INGENIERÍA DE ATRIBUTOS PARA MODELOS DE PREDICCIÓN DEL ABANDONO UNIVERSITARIO EN ARGENTINA

Línea Temática: Factores, tipos y perfiles asociados al abandono – Investigación.

Martin Pustilnik, Universidad Nacional de Hurlingham, martin.pustilnik@unahur.edu.ar
Gianluca Ndukanma, Villa Maria College, ndukanmagianluca14@gmail.com

Resumen

En Argentina, según la estadística del año 2021, se estima que en el sistema universitario sólo el 27,66% de los estudiantes que ingresan se gradúa en tiempo teórico [1], muy por debajo de otros países de la región, como en Chile, que aunque tiene una matriz universitaria diferente (mayoritariamente arancelada), se gradúa alrededor del 57% [2]. Entendemos que el abandono estudiantil es, tal vez, el factor individual más importante que explica este fenómeno.
En la Universidad Nacional de Hurlingham en Argentina (UNAHUR) se desarrollaron modelos de predicción de abandono utilizando técnicas de Aprendizaje Automático [3], con el objetivo de prevenir el abandono estudiantil de manera temprana. Estos modelos se basan en los datos del Sistema de Información Universitaria Guaraní [4]. 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 muestran los resultados de generación de nuevos atributos que permiten realizar la predicción de estudiantes en riesgo de abandono con mayor precisión.

Palabras clave: Modelo de Predicción, Ingeniería de Atributos, Abandono Universitario, Aprendizaje Automático

Contextualización: Según las estadísticas publicadas en [1], en el sistema universitario Argentino, en 2021, se graduaron en universidades públicas y privadas el 27,66% de los alumnos, considerando un tiempo teórico de 5 años. En el período 2016-2021 el porcentaje promedio de graduados en tiempo teórico fue del 28,5%. Este número es incluso más bajo en las carreras de informática, cercano al 20% [5], las cuales son actualmente de interés estratégico para el desarrollo del país. En la Tabla 1 se muestran los nuevos inscritos, Egresados y el Porcentaje de graduados en tiempo teórico:

Tabla 1: Nuevos (Nuevos Inscritos), Egresados y Porcentaje (Porcentaje de egresados en tiempo teórico de 5 años) para las universidades Argentinas (2012-2021). Fuente: Elaboración propia a partir de [1].

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%**

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 cuatro o más semestres sin actividad [6]. Si queremos emitir una alerta temprana en lugar de detectar un "abandono definitivo", consideramos una situación de abandono a aquel alumno que habiendo comenzado sus estudios, no presenta actividad por al menos un semestre ya sea porque no continúa sus estudios o porque cambia de universidad o de carrera, pudiendo volver más adelante (primera deserción [7]).

Objetivo: El objetivo de este trabajo es mejorar los modelos de predicción de abandono que desarrollamos en UNAHUR, incorporando nuevas variables y comparando los resultados con los modelos anteriores. Nuestra intención es generar variables calculadas a partir de las ya existentes mediante técnicas de ingeniería de atributos, y probar hipótesis para identificar qué variables influyen en el abandono, para poder intervenir de manera temprana y asistir a los alumnos antes de que abandonen, permitiéndonos así iniciar la comunicación y explorar los motivos subyacentes sin tener que censar a toda la población estudiantil. Además, pretendemos proporcionar recomendaciones sobre qué variables latentes deberían ser censadas, basándonos en la tabla de resultados que indica los estudiantes con mayor riesgo de abandono y las recomendaciones de la bibliografía,

Método: Implementamos la metodología Cross Industry Standard Process for Data Mining (CRISP-DM) [8], que consiste en sucesivas iteraciones ajustando las variables en cada iteración para cada modelo.
En la Tabla 2 se muestran las variables (o grupos de variables) de SIU-Guaraní obtenidos en la primera iteración obtenidos a partir de la bibliografía o de las sugerencias de actores de UNAHUR, bajo el marco teórico propuesto en [9]:

Tabla 2: Tipo y Descripción de cada Variable del primer modelo, bajo el marco teórico propuesto en [9].

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

Preprocesamiento

Dado que la base tenía valores erróneos como edad con valor 0, realizamos un preprocesamiento. Para variables como 'Edad' y 'Meses Censo' se eliminaron datos atípicos extremos según la siguiente definición: Un dato d es atípico extremo sii $d > Q_3 + 3 * DIQ$ o $d < Q1 - 3 * DIC$, dónde Q_3 es el tercer cuartil y DIQ es la distancia entre Q_1 y Q_3, del test de Tukey [10]. La Figura 1 muestra los histogramas de frecuencias absolutas con y sin datos atípicos para la 'Edad'.

Cuando fue posible se corrigió el ruido, como la variable categórica 'Número de Piso', en donde se unificaron los valores "planta baja" y "piso 0". Los datos alfanuméricos se pasaron a mayúscula. Se descartaron las variables con datos faltantes, excepto cuando la correlación entre registros permitió realizar imputaciones mediante la técnica k-nearest neighbors (KNN) [11]. Cuando no teníamos el barrio, se sustituye por el barrio más cercano al partido y calle informado.

Para los modelos que solo utilizan datos numéricos (i.e Logistic Regression, XgBoost) se transformaron los datos categóricos en numéricos.

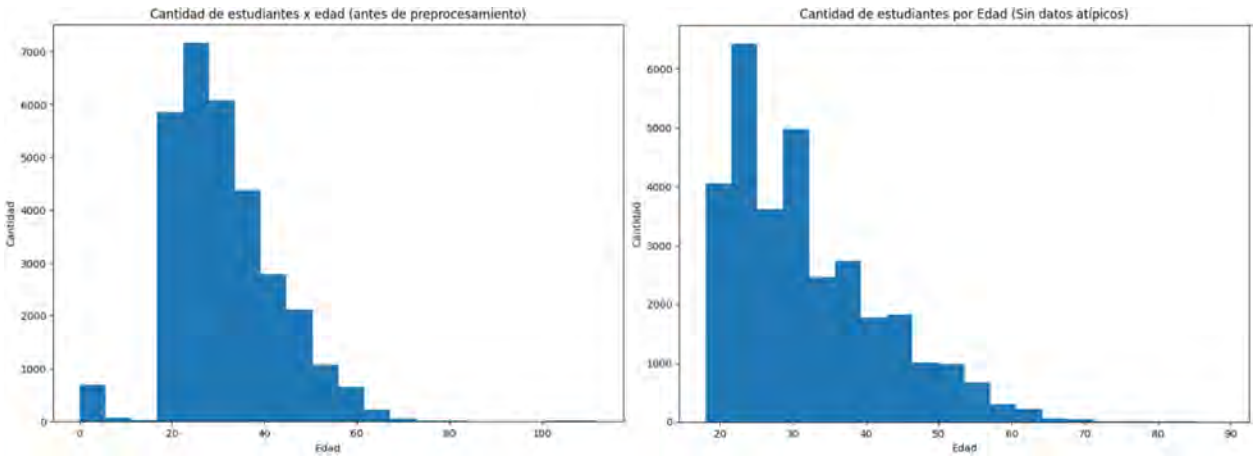

Figura 1: Frecuencia absoluta de la variable 'Edad' antes y después de corregir los datos atípicos. Fuente: Elaboración propia.

Entrenamiento

Utilizamos herramientas de Aprendizaje Automático como XgBoost [12] para entrenar los modelos.

Para el caso de Árboles de Decisión [13] se separó de manera aleatoria el 70% de los datos como entrenamiento, se repitió este experimento 10 veces. Las métricas se calcularon a partir del promedio de las 10 repeticiones. Para XgBoost las repeticiones las hace el propio algoritmo.

Como cada persona puede estar inscripta en varias carreras, resultando en varios estudiantes por persona, la predicción se realizó por persona (i.e la persona abandona todas las carreras) y se consideró un total de 30.000 personas. Una vez entrenado el modelo se le suministra la base de datos de personas de 2021 para generar una predicción para 2022. La predicción consiste en el listado de personas con 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, lo cual se muestra en la Tabla 3. El listado está anonimizado para proteger los datos personales.

En el caso de modelos basados en árboles, la probabilidad surge de la proporción de personas que abandonan en la hoja donde se clasifica cada persona. El umbral (U) de abandono (i.e

U = 0,6) es un parámetro del modelo que los usuarios pueden modificar para agrandar o reducir la lista, para no tener que contactar a todas las personas al mismo tiempo. U 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.

Tabla 3: Resultado del modelo: listado de personas con probabilidad de abandono ≥ 0, 6 (U = 0, 6). Fuente: Elaboración propia.

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

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 (variables latentes) 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.

Ingeniería de Atributos e hipótesis

Para la segunda iteración aplicamos ingeniería de atributos para generar nuevas variables, con el propósito de testear diversas hipótesis. La Tabla 4 muestra las variables calculadas y su descripción. El tiempo de viaje se estima a partir de la dirección informada y la librería de google maps.

Tabla 4: Variables calculadas para la segunda iteración. Fuente: Elaboración propia.

Variable calculada	Tipo	Descripción
#Eval2020S1	Académica	Cantidad de evaluaciones rendidas en 2020, Semestre 1
…		Se consideran los últimos 4 semestres disponibles
#Eval2021S2	Académica	Cantidad de evaluaciones rendidas en 2021, Semestre 2
TiempoViajeSiAuto	Demográfica/ Socioeconómica	Tiempo de viaje si fuera en auto, en minutos.
TiempoViajeSiPúblico	Demográfica/ Socioeconómica	Tiempo de viaje si fuera en transporte público, en minutos.

A partir de los modelos queremos testear si la cantidad de evaluaciones de los alumnos a lo largo de su historia académica y el tiempo que demora en llegar al campus son variables importantes para el abandono.

Vamos a medir el abandono condicional por todas las variables (i.e ¿hay algún turno que tenga más abandono que otro?) y mostraremos los resultados más importantes.

Los 'Tiempos de viaje' 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® y así obtener un tiempo de viaje promedio desde la dirección de la persona hacia la

universidad como muestra el esquema de la Figura 2. Se descartaron valores atípicos o direcciones inválidas.

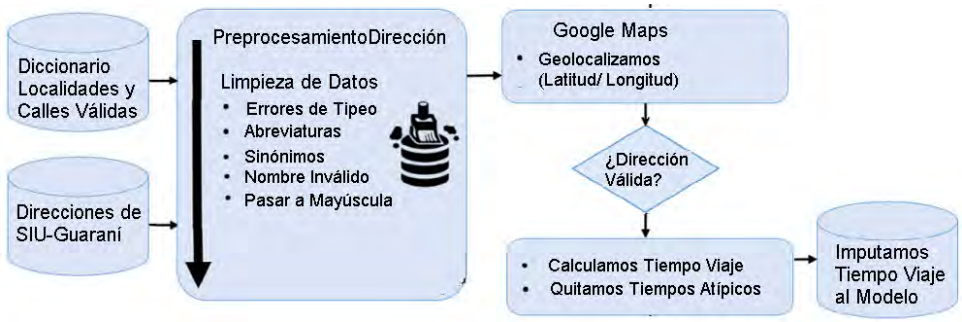

Figura 2: Esquema que muestra el procesamiento del tiempo de viaje a partir de la dirección de la persona. Fuente: Elaboración Propia.

Resultados: Se utilizó el método Mean Decrease in Impurity (MDI) [14] 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 como muestra el gráfico de la Figura 3. Se muestran las variables con MDI > 0,005.

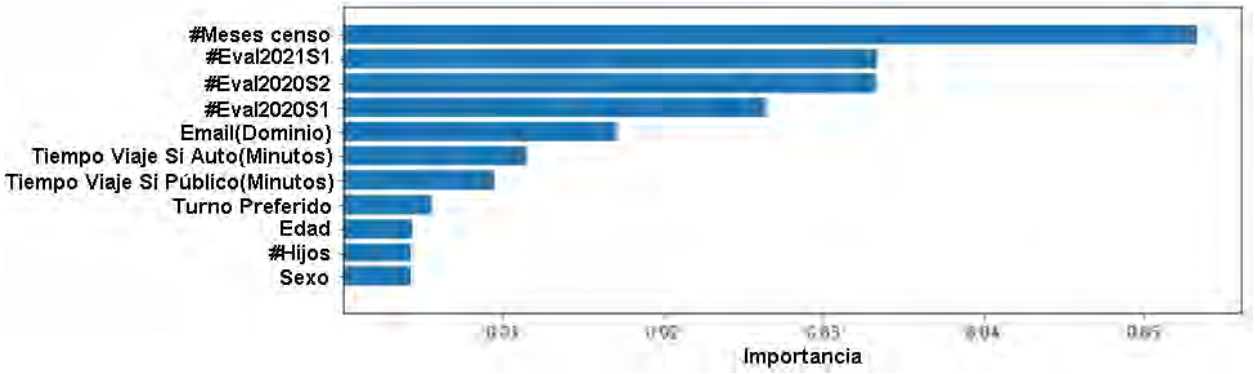

Figura 3. Las 11 variables más importantes (método MDI).

Encontramos que el modelo funciona mejor agregando las variables 'Tiempo de viaje si tuviera auto' (1) y 'Tiempo de viaje si viniese en transporte público' (2) que sin ellas. Cómo estas variables son estimadas a partir de los datos, esperamos obtener mejores resultados conociendo el modo de traslado del alumno. Las variables 'Cantidad de evaluaciones' en los últimos períodos, 'Hace cuánto completo el censo' son importantes para explicar el abandono en estos modelos. Si bien las variables socioeconómicas (i.e. 'Cantidad de hijos', 'Sexo' , 'Edad' y 'Salud (Cobertura)' y 'Tipo Vivienda') son relevantes como se explica en [15, 16, 17], los alumnos agrupados por 'Sexo' o 'Edad' registraron un porcentaje de abandono similar al poblacional en el modelo propuesto. Los alumnos de 20 años o menos registraron un abandono levemente inferior (58,2%) respecto de otras edades. Los que eligieron el turno tarde también registraron un abandono menor (59,1%) respecto de otros turnos, pudiendo responder a variables latentes como la 'Situación laboral', entre otras. Los alumnos con dominio @unahur.edu.ar abandonan menos (57,9%) que los alumnos con otros dominios de cuenta de email (poblacional 60,27%). Creemos que el mail institucional denota una mayor adaptación a la vida académica [7] que no tenerlo. Los alumnos con

dominio @hotmail.com por otro lado abandonan más (63,7%). En este caso notamos que ese dominio es más antiguo que el promedio y está correlacionado con la edad del alumno.

De un total de 33.584 alumnos matriculados en UNAHUR hasta 2021, el 56,93% no trabajaban al comenzar los estudios. En la Figura 4 se muestran los alumnos agrupados por cantidad de horas de trabajo por semana al comenzar su cursada, discretizado en tres grupos para 2021S2. El grupo azul (cantidad de alumnos que no tuvo actividad), el rojo (cantidad de alumnos si tuvieron) y el amarillo (porcentaje de alumnos que tuvo actividad). La cantidad de alumnos que no trabajaba al comienzo de su cursada fue de 14.463. De los cuales 11.956 no presentaron actividad en 2021S2, mientras que 7.165 si (37,47%). Estos valores se muestran en las líneas azul y roja, que se representan al 1% (11.956 → 119 y 7.165 → 71).

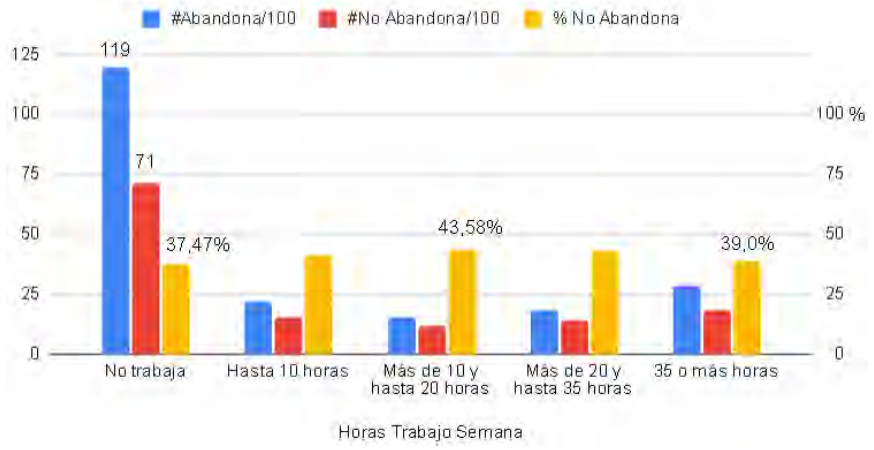

Figura 4: Cantidad de alumnos que abandona, Cantidad que continúa y Porcentaje que continúa, según la cantidad de horas de trabajo. Fuente: Elaboración propia.

Como se sugiere en [18] el mayor porcentaje de abandono no se alcanzó entre los alumnos que trabajaban, a menos que trabajen más de 35 horas por semana. Para los alumnos que trabajan entre 10 y 35 horas por semana el porcentaje de 'No abandono' fue de 43,58%.

La cantidad de becas también tuvo un impacto positivo en el abandono, el porcentaje de abandono para alumnos con becas fue del 58,36% en promedio, mientras que para los alumnos sin becas fue del 61,12%.

Métricas

Para la evaluación, se evitó el uso de métricas como Accuracy o Matriz de Confusión debido al alto índice de abandono (60,27%). Para considerar el desbalance de la variable 'Abandono' y falsos positivos en conjunto con falsos negativos, utilizamos métricas como el Área Bajo la Curva ROC (AUC) [19] y Exactitud Balanceada (EB), que se alcanza cuando: cantidad de falsos positivos = cantidad de falsos negativos.

Durante 2022 se testaron los resultados de los modelos respecto de la inscripción efectiva de los alumnos. El mejor modelo se obtuvo utilizando la herramienta de Aprendizaje Automático XgBoost, lo que resultó en un valor de AUC de 0,83 como muestra la Figura 5. Sin las variables calculadas, el valor del AUC fue de 0,78.

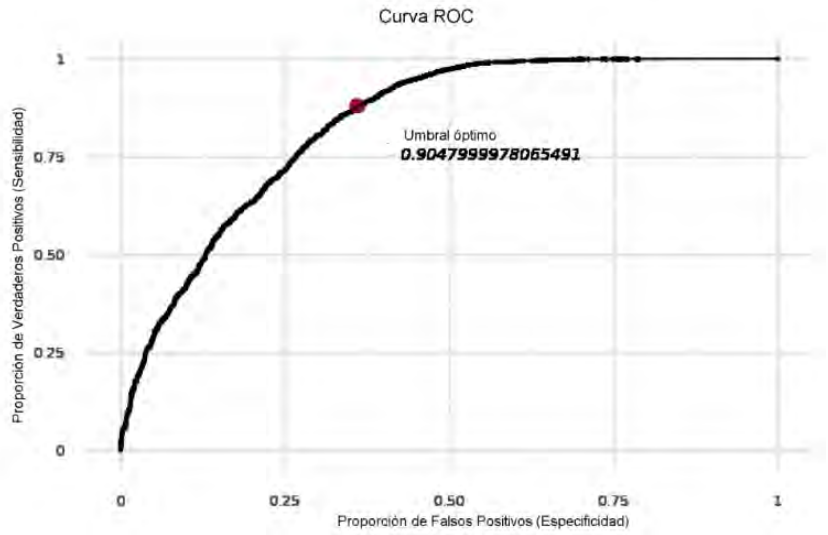

Figura 5: Curva ROC de XgBoost: AUC = 0,83. Fuente: Elaboración propia.

Conclusiones: La identificación temprana de estudiantes en riesgo de abandono permite desarrollar intervenciones más eficaces, y consideramos que el sistema de recomendación existente [20] podría adaptarse para proporcionar orientación a estos estudiantes. En este trabajo, mejoramos el desempeño del modelo de predicción de abandono en UNAHUR, incorporando variables como (1) y (2). La adición de estas variables y la aplicación de estas recomendaciones pueden potenciar aún más la efectividad del modelo.
En futuros trabajos deseamos medir:
-Esperamos obtener resultados aún mejores si conocemos el modo de traslado real de los alumnos.
-El impacto del covid-19 en la tasa de abandono, dado el cambio de cursada virtual y otros factores.
Recomendamos enriquecer el censo anual con preguntas adicionales sobre el modo de transporte, si es "primera generación de universitarios en su familia" como se sugiere en [15] y como cursaron durante el codiv-19.

Agradecimientos
Este trabajo se realizó en el marco del proyecto de investigación "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 Universidad Nacional de Hurlingham, integrado al banco de Proyectos de Desarrollo Tecnológico y Social (Banco PDTS) del Ministerio de Ciencia, Tecnología e Innovación de la Nación.

Referencias:
[1] Síntesis de información : estadísticas universitarias 2020-2021. República Argentina. Disponible en: https://www.argentina.gob.ar/sites/default/files/sintesis_2021-2022_sistema_universitario_argentino_1.pdf
[2] 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. Disponible en http://sedici.unlp.edu.ar/handle/10915/61343
[3] Pustilnik, M., Ndukanma, G. (2023). Modelos Para La Predicción del Abandono en la Universidad Nacional de Hurlingham XVIII Congreso Tecnología en Educación & Educación en Tecnología (TE&ET). ISBN: 978-987-46875-6-2. Disponible en http://sedici.unlp.edu.ar/handle/10915/155526
[4] Sistema de Información Universitaria Guaraní: https://www.siu.edu.ar/

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

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

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

[8] Azevedo, A., Santos, M.: Kdd (2008). Semma and crisp-dm: a parallel overview. In: IADIS European Conf. Data Mining. Disponible en https://api.semanticscholar.org/CorpusID:15309704

[9] 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

[10] Tukey, J.W. (1977): Exploratory Data Analysis. Addison-Wesley.

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

[12] Chen, T., 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. Disponible en https://doi.org/10.1145/2939672.2939785

[13] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). Disponible en

The Elements of Statistical Learning : Data Mining, Inference, and Prediction (PDF)

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

[15] 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. Disponible en https://ri.conicet.gov.ar/handle/11336/51703

[16] 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 Maestría Argentina. Disponible en http://hdl.handle.net/10469/16829

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

[18] 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.

[19] 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. Disponible en https://doi.org/10.1023/a:1010920819831, https://doi.org/10.1023/a:1010920819831

[20] Pustilnik, M. et al. (2022): Estrechando el contacto entre universidades y estudiantes: comunicación ante posibles casos de abandono, propuestas para la inscripción. In: XXIV Workshop de Investigadores en Ciencias de la Computación (WICC 2022, Mendoza). pp. 734–738. Disponible en http://sedici.unlp.edu.ar/handle/10915/145216