# Análisis de la deserción estudiantil universitaria desde una perspectiva analítica

Artículo

**Jaime Miranda**
*Doctor en Sistemas de Ingeniería, Universidad de Chile*
*Director Escuela de Pregrado ESIA, Facultad de Economía y Negocios, Profesor Asociado, Universidad de Chile.*

***Jonathan Vázquez***
*PhD (c) in Computer Science, George Mason University*
*Profesor Adjunto, Universidad de Valparaíso.*

## 1. Introducción

La deserción estudiantil es una de las principales preocupaciones en todas instituciones privadas y públicas de educación superior debido a su impacto negativo. Por un lado, las instituciones educativas tienen una serie de impactos financieros, pues tener menos estudiantes se traduce en un menor ingreso por el pago de matrículas, lo que puede afectar la capacidad de la institución para ofrecer programas de calidad y mantener su infraestructura. Sin embargo, la deserción temprana en programas de pregrado no solo afecta a las instituciones educativas. Desde la perspectiva del estudiante, aquellos que desertan a menudo enfrentan consecuencias económicas negativas a largo plazo, causada por una menor probabilidad de conseguir empleos mejor remunerados y el aumentodel riesgo de desempleo. Además, la falta de formación en educación superior puede limitar su desarrollo personal y profesional a lo largo de su vida. Por otro lado, desde una perspectiva social, la deserción estudiantil disminuye los niveles de educación de la población, lo que puede conducir a tasas más altas de pobreza, problemas de salud, dependencia de asistencia social y tasas de criminalidad.También una fuerza laboral menos educada puede afectar la competitividad económica de un país. Por tanto, la pérdida de estos talentos afecta el progreso en sus comunidades y campos de trabajo, perpetuando ciclos de desigualdad de mayores brechas sociales y económicas [27].

La deserción surge por diferentes factores, tanto académicos como no académicos. Algunos identificados en la literatura son el rendimiento académico [1], la demografía del estudiante, la interacción social, las limitaciones financieras, la motivación y la personalidad [2]. Cada factor de riesgo, puede abordarse de diferentes maneras, como por ejemplo la asistencia académica (tutorías, asesoramiento y mentorías [3,4]), compromiso social y vinculación individual con la institución [5-6], propósito para finalizar la carrera (por ejemplo, vocación educacional, trabajos part time, pasantías) y asistencia financiera. Estos distintos tipos de enfoques de intervención son considerados por las universidades al momento de diseñar las campañas de deserción. En

primer lugar, se deben comprender adecuadamente los factores que gatillan la deserción estudiantil en su contexto educacional y luego, identificar y predecir con precisión a los estudiantes que tienen más probabilidad de responder de forma positiva a estas intervenciones. El presente artículo tiene como objetivo contribuir en este segundo aspecto, entregando una descripción y análisis en el uso de herramientas de analítica, para generar una alta precisión en la predicción de la deserción estudiantil de una institución educativa de la educación superior en Chile.

## Panorama actual y antecedentes claves

La deserción estudiantil ha sido ampliamente estudiada en la literatura [8,9]. Los primeros trabajos desarrollados entre los años 70' y 80' por [8-10], establecieron modelos sobre la deserción estudiantil, que luego fueron ocupados como punto de partida en el desarrollo de nuevos métodos de predicción de la deserción [11-14]. Por ejemplo, en [8], los autores usaron un enfoque interdisciplinario y que consideraba variables psicológicas, para proponer un modelo de deserción estudiantil, explicada como la interacción entre el estudiante y su entorno educativo. Posteriormente, [9] propuso un modelo de deserción que consideraba la relación entre los atributos previos a partir el programa y su interacción con sus entornos académicos y sociales. Finalmente, [10] extendió los modelos anteriores, incorporando nuevos elementos relacionados con la interacción entre los estudiantes y la institución educativa.

Estos trabajos, considerados como seminales en el estudio de la deserción estudiantil, generaron diferentes líneas de investigación que han sido ampliamente discutidos en los últimos años, los cuales han estudiado la incorporación de nuevos factores de la deserción y entregado distintas perspectivas de su análisis. Por ejemplo, una de ellas es el tiempo. En el estudio [13], se analizó el impacto que tiene la duración del programa académico en la deserción [13], mientras que otros estudios analizan la deserción en distintos momentos del programa: En [15], se estudió la deserción en los dos primeros años, en [16] entre el segundo y tercer año, y en [14] a lo largo de todo el programa. Por el contrario, otros estudios no se han centrado en la perspectiva temporal, sino más bien analizan el fenómeno de la deserción desde un punto de vista sistémico [17].

Otros trabajos discuten otros factores asociados con la deserción. Algunos de ellos se han centrado en la influencia de las características socioeconómicos de los estudiantes, arrojando distintos resultados, e incluso, en algunos casos, contradictorios. Por ejemplo, [12] identificó como importante el género en la deserción estudiantil, pero en [18] mencionan que la variable género no es un factor significativo. De manera similar, [14] afirmó que los estudiantes de bajos ingresos tienen menos probabilidades de abandonar su programa de licenciatura, en contraste con [19], que sugirió que este grupo tiene un mayor riesgo de abandono. Estas divergencias pueden indicar que la deserción estudiantil depende, en gran medida, de elementos contextuales, según el caso en que se desarrolla el estudio.

Respecto de la predicción, en los últimos años ha tomado fuerza el uso del aprendizaje automático o Machine Learning, dando lugar a un área nueva de investigación conocida como Minería de Datos Educativos (Educational Data Mining, EDM) [20]. En general, los estudios de deserción estudiantil aplican diferentes técnicas para esta tarea, tales como Semi-supervised learning [21], Unsupervised learning [22], y Ensemble learning [23], y la mayoría se centra en aumentar el desempeño en la detección de los estudiantes con más alta propensión a abandonar. Este estudio sigue esta última línea de investigación.

## Factores que deben ser considerados en la deserción de estudiantes

Existen diferentes estudios que muestran que la deserción estudiantil no puede explicarse solo por factores individuales, y lo asocian a un evento social generado por la ruptura del individuo con su entorno social [24]. Por ejemplo, desde las características de su familia, un estudiante es expuesto a diferentes influencias, expectativas o demandas, las cuales podrían afectar su desempeño académico [14]. La Figura 1 muestra los factores que afectan la deserción y sus relaciones planteado en [24].

**"La deserción estudiantil no puede explicarse solo por factores individuales".**

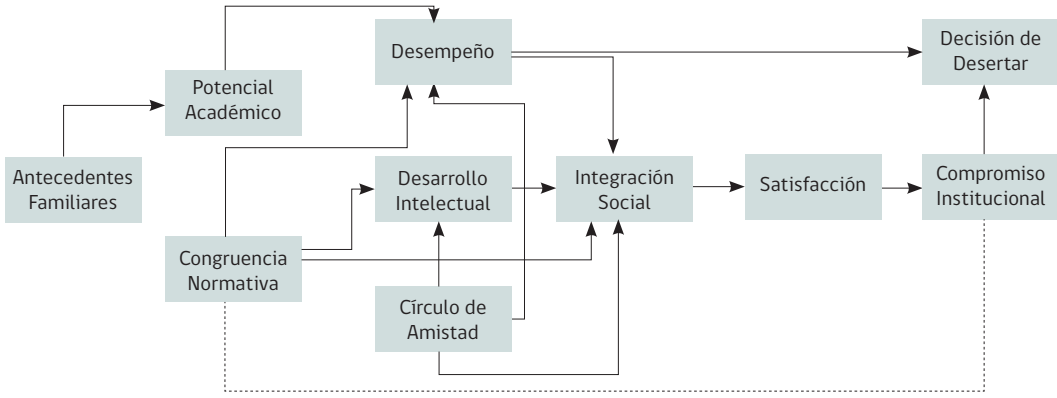


**Figura 1.** Factores que afectan la deserción y sus relaciones [24].

En general, se puede intuir que los estudiantesevitan las situaciones que, desde su percepción y estados emocionales con sus pares, le generan mayores costos que beneficios. Dentro de este ámbito, las características socioeconómicas de un individuo pueden afectar de sobremanera la decisión de deserción. En este contexto, [24] plantea que, asumiendo que no existe otra actividad que le genere mayores beneficios, los estudiantes se mantendrán en el programa siempre y cuando los beneficios percibidos superen el esfuerzo, dedicación y otros costos personales. Además, agrega que los compromisos del estudiante con la institución y sus objetivos personales de formación profesional son afectados por sus antecedentes familiares (por ejemplo, el nivel sociocultural), por sus atributos personales (por ejemplo, edad y género) y por su experiencia académica preuniversitaria (por ejemplo, rendimiento de exámenes de selección universitaria). De esta manera, luego de un tiempo razonable estando en el programa, el estudiante reevalúa sus compromisos iniciales de acuerdo a su integración social y su desempeño académico en la institución, cuyos efectos podrían desencadenar en la deserción estudiantil, si este percibe que los costos son mayores que los beneficios.

También, existen otros factores que gatillan en la deserción, como por ejemplo los antecedentes o desempeños académicos previos y su residencia. Según [26], la variable de desempeño académico previo y estatus socioeconómico impactan en la relación del estudiante con su entorno educacional. De esta manera, aquellos estudiantes que tienen antecedentes de excelencia académica en el colegio, tenderían a obtener mejores desempeños en la universidad, lo que aumentaría el grado de satisfacción, compromiso institucional y su decisión de no desertar. Además, identifica como factores importantes las variables sociodemográficas como el género, edad y etnia del cuerpo estudiantil, pues reflejan la heterogeneidad existente, las que se relacionan con las características propias del estudiante al momento de la deserción.

Aunque la importancia de los factores puede variar según el contexto institucional, estos pueden ser agrupados en al menos tres grupos importantes: Socioeconómico, académico (tanto preuniversitario como universitario) y social. Un resumen de algunas variables y su relación entre ellas, se muestra en la Figura 2.udiante al momento de la deserción.

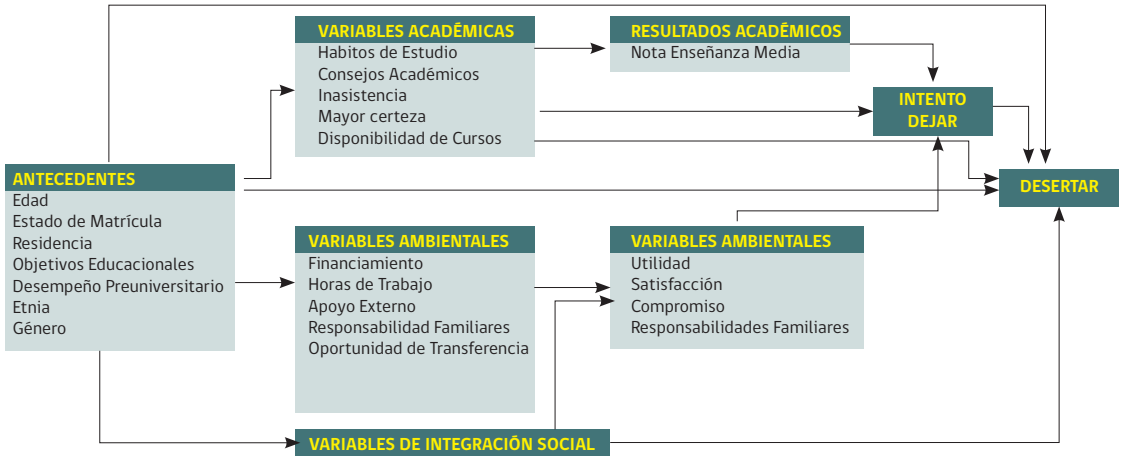


**Figura 2.** Relación entre las variables que afectan la deserción estudiantil [26].

## Enfoques analíticos en educación y sus principales desafíos

Los enfoques analíticos desempeñan un papel crucial en la identificación de predictores de la deserción estudiantil, facilitando la creación de sistemas computacionales que permiten detectar de forma temprana este comportamiento. Originalmente, la identificación de los predictores más relevantes en la detecccción de la deserción se ha centrado en enfoques econométricos. Sin embargo, debido a la diversidad de tipos de variables en distintos contextos educativos, las metodologías econométricas tienden a entregar resultados pocos generalizables. Esta limitación ha fomentado la inclusión de otras disciplinas con enfoques analíticos, tales como el Machine Learning y Deep Learning, las que han entregado en el campo modelos predictivos avanzados para detectar la deserción estudiantil anticipadamente, según el contexto institucional [27].

Los enfoques analíticos han avanzado con el análisis tecnológico y han permitido aplicarlos a distintos contextos y estudios educativos. Este progreso ha dado lugar a la formación de una nueva disciplina conocida como Minería de Datos Educativos o EDM. Esta se enfoca en el desarrollo de técnicas y modelos especializados para el análisis de datos derivados de la gestión educativa en universidades y centros de formación, siendo la creación de modelos para predecir la deserción estudiantil un problema intelectualmente interesante de explorar en el campo.

En los estudios de EDM, destacan tres desafíos principales sobrela deserción. El primer desafío es la generalización de los predictores, es decir, aquellosidentificados como significativos, pero que no siempre son aplicables en todos los contextos educativos. Por ejemplo, se han identificado variables relevantes como aquellas que captan las horas de traslado a la universidad, la residencia (dentro o fuera del Estado o región) y la etnia como predictores importantes para detectar deserción. Sin embargo, estas no están siempre disponibles en todas las instituciones de educación.

El segundo desafío es la identificación precisa de la ocurrencia en el horizonte temporal de la deserción a predecir. Esta ocurrencia no es siempre la misma para todos los programas de estudio, puesto que la duración en semestres y las particularidades de cada programa varían entre instituciones educacionales. De hecho, para un conjunto de programas que forman los mismos profesionales, es posible identificar variaciones a nivel de ciudad, región y país. Esto se traduce en que la ocurrencia de la deserción no siempre se concentra en el mismo semestre, lo que obliga a definir para el contexto institucional, cuál es la temporalidad en términos de semestre o año de la deserción que el modelo analítico predictivo tendrá como tarea predecir.

El tercer desafío surge del desbalance de las clases (deserta/ no deserta) en las bases de datos. Debido a la naturaleza de la deserción estudiantil, hay una preponderancia en el número de estudiantes que no deserta en comparación con los estudiantes que sí lo hacen, resultando en un desbalance significativo en la distribución de las clases. Esto es un problema relevante en las técnicas de Machine Learning, pues los modelos tienden a aprender más de la clase mayoritaria, obteniendo resultados más precisos en esta clase. Sin embargo, en la deserción, es mucho más importante predecir correctamente a estudiantes desertores de los que no lo son. Para tratar este problema, existen diferentes técnicas estudiadas en el campo, tales como el Random Oover Sampling (ROS) y el Random Under Sampling (RUS) [14].

## Modelos de Machine Learning

Las máquinas de aprendizaje o Machine Learning son técnicas que permiten a las computadoras aprender patrones en una base dedatos. De forma más concreta, son algoritmos capaces de generalizar comportamientos, a partir de una información estructurada y no estructurada suministrada en forma de ejemplos. Algunos modelos de Machine Learning utilizados en EDM son el Support Vector Machine, Árboles de Decisiones, Redes Neuronales y Regresiones Logísticas [14].

Las técnicas de Machine Learning permiten construir un indicador, el cual ayuda a estimar la probabilidad de que un estudiante deserte. Este indicador puede ser utilizado para establecer umbrales para clasificar a estudiantes en, donde para todos los estudiantes con un indicador sobre un umbral dado, por ejemplo b, son clasificados como desertores.

A modo de ejemplo, imaginemos que se cuenta con una base de 1.000 estudiantes y, sobre este conjunto, se aplica una técnica de Machine Learning, la que permite estimar la probabilidad de que un estudiante deserte. Con el cálculo de esta probabilidad, es posible ordenar de mayor a menor todos los registros, identificando aquellos estudiantes con un indicador igual o mayor al umbral b=0.5. Todos los estudiantes que cumplen esta condición podrían ser clasificados dentro del grupoque potencialmente podrían desertar. Eventualmente este umbral puede ser más o menos restrictivo, en donde si aumentamos dicho umbral de clasificación se requiera una mayor probabilidad para catalogar a un estudiante en la clase deserta. En general, mientras mayor sea el umbral, más probable es que efectivamente el estudiante deserte y, por tanto, que mayor confianza tengamos de la clasificación. Sin embargo, también aumenta la probabilidad de que un estudiante bajo el umbral definido sea erróneamente identificado como no deserta en circunstancias que sí lo es.

La interpretación y evaluación del desempeño de los modelos debe verificar si los patrones encontrados tienen sentido en el contexto que fueron aplicados. El desempeño de un modelo se puede ver afectado por muchos factores, tales como las

variables de entrada escogidas y el manejo que se hizo de los datos en términos de su pre procesamiento y transformación. En este sentido, se han planteado diferentes métricas para medir el desempeño predictivo de los modelos. Entre los más comunes, están el Error de clasificación y la Precisión de Predicción o Accuracy. Sin embargo, estas métricas miden el desempeño general de los modelos, asumiendo que todos los tipos de errores tienen el mismo costo, lo que no siempre es así en un contexto organizacional.

Dentro de este contexto, es posible identificar dos tipos de errores. El error tipo I, asociado al error cometido cuando se clasifica a un estudiante como desertor siendo que no lo es, y el error tipo II, que es cuando un estudiante se clasifica como no desertor y termina finalmente desertando. Para poder cuantificar ambos errores, en general, se utiliza una matriz de confusión (ver Tabla 1), la cual corresponde a una tabla que posee dos filas y dos columnas, donde las filas representan las predicciones del modelo, mientras que las columnas representan los resultados reales. Por ejemplo, las predicciones de estudiantes como positivo (deserta) son denominados como True Positives (TP, o Verdaderos Positivos) o False Positive (o Falso Positive), si efectivamente eran o no de la clase positiva, respectivamente. Del mismo modo, si un estudiante es asignando como NO deserta y en la realidad no desertó, se clasifica como True Negative (TN o Verdadero Negativo), mientras que se clasifica como False Negative (o Falso Negativo), si es clasificado como NO deserta y finalmente deserta.

| | | Clase Verdadera | |
|---|---|---|---|
| | | **Deserta (+)** | **NO deserta (–)** |
| **Predicción del modelo** | Deserta (+) | True Positive | False Positive |
| | NO deserta (–) | False Negative | True Negative |

**Tabla 1.** Ejemplo de matriz de confusión.

De la matriz presentada en la Tabla 1, se pueden obtener indicadores de desempeño, siendo el más común el índice que mide la exactitud de predicción. Adicionalmente, según la importancia de cada una de las clases, se pueden generar otros indicadores, tales como el Ratio Verdadero Positivo, o True Positive Rate en su versión internacional en inglés. Otros indicadores de desempeño son la Exactitud y el Ratio Verdadero Positivo. Ambos indicadores se calculan según las ecuaciones 1 y 2.

$$Exactitud\ (accuracy) = \frac{TP + TN}{TP + FP + TN + FN}$$

**Ecuación 1.** Fórmula para el cálculo de Exactitud.

$$Ratio\ Verdadero\ Positivo\ (TPR) = \frac{TP}{TP + FN}$$

**Ecuación 2.** Fórmula para el cálculo de ratio verdadero positivo.

## Caso aplicado: Descripción de los datos y resultados

En esta sección, mostraremos la aplicación de un enfoque analítico para la detección temprana de la deserción en una institución de educación superior en Chile. Esto tiene como objetivo facilitar el entendimiento y la aplicación de los conceptos y definiciones previamente discutidos. Adicionalmente, puede ser considerado como una guía a replicar en otra institución educacional, tomando en cuenta que, tal como discutimos anteriormente, los modelos analíticos en la predicción de la deserción son difícilmente generalizables entre instituciones, y deben ser desarrollados considerando los factores contextuales de la institución educacional en donde se desarrollan.

Inicialmente describimos la base de datos y las variables recopiladas, las cuales fueron consideradas en base a la literatura previamente discutida. Luego, considerando el segundo desafío, analizamos la ocurrencia de la deserción para

definir la temporalidad del fenómeno y que será información de entrada relevante del modelo. Adicionalmente, analizamos el impacto del desbalance de clases y consideramos distintas técnicas para tratar este problema en el diseño del modelo predictivo. Finalmente, medimos el desempeño de los distintos modelos diseñados, los cuales son construidos combinando distintos enfoques de Machine Learning.

## Desafíos en Predicción Estudiantil: Variables, temporalidad y desbalance

### Desafío 1: Variables

Hemos recopilado información de 3.627 estudiantes que se inscribieron entre los años 2012 y 2016 en tres programas de licenciatura de una escuela de negocios. Los datos se recopilaron desde tres bases de datos: **(1)** bases de datos internas de la institución educativa, **(2)** base de datos de becas y créditos y **(3)** base otorgada por la institución que lleva a cabo la prueba de selección universitaria.

Respecto de la base de datos interna, estaalmacena toda la información de los estudiantes respecto de la inscripción de cursos, cátedras cursadas, homologación de ramos, desempeño académico, solicitudes estudiantiles y evaluación docente por parte del alumno. Por otro lado, la información almacenada en la base becas y créditos está relacionada con todas las ayudas financieras, tanto el monto como el tipo de ayuda (becas, créditos y mantención). Finalmente, la tercera base de datos es enviada por el Departamento de Evaluación, Medición y Registro Educacional (DEMRE), administrado por la Universidad De Chile, la cual contiene información sociodemográfica, rendimiento académico de preuniversitario, puntajes de pruebas de admisión e historial de postulación de cada estudiante que rindió las pruebas de admisión administradas por el DEMRE. La lista completa de variables se muestra en la Tabla 2. La información almacenada en las distintas bases de datos permite obtener un total de 44 variables:

| Tipo de Variable | Variables |
|---|---|
| Académica | Número de cursos en el 1er semestre, año de ingreso, preferencias de programa de licenciatura, rendimiento en el 1er semestre, rendimiento en Estadísticas, Matemáticas, Economía, Inglés 1er semestre. |
| Antecedentes Familiares | Madre o padre como jefe de familia, número de padres vivos, nivel educativo de la madre y padre, número de miembros de la familia, número de miembros de la familia en universidad, escuela secundaria, número de miembros de la familia que trabajan, número de miembros de la familia en preescolar, número total de miembros de la familia , ocupación del padre y ocupación de la madre. |
| Características del Colegio | Colegio privado, subvencionado o público, colegio masculino, femenino o mixto, año de graduación del colegio. |
| Sociodemográfica | Región central, norte o sur, apoyo de los padres, estudiante independiente, cobertura de salud privada o pública, fuente de financiamiento, becas, apoyo de los padres como primera fuente de financiamiento, género, estado civil, ingreso familiar bruto, horario de trabajo si el estudiante trabaja. |
| Puntajes de Admisión | Puntaje del examen de ingreso, Lenguaje y Comunicación, Matemáticas, Historia y Ciencias, puntaje de ranking, calificaciones de la escuela secundaria y edad durante el 1er semestre. |

**Tabla 2:** Variables utilizadas en los modelos predictivos.

### Desafío 2: Temporalidad

Según el análisis mostrado en la Figura3, los primeros seis semestres de los programas son cruciales para la detección de las deserciones voluntarias, ya que el 97% de ellas ocurren en los tres primeros años, concentrándose en el tercer semestre.

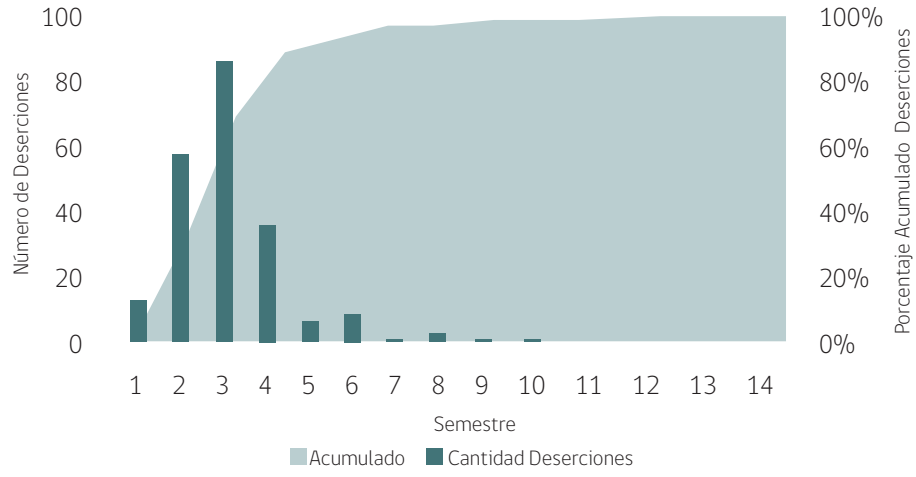


**Figura 3:** Comportamiento deserción por semestre.

### Desafío 3: Desbalance

La Tabla 3 muestra la distribución de las deserciones voluntarias de los estudiantes por semestre. Existe una marcada presencia de la clase No Deserta, tal como fue discutido en las secciones previas.

| Semestre | Deserta | No Deserta | Total |
|---|---|---|---|
| **Sem 1** | 13 | 595 | 608 |
| **Sem 2** | 58 | 591 | 649 |
| **Sem 3** | 86 | 591 | 677 |
| **Sem 4** | 36 | 578 | 614 |
| **Sem 5** | 7 | 486 | 493 |
| **Sem 6** | 9 | 577 | 586 |
| **Total general** | **209** | **3.418** | **3.627** |

**Tabla 3.** Distribución de las deserciones voluntarias de los estudiantes.

### Implementación de modelos

Dado el análisis previo, se construye un modelo que predice la deserción para cada semestre hasta el tercer año. Las variables para cada modelo son todas aquellas disponibles hasta el inicio del semestre a predecir. Por ejemplo, para la predicción del Semestre 1, se usa toda la información hasta el inicio del semestre, tales como desempeño colegio, sociodemográfica, puntajes y postulación; mientras que para el Semestre 2 se usan las variables para la predicción del Semestre 1, más el desempeño del estudiante durante el primer semestre en el programa.

En primera instancia se pre procesaron las bases de datos creadas para cada semestre. En cada base de datos semestral, se consideraron solamente los registros completos. Los atributos categóricos fueron transformados a binominales, generando n nuevas columnas, donde n es la cantidad de distintas categorías

únicas del atributo. De estas n nuevas columnas, se seleccionaron n-1 para evitar problemas de multicolinealidad. Por otro lado, los atributos numéricos fueron normalizados en un rango de 0 a 1.

Para el desarrollo de los modelos, se decidió combinar modelos no supervisados con supervisados de la siguiente manera: primero se aplicó un algoritmo de clusterización, el cual entrega como resultado el número de centroides óptimos, como también la asignación de cada registro a uno de estos centroides. Posteriormente, los cluster fueronentrenados con distintas técnicas de modelos supervisados, calculando el desempeño de cada uno e identificando cuál de todos era el mejor. Los modelos a usar fueron: Support Vector Machine, Árbol de Decisión, Red Neuronal, y Regresión Logística. Adicionalmente, con el objetivo de explorar si la aplicación de clustering mejora la predicción, también se aplicaron los algoritmos de aprendizaje a las bases semestrales sin clusterizar.

El umbral de clasificación b fue determinado usando la base de testeo, donde se identificó aquel umbral de los modelos de Machine Learning, que maximizaba el desempeño medido en la base de testeo

Finalmente, dado que se identificaron problemas en la distribución de las clases, se aplicaron técnicas de desbalance antes de la construcción de cada modelo. Y al igual que con el clustering, también se utilizaron bases de datos no balanceadas y se compararon en términos del desempeño de los modelos.

Combinando las distintas técnicas, se generaron en total 48 procesos para cada semestre, los cuales generaron 288 modelos en total. A modo de resumen, se construyeron modelos combinando: (1) Clusterización, (2) técnicas de desbalance (ROS, RUS y sin técnicas de desbalance), (3) técnicas de Machine Learning (SVM, Red Neuronal, Árboles de Decisión y Regresión Logística)..

### Resultados de los experimentos

La Tabla 4 muestra un resumen de los resultados de cada modelo para cada semestre de predicción. Cada uno, fue evaluado en base a la Exactitud y TPR, según Ecuaciones 1 y 2. Primero se muestra el resultado obtenido por los mejores modelos, luego se presentan las técnicas aplicadas o no para cada uno. Por ejemplo, para el Semestre 1, elmejor modelo se configura con un clustering, sin una técnica de balanceo y con SVM, mostrando una Exactitud y TPR del 90,69% y 100%.

En general, según la Tabla 4, los modelos que mejor desempeño tuvieron fueron aquellos en donde se aplicó clusterización, ninguna técnica de desbalance y SVM. Solamente para los semestres quinto y sexto, el modelo con mejor desempeño fue construido con una técnica de Machine Learning distinta, que fue la Regresión Logística.

| Técnicas | Sem 1 | Sem 2 | Sem 3 | Sem 4 | Sem 5 | Sem 6 | N° usos |
|---|---|---|---|---|---|---|---|
| **Indicadores de desempeño de los Mejores Modelos** | | | | | | | |
| Exactitud | 90.69% | 80.08% | 72.56% | 84.17% | 94.47% | 95.43% | |
| TPR | 100.00% | 77.27% | 76.00% | 85.19% | 100.00% | 87.50% | |
| **Clustering** | | | | | | | |
| **Clustering** | ✓ | ✓ | | ✓ | ✓ | ✓ | 5 |
| **No Clustering** | | | ✓ | | | | 1 |
| **Balance de clases** | | | | | | | |
| **No Balanceado** | ✓ | ✓ | ✓ | | ✓ | | 4 |
| **ROS** | | | | ✓ | | ✓ | 2 |
| **RUS** | | | | | | | 0 |
| **Técnicas de Machine Learning** | | | | | | | |
| **SVM** | ✓ | ✓ | ✓ | ✓ | | | 4 |
| **Árbol de decisión** | | | | | | | 0 |
| **Red Neuronal** | | | | | | | 0 |
| **Regresión Logística** | | | | | ✓ | ✓ | 2 |

**Tabla 4:** Desempeño de los mejores modelos por semestre.

### Variables importantes según mejores modelos

Adicionalmente, considerando el mejor modelo, calculamos el nivel de importancia que cada modelo le daba a las variables e identificamos aquellas clasificadas más importantes, según su peso obtenido en los modelos. De esta manera, es posible identificar que las variables relacionadas con la prueba de ingreso, desempeño académico universitario y del preuniversitario, nivel educacional de los padres, número de integrantes que componen el grupo familiar, cantidad de integrantes de la familia trabajando, participación en semestres de verano y financiamiento son importantes.

Analizando la lista de todos los predictores de cada semestre, se puede constatar que aquellos relacionados con la configuración familiar y el rendimiento durante el transcurso de la carrera son los que más se repiten durante todos los semestres, lo que se condice con los modelos teóricos planteados entre los años 70´ y 80′ y que también muestran que los antecedentes familiares y el desempeño académico del estudiante son variables primordiales, para explicar la deserción del estudiante.

A nivel general, las variables consideradas importantes para todos los semestres son las relacionados con el puntaje de la prueba de admisión del estudiante, seguido del nivel educacional de los padres y el desempeño académico universitario. En un segundo nivel de importancia, destaca el financiamiento y la configuración familiar del estudiante.

### Conclusiones y trabajos futuros

Sin lugar a dudas, es posible describir y predecir con alta precisión el perfil de un estudiante con alta probabilidad de desertar, mediante el uso de técnicas de Machine Learning. Estas técnicas permiten además identificar las variables más relevantes y de mayor impacto, las que pueden ser usadas para mejorar el diseño de políticas personalizadas, con el fin de aumentar la retención de estudiantes.

Desde una perspectiva analítica, con los modelos desarrollados es posible identificar de forma temprana los estudiantes que decidirían dejar de estudiar en un semestre específico. Si consideramos que las variables de rendimiento académico son más importantes en todos los semestres, los administradores y asesores académicos de la institución en estudio podrán prestar mayor atención al rendimiento académico de los estudiantes antes de partir el semestre, y desarrollar políticas que estén acorde a esta información. Además, estas predicciones pueden usarse como insumo para reducir las tasas de deserción, como talleres de contextualización profesional para mejorar el compromiso institucional y programas de apoyo académico o psicológico para mantener los objetivos personales.

**“Los antecedentes familiares y el desempeño académico del estudiante son variables primordiales, para explicar la deserción del estudiante”.**

## *REFERENCIAS BIBLIOGRÁFICAS*

**[1]** L. Thomas, Student retention in higher education: the role of institutional habitus, J. Educ. Pol. 17 (4) (2002) 423-442.

**[2]** T. Dharmawan, H. Ginardi, A. Munif, Dropout detection using non-academic data, 2018 4th International Conference on Science and Technology (ICST), IEEE, 2018, pp. 1-4.

**[3]** C.J. Bland, A.L. Taylor, S.L. Shollen, A.M. Weber-Main, P.A. Mulcahy, Faculty Success through Mentoring: A Guide for Mentors, Mentees, and Leaders, R&L Education, 2009.

**[4]** S. Larose, D. Cyrenne, O. Garceau, M. Harvey, F. Guay, F. Godin, G.M. Tarabulsy, C. Deschênes, Academic mentoring and dropout prevention for students in math, science and technology, Mentor. Tutor. 19 (4) (2011) 419-439.

**[5]** N. Zepke, L. Leach, Improving student engagement: ten proposals for action, Act. Learn. High. Educ. 11 (3) (2010) 167-177.

[**6]** M. Yorke, The development and initial use of a survey of student belongingness, engagement and self-confidence in UK higher education, Assess. Eval. High. Educ. 41 (1) (2016) 154-166.

**[8]** W.G. Spady, Dropouts from higher education: an interdisciplinary review and synthesis, Interchange 1 (1) (1970) 64-85.

**[9]** V. Tinto, Dropout from higher education: a theoretical synthesis of recent research, Rev. Educ. Res. 45 (1) (1975) 89-125.

**[10]** J.P. Bean, Interaction effects based on class level in an explanatory model of college student dropout syndrome, Am. Educ. Res. J. 22 (1) (1985) 35-64.

**[11]** R. Chen, S.L. DesJardins, Investigating the impact of financial aid on student dropout risks: racial and ethnic differences, J. High. Educ. 81 (2) (2010) 179-208.

**[12]** A. Fortin, L. Sauvé, C. Viger, F. Landry, Nontraditional student withdrawal from undergraduate accounting programmes: a holistic perspective, Acc. Educ. 25 (5) (2016) 437-478.

**[13]** B.M. Kehm, M.R. Larsen, H.B. Sommersel, Student dropout from universities in europe: a review of empirical literature, Hungarian Educ. Res. J. 9 (2) (2019) 147-164.

**[14]** J. Vásquez, J. Miranda, Student desertion: What is and how can it be detected on time? Data Science and Digital Business, Springer, 2019, pp. 263-283.

**[15]** A.L. Caison, Analysis of institutionally specific retention research: a comparison between survey and institutional database methods, Res. High. Educ. 48 (4) (2007) 435-451.

**[16]** C.H. Yu, S. DiGangi, A. Jannasch-Pennell, C. Kaprolet, A data mining approach for identifying predictors of student retention from sophomore to junior year, J. Data Sci. 8 (2) (2010) 307-325.

**[17]** G. Johnes, R. McNabb, Never give up on the good times: student attrition in the UK, Oxf. Bull. Econ. Stat. 66 (1) (2004) 23-47.

**[18]** M. Ferreira, Gender issues related to graduate student attrition in two science departments, Int. J. Sci. Educ. 25 (8) (2003) 969-989.

**[19]** M. Saldaña, O. Barriga, An adaptation of tinto's attrition model to the universidad católica de la santísima concepción, chile, Rev. Ciencias Soc. 16 (4) (2016) 616-628.

**[20]** P.L. Peterson, E. Baker, B. McGaw, International Encyclopedia of Education, Elsevier Ltd., 2010.

**[21]** G. Kostopoulos, S. Kotsiantis, P. Pintelas, Estimating student dropout in distance higher education using semi-supervised techniques, Proceedings of the 19th Panhellenic Conference on Informatics, Athens, Greece, 2015, pp. 38-43.

**[22]** N. Iam-On, T. Boongoen, Generating descriptive model for student dropout: a review of clustering approach, Human-Centric Comput. Inf. Sci. 7 (1) (2017) 1.

**[23]** N. Iam-On, T. Boongoen, Improved student dropout prediction in thai universityusing ensemble of mixed-type data clusterings, Int. J. Mach. Learn. Cybern. 8 (2) (2017) 497-510.

**[24]** Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. Interchange, 1, 64-85.

**[25]**Tinto, V. (2007). Taking student retention seriously. Syracuse University.

**[26]** Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. Review of Educational Research, 55, 485-540

**[27]** Olaya, D., Vásquez, J, Maldonado, S., Miranda, J., Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. Decision Support Systems, 134, 485-540