MODELOS PARA LA PREDICCIÓN DEL ABANDONO EN LA UNIVERSIDAD NACIONAL DE HURLINGHAM

Martin Pustilnik[1], Gianluca Ndukanma[2]

[1] CIDIA: Centro de Investigación y Desarrollo en Informática Aplicada, Universidad Nacional de Hurlingham

[2]Villa Maria College, Buffalo, New York

martin.pustilnik@unahur.edu.ar; ndukanmagianluca14@gmail.com;

RESUMEN

En la Universidad Nacional de Hurlingham (UNAHUR) se lleva a cabo desde 2019 un proyecto que busca disminuir el abandono en su población estudiantil.

El mismo tiene dos líneas de trabajo. La primera es generar un "Sistema de recomendación", a través de un software que realiza recomendaciones a los estudiantes de materias a cursar, a partir de la historia académica individual. Una segunda línea de trabajo consiste en el "Estudio de las características de los estudiantes", para la identificación de indicadores de riesgo de abandono, mediante la aplicación de ciencia de datos. Se toma como información la bases de datos del SIU-Guaraní[1] de la UNAHUR y se busca predecir el abandono de manera temprana para intervenir y asistir a los alumnos antes de que se produzca.

Desde 2021 se alcanzan los primeros resultados, a partir de los datos disponibles en el SIU- Guaraní que permitieron generar miles de mensajes personalizados para los estudiantes con recomendaciones de cursada. Durante 2022 se elaboraron modelos de predicción de abandono. Se muestran los resultados obtenidos hasta la fecha.

Palabras clave: Modelo de Predicción, Abandono Universitario, Aprendizaje Automático.

CONTEXTO

Uno de los primeros proyectos impulsados por el CIDIA fue el proyecto de investigación "Estrechando el contacto entre universidades estudiantes: comunicación ante posibles casos de deserción, propuestas para la inscripción", aprobado en la convocatoria PIUNAHUR 6[2]. cuyos resultados se publicaron en [1].

En 2022 se continúa investigando bajo el 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." en el marco de la convocatoria PIUNAHUR 8.

Esta iniciativa fomenta la integración de la comunidad de Informática dentro del ámbito de la UNAHUR.

En 1982 Tinto [2] define el abandono cuando un estudiante aspira a concluir su proyecto educativo pero no lo logra. Considera desertor a aquel estudiante que no presenta actividad académica durante tres semestres

[1] Sistema de gestión universitaria Guaraní: https://www.siu.edu.ar/siu-guarani

[2] Se aprobaron proyectos de la convocatoria PIUNAHUR6 | UNAHUR

consecutivos. Este comportamiento se denomina "primera deserción" o en inglés (*first drop-out*), ya que no se puede determinar si pasado este periodo, el individuo retomará sus estudios o cambiará de universidad.

Para esta situación existe un consenso en definirla como un abandono voluntario que puede ser explicado por diferentes categorías de variables: socioeconómicas, individuales, institucionales y académicas. Sin embargo, la forma de operacionalizar las mismas depende del punto de vista desde el cual se haga el análisis; esto es, individual, institucional y estatal o nacional [3].

En la bibliografía clásica (Pascarella & Terenzin, 1980) [4] suelen utilizar los términos abandono o deserción[3] indistintamente .

En español deserción tiene una connotación marcial y da a entender que el fenómeno es responsabilidad principalmente del estudiante.

Tinto en 1989 [5] afirma que el estudio de la deserción en la educación superior es extremadamente complejo, ya que implica no sólo una variedad de perspectivas, sino que, además, una gama de diferentes tipos de abandono. Adicionalmente, afirma que ninguna definición puede captar en su totalidad la complejidad de este fenómeno, quedando en manos de los investigadores la elección de la definición que mejor se ajuste a sus objetivos y al problema a investigar.

Para el presente trabajo consideramos una situación de ***abandono*** **a aquel alumno que** habiendo comenzado sus estudios, no presenta actividad por al menos un cuatrimestre ya sea porque no continua sus estudios o porque cambió de universidad, pudiendo volver más adelante.

[3] *Deserción* viene del latin ***desertio***: "acción y efecto de abandonar las obligaciones".

El modelo que aquí se presenta, se realiza con colaboración y asistencia de la Secretaría Académica y el área de Orientación estudiantil.

1. INTRODUCCIÓN

En varias universidades nacionales de la Argentina, interesa contar con una gestión que asista y acompañe a cada estudiante en su trayectoria académica. En general, se busca garantizar en la práctica el derecho a la educación y propender al éxito de la mayor cantidad de estudiantes, atendiendo los desafíos que se derivan de las características de la población estudiantil [1].

En la Tabla 1 mostramos la cantidad total de alumnos e inscriptos en 2014 comparado con la cantidad de egresados en 2020 (#Alum. 2014; # Insc. 2014; #Egr. 2020) en las universidades nacionales Argentinas.

#Alum. 2014	#Insc. 2014	#Egr. 2020	%Egr. 2014-2020
1.871.445	445.763	122.679	

Tabla 1. Porcentaje de egresados (%Egr.) en 2020 en un periodo de 6 años. Datos publicados por el departamento de información universitaria (2021) [6].

Con menos del 6% de egresados en períodos de 6 años (Ver Tabla 1), entendemos que el abandono estudiantil es, tal vez, el factor individual que conspira en mayor medida contra el establecimiento de la educación universitaria como un derecho de nuestros jóvenes y adultos, y contra el rol de la universidad como un motorizador de movilidad social para los sectores menos favorecidos. Este fenómeno se manifiesta en

forma generalizada en la mayor parte de las universidades públicas, y en forma particularmente aguda en las del conurbano bonaerense [1].

Una característica relevante es el crecimiento significativo de la población estudiantil, que en algunas universidades reviste características explosivas. UNAHUR, por ejemplo, contaba con casi 5.000 estudiantes en 2017 mientras que en 2022 contaba con casi 35.000.

Otra característica relevante es que una amplia proporción de los estudiantes cuenta con poco, o nulo, conocimiento sobre el funcionamiento de una universidad, y sobre lo que implica cursar una carrera de nivel universitario. Este fenómeno se deriva, al menos en parte, de la gran cantidad de estudiantes que son primera generación de universitarios en su familia [1].

Las dos características mencionadas están fuertemente presentes, en particular, en la gran cantidad de estudiantes que abandonan sus estudios, especialmente antes de completar el primer año de la carrera elegida. Entre las razones de este fenómeno, mencionamos la frustración que genera un bajo desempeño inicial, y la dificultad por adquirir los hábitos necesarios para transitar una carrera universitaria con altas chances de éxito [7;9].

El segundo rasgo es la complejidad que reviste la definición de la oferta de cursos en la universidad. Uno de los motivos es que muchos estudiantes se inscriben en más materias, de las que su situación objetiva les permite cursar correctamente. Esto provoca altas cifras de abandono de cursos, y eventualmente también abandono de los estudios; por lo que hay una vinculación entre las dos cuestiones elegidas.

Estas dos problemáticas pueden ser atenuadas mediante la comunicación directa de distintos actores de la universidad (entre ellos docentes, directivos, personal ligado a la gestión académica, tutores) con cada estudiante, de modo de generar un vínculo de cercanía de los estudiantes con la universidad.

Por otro lado, el fenómeno de crecimiento mencionado antes atenta contra la posibilidad de mantener una comunicación fluida con cada alumno en particular, ya que a medida que la población estudiantil crece, no es acompañado por un crecimiento análogo del personal que puede participar en el fortalecimiento del vínculo.

Un recurso muy valioso al respecto es la gran cantidad de información acumulada sobre la trayectoria y el comportamiento de cada estudiante, tanto desde la plataforma de acompañamiento al aprendizaje(Moodle) como del sistema SIU-Guaraní de donde se obtienen tanto datos personales como la inscripción a las materias y la asistencia a clases. El análisis de este gran volumen de información, usando técnicas de ciencia de datos, puede asistir a la gestión académica en varios aspectos. Asistir en la detección de estudiantes que estén en riesgo de abandonar sus estudios o que los han abandonado recientemente, y también en la generación de propuestas de inscripción personalizadas para cada estudiante, basadas en el historial del rendimiento académico.

Adicionalmente, la posibilidad de enviar mensajes mediante redes sociales, el email o al celular, aumenta en gran medida la capacidad de una institución universitaria para comunicarse efectivamente con su población estudiantil.

Un primer objetivo se logró en 2021 con un software que brinda herramientas para potenciar la comunicación de la Universidad con sus estudiantes, generando mensajes que puedan llegar a una gran cantidad de ellos/as, con contenidos personalizados de acuerdo a la trayectoria académica de cada uno/a.

Un segundo objetivo se desarrolló en 2022.

Implementamos un modelo de predicción del abandono universitario testeado con datos de ese año y se espera poner en funcionamiento en 2023.

El modelo nos brinda una lista **acotada** de estudiantes en riesgo y nos brinda los motivos más probables de abandono para cada alumno. Esto nos permite brindar una comunicación personalizada con cada uno de ellos.

Es importante distinguir que los motivos que predice el modelo están basados en los datos disponibles, y en general existen otros motivos subyacentes a analizar para cada alumno.

El listado generado por el modelo nos sirve como punto de partida para iniciar la comunicación y averiguar los motivos subyacentes, sin tener que analizar la totalidad del alumnado.

El **objetivo** general consiste en definir y desarrollar aplicaciones informáticas que permitan contribuir al abordaje institucional de problemáticas de la población estudiantil de la UNAHUR.

2. MODELOS

Para poder personalizar la comunicación y el seguimiento a los estudiantes, implementamos modelos con cuatro herramientas conocidas de Aprendizaje Automático: XGBoot [14], Logistic Regression [15], Support Vector Machines [16] y Decision Trees [16] como modelo de control, para analizar los resultados [10;12].

Dichos modelos se entrenaron con variables que se obtienen del sistema Guaraní.

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

Tabla 2: Grupos de variables del primer modelo.

Algunas variables surgen de un censo (C_) que los alumnos completan de manera opcional (Ver Tabla 2), por eso tenemos una dirección “inicial” y otra declarada en el censo.

El modelo tiene además como datos de entrada otras variables que fueron calculadas a partir de las evaluaciones realizadas por los alumnos

durante su cursada en cuatrimestres anteriores (Ver Tabla 3).

Variable calculada	Descripción
#Eval 2020C1	Cantidad de evaluaciones rendidas en 2020C1
#Eval 2020C2	C2 = Cuatrimestre 2
#Eval 2021C1	
#Eval 2021C2	

Tabla 3: Variables calculadas.

Variable a predecir (Clase):

Una vez entrenado el modelo se le suministra la base de datos de alumnos de 2021 para generar una predicción para 2022.

La predicción consiste en el listado de alumnos con su probabilidad de abandono para 2022 y la variable más importante para cada caso (Ver Tabla 4). La probabilidad surge de la cantidad de instancias que abandonan en el nodo del modelo entrenado.

Alumno	Proba_ bilidad	Variable más importante
724234	0,91	No rinde hace varios cuatrimestres
008383	0,90	No rinde hace varios cuatrimestres
…		
922524	0,62	No completo censo
629341	0,61	No completo censo

Tabla 4: Listado de alumnos con probabilidad de abandono > 0,6 (P > 0,6).

La Tabla 4 no es suficiente para determinar el motivo de abandono, que además, es multicausal.

La mayoría de los motivos aún no fueron censados, como por ejemplo, **si el alumno consiguió trabajo**.

La lista acotada le sirve a los actores involucrados para contactar al alumno y profundizar su vínculo con la universidad, sin tener que contactar a todos los alumnos al mismo tiempo.

El umbral de abandono (P > 0,6) es un parámetro del modelo que los actores pueden mover para agrandar o achicar la lista.

La definición operativa de cuándo se considera a un estudiante en riesgo de abandono, y el cálculo de la inscripción propuesta, involucran un trabajo en conjunto entre especialistas en educación y de gestión académica.

Preprocesamiento de los datos:

Algunos atributos de la base venían con errores o datos faltantes. Mientras no superaran el 5% y fuesen aleatorios (*Missing Completely At Random*)[4], se optó por reemplazar los datos faltantes por la media. En el caso de datos no aleatorios (*Missing At Random*)[5], se optó por elegir al vecino más cercano. En la Tabla 5 mostramos algunos ejemplos.

Para datos atípicos, como edades fuera del rango [16…75] se decidió reemplazar por la media por ser un dato aleatorio.

Se unificaron datos cuando fue posible, como en caso de la dirección. A veces "planta baja" es informado como "piso nro 0". En ese caso se unifican ambos de la misma manera (se unifica como "piso nro 0"). Para el piso, si además es un dato faltante, se imputó el valor

[4] Missing Completely At Random (MCAR): El dato faltante NO se correlaciona con otros registros o atributos.

[5] Missing At Random (MAR): El dato faltante se correlaciona con otros registros o atributos.

del vecino más cercano utilizando el método k-nearest neighbors(KNN) [17] en términos de atributos más cercanos.

Atributo	**Reemplazo/ Tratamiento**
#Hijos; #Familiares	Media
Dirección(piso)	Unificación
Dirección(piso)	Vecino más cercano
Edad	Atípicos reemplazados por media

Tabla 5: Reemplazo de datos faltantes.

3. RESULTADOS OBTENIDOS/TRABAJO A FUTURO

Como la variable a predecir (Clase) no está balanceada (el 80% de los alumnos abandonan) no se utilizaron métricas como ***Accuracy*** o Matriz de confusión porque darían valores exagerados (i.e **Accuracy>0,80**).

En todo dominio/contexto se fija un criterio de que error queremos minimizar:

Asumimos que el "caso normal" para la **hipótesis nula(H0) es que un alumno no abandone** [18].

En educación es difícil decidir si queremos minimizar el Error de tipo I/Falso positivo (rechazamos H0 pero en realidad no abandona) o el Error de tipo II/Falso negativo (asumimos H0 pero en realidad abandona).

Si predecimos un "falso abandono" la universidad podría destinar recursos a alumnos que en realidad no iban a abandonar (en detrimento de los que sí), mientras que si predecimos un "falso no abandono" no tomaría ninguna acción contingente para evitarlo.

Por esos motivos utilizamos métricas que consideren falsos positivos en conjunto con falsos negativos, como el **Área Bajo la Curva ROC(AUC)** [19] y la **Exactitud Balanceada** de cada modelo.

En cada punto de la curva ROC se representa la sensibilidad para el eje 'Y' y especificidad (Ver Fórmula 1) para el eje 'X'. El área debajo de la curva representa cuán bien se están tratando los falsos negativos/positivos del modelo.

$$Sensibilidad = \frac{VP}{VP + FN} \; Especificidad = \frac{VN}{VN + FP}$$

Fórmula 1. VP:Verdadero positivo; FN:Falso negativo; VN:Verdadero negativo; FP:Falso positivo.

La exactitud Balanceada se obtiene calculando el umbral óptimo en donde #falsos positivos = #falsos negativos.

Existen otras técnicas para balancear la Clase, como el submuestreo , pero se corre el riesgo de sobre ajustarse a los datos por la pérdida de registros sumada a los datos faltantes. Para evitar estos problemas se pueden utilizar técnicas como Bagging o de Bootstrapping [24] que dejaremos como trabajo a futuro.

Durante 2022 se testaron los resultados de los modelos respecto de la inscripción efectiva de los alumnos. En la Tabla 6 se muestra el AUC y el Umbral Óptimo para llegar a la Exactitud Balanceada de cada modelo.

Modelo	**AUC**	**Exactitud Balanceada**	**Umbral Óptimo**
1_Decision Tree (Anonimizado)	0,79	0,73	0,73
2_Decision Tree	0,80	0,74	0,75
3_ Logistic Regression	0,68	0,64	0,46
4_ Support Vector Machine	0,74	0,71	0,83

(SVM)*			
5_ XgBoost**	0,83	0,75	0,89

Tabla 6: Performance de los modelos. *Kernel =Radial Basis Function(RBF); Gamma = 1/#Features; C = 1,0. **Optimización = ROC.

Para proteger la privacidad, el primer modelo (1_) consistió en una base anonimizada de alumnos. Se quitaron los atributos relativos al email, la edad y la dirección(calle, barrio, etc). Como se puede observar, ese recorte generó modelos de menor performance.

Por ese motivo se decidió eliminar la copia de los datos luego de entrenar el modelo para proteger la información y poder trabajar con los datos reales (Modelo 2_ en adelante).

El Modelo 3_ obtuvo el peor desempeño para modelos no anonimizados.

Las regresiones logísticas al igual que en las regresiones lineales, no funcionan bien cuando los datos no se correlacionan linealmente con la *clase* [20].

En nuestro dataset tenemos alumnos que se inscriben a pocas materias y abandonan, otros se anotan a varias y no abandonan, lineal hasta acá. Pero luego hay alumnos que se anotan a varias materias e igual abandonan. Esta falta de linealidad "confunde" al modelo, mientras que otros, basados en árboles son más resistentes a estos cambios.

El modelo 4_ (SVM) se ejecutó con el kernel Radial Basis Function (RBF) [25] ampliamente usado en la bibliografía, cuando la clase no es linealmente separable. Queda pendiente la optimización de hiperparametros(Gamma y C) para lograr una mejor performance. Se realizó la optimización para el modelo más potente.

XgBoost (Modelo 5_) es una herramienta muy usada de aprendizaje automático a partir de 2014. Como los modelos 3_ en adelante trabajan con datos numéricos, se transformaron los datos categóricos(con una **codificación binaria**) para poder entrenarlo. Otras transformaciones como *hot-encoding* generan demasiados atributos para variables "largas" como 'Apellido'.

Comparado con los otros métodos, se observa que fue el de mejor desempeño (Gráfico 1) luego de optimizar los hiperparámetros mediante Random Search [21].

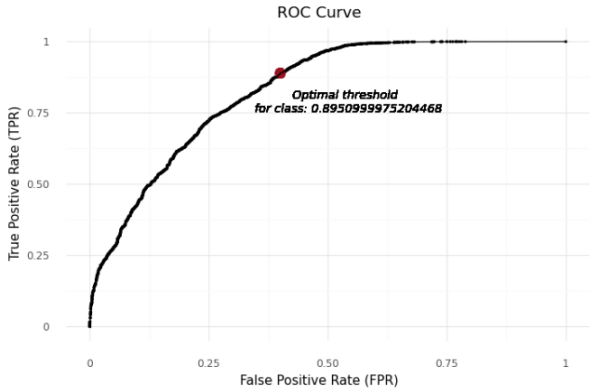

Gráfico 1. Curva ROC XgBoost: AUC = 0,83; BA = 0,75.

Abandono condicional:

Otro resultado fue el análisis exploratorio de los atributos.

Encontramos que los alumnos con dominio @unahur.edu.ar abandonan menos (64%). Creemos que el mail institucional denota una mayor adaptación a la vida académica que no tenerlo. Los alumnos con cuenta @hotmail.com por otro lado abandonan más (82%). En este caso notamos que ese dominio es más antiguo que el promedio y está correlacionado con la edad del alumno.

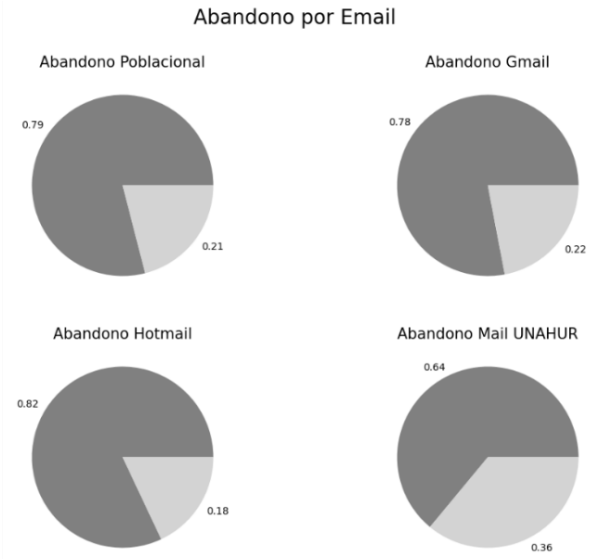

Gráfico 2. Porcentaje de abandono según dominio de email.

En el caso de los turnos también tenemos comportamientos distintos. En el turno noche (80%) hubo más abandono que en turno tarde (78%), pudiendo responder a variables latentes[6] como la 'Situación laboral' entre otras. (Gráficos 2 y 3).

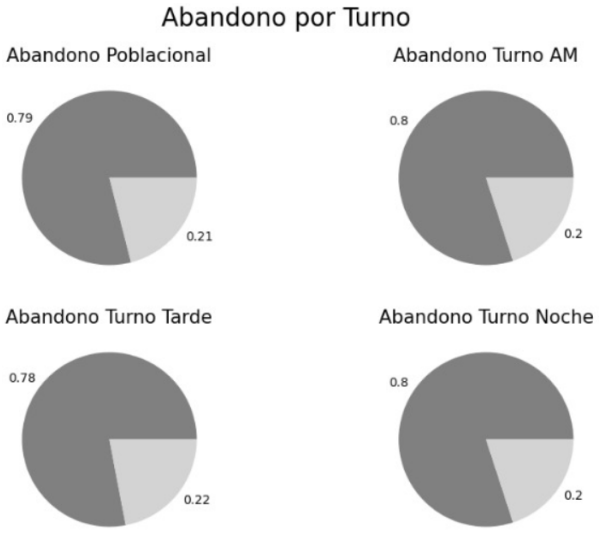

Gráfico 3. Abandono por turno.

Para la edad y la cantidad de hijos también encontramos diferencias. Los alumnos con 3 o más hijos abandonan casi un 2% más que los alumnos con 2 hijos o menos. Si bien hay estudios que demuestran que las tareas del hogar aun recaen en su mayoria sobre las mujeres [13], analizando el sexo, no encontramos que algún género abandone mas que otro en forma estadisticamente significativa (p-valor < 0,05).

Para la edad ocurrió algo similar. Hay casi un 2% más de probabilidad de abandono para los percentiles más altos (Gráficos 4 y 5).

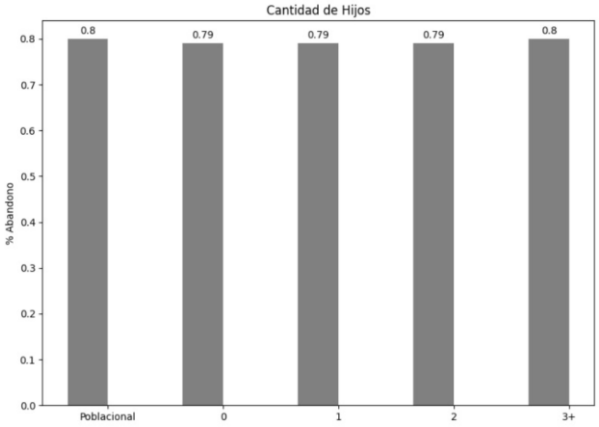

Gráfico 4. Abandono según la cantidad de hijos.

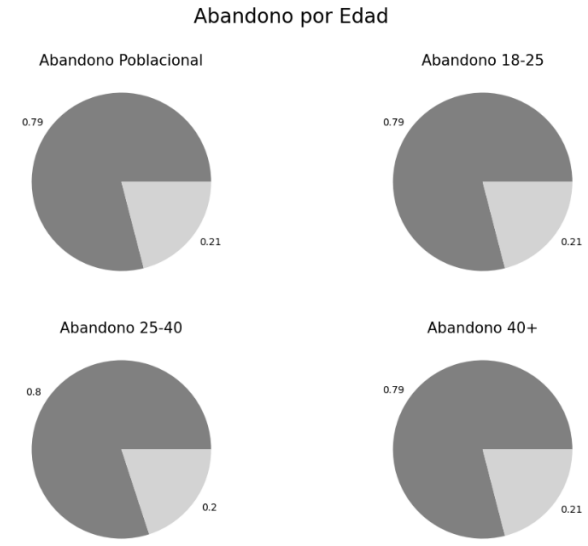

Gráfico 5. Abandono según la edad.

[6] Variables explicativas, que son reflejadas de manera indirecta en el modelo por otras variables.

Variables más importantes:

Se utilizó el método Mean Decrease in Impurity (MDI) [22] para calcular la importancia de las variables para los modelos basados en árboles. MDI cuenta la cantidad de veces que se utiliza dicha variable para partir un nodo ponderado por la cantidad de instancias en esa partición (Gráfico 6).

Encontramos que 'Cantidad de evaluaciones' en los últimos periodos, 'Hace cuánto completo el censo', el 'Turno preferido', la 'Edad', el 'Tipo de vivienda', el 'Estado Civil' y 'Cantidad de hijos' son importantes para explicar el abandono en estos modelos. En algunos casos la importancia se debe al desbalance de la clase. La mayoría de los estudiantes no son celiacos (91%)[7] y esto a su vez correlaciona con abandono (entonces tiene mucha importancia), pero tiene poca importancia para discriminar 'No abandono'.

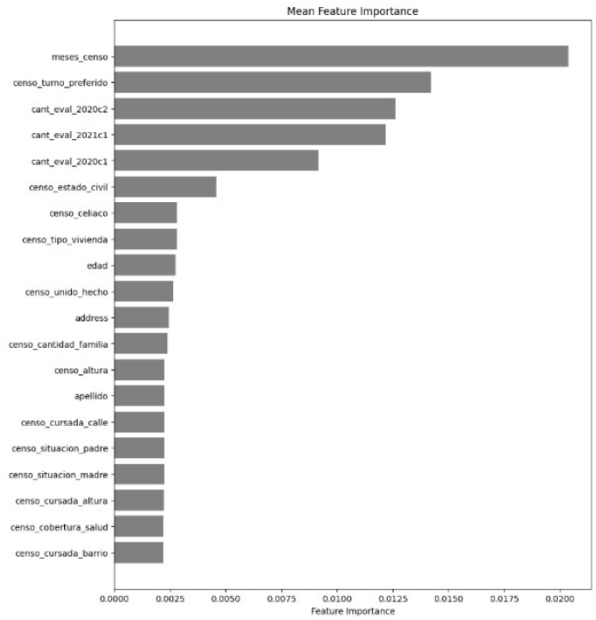

Gráfico 6. Variables más importantes (método MDI: Mean Decrease in Impurity).

5. TRABAJO A FUTURO

En la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [23] se explica que **el modelado es un proceso iterativo e incremental**.

Para la segunda iteración de los modelos queda pendiente:

-Agregar variables de Guaraní y de **Moodle**[8] como 'Cantidad de horas por semana que trabaja', la 'Asistencia a clase' y 'Fecha de inscripción'. Y variables calculadas, como el 'Tiempo de viaje' hasta el campus.

-Entrenar el modelo mediante Bagging y/o Bootstrapping , de manera que la clase quede balanceada.

-Optimizar los hiperparametros, como el C para SVM o el tamaño de los árboles para XgBoost.

Uno de los resultados esperados es generar políticas que permitan disminuir el abandono estudiantil. Los resultados de este proyecto pueden contribuir a profundizar líneas de investigación aplicada sobre ciencia de datos para el análisis de la población estudiantil.

El proyecto resulta un punto de encuentro entre distintas áreas de la UNAHUR, e incluso con actores de otras universidades, siendo un vehículo para establecer lazos que permitan ulteriores iniciativas conjuntas.

Se han realizado varias actividades de difusión del proyecto en la UNAHUR y para ello se ha preparado un video que agiliza la su difusión y resume sus características principales [26].

5. FORMACIÓN DE RECURSOS HUMANOS

La línea de investigación presentada colabora en la formación de varios estudiantes de la carrera Licenciatura en Informática y otras

[7] Valor aproximado. 23,24% de datos faltantes.

[8] Moodle: Acerca de Moodle

universidades a través de la modalidad de pasantía en el CIDIA, HUNAHUR. Durante 2022 Gianluca Ndukanma realizó una pasantía internacional en conjunto con la Universidad de Villa Maria.

6. BIBLIOGRAFÍA

[1] Pustilnik, y otros. (2022). Estrechando el contacto entre universidades y estudiantes: comunicación ante posibles casos de abandono, propuestas para la inscripción. Líneas de investigación y desarrollo del CIDIA. Libro de Actas XXIV del Workshop de investigadores en Ciencias de la Computación (WICC) 2022. *RFUSMA Ediciones, 2022. ISBN: 978-987-48222-3-9:734-738.* https://libros.unlp.edu.ar/index.php/unlp/catalog/book/2015.

[2] Tinto (1982). Limits of theory and practice of student attrition. Journal of Higher Education. Vol. 3, N° 6: 687-700.

[3] Peralta. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. http://dx.doi.org/10.4067/S0718-07052008000200004

[4] Pascarella & Terenzini (1980). Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model. The Journal of Higher Education, Volume 51, 1980 - Issue 1.

[5] Tinto (1989). Definir la deserción: una cuestión de perspectiva. Revista de Educación Superior N° 71, ANUIES, México.

[6] Departamento de información universitaria (DIU) 2021. https://www.argentina.gob.ar/sites/default/files/sintesis_2020-2021_sistema_universitario_argentino.pdf.

[7] Arias, M. F., Mihal, I., Lastra, K., & 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*, 20(64), 47-69.

[8] Chávez, M. J. (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). https://repositorio.flacsoandes.edu.ec/bitstream/10469/16829/2/TFLACSO-2020MJC.pdf

[9] Mendonça, M. (2021). Una aproximación a las estrategias institucionales para lograr la permanencia de los estudiantes en las nuevas universidades del conurbano (2009-2016). Espacios en Blanco. Revista de Educación, 2(31), 275-286.

[10] Hadley Wickham (2019), Advanced R, Second Edition (Chapman & Hall/CRC The R Series).

[11] Brett Lantz (2018). Machine Learning with R: Expert techniques for predictive modeling.

[12] Felie Munizaga, Maria Beatriz Cifuentes Orellana (2018). Retención y Abandono Estudiantil en la Educación Superior Universitaria en América Latina y el Caribe: Una revisión sistemática.

[13] Marquez (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

[14] Chen, Tianqi; Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System". In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM. pp. 785–794. arXiv:1603.02754. doi:10.1145/2939672.2939785.

[15] M. Strano; B.M. Colosimo (2006). "Logistic regression analysis for experimental determination of forming limit diagrams". International Journal of Machine Tools and Manufacture. 46 (6): 673–682. doi:10.1016/j.ijmachtools.2005.07.005.

[16] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning : Data Mining, Inference, and Prediction (PDF)

[17] Mucherino y otros (2009). k-Nearest Neighbor Classification. isbn:978-0-387-88615-2. doi:10.1007/978-0-387-88615-2_4.

[18] Fisher, R.A. (1935), The Design of Experiments. https://home.iitk.ac.in/~shalab/anova/DOE-RAF.pdf

[19] Hand, D.J., & Till, R.J. (2001). A simple generalization of the area under the ROC curve to multiple class classification problems. Machine Learning, 45, 171-186.

[20] When logistic regression simply doesn't work | by Alon Lekhtman | Towards Data Science

[21] Bergstra, J.; Bengio, Y. (2012). "Random search for hyper-parameter optimization". Journal of Machine Learning Research. 13: 281–305.

[22] Perrier (2015). Feature Importance in Random Forests. https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html

[23] Azevedo & Zantos (2008). KDD, SEMMA and CRISP-DM: a parallel overview. https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview

[24] Biswal (2023). Bagging in Machine Learning: Step to Perform And Its Advantages. https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning

[25] Kumar (2020). SVM RBF Kernel Parameters with Code Examples. https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning

[26] Pustilnik (2021). Modelos de deteccion de abandono. https://www.youtube.com/watch?v=ea_wXTBM9KE&t=1145s