

# Descubrimiento de patrones de comportamiento vinculados al abandono en la Universidad Nacional de Luján mediante la aplicación de técnicas de aprendizaje automático

Mario G. Oloriz<sup>1</sup>, Juan M. Fernandez<sup>1</sup>, Carlos Jara<sup>2</sup>, Carla Martinez<sup>2</sup>,  
Ruben Baquel<sup>2</sup>, Sebastián Bertoglio<sup>2</sup>, Hugo Delfino<sup>2</sup>

{moloriz, jmfernandez, cjara, cmartinez, rbaquel, sbertoglio, hdelfino}@unlu.edu.ar,

<sup>1</sup>División Computación, Departamento de Ciencias Básicas, Universidad Nacional de Luján

<sup>2</sup>División Estadística, Departamento de Ciencias Básicas, Universidad Nacional de Luján

## Resumen

El estudio del abandono en la educación superior resulta un tema de relevancia dado el continuo debate que se propicia respecto de los recursos públicos que las naciones afectan a su sostenimiento, principalmente en Argentina, donde el subsistema de gestión pública abarca más del 75 % de la matrícula total del sistema. Las altas tasas de abandono ponen en tensión la política de gratuidad e ingreso irrestricto que se aplica históricamente y que es reconocido como uno de los medios para la movilidad social.

La multicausalidad del fenómeno requiere de su estudio y análisis desde diferentes perspectivas así como de la definición de políticas académicas para tratar de disminuir el fracaso educativo en el nivel superior.

Por su parte, en la actualidad, existen numerosas técnicas de ciencia de datos mediante las cuales es posible extraer conocimiento a partir de bases de datos, las cuales están siendo aplicadas a áreas muy disímiles. Estas técnicas permiten encontrar patrones que no podrían ser descubiertos mediante las técnicas convencionales de análisis de datos multivariantes.

En este trabajo se describen las principales acciones desarrolladas, así como las líneas de I+D comprendidas en el mismo.

## Contexto

Este trabajo se encuentra en el marco del Proyecto de Investigación Interdepar “Generación de herramientas de extracción de información ambiental mediante técnicas de aprendizaje automático, a partir de bases de datos textuales. Su aplicación al caso del Municipio de Luján (Provincia de Buenos Aires)”, aprobado mediante Disposición PCDD-CB N°027/23 en el marco de los Proyectos de Investigación PI2+, Disposición CDD-CB N°139/19, del Departamento de Ciencias Básicas de la Universidad Nacional de Luján.

El objetivo principal de este proyecto consiste en identificar comportamientos vinculados al abandono en el ámbito de la Universidad Nacional de Luján, así como también a nivel nacional, a efectos de generar modelos y acciones que permitan disminuir este fenómeno negativo a partir de la aplicación de técnicas y estrategias propias de aprendizaje automático.

## Introducción

Con la masificación de internet y la digitalización de los procesos organizacionales, surge una creciente necesidad de desarrollar herramientas que permitan extraer información útil (conocimiento) a partir de grandes volúmenes de datos, los cuales están en constante crecimiento. De la mano de esta necesidad, aparece a mediados de la década del 90, un campo emergente denominado descubrimiento de conocimiento en bases de datos (KDD) [13].

Históricamente, la noción de encontrar patrones útiles en los datos ha recibido diversos nombres, incluyendo minería de datos, extracción de conocimiento, descubrimiento de información, recolección de información, arqueología de datos o procesamiento de patrones. Actualmente, se ha acuñado el término ciencia de datos para denominar esta disciplina, la cual se ubica en la intersección entre las disciplinas de ciencias de la computación, estadística y sus ámbitos de aplicación [50].

Existen numerosas técnicas de ciencia de datos mediante las cuales es posible realizar ese proceso de extracción de conocimiento a partir de bases de datos [22]. En este sentido, abordando la problemática desde el punto de vista del objetivo del descubrimiento de conocimiento, existe cierto consenso en categorizar en cuatro tipos principales a estas tareas: clasificación, predicción numérica, asociación y agrupamiento [8].

A su vez, la ciencia de datos es aplicada para la generación de conocimiento en infinidad de áreas totalmente disímiles, permitiendo encontrar patrones

que no podrían haber sido descubiertos mediante las técnicas convencionales de análisis de datos multivariantes. Ejemplos de ello son mejoras de procesos de eficiencia energética [25], detección y tratamiento de enfermedades como COVID-19 [4], medicina del deporte [29], preferencias de investigación en sociología [20], entre otros.

Por su parte, el estudio del abandono en la educación superior es un tema de relevancia dado el continuo debate que se propicia respecto de los recursos públicos que las naciones afectan al sostenimiento de este nivel educativo, principalmente en Argentina, donde el subsistema de gestión pública abarca casi el 75 % de la matrícula total del sistema [12]. Las altas tasas de abandono ponen en tensión la política de gratuidad e ingreso irrestricto que se aplica históricamente y que es reconocido como uno de los medios para la movilidad social [17].

En el contexto descripto, esta investigación se orienta a estudiar la aplicación de técnicas de ciencia de datos para la explotación de información de bases de datos pertenecientes al ámbito de la Educación Superior a efectos de extraer conocimiento relacionado con comportamientos poblacionales, en particular el abandono de los estudios y su estrecha relación con el rendimiento académico, fundamentalmente en la Universidad Nacional de Luján (UNLu).

## Antecedentes

En los últimos años, se han desarrollado diversas investigaciones que aplican técnicas de aprendizaje automático al ámbito educativo, entre las cuales se puede mencionar la detección de fraude en exámenes [24], la clasificación automática de consultas académicas [14], la predicción de preferencias sobre el uso de plataformas de aprendizaje móvil [2] así como la predicción del rendimiento académico en los estudios [54, 55] e incluso estudios del mismo fenómeno a partir del impacto producido por la pandemia originada por COVID-19 [52], entre muchos otros. A su vez existen diversas experiencias alrededor del mundo sobre la aplicación de un enfoque de ciencia de datos para la detección del abandono en educación así como la formulación de modelos que permitan predecir este fenómeno [3, 7, 26, 47, 49].

Así como en la UNLu, en la mayoría de las Universidades públicas de Argentina, el ingreso es irrestricto a las Carreras de grado y pregrado. Las publicaciones oficiales, correspondientes al año 2016, indican que existen alrededor de 1.940.000 estudiantes universitarios de grado y pregrado y 161.000 de carreras de postgrado. A su vez, algo más del 78 % de estas personas cursan carreras de grado y pregrado en instituciones de gestión pública y poco más del 76 % carreras de posgrado en el mismo tipo de instituciones. La misma publicación, da cuenta que ingresan al nivel universitario alrededor de 450.000 estudian-

tes por año y egresan por año, aproximadamente, 120.000 (tasa de egreso del 26 %).

Un problema señalado de manera recurrente en toda Latinoamérica son los comportamientos poblacionales de abandono y el rendimiento académico decreciente como problemáticas que deben ser estudiadas.

El concepto de abandono ha sido muy discutido académicamente, no obstante resulta claro que para los Estados implica un enorme costo, mientras que para el estudiante puede significar algún tipo de frustración y requiere de políticas de estado que posibiliten anticipar este fenómeno y mitigarlo.

Los especialistas presentan diversas hipótesis que intentan explicar la naturaleza de esta problemática. Algunos encuentran explicación en dos teorías sociológicas; por un lado “El modelo de integración del estudiante” [53] donde la integración del estudiante al mundo académico afecta en forma directa a la determinación de abandonar o no los estudios, y por el otro es el “Modelo de desgaste del estudiante” [6] que otorga relevancia a los factores externos a la institución universitaria.

Por su parte, Hackman y Dysinger plantean que el problema fundamental del abandono tiene que ver con la ausencia de interés y no con la imposibilidad por parte del estudiante de cumplir con los requisitos que la Universidad exige [19].

Mientras tanto, otros autores plantean el abordaje desde distintas dimensiones de análisis (Psicológicos, Económicos, Sociológicos, Organizacionales y de Interacción) de acuerdo a la relevancia que se le otorga a las variables que intentan explicar el fenómeno del abandono y retención, sean familiares, individuales o institucionales [9].

Asimismo, existen trabajos en los cuales se concluye que una de las causales principales del abandono tienen origen en temas relacionados con procesos vocacionales [5].

Como se observa, existe una variada cantidad de interpretaciones que intentan explicar el fenómeno del abandono aunque sin embargo, al tratarse de una problemática muy compleja y multicausal, ninguna da por agotado el tema.

En estudios recientes, realizados en el ámbito de la Universidad Nacional de Luján (UNLu), aplicando técnicas estadísticas multivariantes tradicionales, se han observado una serie de cuestiones que sirven de indicio para otras investigaciones. Por un lado, se ha logrado demostrar, parcialmente, que la institución cumple el fin social para el cual fue creada dado que quienes transitan por una de las Carreras que allí se dictan muestran un mejor desempeño en el ámbito laboral que quienes no lo hicieron. [36, 15].

En relación a la cuantificación de la problemática de abandono, objeto de este estudio, se ha encontrado, en un estudio longitudinal de once cohortes de la UNLu, que el 71.3 % de quienes ingresan terminan abandonando la carrera una vez transcurrida su

duración teórica [37]. Al mismo tiempo, en el estudio citado, se comprobó que el 65.5 % del abandono se produce de manera temprana (llegando al 46.7 % de la matrícula) mientras que el 34.5 % restante se produce de manera tardía (el 24.6 % de la matrícula).

En el mismo sentido, intentando generar una taxonomía de abandono para la Universidad Nacional de Luján, las investigaciones previas permiten sostener las siguientes conclusiones:

- La tasa de abandono aumenta en función del aumento de la edad del estudiante al momento de iniciar los estudios universitarios.
- No se han encontrado evidencias sobre que el tipo de gestión del establecimiento secundario del cual egresó el estudiante influya respecto del abandono de los estudios universitarios.
- Tal como lo verificaron otros autores, se encontró que la probabilidad de abandonar los estudios crece en función de la cantidad de horas que trabaja el estudiante [18].
- No se observan diferencias significativas respecto de la relación entre el abandono de los estudios y el género del estudiante.
- Se observa que el estado civil podría influir en la probabilidad que el estudiante abandone la carrera dado que para los estudiantes solteros la tasa de abandono es sustantivamente inferior que la observada para las restantes categorías de la variable estado civil.
- La tasa de abandono aumenta en función de la cantidad de tiempo transcurrido desde el momento en que el estudiante termina el nivel secundario y empieza a cursar el nivel universitario.
- La tasa de abandono es menor para quienes viven en la misma ciudad en que se encuentra la Sede en que cursan o en ciudades que se encuentran a una distancia entre 61 y 200 km.
- Se observó que, para la cohorte 2013, no se verifica que el promedio alcanzado en el nivel secundario pueda utilizarse como un predictor directo del abandono durante el primer año [44].

En cuanto al rendimiento académico, si bien algunos autores sostienen que el concepto de rendimiento nace en las sociedades industriales para referirse a normas, criterios y procedimientos de medida en el ámbito laboral y relacionado más específicamente a la productividad del trabajo [10]. Este concepto, aplicado a quienes cursan estudios universitarios, es lo que se define como rendimiento académico.

En el ámbito de la UNLu, se ha estudiado este fenómeno encontrando que existen algunas modalidades de dictado de las actividades académicas que

favorecen el rendimiento académico de los estudiantes [32]. A su vez, se ha constatado en que no existen prácticamente diferencias significativas en el desempeño académico de los estudiantes que ingresan con título de nivel medio de aquellos que lo hacen mediante la modalidad prevista por el artículo 7° de la Ley de Educación Superior, mayores de 25 años de edad sin estudios secundarios completos [35].

Otro de los aspectos estudiados con técnicas analíticas tradicionales, es el impacto de los programas de becas en el abandono y el rendimiento académico, encontrando que las mismas permiten una disminución del abandono (de hasta el 36 %) y la mejora del rendimiento académico si bien el nivel de efectividad varía de acuerdo al campo disciplinar de la Carrera que cursa el estudiante y el monto que percibe como estipendio incide en la efectividad a medida que avanza en la carrera [45, 39, 43, 38, 33].

También se encontraron efectos positivos de los programas de becas a nivel nacional al correlacionar los fondos que cada institución universitaria aplica a esos programas con el abandono interanual, para el período 2013-2014 [40].

Otro de los aspectos que se abordaron, es la incidencia entre la cantidad de estudiantes por docente en el abandono de los estudios superiores, concluyendo que la relación entre esas dos variables solo pudo validarse en las instituciones clasificadas como nuevas, o sea aquellas que no tienen aún 12 años de funcionamiento. En los demás casos, ya sea agrupando por tamaño o agrupando por región, no se observó correlación estadísticamente significativa entre estos dos indicadores [34].

En las carreras de postgrado, también el abandono es un problema de relevancia dada la baja tasa de graduación que se da en este nivel. En un estudio de las cohortes 2010 a 2014, de la Universidad Nacional de Luján, se calculó un abandono superior al 50 % de la matrícula, dejando de lado el problema de la elaboración de tesis o trabajos finales [41].

Finalmente, dada la sustitución de las clases presenciales por clases remotas o mediadas por las TIC, que se produjo durante los años 2020 y 2021 debido a la pandemia por COVID-19, se abordó el estudio del impacto en el abandono y el rendimiento académico de quienes se encontraban cursando carreras de pregrado y grado [42, 16, 46].

Estos hallazgos sobre abandono y rendimiento académico, obtenidos mediante el abordaje tradicional basado en métodos estadísticos pueden ser enriquecidos con el abordaje que propone la explotación de información para encontrar piezas de conocimiento sin ninguna presunción previa a partir de la masa de información disponible en los sistemas de gestión académica universitaria. Desde esta hipótesis, se plantean las líneas de I+D que se explicitan y abordan a continuación.

## Líneas de I+D

A continuación se presentan y describen brevemente las principales líneas de I+D en las cuales se trabaja actualmente en el marco de este proyecto.

### a. Determinación de las variables más importantes para el análisis del abandono en los estudios

La selección de características, en inglés *feature selection*, tiene como objetivo elegir un subconjunto de características relevantes a partir de las características originales de un conjunto de datos, eliminando las características irrelevantes, redundantes o ruidosas y maximizando la relevancia para la variable objetivo, como las etiquetas de clase en la clasificación [31]. Las técnicas más representativas de este enfoque incluyen la ganancia de información, relieve, puntuación de Fisher y Lazo [51].

La selección de características está comprendida dentro de las técnicas de reducción de dimensionalidad que también comprenden a las técnicas de extracción de características [51]. Complementariamente, estos enfoques de extracción de características proyectan características en un nuevo espacio de características con menor dimensionalidad y las nuevas características construidas suelen ser combinaciones de las originales. Los ejemplos de técnicas de extracción de características incluyen el análisis de componentes principales (PCA), el análisis discriminante lineal (LDA) y el análisis de correlación canónica (CCA).

Esta línea de investigación tiene como objetivo identificar, a partir de los repositorios de datos disponibles relacionados con el abandono y el rendimiento académico en la educación superior, el conjunto de atributos o características más representativas para la predicción del abandono y los grupos de riesgo en la educación superior.

### b. Caracterización de poblaciones estudiantiles

El *clustering* o agrupamiento es el proceso de agrupar objetos similares en diferentes grupos de acuerdo a alguna medida de distancia definida [27]. La agrupación puede considerarse el problema de aprendizaje no supervisado más importante y como tal, se trata de encontrar una estructura en una colección de datos no etiquetados. Por lo tanto, un clúster es una colección de objetos que son similares entre sí y son diferentes a los objetos pertenecientes a otros clústeres.

Dado su potencial, un importante número de estudios de investigación se centran en varios aspectos del análisis de *clustering* por ejemplo, algoritmos de agrupamiento y extensiones para un tipo de da-

tos particular, métricas de disimilitud (o distancia), número óptimo de conglomerados, relevancia de los atributos de datos por conglomerado, evaluación de los resultados del agrupamiento, y conjuntos de grupos [21].

En esta línea de I+D se propone la utilización de técnicas de aprendizaje no supervisado, particularmente de *clustering*, para la caracterización de la población estudiantil a partir de las fuentes de datos disponibles.

### c. Generación de modelos para la predicción del abandono

La predicción del abandono puede identificarse como un problema supervisado, ya sea de predicción o clasificación, de acuerdo al tipo de variable objetivo utilizada.

En un modelo de clasificación, luego de obtener los datos, realizar el preprocesamiento de los mismos para la extracción de características, realizar el etiquetado y avanzar en un esquema de representación, se entrena el clasificador utilizando distintos enfoques o algoritmos [48], como el aprendizaje bayesiano, la regresión logística, redes neuronales, árboles de decisión y máquinas de vectores soporte.

El modelo generado a partir del entrenamiento debe ser capaz de capturar las características distintivas del conjunto de entrenamiento para luego poder analizar otras instancias no observadas previamente, lográndose así la capacidad de generalización del clasificador que se suele evaluar sobre otro conjunto de prueba separado [28].

A la fecha, y debido a la cantidad de algoritmos de aprendizaje existentes, resulta muy complejo sistematizar todos los abordajes posibles. De acuerdo a los antecedentes estudiados, los algoritmos escogidos que se están utilizando en el marco de esta investigación son el clasificador de Naive Bayes [30], la regresión logística [50], las máquinas de vector soporte [23], las redes neuronales recurrentes LSTM (*Long short-term memory*) [1] y XGBoost [11].

Las aspiraciones en torno a esta línea de estudio son generar diferentes modelos de predicción del abandono teniendo en cuenta las estrategias de caracterización de poblaciones estudiantiles y las características más importantes emergentes de la etapa de extracción de características.

### d. Definición de políticas de mitigación del abandono en los estudios

El desarrollo de las líneas de I+D introducidas hasta aquí debieran favorecer la generación de conocimiento asociado al descubrimiento de patrones de comportamiento vinculados al abandono en la Universidad Nacional de Luján, así como en el Sistema de Educación Superior Argentino.

A partir de ese conocimiento, se plantea el objetivo de proponer estrategias que mitiguen el abandono y permitan brindar acompañamiento institucional para mejorar el rendimiento académico de los grupos de riesgo.

Tanto las Instituciones de Educación Superior como los organismos nacionales vinculados con la coordinación del sistema educativo, demandan la generación de información científica que permita la generación de proyectos y la fundamentación de diferentes líneas de acción que propendan a la mitigación del abandono en la educación superior.

## Objetivos

A partir de las líneas de investigación detalladas antes, en el marco de este proyecto se persiguen los siguientes objetivos específicos:

- Identificar, mediante técnicas de ciencia de datos, en el conjunto de variables descriptoras, el subconjunto de variables relacionadas con comportamientos poblacionales en educación superior.
- Identificar patrones relacionados con causales de abandono y disminución en el rendimiento académico de estudiantes de la Universidad Nacional de Luján.
- Proponer una configuración ideal para la aplicación de las técnicas de ciencias de datos a este tipo de problemáticas.
- Desarrollar modelos que permitan modificar estos patrones de comportamiento poblacional vinculados al abandono a efectos de revertir estas situaciones.

Complementariamente, su transferencia a la comunidad universitaria para la generación de políticas académicas que busquen la reducción del abandono resulta de alto interés. Las herramientas a desarrollar pueden ser aplicadas a diversos espacios académicos, lo cual plantea una oportunidad concreta de transferencia tanto en el sector público como en el privado.

## Recursos Humanos

Se espera que este proyecto contribuya a consolidar un grupo de investigación en la temática y brindar un ámbito adecuado para la formación de recursos humanos, en el marco de la Universidad Nacional de Luján, que incorporen saberes y competencias provenientes de la participación en actividades de investigación.

Concretamente, se ha incluido como integrantes a docentes auxiliares y se convocará a estudiantes que participen como pasantes esperando que el proyecto

sea el ámbito para la realización de, al menos, una Tesis de Licenciatura en Sistemas de Información y una tesis de Licenciatura en Gestión Universitaria por año así como la culminación de al menos una Tesis de Maestría en Generación y Análisis de Información Estadística.

## Referencias

- [1] AGGARWAL, C. C., ET AL. Neural networks and deep learning. *Springer 10* (2018), 978–3.
- [2] AKOUR, I., ALSHURIDEH, M., AL KURDI, B., AL ALI, A., SALLOUM, S., ET AL. Using machine learning algorithms to predict people's intention to use mobile learning platforms during the covid-19 pandemic: machine learning approach. *JMIR Medical Education 7*, 1 (2021), e24032.
- [3] ALI, M., AND ABDEL-HAQ, M. K. Bibliographical analysis of artificial intelligence learning in higher education: is the role of the human educator and educated a thing of the past? In *Fostering Communication and Learning With Underutilized Technologies in Higher Education*. IGI Global, 2021, pp. 36–52.
- [4] ALYASSERI, Z. A. A., AL-BETAR, M. A., DOUSH, I. A., AWADALLAH, M. A., ABASI, A. K., MAKHADMEH, S. N., ALOMARI, O. A., ABDULKAREEM, K. H., ADAM, A., DAMASEVICIUS, R., ET AL. Review on covid-19 diagnosis models based on machine learning and deep learning approaches. *Expert systems 39*, 3 (2022), e12759.
- [5] APARICIO, M., AND GARZUZI, V. Dinámicas identitarias, procesos vocacionales y su relación con el abandono de los estudios: Un análisis en alumnos ingresantes a la universidad. *Revista de Orientación Educativa 20*, 37 (2006), 15–36.
- [6] BEAN, J. P. Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in higher education* (1982), 291–320.
- [7] BILQUISE, G., ABDALLAH, S., AND KOBBAEY, T. Predicting student retention among a homogeneous population using data mining. In *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*. Springer, 2020, pp. 243–260.
- [8] BRAMER, M., AND BRAMER, M. Data for data mining. *Principles of data mining* (2016), 9–19.
- [9] BRAXTON, J. M. *Reworking the student departure puzzle*. Vanderbilt University Press, 2000.
- [10] CAMARENA, R., CHÁVEZ, A. M., AND GÓMEZ, J. Reflexiones en torno al rendimiento escolar

- ya la eficiencia terminal. *Revista de la educación superior* 53, 3 (1985).
- [11] CHEN, T., AND GUESTRIN, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.
  - [12] DE EDUCACIÓN DE LA NACIÓN, M. Anuarios estadísticos universitarios, 2020.
  - [13] FAYYAD, U., PIATETSKY-SHAPIO, G., AND SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine* 17, 3 (1996), 37–37.
  - [14] FERNÁNDEZ, J. M., AND ERRECALDE, M. Multi-class e-mail classification with a semi-supervised approach based on automatic feature selection and information retrieval. In *Cloud Computing, Big Data & Emerging Topics: 10th Conference, JCC-BD&ET 2022, La Plata, Argentina, June 28–30, 2022, Proceedings* (2022), Springer, pp. 75–90.
  - [15] FERNÁNDEZ, J. M., AND OLORIZ, M. G. Relación entre el cursado de la tecnicatura en administración y gestión universitaria en la universidad nacional de luján y el avance en la carrera administrativa.
  - [16] FERNÁNDEZ, J. M., OLORIZ, M. G., AND PUGGIONI, N. A. Efectos de la modalidad remota en el desempeño académico según campo disciplinar. In *Congresos CLABES* (2021).
  - [17] GARCÍA DE FANELLI, A. Inclusión social en la educación superior argentina: Indicadores y políticas en torno al acceso ya la graduación. *Páginas de Educación* 7, 2 (2014), 124–151.
  - [18] GIOVAGNOLI, P. I. Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración. *Documentos de trabajo* (2002).
  - [19] HACKMAN, J. R., AND DYSINGER, W. S. Commitment to college as a factor in student attrition. *Sociology of education* (1970), 311–324.
  - [20] HEIBERGER, R. H. Applying machine learning in sociology: How to predict gender and reveal research preferences. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 74, Suppl 1 (2022), 383–406.
  - [21] IAM-ON, N., AND BOONGOEN, T. Generating descriptive model for student dropout: a review of clustering approach. *Human-centric Computing and Information Sciences* 7 (2017), 1–24.
  - [22] IGUAL, L., SEGUÍ, S., IGUAL, L., AND SEGUÍ, S. *Introduction to data science*. Springer, 2017.
  - [23] JOACHIMS, T., ET AL. Transductive inference for text classification using support vector machines. In *Icml* (1999), vol. 99, pp. 200–209.
  - [24] KAMALOV, F., SULIEMAN, H., AND SANTANDREU CALONGE, D. Machine learning based approach to exam cheating detection. *Plos one* 16, 8 (2021), e0254340.
  - [25] KHAYYAM, H., NAEBE, M., MILANI, A. S., FAKHRHOSEINI, S. M., DATE, A., SHABANI, B., ATKISS, S., RAMAKRISHNA, S., FOX, B., AND JAZAR, R. N. Improving energy efficiency of carbon fiber manufacturing through waste heat recovery: A circular economy approach with machine learning. *Energy* 225 (2021), 120113.
  - [26] LUAN, H., AND TSAI, C.-C. A review of using machine learning approaches for precision education. *Educational Technology & Society* 24, 1 (2021), 250–266.
  - [27] MADHULATHA, T. S. An overview on clustering methods. *arXiv preprint arXiv:1205.1117* (2012).
  - [28] MARIÑELARENA-DONDENA, L., ERRECALDE, M. L., AND SOLANO, A. C. Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. *Revista Argentina de Ciencias del Comportamiento* 9, 2 (2017), 65–76.
  - [29] MARTIN, R. K., PAREEK, A., KRYCH, A. J., KREMERS, H. M., AND ENGBRETSSEN, L. Machine learning in sports medicine: need for improvement. *Journal of ISAKOS* 6, 1 (2021), 1–2.
  - [30] MCCALLUM, A., NIGAM, K., ET AL. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (1998), Citeseer, pp. 41–48.
  - [31] MIAO, J., AND NIU, L. A survey on feature selection. *Procedia Computer Science* 91 (2016), 919–926.
  - [32] OLORIZ, M., AND FERNÁNDEZ, J. M. La acreditación de competencias en reemplazo de actividades curriculares: ¿mejora el tránsito de los estudiantes por la propuesta formativa?. el caso de la carrera de ingeniería industrial de la unlu.
  - [33] OLORIZ, M., FERNÁNDEZ, J. M., AND AMADO, V. Impacto del programa de becas de la universidad nacional de luján en la disminución del abandono. In *III Conferencia Latinoamericana sobre el Abandono en la Educación Superior. CLABES* (2013).

- [34] OLORIZ, M., FERNANDEZ, J. M., AND BATTO, M. Relación entre la planta docente y el abandono interanual en las universidades argentinas.
- [35] OLORIZ, M. G., AND FERNANDEZ, J. Rendimiento académico de los estudiantes que ingresaron sin estudios de nivel medio completos en la universidad nacional de luján. *XII Coloquio Internacional sobre Gestión Universitaria de América en Sur. Veracruz, México* (2012).
- [36] OLORIZ, M. G., AND FERNÁNDEZ, J. M. El impacto de transitar por una carrera universitaria en la carrera laboral. el caso de la tecnicatura en administración y gestión universitaria en la universidad nacional de luján.
- [37] OLORIZ, M. G., AND FERNÁNDEZ, J. M. Relación entre las características del estudiante al momento de iniciar estudios superiores y el abandono en la universidad nacional de luján durante el período 2000-2010. In *Congresos CLABES* (2013).
- [38] OLORIZ, M. G., AND FERNANDEZ, J. M. Impacto del programa de becas estudiantiles de la universidad nacional de luján por campo disciplinar. In *Congresos CLABES* (2014).
- [39] OLORIZ, M. G., AND FERNÁNDEZ, J. M. Rendimiento académico de los estudiantes de la universidad nacional de luján que participan en los programas de becas internos y nacionales.
- [40] OLORIZ, M. G., AND FERNÁNDEZ, J. M. Estudio del abandono interanual en el sistema universitario argentino y su relación con los recursos aplicados a los programas de becas. In *Congresos CLABES* (2016).
- [41] OLORIZ, M. G., AND FERNÁNDEZ, J. M. El abandono en las carreras de postgrado de la universidad nacional de luján. In *Congresos CLABES* (2018).
- [42] OLORIZ, M. G., AND FERNÁNDEZ, J. M. El impacto de la pandemia por covid-19 en el abandono temprano de los estudios superiores. In *Congresos CLABES* (2021).
- [43] OLORIZ, M. G., FERNANDEZ, J. M., AND AMADO, M. V. Comparación entre el programa de becas estudiantiles de la universidad nacional de luján y los programas de becas nacionales. In *Congresos CLABES* (2014).
- [44] OLORIZ, M. G., FERNÁNDEZ, J. M., AND RODRÍGUEZ, L. S. Relación entre promedio final en el nivel secundario y el abandono de los estudios superiores. In *Congresos CLABES* (2014).
- [45] OLORIZ, M. G., FERNÁNDEZ, J. M., AND RODRIGUEZ, R. Impacto del programa de “respaldo a estudiantes argentinos” en la disminución del abandono en la universidad nacional de luján. In *Congresos CLABES* (2015).
- [46] OLORIZ, M. G., FERRERO, E. L., AND LUCHINI, M. L. Impacto de la suspensión de correlatividades en el cursado de actividades académicas. In *Congresos CLABES* (2021).
- [47] PALACIOS, C. A., REYES-SUÁREZ, J. A., BEARZOTTI, L. A., LEIVA, V., AND MARCHANT, C. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile. *Entropy* 23, 4 (2021), 485.
- [48] RUSSELL STUART, J., AND NORVIG, P. *Artificial intelligence: a modern approach*. Prentice Hall, 2009.
- [49] SHAFIQ, D. A., MARJANI, M., HABEEB, R. A. A., AND ASIRVATHAM, D. Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access* (2022).
- [50] SKIENA, S. S. *The data science design manual*. Springer, 2017.
- [51] TANG, J., ALELYANI, S., AND LIU, H. Feature selection for classification: A review. *Data classification: Algorithms and applications* (2014), 37.
- [52] TARIK, A., AISSA, H., AND YOUSEF, F. Artificial intelligence and machine learning to predict student performance during the covid-19. *Procedia Computer Science* 184 (2021), 835–840.
- [53] TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research* 45, 1 (1975), 89–125.
- [54] TJANDRA, E., KUSUMAWARDANI, S. S., AND FERDIANA, R. Student performance prediction in higher education: A comprehensive review. In *AIP Conference Proceedings* (2022), vol. 2470, AIP Publishing LLC, p. 050005.
- [55] YAĞCI, M. Educational data mining: prediction of students’ academic performance using machine learning algorithms. *Smart Learning Environments* 9, 1 (2022), 11.