Big Data

¿Superar el primer año?: un acercamiento a las características de los alumnos que aprueban el primer año universitario

Estefanía Walker

Lourdes Gil Deza

Pilar Hüppi Lo Prete

Primavera 2022

Introducción

La deserción de los estudios universitarios supone pérdidas de capital social, muestra ciertas ineficiencias del sistema que provee la educación y también implica costos a niveles familiares y sociales (Parrino, 2014). Este fenómeno social es consecuencia del accionar del mismo sistema educativo que implica cierta complejidad en los mecanismos que operan en dicho acontecimiento (Maccagno y Mangeaud, 2017). En particular, "en Argentina, se estima que más del 30% de los jóvenes que se inscriben en una carrera la abandonan antes del primer año[1]" (Parrino, 2014, p. 2). Es por esto que no sólo nos parece relevante la cifra de deserción universitaria sino también los motivos e implicancias que se vinculan a esta.

En este sentido, la pregunta de investigación que nos motiva es ¿qué variables de los alumnos ingresantes son relevantes o indicativas de que el alumno seguirá estudiando en la Universidad o la abandonará durante el primer año de estudio? Esta cuestión resulta fundamental al momento de definir sobre qué alumnos se debería realizar un seguimiento más cercano de sus trayectorias académicas, especialmente en los inicios de sus estudios universitarios.

El trabajo se propone generar, por medio del método de CART, el árbol de clasificación que indique cuáles son las características de los alumnos que abandonan la Universidad durante el primer año de estudio. De esta manera, en base a los datos presentados por los postulantes a la universidad, se definirá cuáles son los factores relevantes para conocer si en promedio, dadas sus características, el alumno abandona o no durante el primer año.

Este estudio podría ser interesante en nuestro contexto, dado que no está realizado en nuestro país y, particularmente, en la Universidad de San Andrés. Por esta razón, a pesar de que pudiera haber predicciones similares en otros entornos, resulta fundamental adaptarlo a las condiciones específicas de nuestra realidad. Esto se debe a que árboles de decisión correspondientes a otras Universidades no necesariamente son extrapolables a las particularidades de UdeSA. En este sentido, una herramienta con estas características podría colaborar en un mejor entendimiento de las condiciones de éxito de sus estudiantes y favorecer la toma de decisiones a futuro.

Si bien la Universidad cuenta con una estrategia para identificar a aquellos estudiantes que se beneficiarían de un acompañamiento mediante el Centro de Orientación al Alumno, este se lleva a cabo a partir de indicadores del rendimiento del alumno dentro de la Universidad, una vez que el alumno ya fue aceptado. Esto suele implicar que el acompañamiento se profundiza luego de que el alumno ha dado cuenta de un bajo rendimiento tanto en la asistencia como en sus primeros exámenes. Lo distintivo de nuestra propuesta es que dicho acompañamiento se podría realizar desde los primeros días de estudio de los alumnos, a fin de evitar fracasos que lleven a su abandono incluso en el primer año. Esto sería posible a partir de indicadores *ex ante,* definidos a partir de las variables que brindan poder explicativo en el árbol de decisión.

[1] En el caso de querer implementar el modelo propuesto en la práctica, resultaría importante recabar este dato a partir de la información disponible de la Universidad de San Andrés.

Esto resulta particularmente importante durante el primer año de estudio de los alumnos, dado que está demostrado que este y, en particular, las primeras experiencias de evaluación en la Universidad, son fundamentales al momento de definir la continuidad en los estudios (Willcoxson et al., 2011; Tinto, 1993; Johnson, 1994).

Literatura previa

El trabajo de Demeter et al. (2022) es un apropiado antecedente de la presente propuesta de investigación. Este artículo propone predecir si los ingresantes universitarios (por primera vez) se graduarán y cuándo lo harán. La información disponible para este trabajo predictivo son los registros de admisión académicos, solicitudes de ayuda financiera, horas de crédito, promedios del colegio secundario y universitario, aporte financiero de los padres y calificaciones en los cursos de ingreso. El método utilizado para predecir la graduación es Random Forest donde se logró una precisión del 79%. Este trabajo propone una importante referencia de cara a nuestra propuesta de investigación.

Asimismo, resulta relevante el trabajo de Aulck et al. (2017) que busca predecir el abandono universitario en base a datos demográficos (raza, género, fecha de nacimiento, estatus de residencia e identificación como hispano), datos provistos al momento de la inscripción (resultados de exámenes estandarizados y desempeño en la secundaria) y datos del desempeño en la Universidad (clases tomadas, notas, carrera seleccionada). Recurre al método de regresión logística y encuentra que las variables con mayor poder explicativo son las notas en inglés, química y psicología, así como el año de inscripción y año de nacimiento.

Otros trabajos encuentran que uno de los mejores predictores del rendimiento académico es el rendimiento previo, así como la asistencia y participación en clase (García Jiménez, Alvarado Izquierdo y Jiménez Blanco, 2000).

Es importante resaltar que los mencionados antecedentes difieren de nuestra propuesta en el hecho de que utilizan información de los alumnos dentro de la universidad como así también la información previo a su ingreso para predecir si el alumno se gradúa. Nuestro trabajo, tal como mencionamos previamente, constituirá un aporte en el hecho de que intentaremos clasificar el orden de relevancia de las variables que informan si el alumno completa el primer año de universidad, incluso antes de que haya tomado su primera clase.

Base de datos

Utilizaremos como base de datos la información que brindan los estudiantes en el formulario de inscripción a la Universidad de San Andrés. Esta información incluye:

- Datos demográficos sobre el alumno: sexo del DNI, género autopercibido, fecha de nacimiento, localidad de nacimiento, nacionalidad. Hay ciertos datos como nombre, apellido y número de documento que descartamos para el análisis.
- Datos educativos del alumno: año de finalización del nivel secundario, edad al momento de completar la inscripción, promedio del año en curso (si ya finalizó el secundario, promedio del último año), promedio del penúltimo y antepenúltimo año,

título obtenido o por obtener en el nivel secundario, si fue abanderado o escolta de la bandera nacional durante el último año de secundaria, está cursando o ya cursó en otra universidad.
- Datos del ingreso a San Andrés: carrera elegida, en qué semestre comenzaría, sede, si solicita asistencia financiera, qué beca solicitaría. Si bien existe también una presentación personal del alumno, esta sería descartada dado que no realizaremos un análisis de texto.
- Datos del responsable de arancel: fecha de nacimiento, nacionalidad, si vive, si vive con el alumno, estado civil, ciudad de residencia, provincia de residencia, país de residencia, ocupación, empresa, actividad, cargo, nivel de estudios, título.
- Datos del padre del alumno: fecha de nacimiento, nacionalidad, si vive, si vive con el alumno, estado civil, ciudad de residencia, provincia de residencia, país de residencia, ocupación, empresa, actividad, cargo, nivel de estudios, título.
- Datos de la madre del alumno: fecha de nacimiento, nacionalidad, si vive, si vive con el alumno, estado civil, ciudad de residencia, provincia de residencia, país de residencia, ocupación, empresa, actividad, cargo, nivel de estudios, título.

A su vez, como variable dependiente para nuestro modelo utilizaremos como dato si el alumno se dio de baja de la Universidad antes de comenzar a cursar su segundo año de estudios.

La base de datos se compondrá de esas variables, para las cuales cada estudiante representará una observación. Tenemos la intención de que la base de datos alcance el período más largo posible de manera tal que la información esté digitalizada.
Para conseguir esta información deberíamos solicitarla a la Universidad de San Andrés. Consultamos este tema con autoridades de la universidad, y nos comentaron que para utilizarla solo con este fin y en forma anonimizada, conseguirla podría ser factible.

Metodología

La metodología a utilizar buscará generar un árbol de toma de decisiones y, para esto, recurriremos al método de CART. Este devuelve un árbol de decisión con diferentes jerarquías de las variables que indican si un alumno superará el primer año de universidad.

La elección del método radica en que buscamos generar una herramienta fácil de comprender para cualquier persona no especializada en la materia. Es por esto que CART resulta una estrategia óptima para este objetivo. Quien utilice el árbol resultante va a poder observar de forma clara cuáles son las principales variables que se relacionan con abandonar la carrera durante el primer año. Asimismo, van a poder observar cuál es la jerarquía de éstas y cuáles son importantes para ciertos grupos en particular (sub nodos dentro de una rama). En especial, para el presente desafío consideramos relevante conocer las variables que indican a qué características del alumno se le deben prestar más atención. Mientras que si solo predecimos si un ingresante aprobará su primer año no sería posible conocer qué implicancia fue la más importante que definió esa condición.

Por su parte, el método de CART corresponde a una estrategia que consiste en generar subdivisiones que crean nodos que, a su vez, generan particiones homogéneas en función del objetivo de clasificación.

Formalmente, el problema a resolver será:

$$min_{j,s}[min_{c1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + min_{c1} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2]$$

Este implica buscar la variable de partición X_j y el punto de partición s. Con una cantidad de M regiones, el predictor que se obtiene es:

$$\hat{f}(x) = \sum_{m=1}^{M} \widehat{c_m} I(x \in R_m)$$

Donde c_m es el promedio de los y_i para todas las observaciones en la región R_m.

Este método presenta distintas ventajas. En primer lugar, es una buena herramienta para representar no linealidades. Asimismo, constituye una forma sencilla de comunicar y explicar dada su ventaja visual. Por último, se puede utilizar tanto con variables categóricas como numéricas. No obstante, también presenta distintas limitaciones. La principal desventaja es el riesgo de sobreajuste de esta estrategia dado que se pueden crear una cantidad de subdivisiones de manera tal que existan tantas particiones como observaciones de la muestra. Esto se soluciona incluyendo restricciones al modelo de clasificación y con el método de *pruning* (mediante el cual se podan aquellas ramas del árbol que aportan poco poder explicativo al modelo).

Indizando a los diferentes árboles T, considerando que el subárbol $T \in T_0$ se obtiene colapsando los nodos terminales (podando ramas) de otro árbol, y definiendo $[T]$ como el número de nodos terminales del árbol T, obtenemos que la función de complejidad del árbol T se puede escribir como:

$$C_\alpha(T) = \sum_{m=1}^{[T]} n_m Q_m(T) + \alpha[T]$$

Donde:

$Q_m(T) = \frac{1}{n_m} \sum_{x_i \in R_m} (y_i - \widehat{c_m})^2$

n_m = la cantidad de observaciones en cada partición

De esta manera, en la función, $Q_m(T)$ penaliza la impureza o heterogeneidad del árbol dentro de cada región, mientras que el término $\alpha[T]$ penaliza la cantidad de regiones.

Otra limitación de la estrategia corresponde al hecho de que no es óptimo para detectar linealidades dada las diferentes jerarquías de las variables incluidas. Por último, es una

estrategia de poca robustez dado que el árbol puede modificarse ante pequeños cambios en los datos disponibles.

Sin embargo, entendiendo sus limitaciones, consideramos que un árbol de decisión es una buena estrategia para realizar un primer acercamiento al fenómeno de estudio, e identificar cuáles son las variables que mayormente explican el éxito de los alumnos durante su primer año en la Universidad. En particular, su clara visualización y facilidad para la comunicación son características importantes para que los resultados sean compartidos con los actores relevantes.

Para construir el árbol de decisión recurriremos a los datos descritos anteriormente de los alumnos de segundo año en adelante, tomando su información al momento de la inscripción en la Universidad y sus resultados de primer año.

Conclusiones y limitaciones

Como resultados de nuestro modelo esperamos encontrar un árbol de decisión que nos permita visualizar claramente cuáles son las características de los alumnos que son indicativas de que abandonen la Universidad dentro del primer año de estudios. Como así también el orden de prioridad que toman tales variables en la definición.

Esperamos que la primera rama del árbol se base en el promedio general escolar del estudiante, dado que estudios previos encontraron que el desempeño en la escuela secundaria es un buen predictor del desempeño en la universidad (García Jiménez, Alvarado Izquierdo y Jiménez Blanco, 2000). Otras variables que podrían ser explicativas son: la región de procedencia del alumno, si recibe asistencia financiera, edad del alumno o datos académicos/laborales de los padres.

Una posible limitación de nuestro modelo podría ser que, al ir añadiendo datos progresivamente (incorporando las distintas camadas que avanzan a segundo año), el árbol se modifique dada su baja robustez y alta sensibilidad a modificaciones en los datos utilizados para su construcción. A su vez, encontramos que, por un lado, los datos sobre los cuales se elabora el árbol dependen de la forma en que fueron recabados. En este sentido, el cuestionario que brinda la información no fue diseñado para el estudio, sino que son datos administrativos que se encontraban previamente elaborados.

Por otro lado, encontramos que una limitación de la elaboración de un árbol de decisión estaría asociada a su utilización. Entendemos que podría haber cierta resistencia a depender de un algoritmo para identificar temas sensibles como puede ser el éxito o fracaso de los alumnos en la Universidad. Razón por la cual, enfatizaríamos el hecho de que es una herramienta adicional que puede colaborar en un mejor entendimiento de los fenómenos que acontecen en la institución, para así orientar los recursos disponibles a aquellos alumnos que podrían estar en riesgo. Es importante remarcar que, en este sentido, no planteamos este modelo en detrimento de las estrategias actuales de definición de alumnos en riesgo, a partir

de una visión integral sobre sus procesos de aprendizaje, sino como un instrumento más que colabore en la toma de decisiones humanas.

Consideramos relevante tener en consideración que los resultados de este modelo son simplemente una herramienta para simplificar y mejorar la tarea de acompañamiento a los alumnos para evitar que abandonen sus estudios. Se trata de una herramienta que debería ser utilizada como insumo para detectar a qué alumnos prestar más atención, y no como parte de la toma de decisiones de ingreso a la universidad. Cabe destacar la importancia de tratarla con seriedad, ya que se corre el riesgo de que las variables más relevantes que resulten del árbol puedan hacer referencia a factores que, utilizados de forma irresponsable, puedan ser tomados de forma discriminatoria.

Próximos pasos

Una posible continuación del trabajo podría ser intentar predecir qué alumnos van a abandonar la universidad antes del segundo año, en base a la información provista al momento de la inscripción. En ese caso, deberíamos recurrir al método de Random Forest, que constituye una metodología de ensamble de los Árboles de Decisión. Esto implica que se realizan múltiples árboles de decisión y se genera una predicción final por medio de promediar las predicciones individuales de todos los árboles. Consideramos que para esta nueva pregunta de investigación, sería relevante recurrir a una metodología de ensamble ya que permite obtener una predicción con menor varianza y menos posibilidad de sobreajuste que si utilizaramos únicamente CART para predecir.

Sin embargo, consideramos que en el caso de responder la pregunta de predicción, y no de clasificación, sería importante prestar atención al factor humano en la toma de decisiones de acompañamiento. Esto se debe a que particularmente en este caso, qué factores llevan a definir que un alumno sea propenso a abandonar antes del segundo año constituyen una especie de “caja negra”. Por lo tanto, en este caso sería aún más relevante que en el caso del árbol de clasificación tratar los resultados de forma integral, tomándolos simplemente como un indicador y no como un factor determinístico.

Otra posible continuación del proyecto sería realizar una predicción del motivo de abandono de los alumnos, usando como insumo las respuestas al formulario que se envía a los alumnos al decidir desmatricularse de la Universidad. De esta forma, podríamos realizar un mejor acompañamiento en base a los motivos de abandono predichos. Asimismo, si los formularios de salida lo permiten, un análisis más exhaustivo sería distinguir entre estudiantes que se dieron de baja implicando una salida del sistema educativo y aquellos que, en realidad, se cambian de universidad. Esto último podría dividir y profundizar el análisis de manera tal que, por un lado, se realicen las predicciones para alumnos abandonan definitivamente y, por otro lado, para quienes deciden cambiarse de universidad. La estrategia de predicción podría arrojar motivos de abandono distintos para cada grupo lo que también facilitaría un seguimiento más específico.

Referencias bibliográficas

Aulck, L. et al. (2017) Predicting Student Dropout in Higher Education. *DataLab*. The Information School. University of Washington.

Demeter, E. et al. (2022) Predecir los resultados de finalización de grado de los estudiantes que ingresan por primera vez a la universidad. *High Educ.* 84, 589–609. https://doi.org/10.1007/s10734-021-00790-9

García Jiménez, M., Alvarado Izquierdo, J., y Jiménez Blanco, A. (2000) La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema.* (12), 2, 248-252.

Johnson, G. (1994) Undergraduate student attrition: A comparison of the characteristics of students who withdraw and students who persist. *Alberta Journal of Educational Research* (40), 337–53.

Maccagno, A. y Mangeaud A. (2017) La deserción estudiantil en el primer año de la Universidad. Programa de Estadísticas Universitarias. Universidad Nacional de Córdoba (UNC)

Parrino, M. (2014) Factores intervinientes en el Fenómeno de la Deserción Universitaria. *Revista Argentina de Educación Superior.* (8), 39-61.

Tinto, V. (1993) *Leaving college: Rethinking the causes and cures of student attrition*. 2nd ed. University of Chicago Press.

Willcoxson, L., Cotter, J. y Joy, S. (2011) Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities. *Studies in Higher Education* (36), 3, 331-352.