 77,3% tiene cobertura de salud. Queremos testear cómo influye esa variable socioeconómica en el abandono.

H6 Turno: Queremos testear si la elección del turno a cursar (Mañana/Tarde/Noche) censa indirectamente la variable socioeconómica turnos laborales. Y si los turnos tienen diferente grado de incidencia en el abandono.

H7 Piso y Tipo de Vivienda: Queremos testear si "Número de Piso" y el "Tipo de Vivienda" censa indirectamente los ingresos del hogar. Variable socioeconómica, García (2014) . Se testea si un piso más alto indica un acceso diferente a servicios por estar en zonas más edificadas y si influye positivamente en el abandono (variable socioeconómica).

H8 Cantidad de Becas: Queremos testear si el financiamiento en becas que perciben los alumnos a lo largo de su historia académica influyen en el abandono. Autores como Cameron y Taber (2001) y Di Gresia et al. (2007), afirman que las becas tienen una influencia positiva en la retención.

H9 Cantidad de horas de trabajo por semana: Queremos testear cómo se comporta la retención en función de la cantidad de horas de trabajo, en particular, si disminuye. Como mencionaremos en la Sección Resultados, varios autores mencionados en la *Tabla 1 anexo*, consideran que existe una relación.

H10 Carrera: Según la Secretaría de Evaluación y Planeamiento, UNAHUR (2022b) existen diferentes proporciones de alumnos y egresados por carrera. Queremos testear si la tasa de abandono es diferente para cada carrera.

H11 Edad: Varios autores de la *Tabla 1 anexo* como Paz et al. (2011), Giner et al. (2007) y Giovagnoli (2007) sugieren que el abandono es mayor para alumnos de mayor edad. Queremos testear si el abandono es proporcional a la edad.

4. Métodos y Materiales

En la siguiente sección se realizará una breve introducción a las metodologías y algoritmos utilizados: *Cross Industry Standard Process for Data Mining* (CRISP-DM), Árboles de Decisión , *XGBoost* y Máquinas de Soporte Vectorial.

4.1 Cross Industry Standard Process for Data Mining (CRISP-DM)

El proceso de ingeniería de software elegido para la elaboración de los modelos fue el estándar CRISP-DM, explicado en Azevedo y Santos (2008). Este estándar deriva del Knowledge Discovery in Databases (KDD), el primer estándar en Minería de Datos en 1993. El proceso KDD es utilizado para llevar a cabo la extracción automatizada de conocimiento partiendo de grandes volúmenes de datos, y se aplica en diferentes iteraciones hasta que los datos quedan listos para obtener información no trivial a partir de los mismos. En la Figura 24 se ven las cuatro etapas de KDD:

1. Recopilación de datos. Recopilación de datos de diferentes fuentes integrándose en un único repositorio.
2. Selección, Limpieza, transformación. Una vez recopilados todos los datos, durante esta fase, se seleccionan los datos que se entienden como más importantes dentro del modelo y se transforman para poder procesarlos con mayor facilidad. Esta fase se descompone de tres subfases:
    a. Selección de datos: Mediante técnicas de filtrado de registros y de atributos, se eliminan los datos irrelevantes para el análisis posterior.
    b. Limpieza de datos con valores inexistentes o erróneos.
    c. Transformación de datos: Esta tarea consiste en transformar los datos pre procesados ya que en la fase posterior a la Minería de Datos se aplicarán una serie de algoritmos sobre estos datos.
3. Minería de datos. Se aplican algoritmos de estadística sobre los datos transformados con el objeto de obtener variables y generar modelos. Los modelos se alimentan de estas variables como datos de entrada, son entrenados, y pueden generar observaciones nuevas a partir de esos datos o a partir de nuevas observaciones. A todo este proceso se lo suele llamar Minería de Datos.
4. Interpretación de los resultados y evaluación de los modelos. Se evalúa la eficacia de los modelos para predecir o clasificar información. Si esto ocurre se genera un nuevo conocimiento, de lo contrario se tiene que seguir mejorando el modelo.

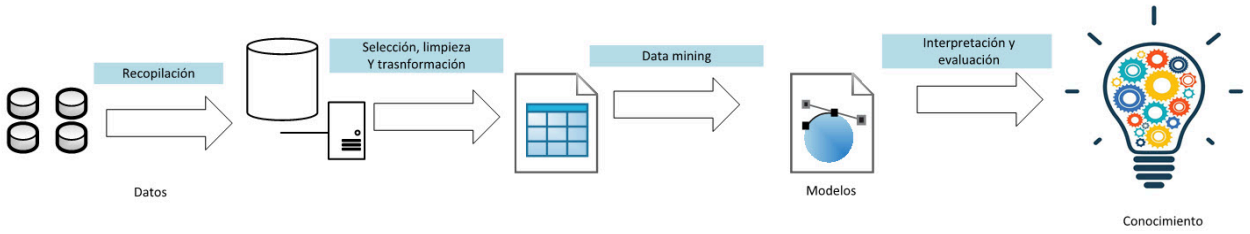


Figura 24: Etapas de KKD. Fuente: Adaptado de Azevedo y Santos (2008); Schröer et al. (2021); Martínez-Plumed et al. (2021).

Como se menciona en Azevedo y Santos (2008), Schröer et al. (2021) y Martínez-Plumed et al. (2021), desde el año 2000 CRISP-DM es un "estándar de facto" para procesos de minería de datos. En la Tabla 8 se muestran los estados de CRISP-DM en comparación a KDD. Crisp-DM es un proceso iterativo en el cual se puede volver a estados anteriores para realizar correcciones:

1. Entendimiento del “Problema”. Junto con los expertos del dominio se evalúa si el problema a resolver es traducible a un modelo de y cómo hacer que produzca un mejor resultado en relación a la solución actual.
2. Entendimiento de los datos. En esta fase tenemos que realizar un análisis exploratorio de los datos para ver datos atípicos, errores y la calidad en general de los mismos, para tenerlos en cuenta en el proceso.
3. Preparación de los datos. Se corrigen los datos que sean posibles, se descartan datos atípicos cuando hay errores y se unifican diferentes fuentes de datos. Se hace una limpieza y transformación (mapeo) de datos para adaptarse a los algoritmos que se utilizarán en la siguiente etapa (i.e. *XGBoost* necesita datos numéricos).
4. Modelado. Se eligen los algoritmos a utilizar según si se trata de un problema de Aprendizaje Supervisado, No Supervisado, por refuerzos o combinaciones de los anteriores. Para detectar grupos se suele combinar algoritmos de clasificación (Supervisado) con agrupamiento (No Supervisado), por ejemplo. Se ajustan los hiper parámetros del modelo para obtener los mejores resultados. Como veremos en nuestro caso, no queremos el mejor accuracy en general, sino un modelo que balancee los falsos positivos con los falsos negativos.
5. Métricas. Una vez evaluado el modelo, podría ser necesaria repetir fases anteriores, ya sea porque se necesitan nuevas variables, para ajustar los hiper parámetros o porque surgieron errores en el preprocesamiento. En esta fase deberemos ser capaces de evaluar los modelos generados hasta el momento.
6. Mantenimiento y Despliegue. Una vez que el modelo queda listo es necesario instalar una versión utilizable por el usuario final del sistema. Además, los modelos envejecen y cada cierto tiempo requieren ser re-entrenados con inputs actualizados y vueltos a desplegar.

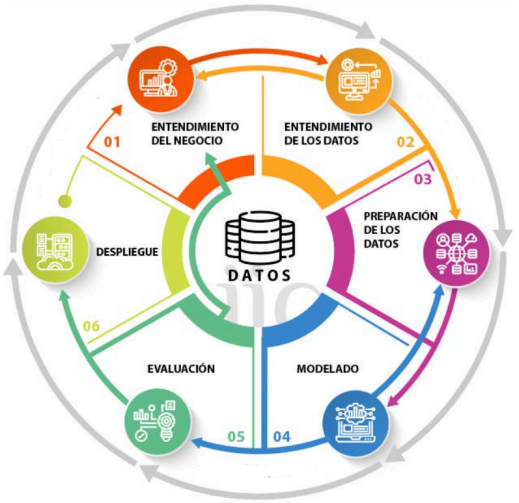
KDD	CRISP-DM	Diagrama de transición de estados CRISP-DM
Pre KDD	Entendimiento del Problema/Negocio	
Selección de los datos	Entendimiento de los datos	
Pre Procesamiento de los datos		
Transformación	Preparación de los datos	
Minería de Datos	Modelado	
Interpretación/ Evaluación	Métricas/Evaluación	
Post KDD	Mantenimiento/Despliegue	

Tabla 8: Comparación de estados de los estándares KDD y CRISP-DM y Diagrama de transición de estados del estándar CRISP-DM. Fuente: Adaptado a partir de: Azevedo y Santos (2008); Schröer et al. (2021); Martínez-Plumed et al. (2021).

En este trabajo se muestran los resultados de dos modelos (con ingeniería de atributos) aplicando la metodología CRISP-DM luego de varias iteraciones. En la Sección 6 se mostrará cómo la incorporación de nuevos atributos permite realizar la predicción de estudiantes en riesgo de abandono con mayor precisión.

4.2. Algoritmos de Inteligencia Artificial/Aprendizaje Automático utilizados

En el contexto de este trabajo de tesis, Inteligencia Artificial es cualquier algoritmo que imite el comportamiento humano. Como una rama dentro de la inteligencia artificial, el aprendizaje automático (ML de su versión en inglés *machine learning* surge en la década de los '80 como una propuesta donde la computadora establece sus propias reglas lo que le permite "aprender" por sí misma sin la necesidad de pautas externas o de un programador. De acuerdo a Mitchell (1997): "Un programa aprende de su experiencia E respecto a una tarea T y una métrica P si el desempeño en la tarea T medido por P mejora con la experiencia E". Existen algoritmos que aprenden reglas nuevas a partir de los datos (Aprendizaje Automático), por ejemplo, las redes neuronales aprenden nuevas reglas configurando el peso relativo de cada neurona. A diferencia del Aprendizaje Automático tradicional, el aprendizaje profundo se enfoca específicamente en el uso de redes neuronales profundas. En la Figura 25 se representan los tres niveles de Inteligencia Artificial. En nuestro trabajo utilizaremos variaciones del Aprendizaje Automático clásico, como *Extreme Gradient Boosting (XGBoost)*.

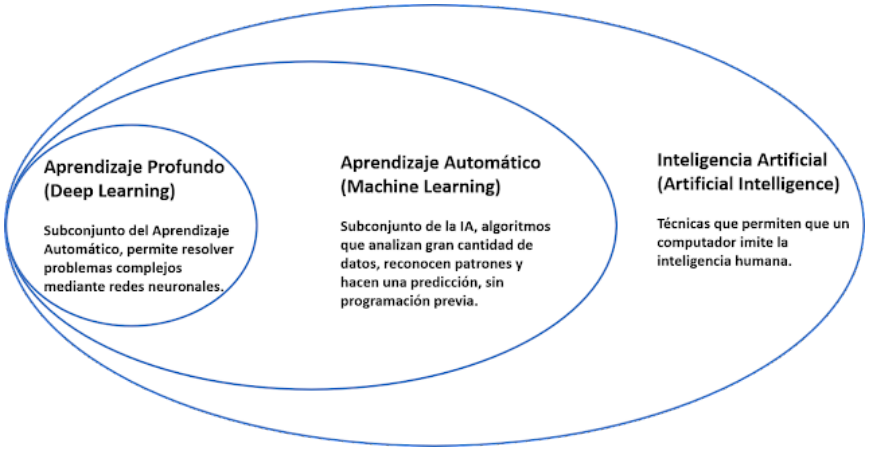


Figura 25: Clasificación de Algoritmos de Inteligencia Artificial. Fuente: Oyarzo (2020).

El objetivo de un proceso de Aprendizaje Automático es utilizar los datos conocidos para generalizar un conjunto de reglas que permitan dar respuesta a nuevos datos aplicando las reglas "aprendidas". Existen tres grupos de problemas que pueden ser tratados: Aprendizaje Supervisado, No Supervisado y Por Refuerzos.

En el Aprendizaje Supervisado, se conoce el resultado esperado para todos los datos de entrada conocidos y se aprenden nuevas reglas a partir de estos datos. A este proceso se lo denomina "Entrenamiento". Cuando surge una observación (dato) nueva se intenta inferir el resultado a partir de las reglas obtenidas durante el entrenamiento. La hipótesis principal es siempre que los datos nuevos se basan en las mismas reglas que en los datos observados previamente. Cuando esto no sucede, los modelos obtenidos a partir de estas reglas no tienen una buena performance. Dentro del Aprendizaje supervisado hay dos grupos (a) cuando la variable es discreta estamos ante un problema de clasificación (i.e. el alumno abandona o termina sus estudios) y (b) cuando la variable es continua se trata de un problema de regresión (i.e cúal es la nota esperada en el próximo examen). En este trabajo calculamos la probabilidad de abandono. Luego, vamos a discretizar esa variable para determinar si abandona a partir de cierto umbral (Si P(abandono) > umbral $\Rightarrow$ Abandona).

El Aprendizaje No Supervisado, particularmente en problemas de agrupamiento o clustering, consiste en agrupar datos a partir de relaciones entre los datos que se quieren testear (i.e. ¿deberían estar dentro del mismo grupo alumnos que compartan la misma edad, notas y nivel socioeconómico?). Una vez

generados estos grupos se contrasta con datos conocidos. En nuestro ejemplo podría ser "cuál es el porcentaje de abandono de cada grupo". El objetivo es buscar evidencia estadística que justifique estos grupos. El aprendizaje son los grupos o patrones que se obtienen y la inferencia consiste en ver si para un nuevo alumno, este se asocia al grupo esperado. También se utiliza para "aprender" grupos nuevos. Podría pasar que si el algoritmo agrupa alumnos por notas promedio estos se agrupen por carrera.

En el Aprendizaje Por Refuerzos el modelo interactúa con su entorno alimentando las reglas por prueba y error. Premiando positivamente a las reglas con los resultados esperados con el objetivo de refinar el modelo en cada "iteración". Este trabajo se focaliza en algoritmos de Aprendizaje Supervisado, dado que clasificaremos a los alumnos en dos grupos cuya existencia ya es conocida antes de modelar el problema.

4.2.1. Árboles de Decisión

Se llama Árboles de Decisión a una familia de algoritmos no paramétricos de Aprendizaje Supervisado. El primero de estos algoritmos (ID3: *Iterative Dichotomiser 3*) fue planteado por Quinlan (1986). Todos estos algoritmos representan las reglas (decisiones) mediante un árbol en donde en cada nodo se particiona el dominio de alguna variable, hasta llegar a las hojas. En cada nodo se acumulan las instancias y la predicción según las decisiones previamente tomadas. En la Figura 26 se muestra un árbol entrenado para clasificar el problema de "salir o no salir a hacer surf". Según ese árbol si hay oleaje (*swell*) pero no hay viento se recomienda salir a hacer surf, mientras que si no hay oleaje no se recomienda hacerlo. Decimos "recomendar" porque cualquier modelo tiene un margen de error, ya sea por las variables que no están incluidas en el modelo o por un entrenamiento deficiente que no logra generar "el mejor árbol posible" para el problema.

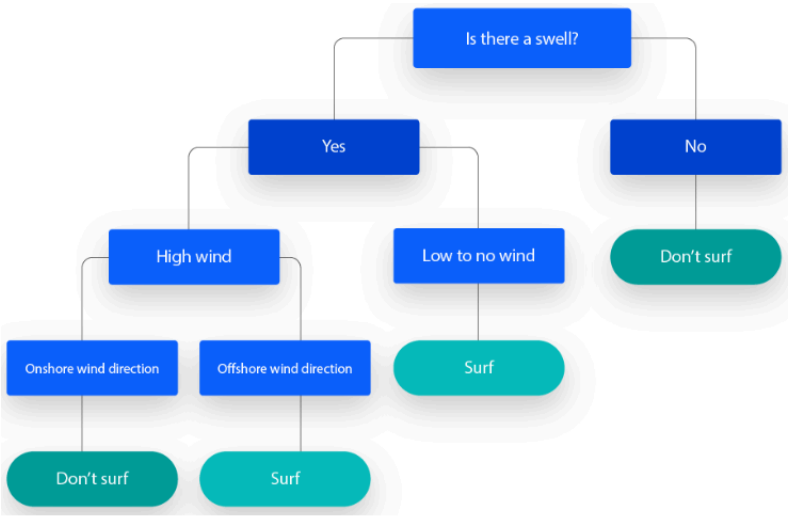


Figura 26: Árbol para el problema de salir a hacer surf. Fuente: IBM (2022).

El mismo atributo puede estar presente en más de un nodo, en principio para poder representar cualquier conjunción y disyunción pero como veremos más adelante, para refinar la discriminacion en las hojas del árbol. Como cada atributo se puede particionar en una cantidad exponencial de subconjuntos (hojas), generar todos los árboles posibles para quedarnos con el mejor no es un algoritmo factible en la práctica. Es por eso que se utilizan heurísticas como en algoritmo C4.5, una implementación de árboles de Quinlan (1993). Los árboles, además, son susceptibles a variables objetivo desbalanceadas (i.e hay más alumnos que abandonan que alumnos que continúan) lo que puede provocar árboles que no capturan la verdadera proporción de esos casos o que se sobreajuste al problema y no generen reglas que permitan predecir correctamente casos no entrenados explícitamente. Como las heurísticas suelen particionar los

nodos de manera probabilística, se introduce un problema de inestabilidad del resultado (i.e. dos ejecuciones con los mismos datos pueden derivar en árboles distintos). Para poder mitigar este problema se pondera el resultado a partir de la ejecución de varios árboles, lo que se conoce como ensamble de árboles de decisión. Las técnicas más comúnmente utilizadas para trabajar con ensambles son B*agging* y *Boosting*.

Bagging (Breiman, 1996) es un ensamble que aplica el mismo clasificador a varios subconjuntos (*Bootstrapping*) de los datos originales tomados aleatoriamente y con reemplazo, para luego agregar el resultado final, por ejemplo por una votación simple como se muestra en la Figura 27.

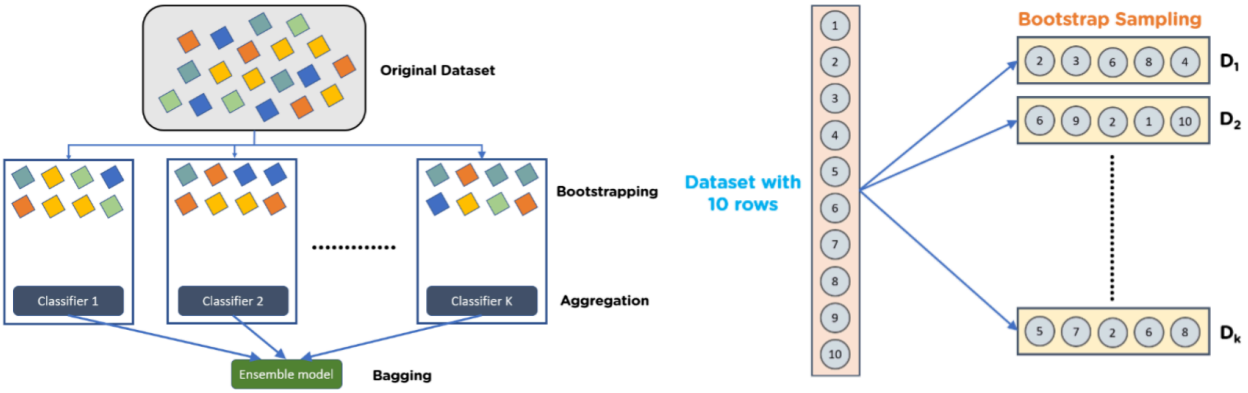


Figura 27: En *Bagging* se aplica el clasificador a K subconjuntos de los datos originales. Fuente: Biswal (2023).

Boosting, propuesto por Freund y Schapire (1996) es un algoritmo general que consiste en aplicar uno o varios clasificadores a la misma muestra como se muestra en el Algoritmo 1:

```
1. Boosting(Clasificador c, Muestra m)
2. Inicializar a todos los individuos de la muestra m con el mismo peso.
3. Entrenar el clasificador c con m, priorizando los individuos con mayor
   peso.
4. Aumentar el peso de los individuos mal clasificados y volver al paso 3.
5. Repetir hasta cierto umbral.
```

Algoritmo 1: Mejoramiento (boosting) para un clasificador y una muestra.

El objetivo principal es dar la oportunidad al clasificador de reclasificar individuos mal clasificados, y así aumentar la performance del clasificador. El algoritmo elige una submuestra aleatoria, priorizando los individuos con mayor peso (paso 3). La implementación varía según qué algoritmo se utilice, para árboles se suele utilizar *AdaBoost (Adaptive Boosting)*, que es similar al Algoritmo 1, mientras que para *XGBoost (Gradient Boosting)* el paso 3 se reemplaza por: `considerar un árbol que tenga menos errores que el árbol de la iteración anterior`. Podríamos decir que el nuevo árbol indirectamente "pesa" de manera diferente a los individuos del árbol previo para lograr un mejor resultado (i.e minimizar un error). Uno de los parámetros de *XGBoost* es la métrica a minimizar (i.e *Accuracy*, AUC, etc).

Los algoritmos heurísticos y probabilísticos que entrenan el árbol tienen varios criterios para elegir el mejor atributo para cada nodo. La mayoría se basan en el criterio de codificación de la información de Shannon (1948), quien define la entropía como la cantidad de información que contiene un conjunto de datos, basado en su distribución. La Entropía (E) representa la cantidad de bits que se necesitan en promedio para representar a un individuo de esa población. En el algoritmo que genera el árbol, queremos elegir cuál es el "mejor" atributo para cada nodo, en especial la raíz. Un criterio es quedarse con el atributo que tenga mayor Ganancia de información (*Information Gain, IG*), es decir, el atributo que más reduce la E, con el

objetivo de dejar las hojas lo más puras posibles. En la Figura 28 se muestra la progresión de la E y la IG. A medida que se profundiza en el árbol disminuye la E y aumenta la IG. En la Ecuación 5.1 se calcula la E para la variable Abandono $\in$ {abandono, noAb}. Para ello se suma la información relativa de cada elemento x_i del conjunto X, donde $1 \leq i \leq n$. La información relativa se calcula multiplicando cuánta información codifica x_i por su probabilidad. Para esta ecuación se asume una codificación binaria ($\log_2$), es decir, cuántos bits se necesitan para representar a x_i tal que no se confunda con otro x_j ($i \neq j$). En las Ecuaciones 5.2 y 5.3 se calcula la E para la variable Abandono (E(Abandono)). La IG para un atributo *a* (*IG(X, a)*) se calcula en las Ecuaciones 6.1, 6.2 y 6.3 a partir de la Entropía relativa de dicho atributo sobre la variable objetivo. *E(X,a)* representa la entropía relativa del atributo a en el conjunto X. *V(a)* es el conjunto de valores que *a* puede tomar y $|S(a)_v|$ representa la cantidad de elementos de *S* con valor *v* para el atributo *a*. Como se puede observar, cuanto más desbalanceada está la variable objetivo o un atributo, hay menos E y IG. En nuestro caso se necesita menos de un bit para codificar "abandono". Esto se debe al desbalance (el 61% abandona). Un método de parada por defecto (en lugar de seguir profundizando el árbol) es parar cuando el nodo no produce una mejor discriminacion.

$$E(X) = \sum_{i=1}^{n} p(x_i) log_2(1/p(x_i)) \quad (5.1)$$

$$E(Abandono) = p(abandono) log_2(1/p(abandono)) + p(noAb) log_2(1/p(noAb)) \quad (5.2)$$

$$E(Abandono) = 0,39\, log_2(2,56) + 0,61\, log_2(1,63) = 0,96 \quad (5.3)$$

$$E(X, a) = \sum_{v \in V(a)} \frac{|S(a)_v|}{|X|} E(S(a)_v) \quad (6.1)$$

$$S(a)_v = \{x_a \in X \mid x_a = v\} \quad (6.2)$$

$$IG(X, a) = E(X) - E(X, a) \quad (6.3)$$

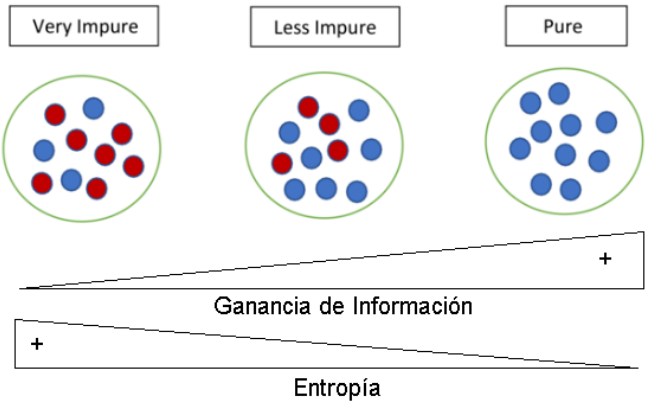


Figura 28: Relación entre Entropía y Ganancia de Información

Pero en la práctica este criterio trae aparejado algunos problemas. El más conocido es el sobreajuste. Eso ocurre cuando el árbol copia literalmente el conjunto X (i.e. cada hoja tiene un solo elemento). El problema con esta "exactitud" es que el árbol no puede generalizar otros datos fuera de X, porque es muy probable que no haya hojas que se ajusten a observaciones nuevas que se quieran predecir. Una de las técnicas utilizadas para evitar este problema se denomina regularización, y consiste en limitar de alguna manera la profundidad, la cantidad de hojas, la cantidad de instancias por hoja o cómo crece el árbol, para poder balancear el aprendizaje respecto de la especificidad. Una forma de regularización es penalizar la cantidad máxima de hojas, comparándola con el error relativo de clasificación para llegar a un balance entre

el error que se comete y el sobreajuste. Como ejemplo, podríamos utilizar la Ecuación 7.1. En ella, para cada elemento x_i se calcula el valor predicho (y_i) y se compara con el valor esperado (c_i) para computar un error. Se busca el árbol A entrenado con X que minimice f(A). alfa es el coeficiente de penalización/regularización.

Luego se penaliza, multiplicando por un coeficiente alfa a la cantidad de hojas y adicionando eso al error. El balance se da porque si se tienen muchas hojas, tendremos poco error pero mucha penalización. Pero si se tienen pocas hojas, tendremos mucho error, porque no podremos clasificar correctamente casi nunca. En la práctica se utilizan fórmulas más insesgadas para computar el error. Como la Ecuación 7.2, en donde se busca reducir la varianza del error.

$$f(A) = \left(\sum_{i=1}^{n} |y_i - c_i|\right) + \alpha\, cant(hojas) \qquad (7.1)$$

$$f(A) = \left(\frac{1}{n}\sum_{i=1}^{n} |y_i - c_i|^2\right) + \alpha\, cant(hojas) \qquad (7.2)$$

4.2.2. *Support Vector Machines (SVM)*

Las Máquinas de Soporte Vectorial (SVM, del inglés *Support Vector Machines*) son una técnica de inferencia estadística introducida en los años 90 por Vapnik y sus colaboradores (Vapnik, 2000). Originalmente fue pensada para resolver problemas binarios (como el problema del surf descrito anteriormente), pero más adelante se adaptó para clasificar variables más complejas e inclusive regresiones. Para clasificaciones binarias y lineales, SVM han probado tener un desempeño superior que los clasificadores tradicionales, como las redes neuronales (Chan et al. 2019). En el problema de la clasificación binaria, cada punto $x_i \in R^n$ del conjunto de entrenamiento X se asocia con una etiqueta de un conjunto binario y, por ejemplo, $y = \{A,B\}$ y x_i es vector de características. Una forma de asociar ambos conjuntos es encontrar un hiperplano h de la forma $wz+b=0$ donde w es un vector en R^n y $b \in R$ de manera que cuando se instancia un punto x_i en h, este se asocia a A cuando el valor es positivo y a B en caso contrario. El objetivo es que el conjunto de test obtenga una clasificación correcta cuando se pase a través de h. Cuando los puntos de entrenamiento no son linealmente separables es poco probable que el conjunto de test sea bien clasificado, dado que la hipótesis principal es siempre que tiene la misma distribución que el entrenamiento. En ese caso se utiliza una función que transforma los puntos x_i originales en otros puntos x_i` de $X$` con el objetivo que X` sea separable. La función que transforma los puntos $X_i$ en $X$` se llama kernel y se dice que "mapea los puntos" porque en muchos casos agrega variables en X` (aumenta la dimensión) como una estrategia para lograr la separabilidad. En la Figura 29 se muestra cómo X, que no es linealmente separable puede separarse a partir de $X$`. Primero se genera el conjunto X' con una nueva dimensión (i.e $x_{i3} = f(x_{i1}, x_{i2})$) de manera que h puede clasificar los puntos de X'. Luego se proyecta ese plano en X para generar la separación en el conjunto original. Existen infinitos tipos de kernels, nombraremos las familias más importantes para explicar cómo se genera la nueva dimensión:

-Cuando los datos son linealmente separables se recomienda utilizar un kernel lineal, en nuestro ejemplo:, x_{i3} se genera con $f(x_{i1}, x_{i2}) = x_{i1} * x_{i2}$.

-Cuando los datos no son linealmente separables pero siguen un patrón, se puede utilizar el kernel polinomial: x_{i3} se genera con $f(x_{i1}, x_{i2}) = (x_{i1} * x_{i2} + d)^c$, donde d y c son constantes.

-Existen funciones más complejas, como el kernel radial, donde $f(x_{i1}, x_{i2}) = exp(-\gamma\, ||x_{i1} * x_{i2}||^2)$, donde γ (gamma) define cuánto peso relativo tiene cada observación individual.

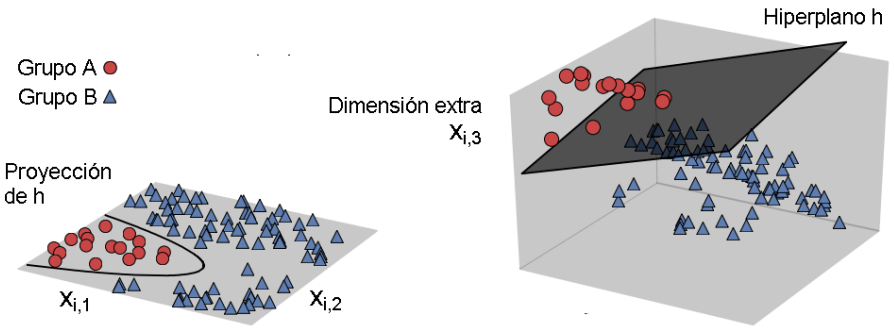


Figura 29: Conjunto X clasificado luego de utilizar una función f(x_{i1}, x_{i2}) de mapeo y proyectar el hiperplano h (derecha) en la dimensión original (izquierda). Fuente: Ntampaka et al. (2015).

En la práctica, no suele haber una separación perfecta y encontrar un kernel que minimice el error de clasificación no es trivial. Como se mostrará en la Sección 5.6.2 (Experimento 2: SVM), en esta tesis se utilizó una función de kernel Radial.

4.2.3. *Gradient Boosting*

A diferencia de *AdaBoost* los predictores por gradiente buscan los mínimos locales para una función que minimiza la suma del error de la predicción para cada instancia del entrenamiento. Estos predictores son métodos numéricos que recalculan la predicción en cada iteración basados en el error de la predicción anterior y eligiendo un gradiente en una dirección que achique el error de la iteración siguiente. Estos algoritmos se basan en clasificaciones débiles, típicamente árboles, y recalculan la predicción (que se encuentra en las hojas), del $árbol_i$ para cada $iteración_i$. La predicción final se calcula sumando de forma ponderada la predicción de cada árbol. La ponderación consiste en asignar un peso relativo a cada árbol, un valor entre 0 y 1, regulando además la velocidad de aprendizaje, para evitar el sobreajuste. En la práctica también se re-entrenan los árboles (nodos internos) con heurísticas que particionan los atributos de diferente manera. Chen y Guestrin (2016) sugieren algoritmos que particionan por cuartil (i.e. un nodo para cada cuartil), pero también sugieren tener en cuenta particiones diferentes en datos dispersos, o cuando hay datos faltantes, en donde los cuartiles no funcionan bien. Otra variante del algoritmo utiliza el submuestreo de filas y/o columnas para evitar el sobreajuste y para dar más posibilidades a registros y atributos que de otra manera no tendrían peso significativo. Por eso mencionamos en la sección anterior que *Gradient Boosting* "pesa" indirectamente los registros como en el paso 3 del Algoritmo 1. En el árbol de la Figura 29 se utilizan estas técnicas, generando así árboles distintos en sucesivas iteraciones. Cuando la variable objetivo es continua, todas las hojas del primer árbol se inicializan con el valor real de la instancia, menos el valor promedio de la variable objetivo. A ese valor se lo denomina **residuo** y lleva un índice que denota el número de árbol o iteración correspondiente siendo el primero el $residuo_1$. Se puede demostrar que tomar el promedio da el gradiente óptimo para la función del error. En la Figura 30 se describen 2 árboles (en 2 iteraciones del algoritmo) para un conjunto de entrenamientos de 5 individuos. También se observa la predicción para los individuos 1 y 3. Para el ejemplo, el peso asignado fue 1. La Ecuación 8.1 representa la predicción para el individuo i, donde $residuo_{ij}$ es la predicción del árbol j para el individuo i. La Ecuación 8.2 representa la predicción para el individuo 1. Se puede ver cómo se calcula la predicción a partir de la suma de predicciones parciales.

$$predicción(x_i) = residuo_{i,1} + \sum_{j=2\ldots n} aprendizaje * residuo_{i,j} \qquad (8.1)$$

$$predicción(x_1) = 2 + 1 * 0,9 = 2,9 \qquad (8.2)$$

Otro aspecto importante es "parar de entrenar" antes de que cada hoja clasifique una sola instancia. Hay regularizaciones para evitar el sobreajuste por este motivo. Para cada nuevo árbol el algoritmo recalcula el nuevo residuo para cada instancia, considerando el valor promedio de la hoja para todas las instancias que la comparten. En el ejemplo, las instancias 2, 3 y 4 obtendrán como $residuo_1$ = -1, que surge como promedio la tercera hoja del primer árbol.

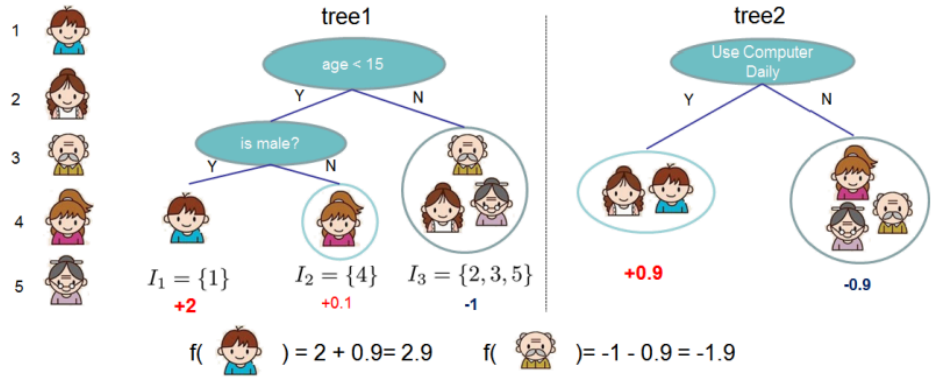


Figura 30: Fuente: Adaptado de Chen y Guestrin (2016).

Cuando la variable es binaria, se modifica la fórmula del residuo por otra ecuación que surge de combinar la probabilidad de la variable objetivo (0 o 1 para una variable binaria) con el **logaritmo natural del ratio de la variable objetivo**[14] ($log_e(odds) = ln(odds)$). Aunque en esta tesis se considera la probabilidad de abandono (variable continua), la consideraremos como binaria para el siguiente ejemplo. Como la probabilidad del abandono es 61%, *ln(odds)* se calcula como *ln(61/39)*, es decir, *0,44*. El primer residuo para todos los individuos se calcula como la diferencia de su valor observado respecto de *ln(odds)*. En nuestro caso, todas las hojas del primer árbol tendrían a (*1 - 0,44 = 0,54*) para "abandono" o bien (*0 - 0,44 = -0,44*) para "no abandono" como $residuo_1$. Dado que las hojas ahora predicen *ln(odds)*, debemos utilizar la Ecuación 9 para transformar el ln(odds) en una probabilidad. Es una transformación similar a la que que utiliza para regresiones logísticas. Por último necesitaremos un umbral, para calcular la probabilidad de abandono. Como una parte del residuo contiene una probabilidad pero las hojas predicen *ln(odds),* necesitaremos una transformación técnica adicional para calcular el siguiente residuo, que se muestra en la Ecuación 10, que define el $Residuo_{j+1,i}$. y *j* es el índice de residuo. La "*probabilidad anterior*" se define como la probabilidad de y_{i-1}. Pero como los individuos pueden "quedar" en diferentes hojas en cada árbol, se toma un promedio. Además, como *ln(odds)* es en definitiva una variable continua, el ensamble es exactamente igual que en el caso continuo. Ponderados de cada árbol por un coeficiente de aprendizaje para calcular el *ln(odds)* predicho para una nueva observación, y así calcular la probabilidad de una nueva observación a partir de la Ecuación 9. En las regresiones logísticas se utiliza una técnica similar.

$$Probabilidad = e^{ln(odds)} / (1 + e^{ln(odds)}) \quad (9)$$

$$residuo_{j+1,i} = \frac{\sum residuo_{j,i}}{\sum probabilidad\ anterior(y_i) * (1 - probabilidad\ anterior(y_i))} \quad (10)$$

Al igual que en las regresiones logísticas, hay que establecer un "umbral de abandono" para convertirla en binaria. Dicho umbral tendrá un sentido semántico (¿cuando un alumno está en peligro de

[14] El ratio de una variable objetivo binaria se define como #casos positivos / #casos negativos

abandono?) y hará variar la eficacia de la predicción (a mayor umbral, mayor la eficacia, pero con mayor sensibilidad), haciendo de este, otro hiper parámetro más. Esto se verá reflejado en la curva ROC (Sección Resultados).

4.2.3.1. *XGBoost (Extreme Gradient Boosting)*

Como explica Chen y Guestrin (2016), XGBoost es una forma más regularizada y eficiente de *Gradient Boosting*. Utiliza dos tipos de regularizaciones. *Lasso*(o L1), que añade una penalización basada en el valor absoluto de los coeficientes al error del modelo. Mientras que Ridge(o L2), añade una penalización basada en el cuadrado de los coeficientes al error del modelo (Hastie et al., 2015).

Está implementado para optimizar los recursos de cómputo, por ejemplo paralelizando y distribuyendo la creación de árboles, además de implementar una caché de árboles cuando no entran todos en memoria. Al igual que en Árboles de Decisión se devuelve la importancia de cada atributo como parte del output del modelo. Chen y Guestrin probaron *XGBost* para diferentes problemas, incluido el cálculo del ratio de abandono para un curso en línea. La implementación utilizada fue la de Colab® mediante la librería *XGBoost.*

4.3. Evaluación de algoritmos

La hipótesis principal es que la muestra elegida (i.e cohorte 2021 semestre 1) es representativa de toda la población (i.e cohorte 20*x* semestre *y*, donde $x \in \{21..99\}$ e $y \in \{1,2\}$). Sin embargo, como la población puede variar en el tiempo, ya sea por la introducción de cursos mixtos a partir del covid-19, por el cambio de tecnologías u otros factores, es recomendable re-entrenar el modelo todos los semestres para compensar el envejecimiento del modelo. Esto significa que para predecir la cohorte 20*x* semestre *y*, es recomendable entrenar con (a) cohorte 20*x* semestre *y-1* o bien con (b) cohorte 20(*x-1)* semestre y si se quieren tomar semestres similares. En este trabajo se entrenó con (b). Luego, es necesario evaluar el rendimiento de cada modelo para poder recomendar cúal se va a utilizar en la implementación.

4.3.1. Conjuntos de datos de entrenamiento y validación

El entrenamiento de los modelos depende de los algoritmos utilizados. Como concepto general, los modelos se entrenan con un subconjunto de los datos y se testean con el remanente con el propósito de calibrar el entrenamiento para prevenir sesgos, una variable objetivo desbalanceada y sobre ajuste. Para cumplir ese objetivo en Árboles de Decisión se utilizó un 70% de los datos elegidos de manera aleatoria como sugiere la literatura, como datos de entrenamiento. Pero esta técnica tiene inestabilidad estocástica, porque genera diferentes resultados en función de cómo elijan los conjuntos. Adicionalmente podría pasar que si elige un conjunto "poco favorable" (por ejemplo no representativo) de entrenamiento el modelo se entrene mal. Para mitigar estos efectos, se implementó una validación cruzada de estilo montecarlo (Xu y Liang, 2001), que consiste en repetir el experimento varias veces, eligiendo los conjuntos de manera aleatoria (de manera similar a mezclar un mazo de naipes). En este trabajo repetimos este experimento 10 veces[15] y nos quedamos con el promedio de cada métrica. No se utilizó *AdaBoost* ni otras técnicas como *RandomForest*, para poder utilizar ese modelo como control con el cual comparar otras técnicas de clasificación. En *XGBoost*, los árboles o clasificadores débiles los genera el propio algoritmo. Las iteraciones sí se realizan para el ajuste de los hiper parámetros. Para ello se utilizó una técnica denominada *Grid Search*. Es una optimización heurística que recorre valores concretos de los hiper parámetros hasta alcanzar la

[15] No se obtuvieron resultados significativamente más estables repitiendo el experimento más veces.

combinación que de la mejor clasificación posible. El alcance de esta técnica permite alcanzar un máximo local, sabiendo que puede haber combinaciones mejores que no fueron testeadas.

4.3.2. Métricas utilizadas

La exactitud (*accuracy*) y la precisión (*precision*) definidas en la Ecuación 11.1 y 11.1, donde VP = #Verdaderos positivos, FP = #Falsos positivos, VN = #Verdaderos negativos y FN = #Falsos negativos, son las métricas más simples y conocidas utilizadas. Pero son poco útiles para variables objetivo desbalanceadas. Para entender este fenómeno podemos tomar como clasificación de abandono un modelo trivial que prediga "abandono" con una probabilidad de 61% sin importar los datos de entrada. Este modelo tendría una exactitud de 61% pero no resulta de utilidad. La precisión responde a la pregunta ¿Qué proporción de casos positivos fue correctamente clasificado?. Y la exactitud representa la proporción de predicciones que el modelo clasificó correctamente.

$$Precisión = \frac{VP}{VP + FP} \tag{11.1}$$

$$Exactitud = \#\, predicciones\ correctas\ /\ \#\, predicciones = \frac{VP + VN}{VP + VN + FP + FN} \tag{11.2}$$

4.3.2.1. Matriz de confusión

Para las clasificaciones binarias se suele usar la matriz de confusión porque representa la proporción de clasificación para los casos positivos y negativos. Si la variable objetivo tiene *k* particiones se puede hacer matriz de *k* x *k* en donde la diagonal representa los casos bien clasificados. Por simplicidad representaremos una matriz de 2x2 asumiendo "abandono" como una variable binaria en la Tabla 9.

Positivo (P) = Abandono Negativo (N) = No abandona		Valores reales	
		Abandono	No
Valores predichos	Abandono	Verdadero Positivo (VP)	Falso Positivo (FP)
	No	Falso Negativo (FN)	Verdadero Negativo (VN)

Tabla 9: Matriz de confusión para la variable objetivo Abandono.

Con esta información se pueden construir además otros indicadores que se muestran en las Ecuaciones 12.1, 12.2, 12.3 y 12.4. La sensibilidad (*Recall*) mide la tasa de verdaderos positivos (*True Positive Rate*, TPR) y responde a la pregunta ¿qué proporción de verdaderos positivos se clasificó correctamente?. La especificidad (*specificity*) mide la tasa de verdaderos negativos (*True Negative Rate*, TNR) que mide la proporción de negativos clasificados correctamente. Ambos tienen la misma dificultad que la exactitud con las variables objetivo desbalanceadas. El puntaje F1 (*F1-Score*) da una puntuación de 0 a 1 y combina la sensibilidad con la precisión. La Exactitud Balanceada promedia la Sensibilidad (recall de la clase positiva) y la Especificidad (recall de la clase negativa):

$$Sensibilidad = \frac{VP}{VP + FN} \tag{12.1}$$

$$Especificidad = \frac{VN}{VN + FP} \tag{12.2}$$

$$Puntaje\ F1\ (F1 - Score) = \frac{2*precisión*sensibilidad}{precisión + sensibilidad} = \frac{2*VP}{2*VP + FP + FN} \tag{12.3}$$

$$Exactitud\ Balanceada\ =\ \frac{Sensibilidad + Especificidad}{2} \qquad (12.4)$$

4.3.2.2. Exactitud Balanceada Óptima

La Exactitud Balanceada se puede ver como una función que depende de un umbral. Para un umbral en particular se calcula la Sensibilidad y la Especificidad para poder obtener "un punto" de esta función. De esta manera se puede calcular el máximo de esta función , al que llamaremos "Exactitud Balanceada Óptima" y será una de las métricas para comparar modelos. Para una variable binaria balanceada se puede demostrar que esta función tiene los mínimos en los extremos (los puntos (0; 05) y (1; 0,5)) y que alcanza un máximo entre esos puntos. En el ejemplo utilizado se alcanza el máximo en el umbral 0,56 como se ve en la Figura 31.

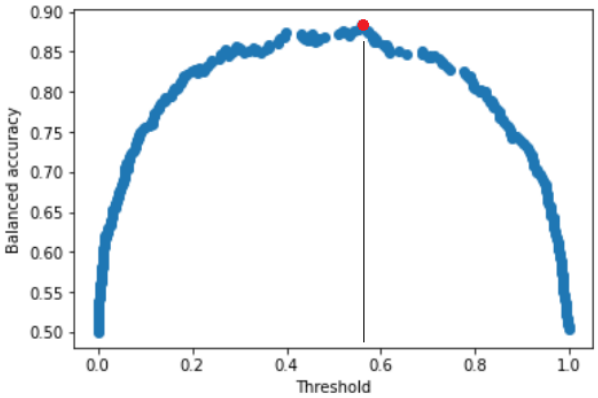


Figura 31: La función de Exactitud Balanceada utilizada en el ejemplo alcanza el máximo en el umbral 0,56 (umbral óptimo). Fuente: se reprodujo el experimento[16].

4.3.2.3. Curva ROC

Como "Abandono" es una variable continua y al mismo tiempo los indicadores enunciados anteriormente varían en función del algoritmo pero también el umbral de abandono elegido, necesitaríamos *n* matrices de confusión para visualizar la clasificación para cada uno de los *n* umbrales elegidos. En lugar de eso utilizaremos un indicador que muestra la sensibilidad y la precisión para cada umbral resumido en sólo gráfico. Por otro lado, al igual que hay más de un algoritmo (criterio) para calcular el abandono, también hay más de un criterio para elegir el umbral de abandono. A continuación se describen algunos de estos criterios para entender el problema:

(1) Si se tienen pocos alumnos y muchos tutores, becas y otros recursos institucionales es posible definir un umbral por debajo de 0,5 para abarcar a la mayor cantidad de alumnos que necesiten ayuda.
(2) Si en cambio se tienen pocos recursos y muchos alumnos, se requiere ser más estricto y poner el umbral por encima de 0,5.
(3) Por último, se puede priorizar una métrica que considere los falsos negativos, los falsos positivos pero que también sea robusta al desbalance. En ese caso, podríamos considerar la **Exactitud Balanceada Óptima** (EBO). La exactitud balanceada se calcula como el promedio de la sensibilidad con la especificidad, como se muestra en la Ecuación 12d. Mediante un algoritmo se obtiene el umbral que maximiza la EBO (Umbral óptimo). En la Sección 6 (Resultados) mostraremos la EBO para cada modelo.

[16] Código funente: https://github.com/gianlucamalato/machinelearning/blob/master/Threshold.ipynb

Como mencionamos en la Sección 5.5 (Salida del Modelo) la elección del umbral en sí no es trivial y requiere el acuerdo de varias partes. En esta tesis elegiremos el criterio (3) para poder optimizar la clasificación mediante un mecanismo que no cambie en función de cada cohorte ni del abandono subyacente de cada semestre. Eventualmente las autoridades pueden elegir otro criterio y reentrenar el problema sin pérdida de generalidad. Con el objetivo de visualizar la eficacia de cada modelo en función del umbral elegido utilizaremos la métrica Área Bajo la Curva ROC (AUC) como se explica en Hand y Till (2001). La curva ROC representa información para un conjunto de umbrales de tamaño *n* donde $0 \leq i \leq n$, que se quieran comparar. Para cada $umbral_i$, se agrega un punto p_i cuyas coordenadas son (x_i = *1* - $especificidad_i$, y_i = $sensibilidad_i$). En el eje *y* se representa la tasa de verdaderos positivos y en el eje *x*, 1 - la tasa de verdaderos negativos. Luego se "unen" los puntos para generar una curva (en realidad es una interpolación de grado 1). Con esa curva construida se puede calcular el área que encierra, que tiene un máximo de 1 para el clasificador perfecto. De esa manera se puede utilizar como métrica para comparar clasificadores y al mismo tiempo se pueden representar puntos estratégicos, como el umbral óptimo que se muestra en la Figura 32. La línea diagonal (línea de no-discriminacion) representa el área del peor clasificador posible (área = 0,5).

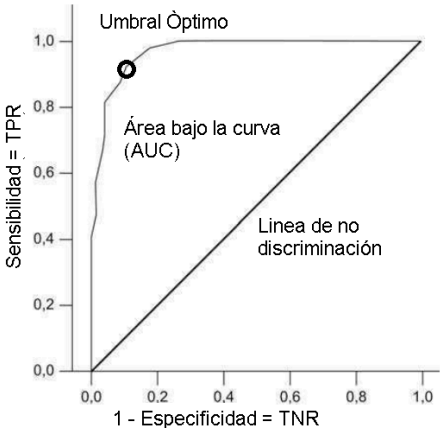


Figura 32: Esquema de una curva ROC.

XGBoost, permite optimizar los árboles para lograr este equilibrio y mostraremos en la Sección Resultados el punto de equilibrio en el gráfico de la curva ROC. Ese umbral será una recomendación para las autoridades de la universidad. Utilizaremos las métricas Puntaje F1 y Área Bajo la Curva, indicando el punto de equilibrio para cada modelo.

4.3.2.4. Correlación entre variables

Se utilizó el método *dataframe.corr(method='pearson')* para calcular la correlación de Pearson entre variables continuas. Todos los resultados se calcularon con el coeficiente de correlación de Pearson, ya que dieron resultados similares que con los coeficientes de Kendall o de Spearman. El coeficiente de Pearson devuelve un valor entre -1 y 1. Los valores cercanos a estos extremos indican una fuerte correlación lineal, mientras que los valores cercanos a 0 indican que no hay correlación lineal, pudiendo haber otro tipo de correlaciones (i.g. cuadrática, etc). En la Ecuación 13 se muestra cómo se calcula el coeficiente de Pearson para las variables X e Y. Una fuerte correlación aporta evidencia empírica de una posible relación entre variables y se utiliza para justificar las variables en un modelo. Adicionalmente tiene que existir más evidencia, típicamente en la bibliografía, para justificar una causalidad.

$$Pearson(X,Y) = \frac{Covarianza(X,Y)}{\sqrt{Varianza(X)Varianza(Y)}} \qquad (13)$$

4.3.2.4. Test Chi Cuadrado de independencia para variables categóricas

Se utilizó el método *scipy.stats.chi2_contingency()* que implementa el Test Chi Cuadrado, para comprobar la dependencia entre variables categóricas como se explica en Solanas et al. (2005). En particular nos interesa demostrar la dependencia de las variables respecto del abandono. El test tiene como hipótesis nula (H_0) que las variables son independientes. Para explicar el test, utilizaremos como ejemplo la población total de alumnos (31.300) y vamos a testear si el abandono es independiente al tipo de email que se utilice. Los datos de entrada consisten en la tabla de frecuencias de las variables a testear, en este caso el "Dominio de Email" respecto del abandono que figuran en la Tabla 10. Primero, calculamos la cantidad "esperada" de alumnos con email institucional que abandona, dada la proporción de alumnos con email institucional general, es decir: *20.278 x 3.005 / 33.383 = 1.825,34*. Luego se calcula el estadístico de prueba a partir de la Ecuación 14.1 agregando el valor para todas las celdas, dando un total de 1140,85. En la Ecuación 14.1 se calcula Estadístico de prueba para Test Chi cuadrado (χ^2). La Ecuación 14.2 calcula el caso particular de la "celda" (Institucional, Abandona). Los grados de libertad de la distribución χ^2 a utilizar se calculan con la fórmula *(número de filas - 1)(número de columnas - 1)*, para el ejemplo: *(2-1)(2-1) = 1*.

Por último se realiza el test con cierto nivel de significancia (alfa), en nuestro caso todos los test se realizaron con alfa = 0,05. El estadístico toma valor 0 cuando hay independencia (H_0) y se rechaza H_0 cuando el p-valor < 0,05 (hay un 95% o más de probabilidad de que las variables sean dependientes), cuando el estadístico supera el umbral establecido para la distribución χ^2 con los grados de libertad y significancia deseadas. La evidencia estadística para el ejemplo, indica que el "Dominio de email" **no** es independiente del abandono.

Email	**No Abandona**	**Abandona**	**Total**
Institucional	2.041	964	3.005
No Institucional	11.064	19.314	30.378
Total	13.105	20.278	**33.383**

Tabla 10: Tabla de frecuencias para "Email" respecto de "Abandono".

$$\frac{(observado - esperado)^2}{esperado} \qquad (14.1)$$

$$\frac{(964 - 1.825{,}34)^2}{1.825{,}34} = 406,44 \qquad (14.2)$$

4.3.2.5. Corrección por cantidad

El test χ^2 presenta hipersensibilidad a partir de n > 500, pudiendo diagnosticar dependencia para variables que podrían ser independientes (Cramér H., 1946). Para reforzar/refutar el grado de dependencia para las variables utilizadas, se agregara la medida de efecto Cramer's V al valor del estadístico χ^2 calculado. Cramer's V es un valor entre 0 y 1 que mide el grado de asociación entre variables, ajustando el estadístico chi cuadrado por el tamaño de la muestra como se muestra en la Ecuación 14.3: Efecto Cramer's V para un χ^2, tamaño de la muestra y dimensiones de la matriz de frecuencias relativas. Cero (0) representa que no hay asociación y uno (1) significa que las dos variables se comportan igual (son totalmente dependientes). Los valores intermedios son relativos a cada problema. En el caso del email (χ^2 = 1.140,85), la medida de Cramer's V calculada es de 0,18. Este valor está fuertemente ponderado por el tamaño de la muestra. Para el Email por ejemplo, solo tenemos el 9% de mails institucionales. Esto hace que dividir por n y aplicar raíz cuadrada "castigue" esta proporción, pudiendo ser más significativa para alumnos del del final de carrera, en

donde hay más alumnos con emails institucionales. En la Tabla 20 se resumen todas las medidas de Cramer's V para las variables e instancias importantes del modelo.

$$Cramer's\ V(x^2,\ n,\ filas,\ columnas)\ =\ \sqrt{\frac{x^2\ /\ n}{min(filas-1,\ columnas-1)}} \quad (14.3)$$

5. Modelado

5.1. Variables consideradas para modelar

Implementamos la metodología CRISP-DM, que consiste en sucesivas mejoras ajustando las variables en cada iteración, como se explicó en la Sección 4 (Métodos). En la Tabla 11 se muestran las variables (o grupos de variables) de SIU-Guaraní obtenidos en la primera iteración a partir de variables de la bibliografía o sugeridas por diferentes actores de UNAHUR, para implementar el primer modelo. Dentro de la base hay variables que surgen de la inscripción y otras que surgen de un censo opcional que la universidad actualiza cada cierto tiempo. Los directores de carrera nos sugirieron registrar hace cuanto tiempo se censó el alumno (*Meses Censo*) como una métrica para evaluar su interés por seguir cursando.

Variable/Grupo	**Tipo**	**Descripción**
Datos Personales	Demográfica (Discreta/ Categórica)	Edad, Género, Nacionalidad
Email	Sugerida (Categórica)	Solo el dominio
Dirección	Demográfica (Categórica)	Localidad, Barrio, Calle, Altura, Piso y Código postal
Variables Censales		
Meses Censo	Sugerida (Discreta)	Hace cuánto completo el censo (opcional)
Estado Civil	Demográfica (Categórica)	(Casado/a; Soltero/a; Viudo/a). Unido de hecho: (Si; No).
Familia	Demográfica (Discreta)	#Hijos. #Familiares. Con quién vive.
Tipo Vivienda	Demográfica/ Socioeconómica (Categórica)	(Casa; Edificio; Otro)
Dirección	Demográfica (Categórica/ Discreta)	Localidad, Barrio, Calle, Altura, Piso y Código postal
Situación Padre, Madre	Socioeconómica (Categórica)	(Vive; No)
Turno Preferido	Académica/ Socioeconómica (Categórica)	(Mañana; Tarde; Noche)
Salud	Socioeconómica (Categórica)	(Cobertura: Privada; Pública). Celíaco/a: (Si; No)

Tabla 11: Grupos y tipos de variables de la primera iteración, respecto de García (2014). Fuente: Elaboración propia.

5.2. Variables Calculadas (Ingeniería de Atributos)

Realizamos una segunda mejora sobre las variables del modelo, con el propósito de generar nuevas variables basadas en recomendaciones de la bibliografía y en observaciones puntuales de UNAHUR que nos permitieron mejorar la performance del modelo original mencionado en Pustilnik y Ndukanma (2023b). En dicho trabajo se había alcanzado un AUC de 0,83.

Cada alumno de UNAHUR puede percibir diferentes tipos de becas, muchas de ellas simultáneamente. Existen becas a nivel internacional, nacional, provincial, municipal así como becas otorgadas por la universidad: apuntes, comedor y fotocopias. En la variable "Becas" se calcula el total de becas percibidas al inicio de la cursada, y como mostraremos más adelante, tiene un impacto positivo en la retención estudiantil. Según autores como Amago (2008); Antoni et al. (2007); Carella (2009); Carella et al. (2007); Fazio (2004) y Patriarca (2012), la situación laboral está correlacionada con el abandono y es una variable que debería ser censada. Fazio (2004) observa que las horas trabajadas se asocian en forma no lineal con el rendimiento. En un tramo de baja cantidad de horas (menos de quince horas semanales), el rendimiento del alumno es positivo, por encima de esta cantidad incide negativamente. A su vez, si el trabajo

está vinculado con la carrera, es posible obtener beneficios del trabajo siempre y cuando la actividad económica se realice en un lapso inferior a 24,5 horas semanales.

Los directores de carrera nos sugirieron censar "Carrera" por tener un comportamiento particular para cada caso. Casi todos los alumnos de UNAHUR están inscriptos en más de una carrera. Existen muchos motivos para explicarlo, a continuación se enumeran algunos de ellos:
-Si hacen un curso de extensión antes del Curso de Preparación Universitaria (CPU). Esos cursos se censan como carreras en el sistema, siendo el CPU una "primera carrera" en el sistema.

-El CPU en sí se considera como carrera. Cuando ingresan a la carrera (i.e. Tecnicatura) ya se registran dos carreras.

-Cambios de planes de carrera generan que un alumno tenga materias aprobadas en "dos carreras". En el plan viejo y en el nuevo.

-Muchas licenciaturas incluyen una tecnicatura como título intermedio. Cuando el alumno se inscribe a una licenciatura ya se censan tres carreras (CPU + Tecnicatura + Licenciatura). Inclusive podría registrar cuatro "carreras" (Taller de extensión + CPU + Tecnicatura + Licenciatura).

-Cambios de carrera reales dentro de la universidad.

-Cursar dos carreras en simultáneo.

-Egresados. Actualmente se encuentran sub registrados en el sistema[17]. Luego de 4 semestres estos alumnos generan falsos positivos (i.e. $\forall\ semestre\ i,\ 0 < i < 4,\ cantidad\ de\ evaluaciones_i = 0$) porque no cursan materias.

Por todos estos casos, se considera la variable "carrera" como la última carrera con actividad de la persona. No obstante, las materias cursadas por semestre contabiliza todas materias de todas las carreras para la misma persona. Como veremos en la Sección Resultados, la carrera resultó ser una variable importante. La Tabla 12 muestra las variables calculadas, el tipo de variable respecto de García (2014) y su descripción:

Grupos de Variables calculada	**Tipo**	**Descripción**
#Eval2020S1	Académica (Discreta)	Cantidad de evaluaciones rendidas en 2020, Semestre 1
…		Se consideran 4 semestres consecutivos
#Eval2021S2	Académica (Discreta)	Cantidad de evaluaciones rendidas en 2021, Semestre 2
DistanciaViaje	Demográfica/ Socioeconómica (Continua)	Distancia (km) y tiempo (hs) de viaje al ingreso de la carrera.
DistanciaViajeCursada	Demográfica/ Socioeconómica (Continua)	Distancia y tiempo de viaje al momento del censo.
#Becas	Recursos (Discreta)	Cantidad de becas que percibe al inicio de la cursada.
#Horas Trabajo	Socioeconómica (Discreta)	Cantidad de horas semanas que trabaja al inicio de la cursada.
Carrera	Sugerida (Categórica)	Última carrera activa.

Tabla 12: Variables calculadas para la segunda iteración del modelo. El tiempo de viaje se estima a partir de la dirección informada y la librería de Google Maps®. Fuente: Elaboración Propia.

Como la distancia de viaje no está censada, agregaremos dos variables nuevas (distancia y tiempo de viaje en transporte público). Las variables se calcularon a partir de la dirección, preprocesado: el barrio, la altura, la localidad y el partido, para obtener una dirección válida que se pueda procesar en Google Maps® e inferir un tiempo de viaje promedio desde la dirección de la persona hacia la universidad como muestra el esquema de la Figura 33. Se descartaron valores atípicos o direcciones inválidas.

[17] Se realizaron entrevistas con autoridades de UNAHUR en 2023 en donde surge una diferencia entre los egresados reportados en el sistema nacional (Araucano: https://datos.produccion.gob.ar/dataset/graduados-universitarios-del-sistema-araucano-2016-2018) y los registrados en SIU-Guaraní.

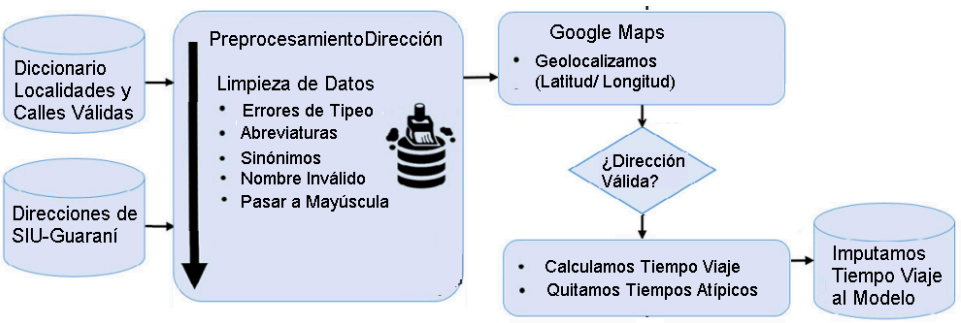


Figura 33: Esquema que muestra el procesamiento del tiempo de viaje a partir de la dirección de la persona. Fuente: Elaboración propia.

5.3. Variables de la bibliografía utilizadas

En esta tesis se utilizaron las siguientes variables: historia académica (cantidad de materias cursada por semestre) y variables demográficas, como género, edad, lugar de residencia, estado civil, cantidad de hijos, personas en el hogar. A partir de los modelos entrenados con estas variables, se testearon hipótesis sobre diversos factores como número de piso de la vivienda, tiempo desde el último censo institucional, cantidad de evaluaciones rendidas en semestres anteriores y tiempo de viaje desde el lugar de la vivienda a la universidad. Si bien las variables "Cantidad de horas de trabajo", "Becas" y "Cantidad de materias cursadas" resultaron relevantes. Los alumnos agrupados por género y edad también presentaron diferencias significativas en cuanto al abandono. En los modelos elaborados para este trabajo predomina el análisis de los factores individuales y de rendimiento académico.

En la Tabla 13 se observan variables influyentes en el abandono propuestas en García (2014) y por diferentes actores de la UNAHUR. También se analiza si pueden ser utilizadas en los modelos. Las variables están agrupadas en factores individuales y organizacionales. Para cada una se incluyen subcategorías y se analiza el grado de utilización (A), indicando distintos tipos de disponibilidad o utilización de los datos. Los datos no utilizados (NU) se refieren principalmente a datos que se reportan y podrían estar disponibles. Los datos utilizados (**DU**) son los ya incluidos en el modelo, mientras que los datos no disponibles (ND) serían datos no relevados o con dificultades para ser integrados.

Factores Individuales					
A	**Demográficas**	**A**	**Socioeconómicas**	**A**	**Académicas**
DU	Sexo	ND	Ingresos del hogar	ND	Promedio escuela secundaria
DU	Edad	NU	Nivel educativo de los padres	NU	Gestión pública-privada esc. secundaria
DU	Nacionalidad	NU	Nivel ocupacional padres	NU	Título de la escuela media
DU	Estado civil	NU	Actividad económica	ND	Horas y esfuerzo dedicados al estudio
DU	Residencia	DU	Cantidad de horas de trabajo	NU	Aspiraciones y motivaciones al ingreso
DU	Cantidad de hijos	NU	Fuente financiamiento de estudios	NU	Rendimiento académico primer año
Factores Organizacionales					
A	**Política académicas**	**A**	**Plan de estudio**	**A**	**Recursos**
NU	Mecanismo de admisión	NU	Duración del programa	ND	Formación y habilidad de los docentes
ND	Orientación vocacional	NU	Flexibilidad de cursado	NU	Relación docente-alumno
ND	Comunicación institucional	NU	Amplitud de oferta horaria	ND	Servicios de bienestar estudiantil
NU	Condición alumno regular	NU	Cantidad de horas de cursado	DU	Becas
ND	Prácticas de enseñanza	NU	Mecanismo de evaluación	ND	Infraestructura y equipamiento
DU	Seguimiento alumno	ND	Estrategias innovadoras 1er año	ND	Gasto por alumno
ND	Tutorías	ND	Dificultad materias primer año	ND	Cultura organizacional

Tabla 13: Principales variables individuales, organizacionales y avance (A) en la utilización de los mismos en el modelo predictivo: datos utilizados (**DU**), datos no utilizados (NU), datos no disponibles actualmente (ND). Fuente: elaboración propia a partir del cuadro propuesto en García (2014).

Los ND son factores variados y hacen referencias a cuestiones difícilmente cuantificables o de corte cualitativo, como la cultura organizacional, el gasto por alumno, las tutorías, las prácticas de enseñanza, las estrategias innovadoras y las dificultades en el 1er año, infraestructura y equipamiento, formación de los docentes, servicios de bienestar estudiantil y orientación vocacional. Los datos del seguimiento del estudiante son considerados DU, se incluyen parcialmente en los modelos, considerando la cantidad de materias cursadas. Los datos NU se refieren principalmente a datos que los estudiantes reportan al momento de la inscripción y que no necesariamente están actualizados como por ejemplo la actividad económica del estudiante. La información sobre actividades extracurriculares, pasantías, tutorías, participación del programa "1 estudiante - 1 compañero", ayudantías u otros programas, permitirían identificar la integración de los estudiantes a la comunidad universitaria. En cuanto al rendimiento académico es relevante diferenciar, por un lado a los estudiantes de primer año y por otro lado, a los recursantes (reinscriptos), tanto como las diferencias de los porcentuales de aprobación según instituto de pertenencia de los estudiantes y por carrera. Dentro de las carreras los coordinadores realizan hipótesis sobre materias específicas que influyen en la retención, que sería interesante analizar.

5.4. Otras variables mencionadas en la bibliografía

Existen variables importantes que no se pudieron censar. La integración con la universidad, las aspiraciones y motivaciones individuales son analizadas por diferentes trabajos. Según Odetti et al. 2010, la cantidad de horas de cursada y la falta de tiempo para estudiar afectan el abandono. Por su parte, en los estudios acotados a las asignaturas, en uno de ellos (Oliver et al. 2011) se expone que una relación alumno-docente menor influye positivamente en la aprobación de la materia. En el otro (Chudnovsky, 2003) se señala que, entre los que abandonan las materias estudiadas predominan aquellos que no tienen claro conocimiento del plan de estudio. La educación de los padres es otro factor a tener en cuenta. Giovagnoli (2002) observa que un estudiante cuyo padre tiene primario incompleto tiene un 70% menos posibilidades de

graduarse que aquel estudiante con padre profesional. Tampoco se censaron los programas como “1 alumno 1 compañero” por no estar registrados en el SIU.

5.5. Salida de los modelos

La alarma de “posible abandono” para una persona se enciende si el modelo predice que no va a cursar ninguna materia (de ninguna carrera) el siguiente semestre. Queda fuera del alcance de los modelos distinguir el “posible abandono” de un cambio de universidad por no poseer esa información. Una vez entrenado el modelo se le suministra la base de datos de personas de 2021 para generar una predicción para 2022. En el caso de modelos basados en árboles, la probabilidad surge de la proporción de personas que abandonan agrupadas en cada hoja. Los usuarios de los modelos (por ejemplo XGBoost) tienen que elegir un umbral (U) a partir del cual un alumno se considera en riesgo de abandono. Como la universidad tiene una cantidad acotada de recursos (i.e. becas, tutores) para asignar a los alumnos en riesgo, tener un listado acotado de alumnos en riesgo es fundamental para poder asignarlos. Si el umbral fuese muy bajo (e.g. $U < 0,1$) el modelo sería muy sensible para considerar a un alumno en riesgo. Tendríamos una lista muy extensa, y no se podrían asignar los recursos eficientemente. Por otro lado, si el umbral fuese muy alto (e.g. $U > 0,9$) podría pasar que alumnos que en realidad están en riesgo no reciban los recursos adecuados. La predicción consiste en el listado de personas y su probabilidad de abandono para 2022. *Esa información se cruza con las variables destacadas y se agrega al listado la variable más importante para cada persona*.

La Tabla 14 se confeccionó con el umbral de 0,6. Como el umbral es un parámetro que los usuario pueden modificar para agrandar o reducir la lista, se elige de acuerdo a la definición de “estudiante en riesgo de abandono”, que involucra un trabajo en conjunto entre especialistas en educación y de gestión académica. En la tabla se asocia la probabilidad de abandono inferida por el modelo XGBoost asociado a la variable más importante, que esté relacionada con la persona. Esto podría dar pistas del “motivo de abandono”. Es importante distinguir que los motivos de abandono que predice el modelo están basados en los datos disponibles, y en general existen otros motivos subyacentes a analizar para cada caso. Por ejemplo, un tiempo de viaje elevado y “no abandono” podría estar midiendo la variable latente “La persona tiene auto”. Y la variable “No rinde hace varios semestres” podría estar reflejando que la persona comenzó a trabajar. El listado se anonimizó para proteger los datos personales.

Persona	Probabilidad	Variable más importante
724234	0,91	No rinde hace varios semestres
008383	0,90	No rinde hace varios semestres
…		
922524	0,62	No completo censo
629341	0,61	No completo censo

Tabla 14: Resultados de ejemplo del modelo XGBoost. Listado de personas con probabilidad de abandono $\geq 0, 6$ ($U = 0, 6$). A la salida del modelo se asocia la variable más importante para cada individuo.

5.6. Experimentos

Se implementaron 3 modelos basados en tres algoritmos diferentes. Se realizó un experimento para cada modelo y se compararon los resultados.

5.6.1. Experimento 1: Árboles de decisión

Se utilizó la implementación “DecisionTreeClassifier” de la librería *sklearn* para entrenar el modelo con Árboles. Se optimizaron los hiper parámetros mediante una búsqueda aleatoria. Se puede obtener el código final de Repositorio Git (2023c). A continuación se describen los parámetros del modelo de Árboles:

“entropy” significa que se usa el criterio de ganancia de información para particionar los nodos, “max_depth = 10” es la profundidad máxima del árbol, “min_samples_split = 5” es la mínima cantidad de instancias en cada nodo, y “min_samples_leaf = 3” es la mínima cantidad de instancias por hoja. Se ejecuta 10 veces el modelo con esa configuración, obteniendo resultados similares en todas las ejecuciones.

5.6.2. Experimento 2: SVM

Se utilizó la librería sklearn.svm para entrenar el modelo con Máquinas de soporte vectorial. Se utilizaron diferentes kernels como el lineal, el polinomial y el radial. Los mejores resultados se obtuvieron con el radial y los parámetros: {C = 1; gamma = 0,1 y sample_weight (con mayor peso para “no abandono”)}. SVM en principio fue pensado como clasificador binario. Calcula la probabilidad de una variable objetivo mediante una validación cruzada de 5 carpetas en la implementación de sklearn. Para ello se debe activar el parámetro “probability”. El parámetro “C” es la penalización para clasificaciones incorrectas (regularización) y comienza en 1 como valor por defecto. Ambos hiper parámetros se optimizaron mediante Grid Search (sklearn.model_selection.GridSearchCV), utilizando los conjuntos {1; 5; 10} y {0,1; 0,5; 1} para C y para gamma respectivamente. Por último, se duplicó el peso inicial del grupo “no abandono” para que (el parámetro “sample_weight” vale 1 para “abandono” y 2 para “no abandono”) por el desbalance.

A diferencia de los modelos de Árboles, las variables en SVN deben ser obligatoriamente numéricas. Cuando las variables categóricas se “convierten” a numéricas (e.g. Tipo de vivienda “Casa/Departamento propio” → 1, “Casa/Departamento alquilado” → 2, “Pensión” → 3, etc) fuerzan un orden artificial (e.g. “Casa propia” < “Casa alquilada” < “Pensión”), que pueden llevar al modelo a cometer clasificaciones incorrectas. Para evitar ese problema, se convirtieron ese tipo de variables (categóricas sin orden), a variables binarias (numéricas). Por ejemplo, “Tipo de vivienda” se transformó en 5 variables binarias (“Casa alquilada: 0/1”, “Casa propia: 0/1”; “Pensión: 0/1”, etc).

5.6.3. Experimento 3: XGBoost

Para XGBoost se utilizó la librería sklearn.xgboost para el entrenamiento del modelo en sí y sklearn.RandomizedSearchCV para la optimización de hiper parámetros. Nos quedamos con los mejores parámetros luego de 10 ejecuciones para el modelo final. Se optimizan los hiper parámetros. Se eligió AUC como métrica de evaluación con el mismo criterio de priorizar la exactitud balanceada. En la Tabla 15 se describe el propósito de cada uno.

Parámetro	**Descripción**	**Valor o rango (dominio)**	**Valor**
objective	Tipo de predicción	binary: (binario), binary:logistic (Para predecir una probabilidad)	binary:logistic
eval-metric(scoring)	Métrica para evaluar el error	Error cuadrático medio, AUC, etc	AUC
booster	clasificador débil utilizado	lineal, tree	tree
max_depth	máxima profundidad del árbol	1…11 (1…∞)	9
learning_rate(eta)	Velocidad de aprendizaje	0,1…1	0,2
subsample	ratio de la cantidad de instancias seleccionadas al azar para cada árbol	0…1	0,6
colsample_bytree	ratio de columnas elegidas para cada árbol	0…1	0,7
colsample_bylevel	ratio de columnas elegidas para cada nivel del árbol	0…1	0,6
gamma	mínima pérdida requerida para hacer un split	0…1 (0…∞)	0,5
reg_lambda	Regularizacion tipo L2 (análogo Ridge)	1,3,5,9 (0…∞)	5
n_estimators(num_boosting_rounds)	Cantidad de iteraciones (árboles)	200…500 (1…∞)	256
scale_pos_weight	aumenta el peso de la variable objetivo desbalanceada en una proporción	Se calcula como #positivos / #negativos	2

Tabla 15: Parámetros utilizados de XGBoost.

6. Resultados

6.1. Resultados de los modelos

En esta sección se muestran los resultados de los tres modelos (Árboles , SVM y XGBoost). En el modelo de Árboles se trabajó con Árboles de profundidad 9 (más allá de ese valor se produce sobreajuste). A efectos ilustrativos se eligió un árbol al azar de profundidad 3 el cual se muestra en la Figura 34. Como nodo raíz el modelo eligió “Hace cuantos meses se completó el censo” como variable más importante, basada en la ganancia de información que proporciona. En cada nodo se ve la variable elegida a particionar, la cantidad de instancias y rango que se particiona y la ganancia de información relativa de esa variable. Como veremos en la Sección 6.3. (Variables más importantes en el modelo XGBoost), las variables “Carrera”, “Distancia/Tiempo de viaje”, “Cantidad de evaluaciones hace un semestres” también resultaron importantes para todos los modelos. El “Código Postal” (censo_cursada_cp) aparece en algunos árboles, indicando cierto grado de discriminacion en la variable objetivo.

En las Tablas 16, 17 y 18 se muestra la Matriz de confusión para la EBO, la Curva ROC y el resumen de medidas. En la Figura 35 se muestra la Exactitud balanceada en función del umbral para el modelo basado en SVM. Se observa que la Exactitud balanceada óptima se alcanza en umbral 0,6. Notar que la Exactitud balanceada para los umbrales mínimo y máximo (0 y 1) es 0,5.

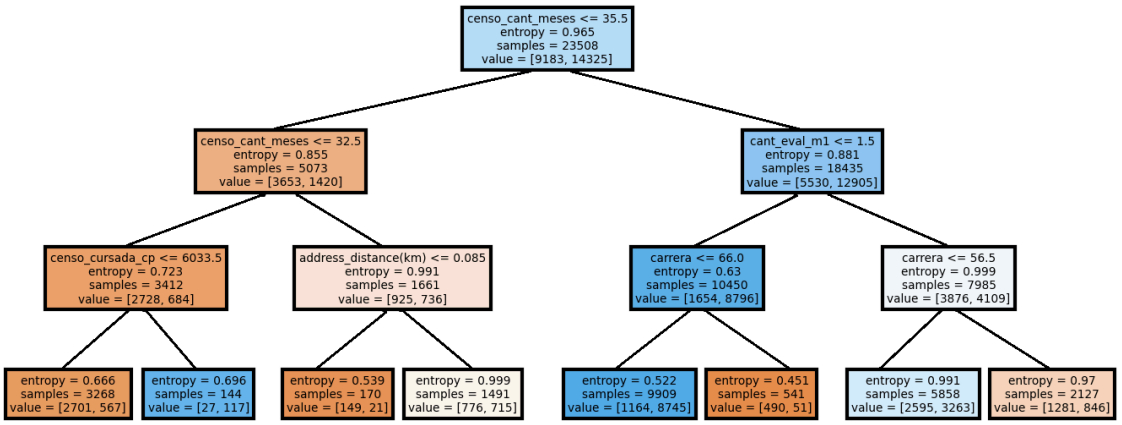


Figura 34: Árbol de tres niveles para el primer modelo. En los recuadros (nodos): Nombre variable, Partición, Cantidad de instancias y Entropía Relativa asociada.

Matriz de confusión		
P = Abandono	Abandono Real	No Abandono
Abandono Predicho	VP = 3532	FP = 430
No Ab.	FN = 2088	VN = 4096
Resumen		
AUC = 0,84		
Optimal Threshold = 0,7		
Optimal Balanced Accuracy = 0,77		

Curva ROC

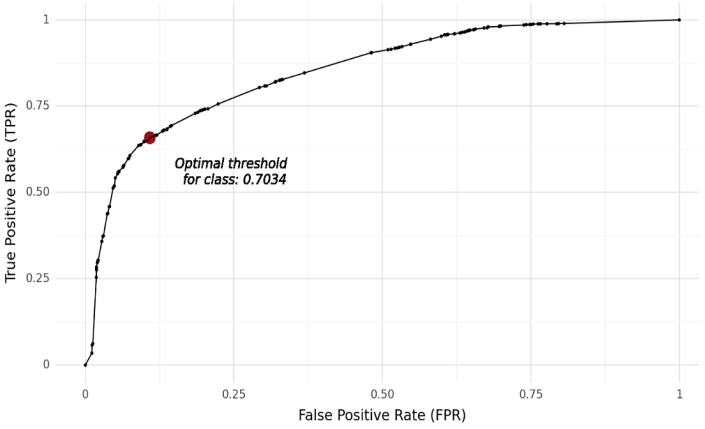


Tabla 16: Matriz de confusión, Curva ROC y resumen de medidas para el modelo de Árboles. Parametros: criterion='entropy'; max_depth = 9; min_samples_split = 5; min_samples_leaf = 3.

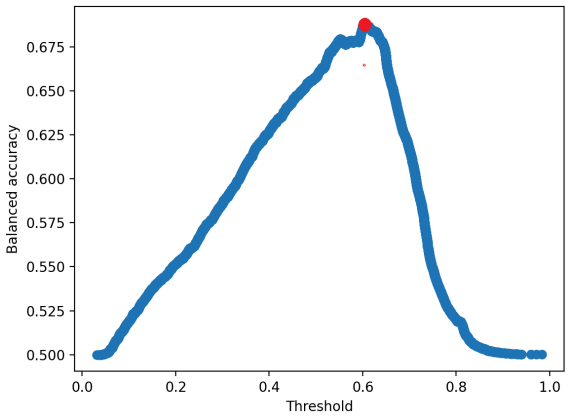


Figura 35: Exactitud Balanceada en función del umbral para el modelo SVM. La EBO (punto rojo) se alcanza para el umbral 0,6

Matriz de confusión		
P = Abandono	Abandono Real	No Abandono
Abandono Predicho	VP = 2805	FP = 466
No Ab.	FN = 2371	VN = 2754
Resumen		
AUC = 0,74		
Optimal Threshold = 0,6		
Optimal Balanced Accuracy = 0,69		

Curva ROC

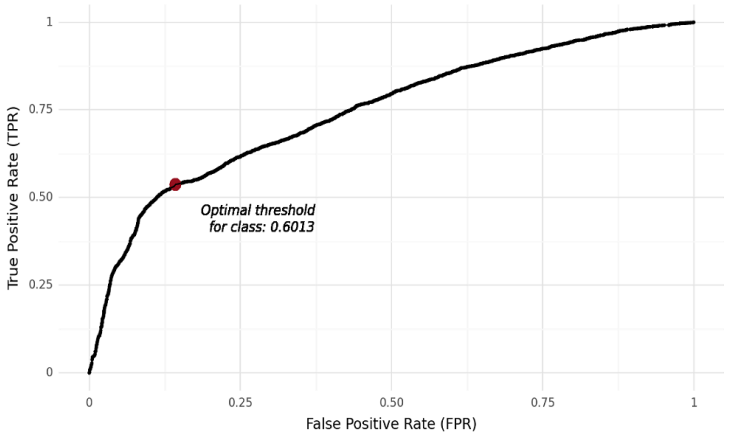


Tabla 17: Matriz de confusión, Curva ROC y estadísticas para el modelo de SVM. Se alcanza la Exactitud balanceada óptima para el umbral óptimo 0,6.

Matriz de confusión		
P = Abandono	Abandono Real	No Abandono
Abandono Predicho	VP = 2957	FP = 380
No Ab.	FN = 1462	VN = 3597
Resumen		
AUC = 0,88		
Optimal Threshold = 0,84		
Optimal Balanced Accuracy = 0,79		

Curva ROC

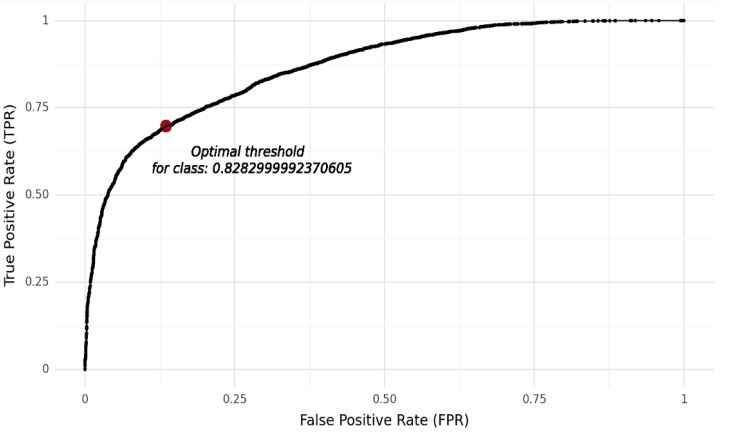


Tabla 18: Matriz de confusión, Curva ROC y estadísticas para el modelo de XGBoost. Se alcanza la Exactitud balanceada óptima para el umbral 0,84.

6.2. Comparación entre modelos

En esta sección se muestran los resultados más importantes de los tres modelos y de los datos en general. En la Tabla 19 se puede apreciar que *XGBoost* fue el modelo con mejor performance con un AUC de 0,88. No obstante esto sólo sucede en el umbral 0,84. Árboles tiene la mejor Especificidad (0,9), basado principalmente en el desbalance de la variable objetivo. Con SVM no se obtuvo una buena performance. Como explicamos en la Sección 4, esta técnica es muy dependiente del kernel y parámetros seleccionados, pudiéndose intentar otras configuraciones que optimicen la performance.

Modelo	Sensibilidad	Especificidad	Puntaje F1	Umbral óptimo (UO)	Exactitud balanceada para (UO)	AUC
Árboles	0,62	**0,9**	0,73	0,7	0,77	0,84
SVM	0,54	0,85	0,66	0,6	0,69	0,74
XGBoost	0,67	0,89	**0,76**	**0,84**	**0,79**	**0,88**

Tabla 19: Resumen de métricas para cada modelo.

6.3. Variables más importantes en el modelo *XGBoost*

Se utilizó la métrica *Gain* provista por *XGBoost* para calcular la importancia de las variables para los modelos basados en árboles. *Gain* se define como la ganancia promedio (average *Information Gain*) obtenida al particionar un atributo. En la Figura 36 se muestran las variables con *Gain* > 4,5. Encontramos que el modelo funciona mejor agregando las variables "Cantidad de meses censo", "Cantidad de evaluaciones hace 1, 2 y 3 semestres", "Carrera", "Género", "Tiempo/distancia de viaje" que sin ellas. Como las variables "Tiempo/distancia de viaje" son estimadas a partir de los datos, esperamos obtener mejores resultados conociendo el modo de traslado del alumno. "Tipo de vivienda", "Número de piso", "Situación de padre/madre" no resultaron variables significativas en este modelo.

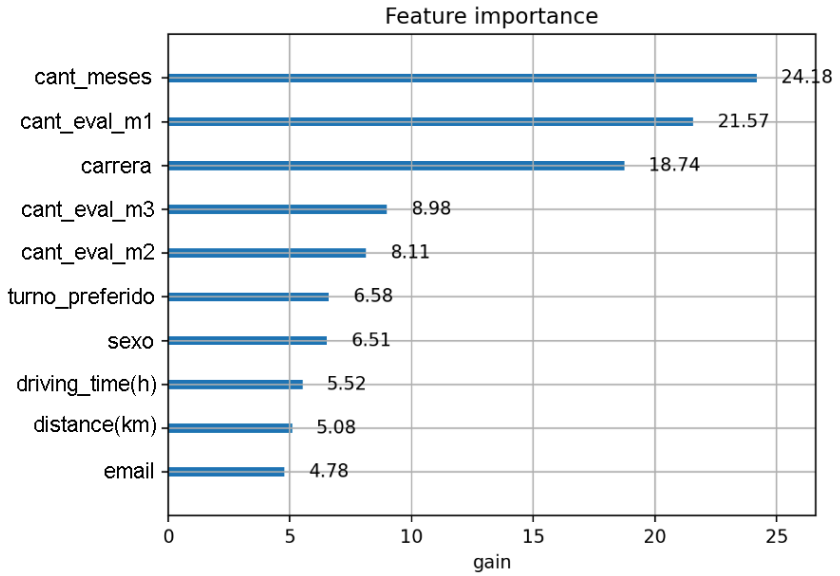


Figura 36. Las 11 variables más importantes, métrica Ganancia de Información (IG, *Information Gain*).

Para relacionar la importancia de las variables con su influencia en la predicción utilizamos la métrica propuesta por Lundberg y Lee (2017). Los valores SHAP (*SHapley Additive exPlanations*) son una forma de explicar el resultado de cualquier modelo de aprendizaje automático. Utiliza un enfoque de la teoría de juegos, que mide la contribución de cada jugador (variable) en el resultado final. En el aprendizaje automático, a cada variable se le asigna un valor de importancia que representa su contribución al resultado

del modelo. En la Figura 37 se muestran los atributos por orden descendente de importancia en el eje *y*. En el eje *x*, cada atributo se representa con un gráfico de violín apaisado. Su longitud representa el impacto total de cada atributo en la predicción, pudiendo impactar más en el abandono (derecha, Shap value > 0) o en el "no abandono" (izquierda, Shap value < 0). Los valores "altos" de cada atributo se presentan en magenta, mientras que los "bajos", en cian. Para variables como "Género" se debe asignar un valor para "Masculino" y otro para "Femenino", pudiendo ser cualquiera de los dos el valor más alto. Por ese motivo, para las variables categóricas estos valores son arbitrarios, teniendo más significado para las variables numéricas. El ancho del violín representa la frecuencia de valores, siendo la parte más ancha la de mayor frecuencia. Además de la longitud del violín, hay que considerar si el punto de mayor frecuencia favorece o no al abandono.

En el gráfico se puede observar cómo los valores altos de "Cantidad de evaluaciones 1, 2 y 3" se correlacionan con "no abandono". Los valores altos de "Cantidad de meses censo" también se correlacionan con "abandono", sugiriendo un menor grado de integración con la institución. Para "Tiempo/distancia de viaje" (*address_distance(km)*, *driving_time(h)*) también se ve la correlación de distancia con "Abandono", sugiriendo que a mayor distancia es más difícil cursar. Para la variable "Género", el gráfico sugiere que hay mayor abandono para sexo masculino (magenta). Pero para muchas otras variables el resultado está dividido. Por ejemplo, no hay resultados concluyentes acerca del turno mañana o noche, ya que como vimos en la Sección 3.5.1 (Abandono condicional), hay ejemplos en ambos turnos con altas tasas de abandono. En esa sección se realizó un gráfico particularizado para algunas de las variables que mejor discriminan el abandono.

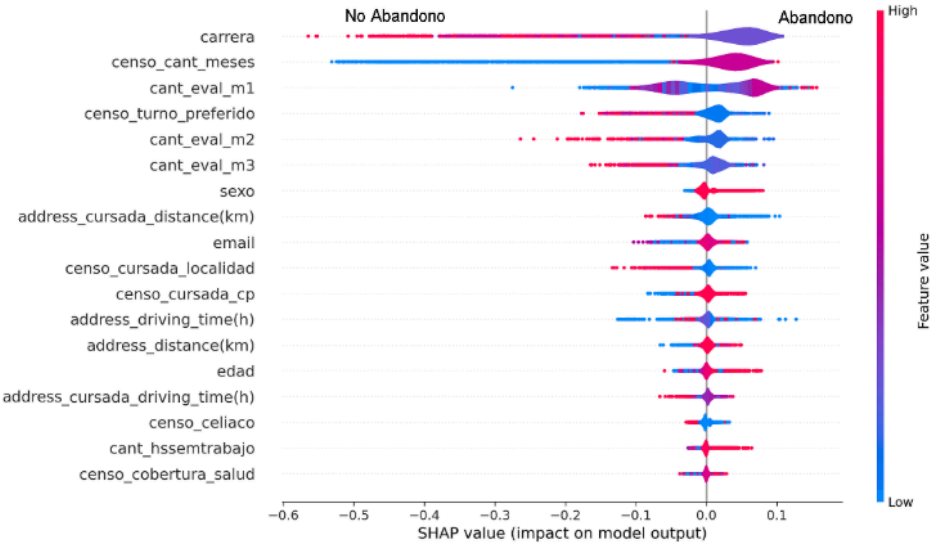


Figura 37: Gráfico de valores SHAPE.

6.3.1. Abandono condicional e importancia de las variables

A continuación mostraremos las hipótesis, sus variables e instancias relacionadas, ordenadas por orden importancia para el modelo *XGBoost*. Las variables importantes mostraron un porcentaje de abandono distinto al porcentaje promedio (61%) para ciertas instancias. En la Tabla 20 se resumen esas instancias. Por ejemplo, La "Cantidad de evaluaciones rendidas hace un semestre" proviene de la hipótesis H1. La proporción de alumnos que rindió al menos una vez es 0,48 (el 48% de la población rindió al menos una vez en el semestre anterior al que se quiere predecir). Este grupo de alumnos tiene un porcentaje de abandono de 45,05%, mucho menor que la media poblacional. La importancia relativa (baja, media, alta) se asigna en

función del porcentaje de población afectada, la distancia al abandono medio, la importancia en el modelo XGBoost y valor de Cramers'V. Por ejemplo, la variable "Cant. Hs. Trabajo" tiene una importancia (IG) de 4.1 en el modelo XGBoost y un valor de Cramers'V de 0,04. Para el caso de "10 a 25 horas por semana" afecta al 8% de la población y tiene una diferencia de 4,58 puntos porcentuales respecto del abandono medio. Por esos motivos se le asignó una importancia "media".

Para todas las variables que superan el umbral de 0,2 de Cramer's V sugerido en Kim (2017), se les asignó valores de importancia entre "medio" y "alto". Dichas variables fueron el "Turno de cursada", "Hace cuantos meses completo el censo", "Carrera" y "Cantidad de evaluaciones rendidas hace un semestre".

Variable (Hipótesis)	Instancia	Proporción del total	Porcentaje Abandono	Dif. Ab. Medio	Cramer's V	Importancia Relativa
Cobertura de Salud (H5)	-			< 1%		Baja
Tipo de Vivienda (H7)	-			< 1%		Baja
Número de Piso (H7)	-			< 1%		Baja
Cantidad de Becas (H8)	Una o más becas	0,04	50,67 %	10,33 %	0,04	Media
Edad (H11)	De 18 a 20 años	0,06	52,78 %	8,22 %	0,05	Media
Cant. Hs. Trabajo Semana (H9)	Entre 10 y 25 horas	0,08	56,42 %	4,58 %	0,04	Media
Cantidad de Familiares Convive (H4)	Más de uno	0,34	57,60 %	3,4 %	0,08	Media
Turno de Cursada (H6)	Tarde	0,08	67,30 %	-6,3 %	**0,24**	Media
Tiempo de Viaje (H2)	De 6 y 18 minutos	0,23	64,53 %	-3,53 %	0,16	Media
Dominio de Email (H3)	Institucional	0,09	32,08 %	28,92 %	0,18	Alta
Género (H4)	Mujer	0,66	58,27 %	2,73 %	0,07	Alta
Cant. Meses Censo (H1)	Hace más de 50 meses	0,35	80,11 %	-19,11 %	**0,32**	Alta
Carrera (H10)	CPU	0,45	70,00 %	-9 %	**0,33**	Alta
Cant. Eval. Rendidas hace un Semestre (H1)	Más de una	0,48	45,05 %	15,95 %	**0,41**	Alta

Tabla 20: Hipótesis, variables asociadas y algunas instancias, ordenadas por: importancia relativa. Se muestra la proporción de la población total, el porcentaje de abandono para la instancia seleccionada y la importancia de la variable. Diferencia respecto de Abandono medio = 61% - Porcentaje Abandono. Y medida de asociación de Cramer's V.

6.3.1.1. Cantidad Meses Censo

La Figura 38 (izquierda) muestra la cantidad de alumnos que completaron el censo hace una cantidad de meses (eje x). La Figura 38 (derecha) muestra el porcentaje de abandono para cada cantidad censada. A partir de los datos se calculó y graficó una regresión lineal (línea negra) con pendiente positiva, indicando una relación proporcional de abandono respecto de hace cuantos meses completo el censo. Los alumnos que nunca completaron el censo se agrupan al final de la serie. El 71,79% de estos 2.896 alumnos están clasificados como "abandono". Muchos grupos, como por ejemplo (x = 62 meses) tienen asociados porcentajes altos (porcentaje > 90). Estos datos son atípicos por tener menos de 50 alumnos agrupados en ese rango.

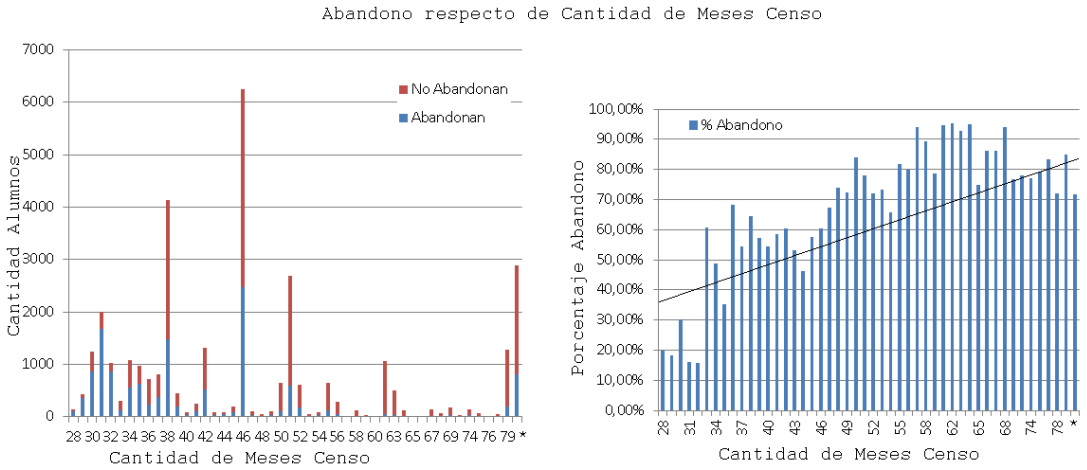


Figura 38: Cantidad de alumnos respecto de hace cuantos meses completaron el censo (izquierda). Porcentaje de alumnos que abandonan (derecha). *Alumnos que nunca completaron el censo, se agrupan al final de la serie.

6.3.1.2. Cantidad Evaluaciones Rendidas Hace un Semestre

La Figura 38 (izquierda) muestra la cantidad de alumnos que rindieron 0,1 o más exámenes hace un semestre (eje x). La Figura 39 (derecha) muestra el porcentaje de abandono para cada cantidad censada. A partir de los datos se calculó y graficó una regresión lineal (línea negra) con pendiente decreciente, indicando una relación inversamente proporcional de abandono respecto de la cantidad de exámenes rendidos. Esta relación también se presenta para la cantidad de exámenes rendidos hace más de un semestre Se censaron los últimos cuatro.

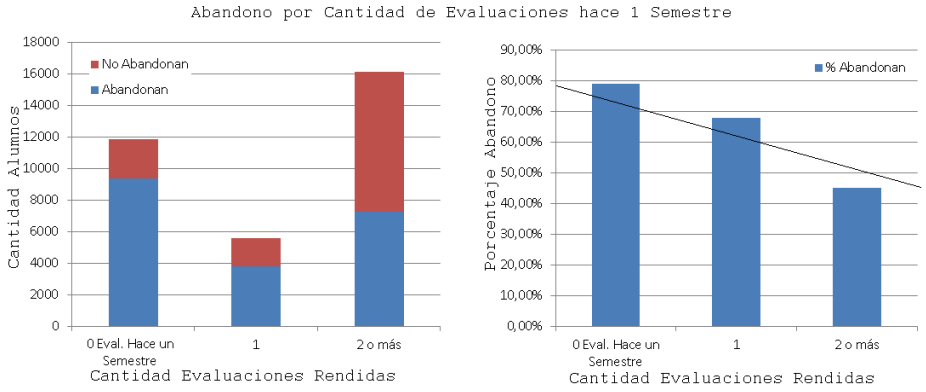


Figura 39: Cantidad de alumnos respecto de cuántos exámenes rindieron el semestre anterior (izquierda). Porcentaje de alumnos que abandonan (derecha).

6.3.1.3. Tiempos y distancias de viaje

En la Figura 40 se ven los tiempos y distancias de viaje (variables continuas) que se calculan a partir de la dirección. Vemos que los valores declarados al inicio de la carrera se correlacionan con los valores censados durante la carrera. Es esperable que cambien las direcciones y por lo tanto los tiempos y distancias de viaje.

	Abanono	address_driving_time(h)	address_distance(km)	address_cursada_driving_time(h)	address_cursada_distance(km)
Abanono	1.000000	0.189836	0.037280	0.000136	0.032795
address_driving_time(h)	0.189836	1.000000	0.128739	-0.050885	0.118013
address_distance(km)	0.037280	0.128739	1.000000	-0.395780	0.594194
address_cursada_driving_time(h)	0.000136	-0.050885	-0.395780	1.000000	-0.593984
address_cursada_distance(km)	0.032795	0.118013	0.594194	-0.593984	1.000000

Figura 40: Correlaciones entre el “Tiempo/Distancia de viaje” de inscripción y de cursada.

7. Discusión y conclusiones

Se estima que en el Sistema Universitario Argentino sólo el 27,66% de los estudiantes que ingresan se gradúa en un tiempo teórico de 5 años[18], generando cada año el abandono o la extensión de la carrera para más de 400.000 alumnos. Esta situación ocurre en todo el mundo en mayor o menor medida y comenzó a censarse desde principios del siglo XX. Los primeros trabajos al respecto[19], en las décadas del '30 al '50, son descriptivos, ya que crean marcos conceptuales que intentan explicar el abandono. En las décadas del '60 y '70 se encuentran los primeros trabajos predictivos del abandono. Estos trabajos se dividen en modelos conceptuales que predicen el abandono cuando se alcanzan ciertas condiciones (e.g. "Si el alumno no se adecua al reglamento universitario, es más propenso a abandonar") y en modelos probabilísticos. Estos modelos combinan regresiones lineales con diagramas que indican causalidad mediante conectores (e.g. los conceptos se representan con rectángulos y estos se relacionan con fechas que indican causa/efecto). Los pesos de cada variable ("betas") de estos modelos se calculan con herramientas estadísticas. En todos los casos, los modelos de esta época se basan en censos. Por la gran cantidad de datos, en la mayoría de los casos los "betas" se calculan por computadora. En la década del '90 encontramos modelos probabilísticos basados en datos provenientes de bases de datos de las universidades. En la primera década del siglo XXI surgen modelos basados en Aprendizaje Automático, dado que a diferencia de la inferencia estadística, estos modelos tienen mayor capacidad para lidiar con datos faltantes y ruido. En particular los Árboles de Decisión son robustos a estos problemas. La otra gran diferencia con los modelos probabilísticos predictivos de las décadas del '70 y '80 es que además de la predicción en sí, dan "explicabilidad" como resultado de los modelos, indicando por ejemplo la importancia de cada variable en la predicción.

En este trabajo se usaron modelos de Aprendizaje Automático para testear hipótesis que surgen de la bibliografía o de actores de UNAHUR y para mejorar los modelos planteados en Pustilnik y Ndukanma (2023a y 2023b). Los datos del SIU constan de aproximadamente de 33.000 registros que no representan un problema de "cantidad" para la computación de hoy en día pero representan un problema para generar las variables dada la cantidad de tablas y variables que hay que generar de manera manual, a partir de las 10.000 variables de siu-guaraní, para cada modelo. Se recomienda re-entrenar el modelo con cada cohorte nueva para captar particularidades de cada año, como el caso de la pandemia en donde no se registró asistencia presencial, las actualizaciones y nuevas carreras y cualquier otro cambio que haga envejecer el modelo. En los repositorios git 2023a, git 2023b y git 2023c se encuentra el código para re-entrenar los modelos cada año.

El aumento de la capacidad de cómputo junto a las mejoras en los algoritmos de Aprendizaje Automático que surgieron en los últimos años (e.g. XGBoost, 2016) permiten entrenar modelos con mejor performance para los mismos datos que en décadas pasadas. El mejor modelo obtuvo un AUC 0,88 para una exactitud balanceada óptima de 0,77. Las variables más explicativas se resumen en la Tabla 20 y son: "Cantidad de Becas" (H8), "Edad" (H11), "Cantidad de Horas Trabajo Semana" (H9), "Cantidad de Familiares Convive" (H4), "Tiempo de Viaje" (H2), "Dominio de Email" (H3), "Género" (H4), "Turno de Cursada" (H6), "Cantidad Meses Censo" (H1), "Cant. Eval. Rendidas hace un Semestre" (H1) y "Carrera" (H10). Estas variables resultaron importantes en todos los modelos, validando las hipótesis que están entre paréntesis. Hubo hipótesis que no se pudieron validar ya sea por los datos actuales o porque no tienen relevancia en general (H5: "Cobertura de salud" y H7: "Tipo de vivienda" H7). Variables como "Cantidad Meses Censo" y "Cantidad de evaluaciones hace un semestre" mostraron tener una relación lineal con el abandono, sugiriendo prestar más atención a los estudiantes en esas condiciones. Se validaron variables poco mencionadas en la literatura, como "Hace cuantos meses completo el censo", "Email" y "El Tiempo/Distancia de viaje".

[18] Estadísticas para el año 2021.
[19] e.g. Iffert (1958).

La identificación temprana de estudiantes en riesgo de abandono permite desarrollar intervenciones más eficaces, y consideramos que el sistema de recomendación existente (Pustilnik et al., 2022) podría adaptarse para proporcionar orientación a estos estudiantes. En este trabajo, mejoramos el desempeño de un modelo de predicción de abandono en UNAHUR propuesto en Pustilnik y Ndukanma (2023a y 2023b), incorporando variables como “Cantidad de horas de trabajo por semana”, “Cantidad de becas” y “Carrera”. La adición de estas variables y la aplicación de estas recomendaciones pueden potenciar aún más la efectividad de los modelos. En futuros trabajos esperamos obtener resultados aún mejores si conocemos:

- El modo de traslado real de los alumnos.
- La situación laboral durante la cursada, como sugieren Amago (2008), Antoni et al. (2007) y otros.
- La percepción de becas durante la cursada, como sugieren Cameron y Taber (2001), Di Gresia et al. (2007) y otros.
- El impacto de los programas como “1 estudiante - 1 compañero” que aún no han sido censados.
- El impacto del covid-19 en la tasa de abandono, dado el cambio de cursada virtual y otros factores.
- La asistencia a clase, que durante el covid-19 no se pudo censar

Recomendamos enriquecer el censo anual con preguntas adicionales sobre el modo de transporte, la percepción de becas, si es “primera generación de universitarios en su familia” como sugiere Arias et al. (2015), como cursaron durante el codiv-19, la inscripción en programas de acompañamiento universitario y la situación laboral de los estudiantes durante la cursada.

8. Anexos:

8.1. Antecedentes personales en el tema.

Desde 2021 empecé a relevar el tema de abandono en UNAHUR, en donde comencé a trabajar como investigador ese mismo año. Me integré al proyecto: *"Reforzando las capacidades de comunicación y abordaje de problemáticas de poblaciones estudiantiles en rápido crecimiento: propuestas de inscripción, detección temprana de riesgo de deserción"* (Resolución CS - 325 / 2022 de la UNAHUR, en el que continuo actualmente. El mismo está en el Banco de Proyectos de Desarrollo Tecnológico y Social (Banco PDTS) del Ministerio de Ciencia, Tecnología e Innovación de la Nación. Desde 2023 estoy colaborando con el programa +Acompañamiento[20] para el desarrollo de modelos de abandono de ingresantes a la FCEN, UBA. A continuación se resumen los antecedentes:

- Presentación del trabajo "Estrechando el contacto entre universidades y estudiantes: comunicación ante posibles casos de abandono en el , propuestas para la inscripción" en el XXIV Workshop de Investigadores en Ciencias de la Computación (WICC, Mendoza), Pustilnik et al. (2022).

- Presentación del trabajo "Modelos Para La Predicción del Abandono en la Universidad Nacional de Hurlingham" en el XVIII Congreso Tecnología en Educación & Educación en Tecnología (TE&ET, Buenos Aires), Pustilnik y Ndukanma (2023a).

- Presentación del trabajo "Ingeniería de atributos para modelos de predicción del abandono universitario en Argentina" en el XII Congreso Latinoamericano sobre el Abandono en la Educación Superior (CLABES, Chile), Pustilnik y Ndukanma (2023b).

- Investigador Externo en las Becas de Iniciación a la Investigación en Ciencias de la Computación[21], para la beca "Predicción de abandono en ingresantes de la FCEN, UBA", durante 2023-2024. En el marco del proyecto *+Acompañamiento* (Aprobado por RESCD-2023-711-E-UBA-DCT#FCEN).

[20] https://exactas.uba.ar/acompanamiento/
[21] https://www.dc.uba.ar/un-primer-acercamiento-a-la-investigacion/

8.2. Tabla 1 anexo

En la Tabla 1 anexo se muestran las características de los trabajos científicos sobre rendimiento académico y abandono en las universidades nacionales de la Argentina, 2002-2012, recopilados por Garcia (2014):

Autor	Nivel de Análisis	Metodo de Analisis	Variable a Explicar	Principales Resultados: Factores asociados con las variables dependientes
Amago (2008)	Universidad	Estadística descriptiva y metodología cualitativa	Abandono	Variable: desgranamiento primer año. Asociado positivamente con el menor ingreso del hogar, la actividad laboral y la experiencia académica. En el plano de la institución afectan la distancia académica con experiencias previas, el ritmo intenso de la programación del plan de estudio y prácticas de enseñanza.
Antoni et al. (2007)	Carreras	Logit Correspondencia múltiples	Rendimiento académico	Éxito superior entre los que aprobaron los cursos de nivelación inicial y el examen de ingreso; mejor no trabajar y ser mujer, favorecido por alto nivel de instrucción de los padres y categorías ocupacionales altas. No influye el haber recibido orientación vocacional ni el contar con un título medio afín con la carrera elegida (contador público).
	Universidad	Estadística descriptiva, Análisis de tablas de contingencia, metodología cualitativa	Graduación y Abandono	Graduados con promedio más alto: ingreso con menor edad, mayoría mujeres y de los estratos más altos, con menos retraso en los estudios, no trabajan. Desertores: la edad no es factor discriminante de mayor logro. Las mujeres que desertaron presentaron un mayor rendimiento que los varones y menores porcentajes de cambios de carrera. Los desertores de estratos más bajos presentaron menores niveles de rendimiento y la herencia cultural influyó en los logros. Los desertores que trabajaban presentaban menor rendimiento, peores promedios y mayor tendencia a cambiar de carrera.
Berges et al. (2007)	Carreras	Regresión por Mínimos cuadrados ordinarios (MCO)	Rendimiento académico	Variable cantidad de materias aprobadas y promedio: significativa en forma positiva el haber asistido a una escuela privada y el promedio del secundario. Las variables género, lugar de residencia familiar, turno en el que cursó y tipo de bachillerato no resultaron significativas. Altos puntajes en los exámenes de admisión y altos promedios polimodal relacionados con rendimiento posterior, aunque esto no es condición suficiente, quizás por la heterogénea calidad del nivel medio.
Carella (2009)	Universidades Nacionales	Estimación de frontera de eficiencia	Rendimiento académico	Son más eficientes los alumnos: que no trabajan, egresados de escuelas medias dependientes de las UUNN, los que trabajan en tareas relacionadas con la carrera, los que ingresaron siendo mayores, los que reciben ayuda familiar para sostener sus gastos y cuyas madres trabajan en el hogar.
Carella et al. (2007)	Carreras	Modelo estadístico Tobit para datos censurados (truncados)	Rendimiento académico	La probabilidad de éxito de concluir el ciclo de formación inicial es mayor para las mujeres, para quienes poseen padres con mayores niveles de educación, no trabajan al inicio y tienen un nivel socioeconómico alto (posee o no obra social).

Tabla 1 anexo (parte I): Autor, Nivel de Análisis, Método de análisis, Variables a explicar y Principales resultados recuperados de la bibliografía por Garcia (2014).

Autor	Nivel de Análisis	Metodo de Analisis	Variable a Explicar	Principales Resultados: Factores asociados con las variables dependientes
Cerro (2007)	Carreras	Regresión por MCO	Rendimiento académico	Variable cantidad de materias aprobadas: inciden positivamente la educación de los padres, el tipo de escuela secundaria (privada), el desempeño en la escuela secundaria, y el ciclo inicial de la facultad
Chudnovsky (2003)	Asignaturas	Estadística descriptiva y metodología cualitativa	Abandono	Variable: desgranamiento primer año: predominan los que tiene mayor edad en el ingreso, los que viven en el GBA, los que trabajan o buscan trabajo, los que no conocen el plan de estudio y la oferta horaria de las carreras, los que presentan problemas vocacionales y siente inseguridad laboral futura.
Di Gresia (2009)	Carreras	Modelo de datos censurados para tiempo de abandono[22] https://core.ac.uk/reader/290002480	Rendimiento académico	Variable cantidad de materias aprobadas: mejor rendimiento si se es mujer; no significativos: estado civil, edad al ingreso, tipo de secundario y lugar de nacimiento, ocupación del padre; la educación de los padres tiene una influencia positiva;los que no trabajan al inicio de la carrera tienen mejor rendimiento, la carrera de economía tiene desempeños mejores que las otras. Hay relación entre el desempeño inicial y el posterior. El 84,2% de los alumnos que no aprobaron ninguna materia en el ciclo inicial no logran aprobar ninguna materia hasta el 6to año de la carrera.
Di Gresia et al. (2007)	Universidades Nacionales	Regresión por MCO	Rendimiento académico	Variable cantidad de materias aprobadas: mejor rendimiento si son mujeres, extranjeros, solteros, asistieron a educación privada, padres más educados, traslado de residencia por estudios, estudiantes que trabajan, financiamiento por becas y contribución familiar. Las horas de estudio tienen efectos positivos en todos los cuantiles en la distribución condicional del rendimiento pero el efecto de tiempo adicional es más fuerte entre los de menor rendimiento.
Fazio (2004)	Universidades Nacionales	Regresión por MCO	Rendimiento académico	En un tramo de baja cantidad de horas de trabajo (14,5 horas) el rendimiento del alumno es positivo, más allá de este punto inciden negativamente. Si el trabajo está vinculado con la carrera se puede aprovechar los beneficios del trabajo hasta 24,5 horas semanales.
Ferreira (2007)	Carreras	Estimación por cuantiles y Tobit[23]	Rendimiento académico	Variable cantidad de materias aprobadas y promedio de notas combinados: La influencia no es homogénea. Entre los alumnos de bajo rendimiento influye positivamente el ser mujer, más joven, tener padres con mayor capacitación y no trabajar al ingreso. Entre los alumnos de mayor rendimiento influyen negativamente: el estado civil (estar casado), el tipo de residencia (con los padres), la escuela secundaria (pública), la situación socioeconómica (no poseer obra social) y la condición de actividad de la madre (activa). Tanto el rendimiento en la escuela secundaria como en los primeros meses de la carrera resultaron relevantes y con efectos similares para la mayoría de los estudiantes.

Tabla 1 anexo (parte II): Autor, Nivel de Análisis, Método de análisis, Variables a explicar y Principales resultados recuperados de la bibliografía por Garcia (2014).

[22] Se adaptan modelos que calculan “la probabilidad de un evento para un tiempo t” como “la probabilidad de abandono para un tiempo t” (e.i distribución poisson).

[23] Tobit: Adaptación de cuadrados mínimos cuando hay datos no observables o truncados en la serie.

Autor	Nivel de Análisis	Metodo de Analisis	Variable a Explicar	Principales Resultados: Factores asociados con las variables dependientes
Gertel et al. (2007)	Carreras	Regresión por MCO	Eficiencia y Rendimiento académico	Variable eficiencia: significativa en términos negativos si costea sus estudios con el aporte familiar únicamente y positivo a más alto nivel de educación del padre o la madre, nota promedio del secundario, especialidad del secundario (no afín) y las tres asignaturas del curso de nivelación. No fueron significativas sexo y tipo de dependencia del nivel medio. Rendimiento: igual resultado que en eficiencia excepto el efecto del curso de nivelación.
Giner et al. (2007)	Carreras	Regresión por MCO y Probit[24]	Rendimiento académico	Variable promedio con aplazo: significativo en forma positiva según sexo (ser mujer), edad (ser joven), educación del padres y de la madre. Variable cantidad de materias aprobadas y promedio: significativo en forma positiva si sexo (ser mujer) y la educación del padre y de la madre.
Giovagnoli (2002)	Carreras	Modelos de riesgo proporcional no paramétricos	Rendimiento académico	Un estudiante cuyo padre tiene primaria incompleta tiene un 70% menos posibilidades de graduarse que aquel estudiante con padre profesional. Controlando por heterogeneidad no observable, el riesgo de desertar es 2.86 veces mayor para el alumno cuya madre cuenta con primaria incompleta con respecto a otro cuya madre tiene superior completo. El riesgo de deserción es menor si los padres son jefes, directores o altos jefes en comparación con hijos de obreros o empleados. Si al iniciar la carrera está trabajando tiene 3.4 veces más riesgo de desertar. El riesgo de abandono de los varones es 1.36 veces mayor que el de una mujer. Afectan también el abandono ser residente de Rosario (lugar de la universidad), vivir con la familia, ser soltero y tener más edad al iniciar la carrera.
Giovagnoli (2007)	Carreras	Regresión por MCO y Tobit	Rendimiento académico	Variable cantidad de materias aprobadas: significativo en forma positiva: mujeres, menor edad al ingreso, no trabajar al inicio, comenzar inmediatamente la carrera, no residir con sus padres y hermanos, asistir a colegios universitarios, ser no rosarinos y padres de alta educación. Aquellos que aprobaron más materias al inicio se graduaron
Giuliodori et al. (2010)	Carreras	Regresión logística y por cuartiles	Rendimiento académico	El desempeño en el nivel secundario y en el curso de ingreso son factores de gran poder predictivo del rendimiento académico posterior en la carrera universitaria. La nota promedio del secundario ejerce un efecto más fuerte cuando se trata de un estudiante de rendimiento intermedio o alto. El efecto es menor cuando el estudiante es de bajo rendimiento. Los estudiantes de alto rendimiento generalmente tienen una buena formación previa que se expresa en un promedio más alto en el secundario. Entre los alumnos de rendimiento bajo se aprecia un nivel heterogéneo de calificaciones en el secundario.
Goldenhersh et al. (2011)	Carreras	Regresión logística y metodologías cualitativas	Abandono	Es crítico sortear el primer año. El mejor rendimiento en el primer año está correlacionado con la dependencia de la escuela secundaria, situación de actividad y ocupación del padre, sexo de la persona con quién vive, cómo costea sus estudios y estudios de los padres.
Jaime (2004)	Carreras	Análisis correlacional y metodología cualitativa	Abandono	La repitencia y la deserción están vinculadas con la baja formación en el nivel medio y con la falta de orientación vocacional. También incide la reprobación en algunas materias. Hay más desertores varones que mujeres y son más entre los que residen en el mismo lugar de la universidad.

Tabla 1 anexo (parte III): Autor, Nivel de Análisis, Método de análisis, Variables a explicar y Principales resultados recuperados de la bibliografía por Garcia (2014).

[24] Los modelos Logit y Probit son modelos econométricos no lineales que se utilizan cuando la variable dependiente es binaria.

Autor	Nivel de Análisis	Metodo de Analisis	Variable a Explicar	Principales Resultados: Factores asociados con las variables dependientes
Kuna et al. (2011)	Carreras	Proceso de explotación de información estandarizada	Abandono	Variable: alumnos sin actividad en el segundo año. Factores de riesgo: los que regularizaron sólo una o cero materias en el primer año; dentro de este universo: los que costearon sus estudio con su trabajo, que han dejado pasar mayor lapso de tiempo entre el fin del secundario y el ingreso a la universidad, bachilleres (la facultad brinda carreras técnicas), los que debían viajar más de 10 Km. para asistir a clases.
López et al. (2012)	Carreras	Modelo de Regresión Logística Multinomial y dos modelos de Redes Neuronales	Rendimiento académico	Variable: aprobación de exámenes parciales de las asignaturas del primer cuatrimestre del primer año. En la carrera de bioquímica, el análisis de sensibilidad mostró que la variable más importante fue el título del secundario. En el conjunto de las carreras de perfil profesional de esta facultad resultaron relevantes para explicar el rendimiento académico el estudio de los padres y el año de ingreso a la Facultad.
Odetti et al. (2010)	Carreras	Correlación y metodología cualitativa	Rendimiento académico y abandono	Entre los problemas detectados se encuentran la adaptación a la universidad y el desarraigo emocional y para los estudiantes de química inorgánica, la cantidad de horas de cursado y la falta de tiempo para estudiar.
Oliver et al. (2011)	Asignatura	Estadística descriptiva	Rendimiento académico	Se mide el rendimiento académico y el abandono en la materia de química dentro de distintas especialidades de la ingeniería. Analizando este resultado se concluye que la variabilidad de estos indicadores se asocia con una distinta relación alumno-docente, obteniendo mejores resultados cuando ésta es menor.
Parrino (2012)	Carreras	Análisis factorial y metodología cualitativa	Abandono	Se realiza un análisis descriptivo encontrando como factores que explican la deserción: el contexto (la masividad, el mercado laboral y las exigencias de calidad) como inductores; el sistema de educación superior (política de baja selectividad, alta regulación de la permanencia y falta de acreditación) y la institución (sistema de acceso, condiciones académicas y normativa) como favorecedores de la deserción. En el plano personal incide el capital económico, cultural y escolar como promotores de la deserción.
Patriarca (2012)	Carreras	Estadística descriptiva y metodología cualitativa	Abandono	La encuesta a alumnos que desertaron antes de iniciar el curso de preparación universitaria revela como principales factores los laborales y en menor medida los académicos (elección de otra institución). Respecto de los alumnos que están cursando el ingreso, destacan dificultades relacionadas con la formación previa y con las condiciones propias del aprendizaje. Las entrevistas a informantes clave permiten destacar la importancia de las políticas institucionales para evitar la deserción, especialmente durante el primer año.

Tabla 1 anexo (parte IV): Autor, Nivel de Análisis, Método de análisis, Variables a explicar y Principales resultados recuperados de la bibliografía por Garcia (2014).

Autor	Nivel de Análisis	Metodo de Analisis	Variable a Explicar	Principales Resultados: Factores asociados con las variables dependientes
Paz et al. (2011)	Universidad	Regresión por MCO y micro descomposiciones	Rendimiento académico	Tres indicadores de desempeño: la calificación promedio, el número de aplazos acumulados en la carrera y la duración relativa. Las mujeres aventajan a los varones y existen fuertes diferencias según facultades y según carreras y entre planes de estudio dentro de las carreras. No se encontraron razones de composición demográfica de los egresados que expliquen estas disparidades. Algunos indicios sugieren que la resolución que elimina los aplazos del cómputo de los promedios provocó una caída del desempeño académico global. El pertenecer a Exactas, Económicas, Salud y Sedes regionales, el tener edad más avanzada y el ser extranjero.
Pron (2007)	Carreras	Modelo de regresión para variables enteras	Rendimiento académico	Variable cantidad de materias aprobadas a 5 años de iniciados los estudios de la cohorte de ingresantes 2001: en el grupo que aprobó más de 13 materias, los porcentajes son mayores entre las mujeres, los solteros, argentinos, nacidos en La Plata, que no viven con sus padres, asistieron a escuela privada, obtuvieron mayor promedio en el nivel medio y cuyos padres tienen mayor educación y son activos. La única variable asociada exclusivamente a una mayor probabilidad de pertenecer al grupo que no aprueba materias es trabajar al inicio. Las variables cuya variación marginal se asocia con una reducción en el número esperado de materias son: nivel educativo del padre, residencia independiente, número de materias aprobadas en el ciclo inicial.
Ríos (2010)	Carreras	MCO y metodología cualitativa	Rendimiento académico	Variables cantidad de materias aprobadas, promedio y coeficientes que miden la eficiencia según número de materias aprobadas respecto a materias que se deberían haber aprobado en el 3º y 6º año. El promedio obtenido en la escuela media y el máximo nivel alcanzado por el padre y la madre son los mejores predictores del rendimiento académico. Existe relación significativa entre los estudiantes que no trabajan y mejores promedios pero no mostró asociación con el avance en la carrera o con la eficiencia. No existe asociación con variables tales como convivencia con los padres y ayuda familiar para costear los estudios, escuela privada u orientación de la escuela media. No resultaron significativas las variables edad al ingreso y sexo. En las encuestas se observó que entre los que presentan mayor atraso en el número de materias aprobadas, existía indefinición vocacional.
Sosa Escudero et al. (2009)	Carreras	Regresión por cuantiles y por MCO	Rendimiento académico	Variable cantidad de materias aprobadas: efecto heterogéneo de los distintos factores observados a lo largo de los distintos cuantiles de la distribución condicional del rendimiento. Por ejemplo, entre los contadores, el rendimiento esperado de los varones es aproximadamente 6 por ciento menor que el de las mujeres, siendo este efecto más fuerte en los niveles más bajos de rendimiento académico y prácticamente desaparece en los más altos. En particular, algunas variables (por ejemplo, la educación de la familia y trabajar y hacerlo en tareas no relacionadas con el estudio) ejercen mayor efecto entre los alumnos con menor rendimiento promedio. El tipo de secundario (comercial, bachiller, técnico) no ejerce efecto sobre el rendimiento de los contadores y abogados.

Tabla 1 anexo (parte V): Autor, Nivel de Análisis, Variables a explicar y Principales resultados recuperados de la bibliografía por Garcia (2014).

9. Bibliografía

Almeida L., Ferreira J., Soares A. (1999). Apresentação do Questionário de Vivências Académicas (QVA). Psicologia, 19(2), 149–160.

Amago L. (2008). Desgranamiento en el primer año de la Universidad. La cohorte 2005 en la Universidad Nacional de General Sarmiento. (Tesis inédita de maestría). Universidad Nacional de Luján, Pcia. de Buenos Aires.

Antoni E.J., Pagura J.A., Quaglino M.B. (2007). El rendimiento universitario. Un estudio de posibles factores causales en una facultad de la Universidad Nacional de Rosario. En Porto, A. (Ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas (pp. 177-191). La Plata: Editorial de la Universidad de La Plata.

Arias M., Mihal I., Gorostiaga, J. (2015). El problema de la equidad en las universidades del conurbano bonaerense en Argentina: Un análisis de políticas institucionales para favorecer la retención. Revista mexicana de investigación educativa 51(20), 47–69, https://ri.conicet.gov.ar/handle/11336/51703

Azevedo A. y Santos M.F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conf. Data Mining, https://api.semanticscholar.org/CorpusID:15309704

Bean J. y Metzner B. (1985). Conceptual model of nontraditional undergraduate student attrition. Review of Educational Research, 55(4), 485–540. https://doi.org/10.3102/00346543055004485

Bean J. (1980). Student Attrition, Intensions and Confidence: Research in Higher Education. Nº 17: 291-320.

Bean J. P. y Vesper N. (1990). Qualitative approaches to grounding theory in data: Using LISREL to develop a local model and theory of student attrition. Communication presented in the annual meeting of the American Educational Research Association, Boston, Ma.

Bean J. P. y Eaton S. B. (2001). The psychology underlying successful retention practices. Journal of College Student Retention Research, Theory & Practice Vol. 3, N° 1: 73-89.

Berger J. B. y Braxton J. M. (1998). Revising Tinto's Interactionalist Theory of Student Departure through Theory Elaboration: Examining the Role of Organizational Attributes in the Persistence Process. Research in Higher Education. 39 (2), pp. 103-119.

Bernal E., A., Cabrera A.A., Terenzini P. (2000). The relationship between race and socioeconomic status (SES): Implications for institutional research and admissions policies. Removing Vestiges: Research-based strategies to promote inclusion. A publication of the American Association of Community Colleges. Nº 3: 6-19.

Berger J. y Milem J. (2000). Organizational Behavior in Higher Education and Student Outcomes. In: J. Smart (Ed.), Higher Education: Handbook of theory and research. Vol. 15: 268-338.

Berger J. (2002). Understanding the Organizational Nature of Student Persistence: Empirically based Recommendations for Practice. Journal of College Student Retention: Research, Theory and Practice. Vol. 3, N° 1: 3-21.

Berges M., Pérez Rojas M., Malamud C., Pesciarelli S. (2007). Mecanismos de ingreso a la Facultad y rendimiento de los alumnos durante el primer año. En Porto, A. (ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas (pp. 141-158). La Plata: Editorial de la Universidad de La Plata.

Biswal A. (2023). Bagging in Machine Learning: Step to Perform And Its Advantages.

https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning

Bourdieu P. y Passeron J.C. (1998). La reproducción. Elementos para una teoría del sistema de enseñanza, Fontamara, México, D. F.

Braxton J., Sullivan A., Johnson Jr. (1997). Appraising Tinto's theory of college student departure. In: J.C. Smart (Ed.), Higher education: Handbook of theory and research. New York. USA. pp. 107-164.

Breiman L. (1996). Bagging predictors. Mach Learn, 24(2):123–140.

Cabrera A.A., Nora M., Castañeda (1992). The role of finances in the persistence process: A structural model. Research in Higher Education. Vol 33, N° 5: 303-336.

Cabrera A.A., Nora M., Castañeda (1993). College Persistence: Structural Equations Modelling Test of Integrated Model of Student Retention. Journal of Higher Education. Vol. 64, N° 2: 123-320.

Cabrera A.F. y La Nasa S. (2001). Three Critical Tasks America's Disadvantaged Face on Their Path to College. En Cabrera, A. F. & La Nasa, S. (eds). Understanding the College Choice of Disadvantaged Students. San Francisco: Jossey-Bass Publishers.

Cameron y Taber C. (2001). Estimation of Education Borrowing constraint using Returns of Schoolong. Journal of Political Economy, 2004, vol. 112, n. ° 1.

Carella L. (2009). Educación universitaria: medición del rendimiento académico a través de fronteras de eficiencia. (Tesis inédita de maestría). Universidad Nacional de La Plata.

Carella L., Ferreira G.; Pron J. (2007). Desempeño en el ciclo de formación inicial: Análisis de cohortes de la Facultad de Ciencia Económica de UNLP. En Porto, A. (ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas pp. 117-140. La Plata: Editorial de la Universidad de La Plata.

Castillo Diaz M. (2022). Perfil de ingreso y vivencias académicas tempranas implicadas en la intención de abandono en educación superior. Análisis en contexto hondureño. XI Congreso Latinoamericano Sobre Abandono en Educación Superior (XI Clabes 2022). Recuperado de https://doity.com.br/anais/trabalhos-apresentados/trabalho/256032

Cerro A. M. (2007). Estudio del rendimiento estudiantil en la Facultad de Ciencias Económicas UNT. En Porto, A. (ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas pp. 193-216. La Plata: Editorial de la Universidad de La Plata.

Chan D., Badano C., Rey A.A. (2019). Análisis inteligente de datos. ISBN 978-987-4998-22-4. Recuperado de https://www.scribd.com/document/667751672/Analisis-Inteligente-de-Datos

Chavez M. (2020). Somos mujeres de sectores populares : ¿llegamos a la universidad? : aproximaciones al acceso de mujeres de sectores populares a las universidades públicas del conurbano bonaerense : el caso de la universidad nacional de hurlingham (unahur). Tesis Maestria Argentina, http://hdl.handle.net/10469/16829

Chen T. y Guestrin C. (2016). Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2939672.2939785

Chudnovsky M. (2003). Causas de la deserción universitaria: un estudio exploratorio en la Universidad de Buenos Aires entre los años 2000-2002. (Tesis inédita de maestría). Universidad de San Andrés.

Collen A. y Gasparski W. (1995). Design & Systems: General Applications of Methodology (Praxiology). Transaction Publishers, U.S. 480 pp.

Cramér H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press, page 282 (Chapter 21. The two-dimensional case). ISBN 0-691-08004-6

Departamento de Información Universitaria (2022). Estadísticas Universitarias 2021-2022. República Argentina. https://www.argentina.gob.ar/sites/default/files/sintesis_2021-2022_sistema_universitario_argentino_1.pdf

Devore J.L. (2016). "Probabilidad y estadística para ingenieros y ciencias." Editorial: Cenpage Learning Editores S.A.

Di Gresia L., Fazio M. V., Porto A., Ripani L., Sosa Escudero W. (2007). Academic performance of public university students in Argentina. Well-Beingand Social Policy, 3(2), pp. 67-100

Díaz Peralta C. (2008). Modelo Conceptual para la Deserción Estudiantil Universitaria Chilena. Estudios pedagógicos (Valdivia), 34(2), 65-86. https://dx.doi.org/10.4067/S0718-07052008000200004

Dirección General de Educación Superior Universitaria, Perú (2021). Autodiagnóstico de las capacidades institucionales. Implementación de la educación remota en las universidades. https://hdl.handle.net/20.500.12799/7643

Durkheim E. (1951). Suicide: A study in sociology (G. Simpson, Ed. J.A. Spaulding $ G Simpson, Trans.). New York: Free Press

Ethington C. (1990). A psychological model of student persistence. Research in Higher Education. N° 31, Vol. 31: 279-293.

Fazio M. V. (2004). Incidencia de las horas trabajadas en el rendimiento académico de estudiantes universitarios argentinos (Tesis inédita de Maestría en Economía). Universidad Nacional de La Plata.

Ferreira M. G. (2007). Determinantes del desempeño universitario: efectos heterogéneos en un modelo censurado. (Tesis inédita de Maestría en Economía). Universidad Nacional de La Plata.

Fishbein M. y Ajzen I. (1975). Attitudes toward objects as predictors of simple and multiple behavioural criteria. Psycological Review. N° 81: 59-74.

Freund Y. y Schapire R. (1996). Experiments with a New Boosting Algorithm.

Gansemer-Tof A.M. y Schuh A.J. (2006). Institutional selectivity and Institutional Expenditures: Examining organizational factors that contribute to retention and graduation. Research in Higher Education. 47(6), pp. 613-640.

García A.M. (2014). Rendimiento académico y abandono universitario modelos, resultados y alcances de la producción académica en la Argentina. Revista Argentina de Educación Superior. Editorial Universidad Tres de Febrero. https://ri.conicet.gov.ar/handle/11336/35674

García de Fanelli A. (2013). Graduación y equidad en las universidades argentinas. Ponencia en el VII Encuentro Nacional y IV Latinoamericano: La universidad como objeto de investigación, 29-31 de agosto, Universidad Nacional de San Luis.

Gertel H.R., Giuliodori R. F., Casini R., González, M.V. (2007). Rendimiento y éxito académico de los estudiantes de la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba. Un análisis para dos cohortes de ingresantes. En Porto, A. (ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas pp. 19-48. La Plata: Editorial de la Universidad de La Plata.

Giner M. E., Lara de Ricci M. I., García Schilardi M.E. (2007). Influencia de características socioeconómicas en el ingreso y en el rendimiento académico. En Porto, A. (ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas pp. 49-90. La Plata: Editorial de la Universidad de La Plata.

Giovagnoli P. (2002). Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración. (Tesis de Maestría en Economía). Universidad Nacional de La Plata. https://sedici.unlp.edu.ar/bitstream/handle/10915/3436/Documento_completo.pdf?sequence=1&isAllowed=y

Giovagnoli P. (2007). Factores asociados al desempeño académico universitario. En Porto, A. (ed). Mecanismos de admisión y rendimiento académico de los estudiantes universitarios. Estudio comparativo para estudiantes de Ciencias Económicas pp. 159-176. La Plata: Editorial de la Universidad de La Plata.

Giuliano M. y Pustilnik M. (2024). "Características de abandono por facultad en la Universidad Nacional de Hurlingham". XIX Congreso de Tecnología en Educación y Educación en Tecnología. https://sedici.unlp.edu.ar/handle/10915/171424

Giuliodori R., Gertel H., Casini R., González M.V. (2010). Desempeño académico de los estudiantes de las Facultades de Ciencias Económicas y de Arquitectura, Urbanismo y Diseño de la Universidad Nacional de Córdoba. Córdoba: Facultad de Ciencias Económicas, Universidad Nacional de Córdoba.

Goldenhersh H., Coria A., Saino M. (2011). Deserción estudiantil: desafíos de la universidad pública en un horizonte de inclusión. RAES Revista Argentina de Educación Superior 3 (3), pp. 96-120.

Hand D.J., Till R.J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. Machine Learning 45(2), 171–186. https://doi.org/10.1023/a:1010920819831, https://doi.org/10.1023/a:1010920819831

Hastie T., Tibshirani R., & Wainwright M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18401

Hutagaol N. y Suharjito S. (2019). Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education. Advances in Science, Technology and Engineering Systems Journal, 4(4), 206–211. https://doi.org/10.25046/aj040425

IBM (2022). What is a Decision Tree?. https://www.ibm.com/topics/decision-trees

Iffert R. E. (1958). Retention and withdrawal of college students. U.S.Dept. of Health, Education and Welfare, Bulletin No.1.

Istvan R.M., Falco M., Antonini S.A.(2022). Análisis de los modelos educativos de universidades estatales chilenas. Revista Latinoamericana de Educación Comparada 53 (19), 1109–1113. http://sedici.unlp.edu.ar/handle/10915/61343

Jaime D.E. (2004). Deserción estudiantil en la Facultad de Agronomía y Zootecnia de la Universidad Nacional de Tucumán (1991-2001). (Tesis inédita de maestría). Universidad Nacional de Tucumán.

John E., Cabrera A., Nora A., Asker E. (2000). Economic influences on persistence. In: J. M. Braxton. Reworking the student departure puzzle: New theory and research on college student retention. Nashville: Vanderbilt University Press. pp. 29-47

Kim HY (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor Dent Endod. 42(2):152-155. https://doi.org/10.5395/rde.2017.42.2.152

Kuna H., García R., Martínez F., Villatoro R. (2011). Identificación de causales de abandono de estudios universitarios. Uso de procesos de explotación de información. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología (TE & ET), Nro. 6, pp. 39-44.

Landi J.A. y Giuliodori R. F. (2001). Graduación y deserción en las universidades nacionales. En Jozami, A. & Sánchez Martínez, E. (Comps.) Estudiantes y profesionales en la Argentina. Una mirada desde la Encuesta Permanente de Hogares pp. 79-103. Tres de Febrero: EDUNTREF

Likert R. (1932). "A Technique for the Measurement of Attitudes". Archives of Psychology 140: 1-55.

López M., Longoni M., Porcel E. (2012). Modelos estadísticos y conexionistas para predecir el rendimiento académico de alumnos universitarios. Investigación operativa, XX (33), pp. 135-157.

Losio M. y Macri A. (2015). Deserción y Rezago en la Universidad. Indicadores para la Autoevaluación. Revista Latinoamericana de Políticas y Administración de la Educación, 114-126.

Lundberg S.M. y Lee S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. https://arxiv.org/pdf/1705.07874.pdf

Márquez D., Ortiz S., Rendón M. (2009). Cuestionario de Vivencias Académicas en su versión reducida (QVA-r): un análisis psicométrico. Revista Colombiana de Psicología, 18(1),33-52. Recuperado de https://www.redalyc.org/articulo.oa?id=80412413004

Marino V., Sustas S., Quartulli D., Curcio J. (2023). Por qué estudiamos informática. indagación sobre trayectorias universitarias: instituciones, estudiantes, género y trabajo. https://program.ar/por-que-estudiamos-informatica/

Marquez E. (2022). Recrudecen las desigualdades económicas de género en el conurbano bonaerense en el escenario pos-covid-19. Universidad Nacional General Sarmiento. http://observatorioconurbano.ungs.edu.ar/?p=17483

Martínez-Plumed F., Contreras-Ochando L., Ferri C., Hernández-Orallo J., Kull M., Lachiche N., Ramírez-Quintana M., Flach P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3048-3061. https://doi.org/10.1109/TKDE.2019.2962680

Mendoza Santamaria J.V., Contreraz Ortiz M.S., Toro Cruz A. (2022). Modelo predictivo de la mortalidad académica de un programa de ingeniería de sistemas basado en técnicas de Machine Learning. XI Congreso Latinoamericano Sobre Abandono en Educación Superior (Clabes 2022). Recuperado de https://doity.com.br/anais/trabalhos-apresentados/trabalho/256042

Mitchell T. (1997). Machine learning. McGraw-Hill International Editions Computer Science Series.

Mucherino A., Papajorgji P., Pardalos P. (2009). k-Nearest Neighbor Classification, pp.83–106. Springer New York.

Ntampaka M., Trac H., Sutherland D.J., Battaglia N., Póczos B., Schneider J. (2015). A machine learning approach for dynamical mass measurements of galaxy clusters. DOI 10.1088/0004-637X/803/2/50

Nye J. (1976). Independence and Interdependence. Foreign Policy. Spring, Nº 22: 130-161.

Oliver M.C.; Eimer G.A., Bálsamo N. F., Crivello M.E. (2011). Permanencia y abandono en Química General en las carreras de ingeniería de la Universidad Tecnológica Nacional-Facultad Regional Córdoba. Avances en Ciencias e Ingeniería, 2 (2), pp. 117-129.

Ortiz M. S. y Fernández-Pera M. (2018). Modelo de Ecuaciones Estructurales: Una guía para ciencias médicas y ciencias de la salud. Terapia psicológica, 36(1), 51-57. https://dx.doi.org/10.4067/s0718-48082017000300047

Oyarzo J. (2020). ¿Qué es la Inteligencia Artificial (Artificial Intelligence)?. https://jaimeoyarzo.blogspot.com/2020/01/que-es-la-inteligencia-artificial.html

Parrino M. (2012). ¿Evasión o expulsión? Los mecanismos de deserción en el primer año universitario. (Tesis inédita de doctorado), Programa Interuniversitario Doctorado en Educación.

Parrino M. (2016). Permanencia y abandono en la universidad. Referentes e indicadores.. Revista Gestão Universitária na América Latina - GUAL 2016, 9 (1). https://www.redalyc.org/articulo.oa?id=319345197011

Pascarella E. y Terenzini P. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. Journal of Higher Education. Vol. 51, Nº 1: 60-75.

Pascarella E. y Chapman D. (1983a). Validation of a theoretical model of college withdrawal: Interaction effects in a multi-institutional sample. Research in Higher Education. Vol. 19, Nº 1: 25-48.

Pascarella E. y Chapman D. (1983b). A multi-institutional, path analytic validation of Tinto's model of college withdrawal. American Educational Research Journal. Vol. 20, Nº 1: 87-102.

Pascarella E. (1985). College environmental influences on learning and development: A critical review and synthesis. In: J. C. Smart (Ed.), Higher education: Handbook of theory and research. Vol. 1. New York: Agathon.

Patriarca C. (2012). La Deserción en el Inicio de la Vida Universitaria. Estudio contextualizado en la Escuela de Economía y Negocios de la Universidad Nacional de San Martín. (Tesis inédita de maestría). Universidad Nacional de San Martín.

Paz J., Antacle, C., Morales G., Rubio C. (2011). Análisis histórico del rendimiento académico de los graduados de la Universidad Nacional de Salta 1995-2001. En Trayectorias Educativas e Inserción Laboral (pp.89-134). Salta: Mundo Gráfico Editorial.

Peltier, G., Laden R., Matranga M. (1999). Student Persistence in Collage: a Review of Research. Journal College Student Retention, 1 (4), pp. 357-375.

Pustilnik M., Giuliano M., Puricelli F., Lombardi C., González Tulián G., Pagliari F., Saldivia C., Ybarra J., Gaiani M. (2022). Estrechando el contacto entre universidades y estudiantes: comunicación ante posibles casos de abandono, propuestas para la inscripción. XXIV Workshop de Investigadores en Ciencias de la Computación (WICC, Mendoza). pp. 734–738, http://sedici.unlp.edu.ar/handle/10915/145216

Pustilnik M. y Ndukanma G. (2023a). Modelos Para La Predicción del Abandono en la Universidad Nacional de Hurlingham. XVIII Congreso Tecnología en Educación & Educación en Tecnología (TE & ET, Hurlingham). ISBN: 978-987-46875-6-2. http://sedici.unlp.edu.ar/handle/10915/155526

Pustilnik M. y Ndukanma G. (2023b). Ingeniería de atributos para modelos de predicción del abandono universitario en Argentina. XII Congreso Latinoamericano sobre el Abandono en la Educación Superior (CLABES 2023). ISBN: 978-956-6224-39-6 URL: https://clabes.uct.cl/wp-content/uploads/2024/06/Acta-XII-CLABES-Revision-final.pdf

Pron J. (2007). Análisis del desempeño universitario utilizando modelos para variables enteras. (Tesis inédita de maestría). Universidad Nacional de La Plata.

Quinlan J.R. (1986). Induction of deciscion trees. In Machine learning, volume 1, pages 81-106. Kluwer Academic Publishers.

Quinlan J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo.

Repositorio git (2023a). Query, ConsultaModelo.sql
https://github.com/mpustil/modeloAbandono/blob/9da939b8cc31d79d749f606ba4ffda29920758bd/consultaModelo.sql

Repositorio git (2023b). Código, latest_version_data2.py
https://github.com/mpustil/modeloAbandono/blob/a952843a5e01212182eaf4080d4d18abd98c3d1a/latest_version_data2.py

Repositorio git (2023c). Código, modelos_y_metricas.py
https://github.com/mpustil/modeloAbandono/blob/6b945bde4a8055052557ba635208e22570316995/modelos_y_metricas.py

Ríos G. (2010). Factores sociodemográficos y rendimiento académico en la Universidad: el caso de estudiantes de abogacía de la Universidad Nacional de Córdoba. (Tesis inédita de doctorado). Universidad Nacional de Córdoba.

Santos G. A. S., Belloze K. T., Tarrataca L., Haddad D. B., Bordignon A. L., Brandao D. N. (2020). EvolveDTree: Analyzing Student Dropout in Universities. International Conference on Systems, Signals, and Image Processing, 2020-July, 173–178. https://doi.org/10.1109/IWSSIP48289.2020.9145203

Schröer C., Kruse F., Gómez J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181, 526–534.

Schumacker R. y Lomax R. (2018). A beginner's guide to structural equation modeling (4th ed.). Routledge.

Secretaría de Evaluación y Planeamiento, UNAHUR (2022a). Informe estudiantes UNAHUR 2020-2021. https://unahur.edu.ar/wp-content/uploads/2022/07/Informe-estudiantes-UNAHUR-2020-2021.pdf

Secretaría de Evaluación y Planeamiento, UNAHUR (2022b). Informe sociodemográfico ingresantes UNAHUR 2016-2021. https://unahur.edu.ar/wp-content/uploads/2022/07/Informe-perfiles-ingresantes-1_2021.pdf

Secretaría de Evaluación y Planeamiento (2023). Informe estudiantes UNAHUR 2022. https://unahur.edu.ar/wp-content/uploads/2023/09/Informe-estudiantes-UNAHUR-2022-2023.pdf

Shannon C.E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal. 27 (4): 623–656. doi:10.1002/j.1538-7305.1948.tb00917

Solanas A., Salafranca L., Fauquet J., Núñez M. I. (2005). Estadística descriptiva en Ciencias del Comportamiento. Madrid: Thompson.

Sosa Escudero W., Giovagnoli P., Porto A. (2009). The effects of individual characteristics on the distribution of college performance. Económica, La Plata, Vol. LV, enero-diciembre, pp. 99-130.

Spady W. (1970). Dropouts from higher education. An interdisciplinary review and synthesis. Interchange. Vol. 19, Nº 1: 109-121.

Terenzini P.T., Ro H.K., Yin A.C. (2010). Between-college effects on students reconsidered. Ponencia presentada en el congreso de The Association for the Study of Higher education. Indianapolis. Recuperado de: http://www.usc.edu/programs/cerpp/docs/ASHEPAPER-BetweenCollegeEffectsTerenzini.pdf

Tinto V. (1975). Dropout From Higher Education: A Theoretical Synthesis of Recent Research, Journal of Higher Education. Nº 45: 89-125.

Tinto V. (1982). Limits of theory and practice in student attrition. The Journal of Higher Education 53(6), 687–700, http://www.jstor.org/stable/1981525

Tinto V. (1987). El abandono de los estudios superiores: una nueva perspectiva de las causas del abandono y su tratamiento. Universidad Nacional Autónoma de México, Asociación Nacional de Universidades e Instituciones de Educación Superior. 55 pp.

Tinto V. (1989). Definir la deserción: una cuestión de perspectiva. Revista de Educación Superior Nº 71, ANUIES, México.

Tinto V. (1993). Leaving College: Rethinking the Causes and Cures of Student Attrition, 2a Edition, Chicago: University of Chicago Press. 246 pp.

Tukey J.W. (1977). Exploratory Data Analysis. Addison-Wesley.

Universidades Hoy (2021). Inteligencia Artificial aplicada al marketing: una fórmula para potenciar la Educación Superior. https://universidadeshoy.com.ar/nota/72250/inteligencia-artificial-aplicada-al-marketing-una-formula-para-potenciar-la-educacion-superior/

Vapnik V. (2000). Statistical Learning Theory. Willey.

Xu Q.-S., y Liang Y.-Z. (2001). Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems, 56(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2