

## Detección de patrones y tendencias en estudiantes universitarios de carreras de ingeniería para determinar el éxito académico aplicando Machine Learning

José Federico Medrano<sup>1</sup>, Octavio Daniel Coro<sup>1</sup>, Valeria Barriento<sup>2</sup>

jfmedrano@fi.unju.edu.ar, odcoro@fi.unju.edu.ar, barriento.valeria@gmail.com

<sup>1</sup>VRAln / Visualización y Recuperación Avanzada de Información / Facultad de Ingeniería

Universidad Nacional de Jujuy (UNJu) - Ítalo Palanca 10, +54 (388) 4221587

<sup>2</sup> Escuela Superior de Ciencias Jurídicas y Políticas / Universidad Nacional de Jujuy (UNJu)

### RESUMEN

Uno de los indicadores por el cual son medidas y evaluadas las universidades es el número o cantidad de egresados por año. Si bien uno de los tantos objetivos de las instituciones académicas es la formación de profesionales, en las universidades se evidencia que no existe una relación directa entre el número de ingresantes y el número de egresados, es decir, la cantidad de egresados no necesariamente se incrementa al aumentar la cantidad de ingresantes. Esto representa un gran problema puesto que los presupuestos con los que cuentan las universidades, sobre todo las de gestión pública, muchas veces se ven reducidos, es más, cuando un alumno logra finalizar sus estudios, el costo de la inversión para formarlo es altísimo. Claro está que obtener una mayor cantidad de egresados no solo reduciría la inversión mencionada sino que mejoraría los indicadores empleados y posicionaría mejor a cualquier universidad. Lograr identificar algún patrón o tendencia en el modo en que los alumnos cursan su respectiva carrera, o identificar las variables que influyen tanto en el éxito como en la deserción académica, sería beneficioso para que las autoridades puedan tomar las

decisiones correctas. En este proyecto se tomarán las historias académicas de estudiantes de carreras de ingeniería de la Universidad Nacional de Jujuy, para identificar patrones o tendencias en el cursado que permitan determinar las variables que influyen para que un estudiante logre concluir la carrera. Se empleará *Machine Learning* como técnica para conducir este trabajo, encarándolo como un problema de clasificación y/o predicción.

**Palabras clave:** *Machine Learning; Detección de patrones; Tendencias; Rendimiento académico; Deserción*

### CONTEXTO

La línea de investigación aquí presentada se encuadra dentro del Proyecto Bidual 2020-2021 denominado “*Detección de patrones y tendencias en estudiantes universitarios para identificar causas de deserción y éxito académico empleando Machine Learning*”, aprobado y financiado por la Secretaría de Ciencia y Técnica de la Facultad de Ingeniería

de la Universidad Nacional de Jujuy, aprobado por resolución CAFI 449/19 .

## 1. INTRODUCCIÓN

La permanencia y la graduación de los estudiantes suele ser una problemática relevante compartida por la mayoría de los sistemas universitarios e instituciones que lo componen, y de acuerdo al nivel alcanzado en sus indicadores se constituye en condición de existencia o desaparición de aquellas universidades que dependen en forma casi exclusiva de su matrícula, como es el caso de las universidades privadas, o de condicionamiento presupuestario en el caso de las universidades estatales (Lattuada, 2017).

El rendimiento académico es un fenómeno complejo que suele abordarse en las universidades desde distintos indicadores. Con frecuencia se consideran las calificaciones promedio obtenidas en distintos períodos de la estancia en la universidad, tasas de aprobación de materias y créditos, tasas de retención-deserción, tasas de graduación, períodos de graduación respecto de planes de estudio (Sánchez & Chinchilla Brenes, 2005). Estos indicadores son relativamente sencillos de calcular y medir, dado que se basan en recuentos estadísticos, y los informes y reportes que se arman en base a estos números solo reflejan una parte de la historia, la que la estadística descriptiva puede analizar (Fernández, Sánchez, Córdoba, & Largo, 2002).

Ahora bien, si el número de estudiantes que egresan cada año de una universidad fuera el indicador más relevante para estimar la calidad de una universidad, sería deseable no solo identificar las causas que influyen en la deserción de los estudiantes sino también identificar las causas que determinan que a un estudiante le demande mayor o menor

cantidad de años finalizar dicha carrera, ya sea una carrera de grado o pregrado. Por ello existen modelos, técnicas y herramientas avanzadas que emplean Inteligencia Artificial para poder identificar patrones y tendencias en los datos analizados (Wang, Rudin, Wagner, & Sevieri, 2015), del mismo modo estos modelos pueden realizar predicciones o pronósticos partiendo del análisis datos históricos, lo que se conoce comúnmente como Aprendizaje Automático o *Machine Learning* (Witten, Frank, Hall, & Pal, 2016).

El mecanismo que emplea el *Machine Learning* es muy sencillo, a partir de un conjunto de datos se realiza una división del mismo en tres grupos, uno de entrenamiento, uno de prueba y uno de evaluación. El conjunto de datos de entrenamiento es empleado por el modelo/algoritmo que se desee entrenar, esto es, el modelo aprende la lógica de los datos para realizar predicciones, para clasificar o agrupar datos, el modelo se ajusta a los datos. El conjunto de prueba sirve para testear y ajustar el modelo diseñado. Por último el conjunto de datos de validación es empleado para obtener indicadores de la eficiencia del modelo planteado. Una vez ajustado y validado el modelo, el mismo es utilizado con datos nuevos y desconocidos para obtener “nuevo conocimiento”.

Dentro de esta área del conocimiento (la Inteligencia Artificial), el *Machine Learning* ha recibido muchísima atención en esta última década, debido principalmente a la reducción de costos de los equipos de cómputo y a la proliferación de librerías y herramientas que automatizan gran parte del trabajo. En este sentido se ha aplicado efectivamente al ámbito académico para analizar la retención de estudiantes (Nandeshwar, Menzies, & Nelson, 2011) o para predecir la deserción

de estos (Kuna, García-Martínez, & Villatoro, 2010; Solis, Moreira, Gonzalez, Fernandez, & Hernandez, 2018; Bowen & Agustín, 2016).

Una ventaja adicional de este tipo de análisis serviría para producir visualizaciones y representaciones de información con técnicas novedosas que permiten encontrar nuevas relaciones e información a partir de dichas representaciones. Para ello se emplearán técnicas de Visualización de Información (InfoVis) que tienen por objeto presentar la información visualmente, en esencia, para reducir la carga de trabajo cognitivo al sistema perceptivo visual humano (Ware, 2004; Ware, 2008).

Otro estudio por ejemplo se ha centrado en la construcción de un modelo predictivo a través de técnicas de minería de datos que permita, por un lado, determinar la cantidad de alumnos que pueden graduarse en carreras de ingenierías de UTN Facultad Regional San Francisco, y por otro lado, identificar patrones que puedan incidir en la graduación (Carrizo, 2019). Del mismo modo, en (Istvan & Lasagna, 2019) se plantea la identificación de los diferentes factores que influyen en la deserción estudiantil, con el fin de caracterizar y extraer los perfiles de los estudiantes desertores empleando minería de datos como una alternativa útil para encontrar información derivada a partir de la detección de patrones de atributos individuales.

La Universidad Nacional de Jujuy posee un Sistema de Gestión Académica llamado SIU Guaraní, cada facultad posee una configuración e instalación individual, es decir, el sistema no es centralizado. Cada instancia en cada facultad tiene una base de datos única en donde se gestiona y almacena la “historia” académica de los alumnos, desde el momento en que ingresan a la facultad hasta que egresan de la misma. Partiendo de

esto, se plantea en este proyecto analizar la base de datos del sistema SIU Guaraní de la Facultad de Ingeniería para identificar las variables y causas que influyen tanto en el éxito como en el fracaso académico.

El presente proyecto tiene como objetivo general la búsqueda e identificación de patrones en la conducta de cursado de los estudiantes universitarios a lo largo de su vida académica.

De este objetivo general se desprenden los siguientes Objetivos Particulares:

- Analizar la base de datos actual del sistema SIU Guaraní.
- Determinar los atributos y variables relevantes para el estudio.
- Determinar la muestra que servirá para el análisis.
- Determinar el conjunto de datos que se utilizará para entrenamiento, prueba y validación.
- Verificar el balance de las clases a emplear en los modelos.
- Diseñar modelos de clasificación y agrupación de información mediante la aplicación de técnicas de Machine Learning.
- Representar las distintas relaciones y/o patrones entre los datos procesados mediante el empleo de técnicas de visualización de información.

## 2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación principal se enmarca en un proyecto de la Facultad de Ingeniería de la UNJu que aborda el estudio de las variables que influyen tanto en el éxito como en la deserción académica. Mediante el empleo de *Machine Learning* pretende predecir cuales

son los atributos que influyen para que un estudiante logre finalizar la carrera. De este modo, con el nuevo conocimiento adquirido, se podrán tomar distintas decisiones para atacar el problema y garantizar bajo ciertos parámetros una cursada exitosa.

El equipo de Investigación que llevará a cabo este trabajo centra sus líneas de investigación en la Inteligencia Artificial, Minería de Datos/ Textos, y Recuperación de Información.

### 3. RESULTADOS OBTENIDOS/ESPERADOS

A partir de los objetivos planteados en este trabajo, se espera diseñar ciertos modelos de aprendizaje automático que sean capaces, luego del correcto entrenamiento, ajuste y validación, predecir cuales son las variables y atributos del conjunto de datos (historia académica de los alumnos) que determinan el éxito en la cursada o por el contrario, detectar los factores que influyen en el re-cursado o abandono de la carrera.

Luego de este análisis, se espera conformar un conjunto de visualizaciones que ayuden a tener una visión más genérica de los resultados hallados (Lanzarini, Charnelli, Baldino, & Díaz, 2015).

### 4. FORMACIÓN DE RECURSOS HUMANOS

El Equipo de Trabajo está conformado por docentes investigadores de la Universidad Nacional de Jujuy. Los mismos llevan adelante esta línea de investigación desde hace años. Cada año se incorporan al proyecto alumnos avanzados de distintas carreras, quienes trabajan en temas relacionados con las temáticas planteadas. Del mismo modo, los integrantes del equipo participan en el

dictado de asignaturas/cursos de grado y postgrado de la UNLP, UNJu y UCSEDASS.

### 5. BIBLIOGRAFÍA

- Bowen, O., & Agustín, A. (2016). *Análisis de la deserción y permanencia académica en la educación superior aplicando minería de datos*. Tesis Doctoral, Universidad Nacional de Colombia-Sede Bogotá.
- Carrizo, C. J. (2019). Modelo predictivo para determinar la graduación de alumnos de carreras de ingenierías aplicando técnicas de minería de datos. *I Congreso Internacional de Ingeniería de Sistemas*.
- Fernández, S. F., Sánchez, J. M., Córdoba, A., & Largo, A. C. (2002). *Estadística descriptiva*. Esic Editorial.
- Istvan, R. M., & Lasagna, V. (2019). Aplicación de minería de datos en la extracción de perfiles de deserción estudiantil sobre ingresantes de la UTN FRLP. *XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan)*. San Juan.
- Kuna, H., García-Martínez, R. V., & Villatoro, F. (2010). Pattern Discovery in University Students Desertion Based on Data Mining. *IV Meeting on Dynamics of Social and Economic Systems, 2*, pp. 275-285.
- Lanzarini, L. C., Charnelli, M. E., Baldino, G., & Díaz, F. J. (2015). Selección de atributos representativos del avance académico de los alumnos universitarios usando técnicas de visualización. *TE & ET(15)*, 42-50.

- Lattuada, M. (2017). Deserción y retención en las unidades académicas de educación superior. Una aproximación a las causas, instrumentos y estrategias que contribuyen a conocer y morigerar su impacto. *Debate Universitario*, 5(10), 100-113.
- Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, 38(12), 14984-14996.
- Sánchez, E. G., & Chinchilla Brenes, S. (2005). Detección de estudiantes en riesgo académico en el Instituto Tecnológico de Costa Rica. *Revista Educación*, 29(2), 123-138.
- Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, (pp. 1-6).
- Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2015). Finding patterns with a rotten core: Data mining for crime series with cores. *Big Data*, 3(1), 3-21.
- Ware, C. (2004). *Information Visualization - Perception for Design*. Morgan-Kaufmann.
- Ware, C. (2008). *Visual Thinking for Design*. Morgan Kaufman/Elsevier.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.