# PREDICCIÓN DEL ABANDONO ESTUDIANTIL
## UN CASO APLICADO DE DESCUBRIMIENTO DE INFORMACIÓN A PARTIR DE DATOS EN LA FACULTAD DE CIENCIAS ECONÓMICAS DE LA UNICEN

**IGNACIO A. CARRERAS**[1] **- MARÍA DEL CARMEN ROMERO**[2]
[1] **Facultad de Ciencias Económicas - Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Argentina**
[2] **Centro de Estudios en Administración (CEA) - Facultad de Ciencias Económicas - Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Argentina**
***ignacio.carreras@econ.unicen.edu.ar - maria.romero@econ.unicen.edu.ar***

Fecha recepción: octubre 2024 Fecha aprobación: noviembre 2024

ARK CAICYT: https://id.caicyt.gov.ar/ark:/s18539777/aidkaxl3k

**RESUMEN**

En el presente trabajo se aborda la problemática del abandono estudiantil en la Facultad de Ciencias Económicas de la Universidad Nacional del Centro de la Provincia de Buenos Aires para las cohortes 2009 a 2019 para las carreras de Contador Público y Licenciatura en Administración. Resulta una aplicación en la cual se resalta el proceso completo de descubrimiento de información y conocimiento a partir de datos. Los hallazgos de este estudio de carácter cuantitativo, proporcionan un valioso instrumento para la gestión académica ya que, mediante el análisis discriminante, es posible prever potenciales casos de abandono estudiantil. Esta capacidad de anticipación permite a las autoridades implementar medidas preventivas oportunas, interviniendo antes de que el abandono se concrete.

**PALABRAS CLAVE**: ESTADÍSTICA - ABANDONO ESTUDIANTIL - DATOS - INFORMACIÓN - CONOCIMIENTO

**ABSTRACT**

This work addresses the problem of student dropout at the Faculty of Economic Sciences at the National University of the Center of the Province of Buenos Aires for the cohorts 2009 to 2019 for the Public Accountant and Bachelor of Administration careers. It is an application in which the complete process of discovery of information and knowledge from data is highlighted. The findings of this quantitative study provide a valuable tool for academic management since, through discriminant analysis, it is possible to foresee potential cases of student dropout. This ability to anticipate allows authorities to implement timely preventive measures, intervening before dropout occurs.

**KEYWORDS:** STATISTICS - STUDENT DROPOUT - DATA - INFORMATION - KNOWLEDGE

## 1. INTRODUCCIÓN

El acceso, la permanencia y la graduación son pilares fundamentales para mejorar la calidad educativa en las instituciones de educación superior. Aunque en Argentina se han implementado políticas para ampliar el acceso universitario, se enfrentan altas tasas de abandono y bajas tasas de graduación. Este fenómeno, conocido como deserción, es un problema crítico que requiere investigación y atención, especialmente considerando que Argentina tiene una de las tasas brutas de escolarización superior más altas de América Latina (Accinelli, Losio y Macri, 2016).

La deserción puede describirse como el "fenómeno de carácter colectivo, en el cual los individuos, una vez que logran insertarse en el sistema de educación, abandonan el proceso formal sin completar el ciclo respectivo, debido a causas endógenas y exógenas al mismo sistema" (Vásquez Martínez y Rodríguez Pérez, 2007). Tinto (1989) propone una visión matizada de la deserción universitaria, argumentando que no todo abandono debe considerarse un fracaso institucional o personal. Señala que existen metas individuales que pueden cumplirse antes de la graduación, y que algunos estudiantes pueden no tener el interés suficiente para comprometerse con el esfuerzo requerido. Sugiere que, si bien es posible reducir la tasa de deserción, eliminarla por completo es irrealista. Esta perspectiva implica que el término "abandono" puede ser más apropiado que "deserción" en algunos contextos, especialmente al calcular indicadores donde no se indagan las causas específicas de la no rematriculación de los estudiantes.

La existencia de grandes cantidades de datos trajo consigo la necesidad de aplicar técnicas computacionales para extraer información a partir de ellos. Fayyad *et al.* (1996) rotularon al proceso de descubrimiento de conocimiento a partir de los datos como KDD (*Knowledge Discovery in Databases*). Lo definieron como el "proceso no trivial de identificar, en los datos, patrones válidos, nuevos y potencialmente útiles, susceptibles de ser comprendidos", señalando que un patrón es una expresión que describe un subconjunto de datos, un modelo aplicable al subconjunto. Plantearon, inicialmente, una serie de pasos que, pueden resumirse en: identificación del objetivo (requerimiento), selección del conjunto de datos potencialmente útiles, preprocesamiento de los datos (limpieza y transformación de los mismos para, entre otras cosas, eliminar ruido, trabajar con campos faltantes), selección y aplicación de métodos de *data mining* para dar respuesta al objetivo (aplicar algoritmos de análisis para describir y producir modelos sobre los datos), interpretación de los resultados obtenidos y utilización de la información obtenida.

La propuesta metodológica planteada en este enfoque puede relacionarse con los conceptos dato, información y conocimiento. Estos conceptos no son fácilmente separables en la práctica; en el mejor de los casos, se puede construir un continuo utilizándolos (Davenport, 1997). Los datos son hechos crudos y por sí solos tienen poca relevancia o significado (Davenport y Prusak, 1998), pueden describirse como registros estructurados de transacciones. La información es un conjunto de datos procesados que tienen un significado (relevancia, propósito y

contexto), datos útiles y con sentido. El conocimiento es información combinada con experiencia, contexto, interpretación y reflexión (Tippins y Sohi, 2003) y resulta en una mayor capacidad para la toma de decisiones y la acción para lograr algún propósito particular.

Dado el requerimiento específico, los datos se obtienen mediante la etapa de selección de datos que proponen Fayyad *et al.* (1996), la información al preprocesar los datos y aplicar técnicas de *data mining*, y el conocimiento al interpretar y utilizar la información obtenida.

El presente trabajo aborda la problemática del abandono estudiantil en la Facultad de Ciencias Económicas (FCE-UNICEN) de la Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN) para las cohortes 2009 a 2019 para las carreras de Contador Público y Licenciatura en Administración. Resulta una aplicación del proceso de descubrimiento de conocimiento a partir de bases de datos y se describen todos los pasos a seguir, concluyendo con la construcción de un modelo para predecir, en función del rendimiento académico, la probabilidad de que un alumno abandone (para una carrera y año determinado). La información brindada por los modelos obtenidos, y combinada con experiencia, contexto y reflexión generará conocimiento que podría resultar en una contribución a la gestión y toma de decisiones sobre los programas y acciones de retención estudiantil implementados.

## 2. METODOLOGÍA

El requerimiento de construir un modelo para predecir la probabilidad de que un alumno abandone, en un año y carrera determinada, se llevó a cabo desarrollando el proceso KDD que comprende la identificación de los datos que darían respuesta a la problemática, su preprocesamiento, la aplicación de métodos de análisis (métodos de *data mining*) para obtener información a partir de los mismos, hasta el post procesamiento para convertir dicha información en conocimiento.

### 2.1 Selección del conjunto de datos

El análisis se basó en datos extraídos del sistema SIU-GUARANI de la Facultad de Ciencias Económicas de la UNICEN. Se seleccionaron los alumnos registrados dentro del año académico solicitado para las cohortes de 2009 a 2019 de las carreras Licenciado en Administración y Contador Público Nacional.

Se extrajeron los siguientes datos para cada alumno: Nro de legajo, DNI, Carrera (Contador Público (CP) y Licenciatura en Administración (LA)), Cohorte (año académico de la primera matriculación), Matriculación (año académico en que realizó la matriculación), Fecha de cursada, Código de materia (Cursada), Resultado de cursada (aprobado, libre, promocionado, reprobado, ausente), Nota de cursada, Fecha de finales, Código de materia (Final), Resultado de finales (aprobado, reprobado), Nota de final, Fecha de egreso, Promedio general, Promedio sin aplazos.

Tras el proceso de extracción, el conjunto de datos resultante incluyó 2861 legajos (1907 de la carrera Contador Público y 954 de Licenciado en

Administración). Respecto a la actividad académica, se conservaron 61714 registros de cursadas y 46251 registros de exámenes finales.

### 2.2 Preprocesamiento de los datos

Considerando los datos obtenidos a partir de la selección de datos, se generaron variables cuantitativas para caracterizar la actividad académica: Cantidad total de cursadas (CT), Cantidad de cursadas promocionadas (CP), Cantidad de cursadas aprobadas (CA), Cantidad de cursadas reprobadas (CR), Cantidad de cursadas abandonadas (CAB), Promedio de cursadas con Aplazos por año académico (CPROA), Cantidad total de finales (FT), Cantidad de finales aprobados (FA), Cantidad de Finales reprobados (FR). Dado que son variables que cambian año a año, para cada una de ellas se definieron tantas variables como años académicos en los que el alumno esté activo (por ejemplo, si el alumno está activo durante 10 períodos académicos, la característica referida al promedio, se relevará con 10 variables, una por cada año). Es decir que, para cada una de las características, se esperaría tener tantos datos como los que resulten de multiplicar los 2861 estudiantes por la cantidad de años académicos en los que hayan estado activos.

De esta manera, la tabla de datos con los cuales se realizó el análisis tiene 2861 filas (cada una se corresponde con un alumno) y las siguientes columnas:

- Legajo-Carrera: identificador clave (combinación de “legajo” y “carrera”).
- $D_p$: variable binaria que indica si el alumno abandonó en el período académico p. $D_p = 1$: se considera abandono en el año p si se matriculó en el año *p* y no se matriculó en el año *p+1* ni egresó durante el año *p*, $D_p = 0$, caso contrario.
- $CT_p$: Cantidad total de cursadas en el año académico p.
- $CP_p$: Cantidad de cursadas promocionadas en el año académico p.
- $CA_p$: Cantidad de cursadas aprobadas en el año académico p.
- $CR_p$: Cantidad de cursadas reprobadas en el año académico p.
- $CAB_p$: Cantidad de cursadas abandonadas en el año académico p.
- $CPROA_p$: Promedio de cursadas con Aplazos en el año académico p.
- $FT_p$: Cantidad total de finales en el año académico p.
- $FA_p$: Cantidad de finales aprobados en el año académico p.
- $FR_p$: Cantidad de finales reprobados en el año académico p.

donde p refiere al período académico (p = 1…10).

### 2.3 Selección y aplicación de métodos de *data mining*

Dado que el objetivo es encontrar un modelo que permita predecir la probabilidad que tiene un alumno para abandonar o no la carrera en función del rendimiento académico, se identificaron, en primera instancia las variables más representativas de dicho rendimiento.

Para ello se seleccionaron las variables con comportamientos diferenciales entre ambos grupos (abandono y no abandono) mediante la prueba diferencia de medias o de medianas (en caso de que las medias no sean representativas) con un nivel de confianza del 95%. Estas pruebas establecen como hipótesis nula

que las medias (medianas) son iguales. Para el grupo de variables resultante se analizó la asociación entre las mismas mediante el coeficiente de correlación de Pearson con el objetivo de descartar variables redundantes.

El modelo para predecir la probabilidad de que un alumno abandone en función de variables de rendimiento académico plantea una variable de respuesta dicotómica y variables predictoras cuantitativas. Se trabajó con el método multivariado análisis discriminante por ser un método de clasificación supervisada que detecta las variables que permiten discriminar entre grupos conocidos a priori, construyendo una regla de clasificación para predecir la probabilidad de pertenencia a un grupo. Puede realizarse con fines predictivos relacionados a la clasificación que permitirá que una observación nueva, que no fue utilizada para la construcción de la regla de clasificación, se asigne al grupo en el cual tienen más probabilidad de pertenecer en base a sus características medidas. Tal asignación se realiza con una regla de clasificación, en este caso, la función discriminante lineal. (Balzarini *et al*., 2008).

En este caso, dado que sólo se tienen dos grupos (compuestos por los que abandonan y los que no), la clasificación se realiza con una sola función discriminante que resulta una combinación lineal de variables que sirven para evaluar la pertenencia de un alumno a uno de los grupos determinados a priori. Este método ofrece tasas de errores de clasificación que brindan una idea de la capacidad predictiva de la función. Para evaluar cuán bien estas funciones podrán clasificar nuevas observaciones, se consideran los porcentajes de falsos positivos y falsos negativos. El porcentaje de falsos positivos refiere a aquellos clasificados como alumnos que abandonan cuando en realidad no lo hacen, y el porcentaje de falsos negativos a los que abandonan que fueron clasificados como que no abandonan. En términos de decisiones a tomar se opta por minimizar los falsos negativos, es decir, tener el menor porcentaje de alumnos detectados como que no abandonan, cuando en realidad sí lo harán.

Si bien este método tiene supuestos respecto a la distribución y a la homogeneidad de la estructura de variación y covarianza de los grupos, no son verificados en este trabajo dada la naturaleza descriptiva de las conclusiones.

La tabulación, graficación y análisis estadístico de los datos se trabajó con *Google Sheets* y con el software estadístico InfoStat (Di Rienzo *et al*., 2020).

## 3. RESULTADOS

Dada la gran cantidad de variables relacionadas con el rendimiento académico, y que muchas de ellas están asociadas, se seleccionaron en primer lugar aquéllas que mejor distinguen entre grupos, y, del conjunto resultante, una variable representante de cada conjunto de variables asociadas. En la selección de las variables que más distinguen entre grupos, se trabajó con diferencia de medianas (ya que la mayoría no posee una media representativa). Estas pruebas se realizaron por carrera y considerando los grupos de los alumnos que abandonan y los que no.

De las variables relacionadas a las cursadas y a los finales, se eliminan Cantidad de cursadas reprobadas, Cantidad de Finales reprobados y Promedio de cursadas con Aplazos por año académico por no presentar diferencias significativas de medianas entre ambos grupos. Luego, y mediante el análisis de

correlación lineal de Pearson (considerando asociaciones superiores a 0.75), se quita Cantidad de cursadas aprobadas por estar asociada con Cantidad de Cursadas Totales, Cantidad de cursadas promocionadas y Cantidad total de finales por estar ambas asociadas con Finales Aprobados. Las variables para armar el modelo son, entonces: Cantidad total de cursadas, Cantidad de cursadas abandonadas y Cantidad de finales aprobados.

El conjunto de datos con los que se trabajó tiene a cada uno de los alumnos clasificados en función de si abandonaron o no. Como sólo se tienen dos grupos, el método brindará una única función discriminante (función canónica).

Se hallaron las funciones discriminantes canónicas para todos los años académicos (las variables de clasificación fueron D1, D2, D3, …, D10) para cada una de las carreras. Hasta el quinto año se expone con apertura por carrera por contar con registros suficientes para determinar la función. Dado que a partir del sexto año se reduce significativamente la cantidad de registros, se calcula la función discriminante sin desagregar por carrera.

Respecto de los porcentajes de falsos negativos se observan valores entre 5 y 20%. A partir del sexto año, la definición de minimizar los falsos negativos es más evidente. Se permiten grados de error total mayor, pero se obtienen tasas de falsos negativos entre el 4 y el 14% (TABLA 1). Esto podría implicar que se haga seguimiento de estudiantes que no abandonarán, y se destinen recursos humanos especializados innecesarios. Como atenuante puede identificarse que la cantidad de alumnos matriculados a partir del año seis es muy menor a la de los primeros años.

## 4. CONCLUSIONES

El presente trabajo resulta ser una aplicación del proceso de descubrimiento de conocimiento a partir de bases de datos. Frente a un requerimiento concreto referido al abandono estudiantil, se seleccionan los datos, se los transforma en información mediante preprocesamiento y métodos de *data mining*, y dicha información constituye una base de conocimiento para la toma de decisiones.

El requerimiento consiste en la construcción de un modelo para predecir la probabilidad de abandono de estudiantes de las carreras Licenciado en Administración y Contador Pública de la Facultad de Ciencias Económicas de la UNICEN. Los datos fueron extraídos principalmente del sistema SIU-Guaraní, y se han estudiado las cohortes 2009 a 2019 de las carreras especificadas. Luego de una etapa de preprocesamiento de datos, se aplicó análisis discriminante para generar información. Esta información consiste en funciones discriminantes que permiten clasificar a un alumno de una carrera y de un año dado, como potencial alumno que abandonará (o no). En esta clasificación se consideraron tres variables de rendimiento académico: cantidad total de cursadas, cantidad de cursadas abandonadas y cantidad de finales aprobados, y se intentó minimizar el porcentaje de falsos negativos (el error de clasificar a priori dentro del grupo de no abandono a estudiantes que tengan intenciones de abandonar). Esto implica que se consideró preferible “usar recursos” adicionales en realizar el seguimiento de potenciales alumnos que podrían abandonar, aunque no tuvieran la intención de hacerlo. Se obtuvo un máximo porcentaje de falsos negativos de 20.47%.

La información generada constituye una base de conocimiento para la toma de decisiones acerca de la adopción de medidas de gestión académica tendientes a incrementar los índices de permanencia. A la vez, permiten plantear lineamientos y posibles campos de acción, que contribuyan con la formulación de políticas y programas llevados a cabo por el equipo de gestión de la Facultad.

| Año académico | Carrera | Función | % FP | % FN | Error Total |
|---|---|---|---|---|---|
| 1 | CP | - 3.02 + 0.50 CT - 0.48 CAB + 0.05 FA | 13.52 | 18.07 | 14.29 |
| | LA | 2.43 - 0.46 CT + 0.48 CAB - 0.02 FA | 17.37 | 16.36 | 17.14 |
| 2 | CP | 2.36 - 0.41 CT + 0.57 CAB - 0.06 FA | 13.79 | 15.92 | 14.02 |
| | LA | 2.13 - 0.39 CT + 0.49 CAB - 0.08 FA | 15.66 | 15.50 | 15.52 |
| 3 | CP | - 2.46 + 0.46 CT - 0.39 CAB + 0.06 FA | 19.76 | 11.11 | 19.09 |
| | LA | 1.94 - 0.38 CT + 0.42 CAB - 0.12 FA | 20.32 | 11.63 | 19.55 |
| 4 | CP | - 2.71 + 0.53 CT - 0.32 CAB + 0.03 FA | 16.36 | 5.88 | 15.82 |
| | LA | - 2.22 + 0.50 CT - 0.34 CAB - 0.01 FA | 17.71 | 5.26 | 16.54 |
| 5 | CP | - 2.63 + 0.54 CT - 0.49 CAB + 3.3E-03 FA | 11.85 | 11.76 | 11.85 |
| | LA | - 2.22 + 0.38 CT - 0.13 CAB + 0.23 FA | 13.79 | 20.47 | 19.88 |
| 6 | Todas | - 1.42 + 0.15 CT - 0.06 CAB + 0.44 FA | 28.64 | 5.77 | 27.28 |
| 7 | Todas | - 1.09 + 0.15 CT - 0.08 CAB + 0.50 FA | 43.24 | 4.26 | 40.03 |
| 8 | Todas | 0.57 - 0.12 CT + 0.67 CAB - 0.60 FA | 47.46 | 7.32 | 43.83 |
| 9 | Todas | - 0.76 + 0.16 CT - 0.38 CAB + 0.65 FA | 40.53 | 9.38 | 36.04 |
| 10 | Todas | 0.66 - 0.46 CT + 0.80 CAB - 0.50 FA | 40.71 | 14.29 | 36.57 |

**TABLA 1.** Funciones discriminantes canónicas con apertura por carrera. Porcentaje de falsos positivos (% FP), falsos negativos (% FN) y totales - Fuente: Elaboración propia

## 5. REFERENCIAS

Accinelli, A.; Losio, M.; Macri, A. (2016): "Acceso, rezago, deserción y permanencia de estudiantes en las universidades del conurbano bonaerense". *Debate Universitario*, vol. 5 (9), pp. 33-52.

Balzarini, M.; González, L.; Tablada, M., Casanoves F.; Di Rienzo, J.; Robledo, C. (2008): "Infostat. Manual Del Usuario". Ed. Brujas, Córdoba, Argentina.

Davenport, T. H. (1997): "Information ecology: Mastering the information and knowledge environment". Oxford University Press, 9.

Davenport, T. H.; Prusak, L. (1998): "Learn how valuable knowledge is acquired,

created, bought and bartered". The Australian Library Journal, vol. 47 (3), pp. 268-272.

Di Rienzo, J. A.; Casanoves F.; Balzarini, M. G.; González, L.; Tablada, M.; Robledo, C. W. (2020): "InfoStat versión 2020". Centro de Transferencia InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996): "From Data Mining to Knowledge Discovery in Databases". *Artificial Intelligence Magazine*, vol. 17 (3), pp. 37-54.

Tinto, V. (1989): "Definir la deserción: Una cuestión de perspectiva". *Revista de Educación Superior*, vol. 18 (71), pp. 1-9.

Tippins, M.; Sohi, R. (2003): "IT competency and firm performance: is Organizational Learning a missing link?" *Strategic Management Journal*, vol. 24 (8), pp. 745-761.

Vásquez Martínez, C.; Rodríguez Pérez, M. (2007): "La deserción estudiantil en educación superior a distancia: perspectiva teórica y factores de incidencia". *Revista Latinoamericana de Estudios Educativos*, vol. 37 (3), pp. 107-122.