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

Mario Oloriz[1], Juan M. Fernandez[1], Carlos Jara[2], Carla Martinez[2], Ruben Baquel[2],Sebastián Bertoglio[2], Hugo Delfino[2]

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

[1]División Computación, Departamento de Ciencias Básicas, Universidad Nacional de Luján

[2]División Estadística, Departamento de Ciencias Básicas, Universidad Nacional de Luján

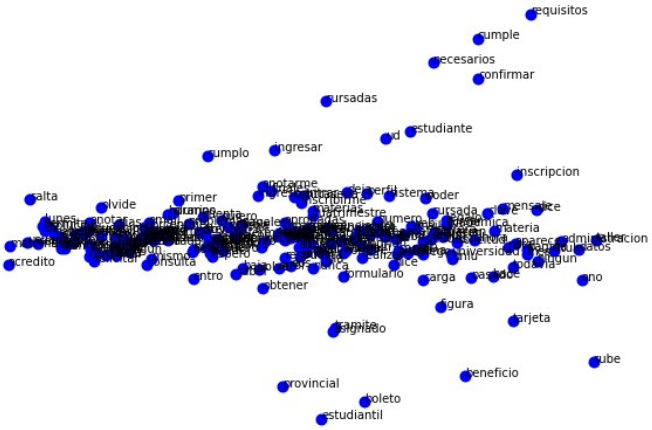
Contexto

Este trabajo se encuentra en el marco del Proyecto de Investigación "Descubrimiento de patrones de comportamiento vinculados al abandono en la Universidad Nacional de Luján mediante la aplicación de técnicas de aprendizaje automático", aprobado mediante Disposición PCDD-CB N° 027/23 en el marco de los Proyectos de Investigación PI^2+A del Departamento de Ciencias Básicas de la Universidad Nacional de Luján que tiene como objetivo principal 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.

LÍNEAS DE I+D

DETERMINACIÓN DE VARIABLES IMPORTANTES

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 maximizando la relevancia para la variable objetivo.
Esta línea de I+D busca identificar, a partir de los repositorios de datos disponibles, el conjunto de características más representativas para la predicción del abandono y los grupos de riesgo en la educación superior.



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. Es considerado el problema de aprendizaje no supervisado más importante y como tal, se trata de encontrar una estructura en una colección de datos no etiquetados.
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.

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, como el aprendizaje bayesiano, la regresión logística, redes neuronales, árboles de decisión y máquinas de vectores soporte. 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.

DEFINICIÓN DE POLÍTICAS DE MITIGACIÓN DEL ABANDONO

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.

Formación de 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. 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

Aggarwal, C. C., et al. Neural networks and deep learning. Springer 10 (2018), 978–3.

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.

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).

Palacios, C. A., Reyes-Suarez, 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.