

Procedimiento de agrupación de estudiantes según riesgo de abandono para mejorar la gestión estudiantil en educación superior

Procedimento de agrupamento de alunos de acordo com o risco de evasão para melhorar a gestão estudantil no ensino superior
Student clustering procedure according to dropout risk to improve student management in higher education

Mauricio Hinojosa  *¹, Iván Derpich  †¹, Miguel Alfaro  ‡¹,
David Ruete  §², Alejandro Caroca  ¶² e Gustavo Gatica  ||²

¹Universidad de Santiago de Chile, Facultad de Ingeniería, Departamento de Ingeniería Industrial, Santiago, Región Metropolitana, Chile.

²Universidad Andrés Bello, Facultad de Ingeniería, Santiago, Región Metropolitana, Chile.

Resumo

La compleja problemática del abandono estudiantil representa una oportunidad para la aplicación de la tecnología y métodos de la minería de datos en educación superior. El objetivo de esta investigación es obtener el perfil de los estudiantes en riesgo de abandono y así generar planes de gestión estudiantil que impacten sobre las variables que explican esta situación. Para esto se propone utilizar una estructura metodológica CRISP-DM, aplicando herramientas estadísticas y del aprendizaje automático no supervisado. El análisis, de tipo transversal, se realizó sobre un universo de estudiantes de primer año en jornada diurna, en una universidad privada chilena. Las variables sociodemográficas y de comportamiento utilizadas, se basaron en la teoría sobre deserción y juicio de expertos, y los datos se obtuvieron de los registros históricos disponibles en la Institución. Para obtener las variables que más influían en el abandono se realizaron análisis de correlación y de componentes principales. La aplicación del clustering jerárquico aglomerativo y la técnica de conjuntos aproximados produjeron cuatro perfiles de estudiantes con sus reglas de asociación respectivas y cinco variables académicas que permitieron diseñar un sistema de apoyo para disminuir el abandono y fomentar la retención.

Palavras-chave: Abandono estudiantil. CRISP-DM. Análisis de componentes principales. Clustering jerárquico aglomerativo. Conjuntos aproximados.

Resumo

O complexo problema da evasão de alunos representa uma oportunidade para a aplicação de tecnologia e métodos de mineração de dados no ensino superior. O objetivo desta pesquisa é obter o perfil dos alunos em risco de evasão e, assim, gerar planos de gestão estudantil que impactem nas variáveis que explicam essa situação. Para isso, propõe-se a utilização de uma estrutura metodológica CRISP-DM, aplicando ferramentas estatísticas e aprendizado de máquina não supervisionado. A análise transversal foi realizada em um universo de alunos do primeiro ano do turno diurno de uma universidade privada chilena. As variáveis sociodemográficas e comportamentais utilizadas foram baseadas na teoria da desercção e no julgamento de especialistas, e os dados foram obtidos nos registros históricos disponíveis na Instituição. Para obter as variáveis que mais influenciaram o abandono, foram realizadas análises de correlação e de componentes principais. A aplicação do agrupamento hierárquico aglomerativo e da técnica de conjuntos aproximados produziu quatro perfis de alunos com suas respectivas regras de associação e cinco variáveis acadêmicas que permitiram desenhar um sistema de apoio para reduzir o abandono e promover a retenção.

Palavras-chave: Evasão escolar. CRISP-DM. Análise de componentes principais. Agrupamento hierárquico aglomerativo. Conjuntos aproximados.

Abstract

*Email: mauricio.hinojosa@usach.cl

†Email: ivan.derpich@usach.cl

‡Email: miguel.alfaro@usach.cl

§Email: druete@unab.cl

¶Email: acaroca@unab.cl

||Email: ggatica@unab.cl

Textolivre
Linguagem e Tecnologia

DOI: 10.35699/1983-3652.2022.37275

Seção:
Artigos

Autor Correspondente:
Mauricio Hinojosa

Editor de seção:
Hugo Heredia Ponce
Editor de layout:
Leonado Araújo

Recebido em:
26 de novembro de 2021
Aceito em:
26 de janeiro de 2022
Publicado em:
2 de março de 2022

Essa obra tem a licença
"CC BY 4.0".



The complex problem of student dropout represents an opportunity for the application of data mining technology and methods in higher education. The objective of this research is to obtain the profile of students at risk of dropping out and thus generate student management plans that impact on the variables that explain this situation. For this, it is proposed to use a CRISP-DM methodological structure, applying statistical tools and unsupervised machine learning. The cross-sectional analysis was carried out on a universe of freshmen day students at a private Chilean university. The sociodemographic and behavioural variables used were based on attrition theory and expert judgment, and the data were obtained from the historical records available at the Institution. To obtain the variables that most influenced dropout, correlation and principal component analyses were performed. The application of agglomerative hierarchical clustering and rough sets technique produced four profiles of students with their respective association rules and five academic variables that allowed the design of a support system to reduce dropout and promote retention.

Keywords: Student dropout. CRISP-DM. Principal component analysis. Agglomerative hierarchical clustering. Rough sets.

1 Introducción

La deserción o abandono estudiantil es un problema reconocido y estudiado mundialmente en la educación (ARRIAGA; VELÁSQUEZ, 2013), impacta en la efectividad, eficiencia y prestigio de los sistemas e instituciones educacionales, y puede generar consecuencias económicas y/o psicosociales negativas en los estudiantes y sus familias (SCHREIBER; LUDEMAN, 2020; OECD, 2009).

Con la tendencia hacia la universalización de la Educación Superior (ES), una mayor proporción del alumnado proviene de los sectores socioeconómicos más vulnerables, cuya preparación académica muchas veces es deficiente, y deben compatibilizar estudios y trabajo (HINOJOSA V., 2021). Esto genera un mayor abandono en este grupo socioeconómico, que se refleja en una menor tasa de finalización que perjudica el acceso con equidad (FERREYRA et al., 2017; SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2014). Además, la pandemia del COVID-19 aumenta la probabilidad de abandonar de estos estudiantes debido a una inevitable reducción en la calidad del proceso de enseñanza/aprendizaje (OECD, 2021).

En Latinoamérica, los sistemas de acreditación y/o aseguramiento de la calidad ya consideran este fenómeno (LEMAITRE, 2017). En el caso de Chile, su Sistema Nacional de Aseguramiento de la Calidad (MINISTERIO DE EDUCACIÓN, 2006) incluye un proceso de acreditación, el cual, según la nueva Ley de Educación Superior, exige a las instituciones de educación superior (IES) acreditarse, entre otros, en el ámbito de la docencia y resultados del proceso formativo (MINISTERIO DE EDUCACIÓN, 2018). En base a lo anterior, la Comisión Nacional de Acreditación (CNA) propone criterios y estándares de evaluación, donde destacan que IES posean procedimientos definidos de seguimiento y análisis de retención. Para pregrado (CINE 5 y 6) (UNESCO-UIS, 2012) definen como estándar mínimo una tasa de retención institucional de primer año mayor o igual a 65 %, o que la institución presente una tasa de retención de cohorte al término del tercer año, mayor o igual a 60 % (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2020; COMISIÓN NACIONAL DE ACREDITACIÓN (CNA), 2020).

Se estima que en Latinoamérica se gradúan anualmente un 43 % de quienes ingresan a ES (CALDERÓN et al., 2017). En promedio, el año 2013, solo el 46 % de los estudiantes entre 25 y 29 años que estaban matriculadas completaron sus estudios, el 22 % abandonó y el 32 % aún continuaba estudiando (FERREYRA et al., 2017).

En Chile, al 2014, un 30 % de los estudiantes que ingresaban a primer año no continuaban, de los cuales el 44 % reingresaba al sistema de ES en los siguientes tres años (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2014). Los niveles de abandono son más altos en ciertos grupos de estudiantes, en particular para jornada vespertina o quienes provienen de establecimientos públicos (UNESCO-UIS / OECD / EUROSTAT, 2005). Poseer financiamiento, ya sea propio o externo (gratuidad, becas y/o créditos), reduce la probabilidad de abandono de los alumnos (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2020), y a mayor ingreso de estos, menor es el impacto de las ayudas financieras (BARRIOS RUBIO, 2011). Se estima que el abandono es 22 puntos porcentuales mayor para estudiantes que no cuentan con ayudas financieras. Al año 2012 se

mantenía en las universidades privadas independientes (UNESCO-UIS / OECD / EUROSTAT, 2005) un 50 % más de abandono promedio en primer año (30 %) que en las privadas dependientes y las públicas (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2014).

En cuanto a la influencia que puede ejercer la IES, se observa un elevado porcentaje de abandono en universidades privadas independientes (UNESCO-UIS / OECD / EUROSTAT, 2005) de baja selectividad (45 % versus el 30 % del sistema), en universidades no acreditadas o acreditadas entre 1 y 3 años (67 %) y aquellas universidades que presentan un puntaje promedio menor a 500 puntos (62 %) en las pruebas de selección universitaria¹ de lenguaje y matemática (LARROUCAU, 2015).

El momento del abandono es una variable de control importante (KEHM; LARSEN; SOMMERSEL, 2020), la mayoría lo hace en el primer año de su carrera (FERREYRA et al., 2017). Así, la Tasa de retención de primer año es uno de los indicadores más utilizados a nivel internacional para evaluar la eficiencia de las IES (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2020). En Chile, la Tasa de retención de primer año cohorte 2019 fue de 74 %, valor que representa una disminución después de seis años de constante crecimiento, provocada por la crisis social y sanitaria que impacta desde el 2018 (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2020).

El abandono es un fenómeno complejo, multidimensional y dinámico que implica el retiro del estudiante de un programa de ES antes de alcanzar su titulación (ARRIAGA; VELÁSQUEZ, 2013; HIMMEL, 2002; SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2019; CALVACHE F. et al., 2018; DELEN, 2010; MUNIZAGA; CIFUENTES, 2018). A los factores académicos, psicosociales, familiares y económicos que impactan en la deserción (HOSSLER; BEAN, 1990; TINTO, 1998; ISHITANI; DESJARDINS, 2002; ST. JOHN et al., 2000), la literatura reporta el énfasis en factores psicológicos (MEDRANO et al., 2010; CABRERA; BETHENCOURT BENÍTEZ et al., 2006; ROSÁRIO et al., 2012), así como también en el rol fundamental que tienen las IES de generar las condiciones que aseguren la persistencia de los alumnos (TINTO, 1998; KUH et al., 2010; FRANCISCO; MUÑOZ; SALDAÑA, 2011; BRAXTON; MCCLENDON, 2002).

Existe en la literatura un conjunto de modelos y teorías explicativas que permiten analizar este fenómeno (SPADY, 1970; TINTO, 1975, 2005; BEAN, 1985; ETHINGTON, 1990; NORA; RENDON, 1990; BRAXTON; JOHNSON; SHAW-SULLIVAN, 1997). Según (CABRERA; BETHENCOURT et al., 2006), dichas propuestas se pueden categorizar bajo cuatro categorías de modelos: de adaptación, estructural, economicista y psicopedagógico. Estas se resumen en la tabla 1.

La literatura en Chile está compuesta mayoritariamente por estudios que buscan identificar los factores que influyen en el abandono, ya sea a través de regresión (BARRIOS RUBIO, 2011; LARROUCAU, 2015; HENRÍQUEZ; ESCOBAR-RIFF, 2016; CARVAJAL; GONZÁLEZ, 2018), o aprendizaje automático supervisado (MIRANDA; GUZMÁN, 2017; MALDONADO et al., 2021). También están los estudios descriptivos (GONZÁLEZ F; URIBE JORQUERA; GONZÁLEZ VIDAL, 2005; GONZÁLEZ; URIBE, 2018), predictivos (HOFFLINGER, 2020; VELÁSQUEZ; HITPASS, 2014), y los conceptuales (HIMMEL, 2002; DÍAZ PERALTA, 2008).

En cuanto a las medidas que pueden implementar las IES para mitigar este problema, el Centro Interuniversitario de Desarrollo (CINDA) propone una serie de acciones específicas tendientes a apoyar al estudiante, y a un mejoramiento en el currículo, en el proceso de enseñanza-aprendizaje, en la atención al alumno y en la gestión docente (CENTRO INTERUNIVERSITARIO DE DESARROLLO (CINDA), 2006). En (HUESCA RAMÍREZ; CASTAÑO CORVO, 2007) se propone que debe existir, además, un sistema de control de la deserción.

En base a lo expuesto, el propósito de este estudio es formular un procedimiento para agrupar a los estudiantes de pregrado de primer año en una universidad privada según el riesgo de abandono, a partir de la identificación y análisis de las causas, cuyos resultados servirán de base para generar planes de mejora en la gestión institucional.

La metodología empleada se estructura sobre el proceso CRISP-DM (CHAPMAN et al., 1999), que permite sistematizar y estandarizar el trabajo de minería de datos. Las variables explicativas se

1 La batería de pruebas de selección universitaria es uno de los tres factores de selección de estudiantes para ingresar a las universidades chilenas

Tabela 1. Factores y teorías en el abandono estudiantil, estructurados en base a (CABRERA; BETHENCOURT et al., 2006)

| Modelo | Autores | Factores |
|-----------------|---|---|
| Adaptación | (TINTO, 1993), (BEAN; METZNER, 1985), (NORA, 2002) | Atributos de preingreso Experiencias institucionales |
| Estructural | (THOMAS, 2002), (LUJAN; RESENDIZ, 1981) | Políticos Económicos Sociales |
| Economicista | (SCHULTZ, 1962), (BECKER, 1962), (THUROW, 1973) | Capital humano Beneficio/Costo Rentabilidad Gasto en formación versus satisfacción personal Costos versus futuros ingresos Proyectos alternativos y su evaluación Periodo beneficios de formación |
| Psicopedagógico | (RYAN; GLENN, 2003), (GONZÁLEZ F; URIBE JORQUERA; GONZÁLEZ VIDAL, 2005) | Personales Institucionales Sociales Psicológicos Educativos |

Fonte: Elaboración propia

obtienen a partir de la recopilación de datos de registros históricos, de una síntesis de los factores explicativos, categorizada en los cuatro enfoques de (CABRERA; BETHENCOURT et al., 2006), y en el juicio de expertos. Para identificar las variables definitivas se utilizan herramientas estadísticas como el análisis de componente principales (FÉNYES; MOHÁCSI; PALLAY, 2021) y el de correlación (HERNÁNDEZ et al., 2018). Mediante clustering jerárquico aglomerativo (BEHR et al., 2021) se generan grupos de observación con un orden de prioridad, a partir de lo cual se definen los distintos perfiles de estudiantes. Con el método de conjuntos aproximados (BOUZAYANE; SAAD, 2017), se obtienen las reglas de decisión que permiten caracterizar los distintos conjuntos de alumnos. Y, a partir de estos resultados, se proponen acciones que propenden a mitigar la deserción y a fomentar la retención estudiantil.

En los siguientes apartados se describe la metodología, se discuten los resultados y, finalmente, se presentan las conclusiones.

2 Metodología

Se realiza un estudio cuantitativo de tipo exploratorio y descriptivo, basado en la aplicación de la metodología CRISP-DM (CHAPMAN et al., 1999) a la problemática de la deserción universitaria, mediante el Análisis de Componentes Principales (ACP) (FÉNYES; MOHÁCSI; PALLAY, 2021), Clustering Jerárquico Aglomerativo (CJA) (BEHR et al., 2021) y Conjuntos Aproximados (CA) (BOUZAYANE; SAAD, 2017).

La investigación considera una universidad privada de Chile, con cuatro sedes y distintos campus distribuidos en el país. La población objetivo corresponde a estudiantes en primer año de pregrado, matriculados entre los años 2012 y 2017 (VARGAS et al., 2019). La existencia y posibilidad de acceso a estos datos de la Institución determinan la utilización de variables pertinentes al estudio. Se define como el abandono de un programa de estudios de pregrado después del primer año académico (SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES), 2020).

El análisis de minería de datos es mediante el Proceso Transindustrial Estándar o CRISP-DM (Cross Industry Standard Process for Data Mining) (CHAPMAN et al., 1999). Por lo anterior, el proceso se dividió en 6 etapas, cada una con sus tareas y entregables, como se expone en la Figura 1, cada etapa se describe como sigue.

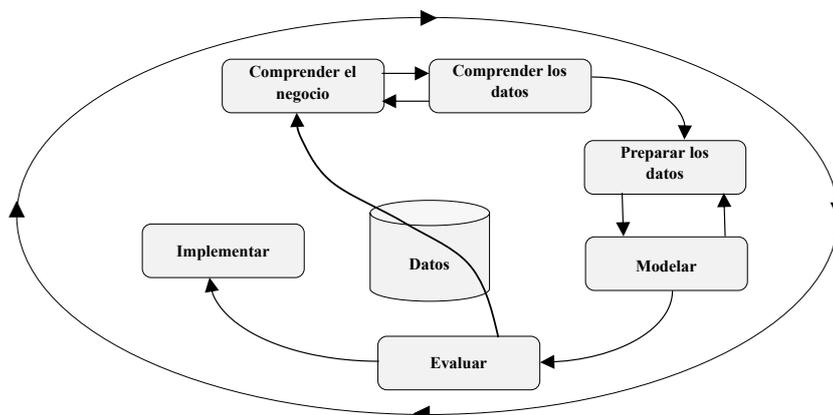


Figura 1. Modelo de proceso CRISP-DM

Fonte: (WIRTH, 2000)

2.1 Comprender el negocio

Se definen los criterios de éxito cualitativos o cuantitativos. Luego, se evalúa el estado de la situación antes de iniciar el proceso de minería de datos, que involucra el nivel de conocimiento disponible acerca del problema. Se analiza la cantidad de datos disponibles y requerida, la relación costo beneficio, entre otros, para definir los requisitos del problema, tanto en términos del negocio como de minería de datos. Y, finalmente, se genera el plan del proyecto, el cual describe los pasos a seguir y las técnicas a emplear.

2.2 Comprender los datos

La segunda fase (fig. 2) se inicia con la recopilación de los datos iniciales, extraídos desde una plataforma informática. Además, para establecer las variables que impactan en la deserción, se considera previamente lo expuesto en (CABRERA; BETHENCOURT et al., 2006) (ver Tabela 2).

Tabela 2. Lista potencial de factores considerados.

| Autor | Factores considerados |
|--|---|
| (LUJAN; RESENDIZ, 1981) | Políticos Económicos Sociales |
| (RYAN; GLENN, 2003); (WASSERMAN, 2001); (KIRTON, 2000); Yip (2019) | Personales Institucionales Sociales Psicológicos Educativos |

Fonte: Elaboración propia

Luego, mediante entrevistas, los factores se validan por expertos internos y externos a la Institución, cuya opinión acotó a cuatro factores: económicos, educacionales, personales y sociales.

En la descripción de las variables se dividen en sociodemográficas y de comportamiento. Las seis variables sociodemográficas consideran el año y la edad al ingresar a la Institución (Año proceso y Edad al ingreso), género (Sexo), sede (Sede), jornada diurna o vespertina que asiste (Jornada) y país donde realizó sus estudios previos (País).



Figura 2. Actividades de la etapa de comprensión de los datos
Fonte: Elaboración propia en base a (WIRTH, 2000)

Las diez variables de comportamiento se subdividen en dos grupos, las de comportamiento previo y posterior al ingreso a estudiar, como se describe en la Tabela 3.

Una tercera actividad es la exploración de los datos. Se analizan las distribuciones para así examinar subconjuntos de datos que representan las subpoblaciones a caracterizar.

Preliminarmente, desde el ámbito demográfico, se encontró que a medida que aumenta la edad, disminuye la frecuencia de estudiantes en el rango etario. Las frecuencias relativas mostraron que la mayoría, un 61 %, son hombres. Casi la totalidad, un 99,6 %, realizaron sus estudios previos en establecimientos nacionales y no en el extranjero. Un 86 %, están inscritos en jornada diurna y el resto en vespertina. Y, por último, un 86 %, inició su carrera desde un principio, es decir, no se realizó alguna convalidación u homologación de asignaturas.

Desde una perspectiva institucional, las sedes con mayor captación de alumnos son La Serena y Santiago, con un 34 % y 28 % respectivamente. Y las facultades con mayor ingreso son Ciencias de la Salud (30 %) y Educación (20 %), que suman la mitad del total de casos estudiados (las facultades con menor cantidad de casos son la de Odontología y Ciencias Agropecuarias).

En el ámbito académico, existen estudiantes que provienen de colegios más vulnerables, que optan por el ingreso vía ranking, y alcanzan el 36 %. La mayoría de las notas NEM se concentran entre 5,0 y 6,0, y representan el 73 % del total. El 37 % de los casos estudiados abandonó.

Como cuarta actividad, se verifica la calidad de los datos, es decir, revisar que estén completos y sin errores o valores fuera de rango. Por un lado, se descubren registros con valores inconsistentes, principalmente por ausencia de datos en NEM y puntajes de prueba (lenguaje, matemática, ciencia e historia), alumnos sin asignaturas inscritas, entre otros. Además, se identifican registros con estudiantes de la sede Concepción y/o la Facultad de Arquitectura, cuyos conflictos administrativos pueden generar un sesgo en el análisis.

2.3 Preparar los datos

Se integran datos mediante cruce de información (fig. 3). Luego, se eliminan todos los registros inconsistentes, identificados en la etapa anterior. Así, se obtiene una base de 6299 casos de un total de 9523, que implica una limpieza del 34 % de los datos. La tercera tarea es la construcción de los datos. Para permitir una mejor representación de la información, se expresa en porcentaje la

Tabela 3. Variables de comportamiento previo y posterior al ingreso a estudiar.

| Variables preingreso | Variables posingreso |
|--|--|
| NEM - factor de selección para ingresar a ES, correspondiente al promedio de notas de la enseñanza media (CINE 3 y 4) (UNESCO-UIS, 2012). | Asignaturas inscritas – número de materias inscritas en el año, considerando los dos semestres. |
| Facultad - área del conocimiento que contiene a la carrera elegida por el estudiante. | Asignaturas aprobadas – número de materias aprobadas en el primer año, considerando los dos semestres. |
| Puntaje Ranking - factor de selección para ingresar a ES, cuyo puntaje representa el lugar que ocupa el puntaje NEM del estudiante en relación con el promedio de notas de todos los alumnos de su contexto educativo de las tres generaciones anteriores a la suya. | Asistencia – resultado de la división del número de horas asistidas y la cantidad total de horas de clase correspondiente a las asignaturas inscritas en el primer año, expresado como porcentaje. |
| Promedio puntaje Lenguaje y Matemática - factor de selección para ingresar a ES, definido como el puntaje promedio entre los resultados obtenidos en las pruebas de Lenguaje y Comunicación, y Matemática. | Abandono - variable objetivo que refleja la situación de seguir estudiando o no durante el segundo año. |
| Puntaje Ciencias o Historia - factor de selección para ingresar a ES, que equivale al puntaje alcanzado en la prueba de Ciencias o en la de Historia, Geografía y Ciencias Sociales. | |
| Vía admisión a estudios superiores – representa el modo de ingreso del estudiante a la educación superior. | |

Fonte: Elaboración propia

variable Asignaturas aprobadas, a partir del cociente entre asignaturas aprobadas e inscritas; además, se generó la variable dicotómica Abandono, obtenida de la misma base, que verifica si el alumno se matriculó o no en el año posterior a su ingreso.

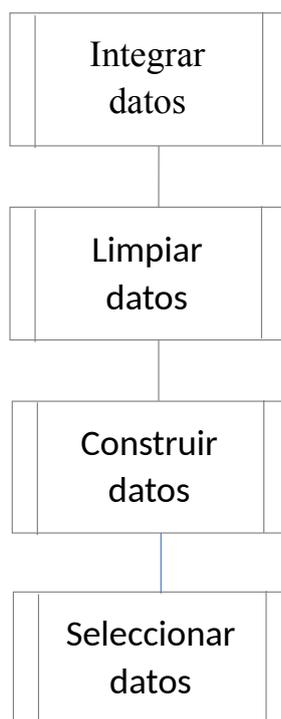


Figura 3. Actividades de la etapa de preparación de los datos

Fonte: Elaboración propia en base a (WIRTH, 2000)

Dicho lo anterior, las variables se reducen a trece, descartándose Año proceso, Sexo y Jornada, porque entre sus posibles valores no se puede atribuir mayor importancia a uno por sobre otro respecto al fenómeno de deserción. Además, se renombran y codifican las variables cualitativas factibles de jerarquizar, ya que el programa XLSTAT (ADDINSOFT, 2021) no reconoce este tipo de variables. Se aplica una codificación entera a Riesgo país de procedencia (antes llamada Nacionalidad), Riesgo facultad (antes llamada Facultad), Riesgo sede (antes llamada Sede), Riesgo ranking (antes llamada Puntaje Ranking) según (DONOSO; SCHIEFELBEIN, 2007), Riesgo vía admisión (antes llamada Vía admisión) según (DONOSO; SCHIEFELBEIN, 2007), y Abandono. Finalmente, se aplican dos métodos para seleccionar las variables más predominantes en la deserción. En primera instancia se realiza un análisis de correlación, con el cual se obtienen las variables más correlacionadas con la deserción (HERNÁNDEZ et al., 2018).

Para reforzar la selección de variables se lleva a cabo el ACP (FÉNYES; MOHÁCSI; PALLAY, 2021). Así, se obtienen los valores propios de la matriz de correlaciones, que corresponden a las componentes principales, combinaciones lineales de las variables estandarizadas. Se trabajó, además, con el gráfico de sedimentación y el círculo de correlaciones para determinar los factores a considerar.

2.4 Modelar

A partir de las variables más correlacionadas con la deserción, se definen diversos perfiles de estudiante, agrupados según la similitud de sus comportamientos. En un primer paso (fig. 4), se propone aplicar CJA, definir las distintas clases o grupos, analizar estadísticamente los resultados obtenidos para establecer las variables que explican el abandono y así, definir los perfiles de estudiantes (ANTONENKO; TOY; NIEDERHAUSER, 2012).

Para determinar la medida de disimilaridad se utilizó el método de Ward (VIJAYA; SHARMA; BATRA, 2019). Consiste en unir los dos clúster para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada clúster, de cada individuo al centroide del clúster. Así, se obtiene la ventaja de no dejar grupos que tengan pocos elementos,

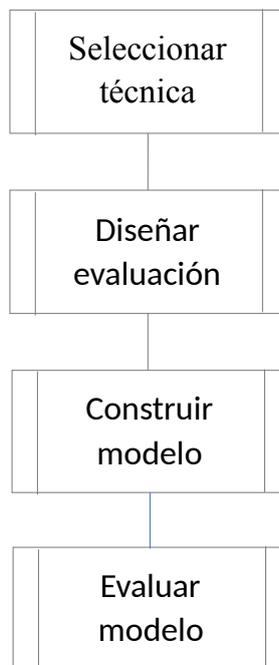


Figura 4. Actividades de la etapa de modelación

Fonte: Elaboración propia en base a (WIRTH, 2000)

se facilita la formación de clúster más compactos y de tamaño relativamente similar, y se minimiza la pérdida de información en el proceso de organización de conglomerados (TSENG et al., 2016).

Para detectar las diferencias entre clases se obtuvo el comportamiento promedio según variables de entrada. Para verificar si la media aritmética de los factores es una herramienta adecuada de comparación entre clases, se calcula la moda, la mediana y la desviación estándar para cada factor en cada clase. Para evaluar la dispersión en cada variable de entrada seleccionada se utiliza el gráfico de caja y bigote. De esta manera se puede observar en cuáles clases y en qué cuartil se presenta una oportunidad de hacer seguimiento y mejorar los índices obtenidos, al indagar en las posibles causas que originan esta situación.

Los perfiles obtenidos a partir del CJA (BURITICÁ et al., 2019), según el comportamiento de las variables definidas como de mayor preponderancia, se validan o refutan. Se busca identificar y definir las principales reglas de asociación de los estudiantes de cada grupo, en relación con las variables en estudio y el fenómeno de deserción. Para esto se utiliza el método de conjuntos aproximados con un enfoque basado en dominancia (DRSA) (BOUZAYANE; SAAD, 2017; GRECO; MATARAZZO; SLOWINSKI, 1999, 2002), el cual aplica a problemas de decisión multicriterio basado en relaciones de dominancia, para obtener la aproximación de una relación de preferencia. El software 4eMka2 (GRECO; MATARAZZO; SLOWINSKI, 1999; LABORATORY OF INTELLIGENT DECISION SUPPORT SYSTEMS IDSS, 2021) entrega diversas reglas de decisión basadas en condiciones y los siguientes criterios de medición de las reglas: coverage, support y strength (razón porcentual entre support y coverage que indica la consistencia). Así, se escogen las reglas de mayor relevancia respecto a los valores alcanzados en estos índices, al considerar como variable decisional el hecho de abandonar o no.

Para obtener reglas confiables se realizó el mismo análisis anterior por clases, pero se estudian de manera independiente las reglas obtenidas para los alumnos que abandonan y quienes no. Luego, se resumen las reglas según mayor consistencia presentada, independientemente de la clase de pertenencia. Con esto se obtienen variables que permiten explicar transversalmente el fenómeno de deserción para el universo de estudiantes en cada una de las clases definidas.

Finalmente, se obtienen las reglas de asociación, con el número de clase del estudiante como variable decisional, sin considerar la variable Abandono. Así, se obtienen las principales reglas que rigen para cada clase, además de verificar la complementariedad existente entre los distintos grupos.

Con el análisis de clúster, entonces, se obtienen las variables que mayor incidencia podrían tener sobre la deserción, y si tienen la incidencia suficiente como para explicar este fenómeno por sí solas.

2.5 Evaluar

A partir de las variables más correlacionadas con el abandono y de los perfiles de estudiantes, se generan dos propuestas de mejora al modelo de gestión estudiantil con el fin de mitigar la deserción y fomentar la retención.

3 Resultados y discusión

En la preparación de los datos, tercera etapa del proceso CRISP-DM, se obtienen los resultados del análisis de correlación (Tabla 4), del cual solo siete variables de las trece consideradas presentan un coeficiente de correlación mayor a 0,1 (HERNÁNDEZ et al., 2018), descartándose las otras seis.

Tabla 4. Coeficiente de correlación entre las 12 variables consideradas y la deserción.

| Variables de entrada | Abandono (Sí/No) |
|--|------------------|
| Edad al ingreso | -0,002 |
| Riesgo país de procedencia | 0,009 |
| Riesgo facultad | 0,062 |
| Riesgo sede | -0,006 |
| NEM | -0,121 |
| Riesgo Ranking | -0,139 |
| Promedio puntaje Lenguaje y Matemática | -0,114 |
| Puntaje Ciencias o Historia | -0,104 |
| Asignaturas inscritas | -0,436 |
| Asignaturas aprobadas | -0,598 |
| Asistencia | -0,476 |
| Riesgo vía admisión | -0,046 |

Fonte: Elaboración propia

Del ACP aplicado a las trece variables originales se obtiene el gráfico de sedimentación (fig. 5) y el círculo de correlaciones (fig. 6). En el primero se observa que las dos primeras componentes acumulan el 39 % de la variabilidad, que puede considerarse en un principio un nivel de representación aceptable dada la tasa de acumulación de variabilidad menor a partir de la tercera componente.

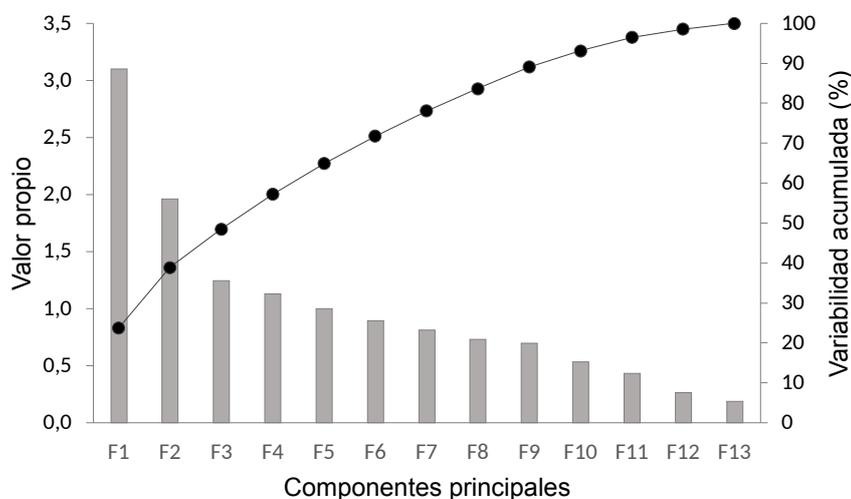


Figura 5. Valores propios y porcentaje de la inercia explicada, de las componentes principales

Fonte: Elaboración propia.

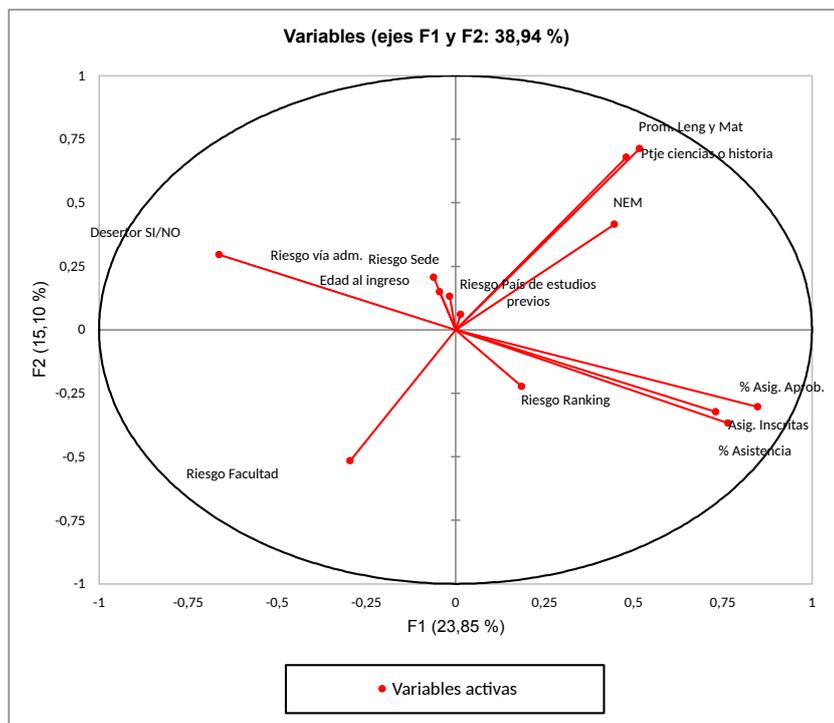


Figura 6. Círculo de correlaciones para las trece variables

Fonte: Elaboración propia.

El círculo de correlaciones muestra que solo las variables académicas posingreso (Asignaturas aprobadas, Asignaturas inscritas y Asistencia) están fuertemente correlacionadas con la deserción. De las variables preingreso, cuatro no están correlacionadas (Promedio puntaje Lenguaje y Matemática, Puntaje Ciencia o Historia, NEM, Riesgo facultad) y solo Riesgo Ranking presenta cierta correlación. Para el resto de las variables esta gráfica no permite generar inferencias. Por esta razón, se trabaja con las siete variables resultantes del análisis de correlación.

A través de CJA se obtuvieron cuatro grupos o clases, cuyas características permiten su diferenciación (Tabla 5). Se observa que a medida que un grupo presenta una mayor deserción promedio, menor es el valor promedio en las variables de entrada, excepto para la variable descartada Riesgo Ranking. Aun así, el grupo que presenta la mayor tendencia al abandono (grupo 3) no alcanza a presentarlo en la mayoría de los estudiantes (43,7 %).

Tabla 5. Promedio de variables para los 4 grupos obtenidos, ordenados por variable Abandono.

| Variable | Clase | | | |
|--|--------|--------|--------|--------|
| | 4 | 1 | 2 | 3 |
| NEM | 5,98 | 5,62 | 5,47 | 5,35 |
| Riesgo Ranking | 1,44 | 1,27 | 1,34 | 1,41 |
| Promedio puntaje Lenguaje y Matemática | 631 | 541 | 467 | 392 |
| Puntaje Ciencias o Historia | 631 | 560 | 481 | 373 |
| Asignaturas inscritas | 12,8 | 11,4 | 10,9 | 10,6 |
| Asignaturas aprobadas | 86,7 % | 79,7 % | 70,5 % | 62,7 % |
| Asistencia | 85,4 % | 82,2 % | 80,2 % | 78,0 % |
| Abandono | 0,250 | 0,314 | 0,371 | 0,437 |

Fonte: Elaboración propia

Para llegar a resultados más concluyentes, se profundizó el análisis anterior dividiendo los grupos en subgrupos de estudiantes que abandonan y los que no (Tabla 6).

Al comparar las variables de entrada por subgrupos y grupos que abandonan y quienes no, se

Tabela 6. Promedio de variables de entrada para los 4 grupos obtenidos, subdivididos según condición de abandono y ordenados por variable Abandono.

| Variable | Abandono | | | | | | | |
|-----------------------|----------|--------|---------|--------|---------|--------|---------|--------|
| | Clase 4 | | Clase 1 | | Clase 2 | | Clase 3 | |
| | No | Sí | No | Sí | No | Sí | No | Sí |
| NEM | 6,01 | 5,91 | 5,63 | 5,62 | 5,51 | 5,42 | 5,39 | 5,30 |
| Promedio pje. L. y M. | 631 | 630 | 541 | 540 | 468 | 466 | 394 | 388 |
| Puntaje Cs. o Hist. | 632 | 629 | 560 | 561 | 482 | 480 | 373 | 372 |
| Asignaturas inscritas | 13,5 | 10,6 | 12,0 | 10,1 | 12,0 | 9,0 | 12,0 | 8,8 |
| Asignaturas aprobadas | 0,94 | 0,64 | 0,89 | 0,59 | 0,86 | 0,43 | 0,83 | 0,37 |
| Asistencia | 88,85 | 74,97 | 85,71 | 74,58 | 86,38 | 69,74 | 85,96 | 67,61 |
| Abandono | 25,0 % | 75,0 % | 31,4 % | 68,6 % | 37,1 % | 62,9 % | 43,7 % | 56,4 % |

Fonte: Elaboración propia

puede inferir que:

Las variables NEM, Promedio puntaje Lenguaje y Matemática, y Puntaje Ciencias o Historia, no presentan una diferencia que permita explicar una tendencia al abandono en alguno de los grupos.

Las variables Asignaturas inscritas, Asignaturas aprobadas y Asistencia, muestran una clara diferencia entre subgrupos. En los estudiantes que abandonan, los valores promedio de estas variables son menores a quienes no lo hacen, independiente del grupo de pertenencia.

Con este resultado, se procedió a caracterizar los perfiles de los estudiantes, y se considera finalmente las variables Asignaturas aprobadas y Asistencia. Se descartó Asignaturas inscritas ya que al ingresar a un programa de estudio las asignaturas están predefinidas para el primer año, y en el caso que el alumno llegase a reprobado alguna, probablemente deba tomar menos asignaturas durante el segundo semestre, aspecto explicable con la variable Asignaturas aprobadas.

Para evaluar la dispersión de los datos en las dos variables de entrada seleccionadas se utilizó el gráfico caja y bigote. Se observan, para Asignaturas aprobadas (fig. 7), rendimientos académicos decrecientes en los alumnos, reflejados en la mediana, con porcentajes de aprobación de 100 %, 92 %, 85 % y 75 %, correspondientes a las clases 4, 1, 2 y 3, respectivamente. Así, el menor rendimiento de las clases 1, 2 y 3 representa una oportunidad de mejora, en especial en los alumnos de la clase 3, que poseen mayor riesgo. En cuanto a la concentración de los datos, es mucho mayor para la clase 4, en la cual la mediana y el cuartil 3 coinciden con el valor máximo, y es mucho menor en la clase 3.

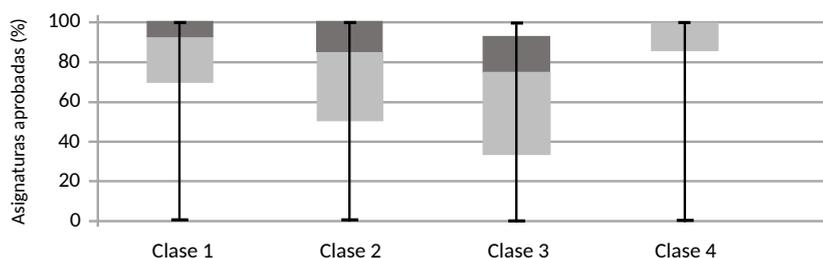


Figura 7. Caja y bigote de variable Asignaturas aprobadas, por cada grupo de estudiantes

Fonte: Elaboración propia.

Para la variable Asistencia se observa la misma situación anterior, pero mucho más leve (fig. 8). La asistencia tiende a decrecer con una mediana de 89 %, 86 %, 86 % y 84 % para los grupos 4, 1, 2 y 3, respectivamente. Así también, la dispersión tiende a aumentar, con un rango intercuartil de 9, 12, 13 y 15 puntos porcentuales para los grupos 4, 1, 2 y 3, respectivamente. De esta forma, en las cuatro clases se presenta una oportunidad de hacer seguimiento y mejorar los índices obtenidos, en especial con los alumnos del cuartil 1, en donde se deberán indagar las posibles causas de la baja asistencia.

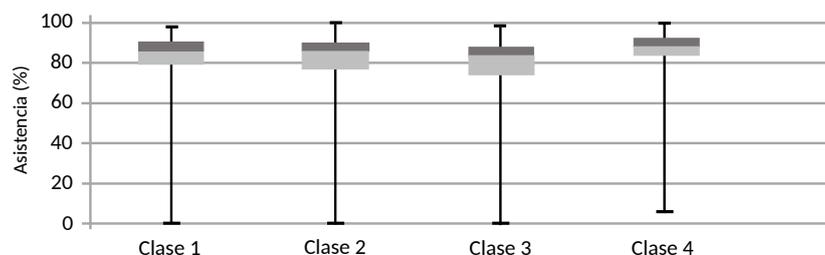


Figura 8. Caja y bigote de variable Asistencia, por cada grupo de estudiantes

Fonte: Elaboración propia.

De acuerdo con lo anterior, se agruparon los alumnos según el comportamiento de las variables definidas como de mayor preponderancia: aquellos que no abandonan ya que poseen porcentajes adecuados de asistencia y asignaturas aprobadas, y los que presentan un nivel insuficiente en uno o los dos factores mencionados, llevándolos a abandonar.

Así, al utilizar los resultados obtenidos con el CJA, se establecen los perfiles descritos en la Tabela 7.

Tabela 7. Perfiles de estudiantes.

| Perfil alumno | Descripción |
|---------------|--|
| Bajo riesgo 4 | Alta probabilidad de ser retenidos de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 4, poseen un riesgo bajo. |
| Bajo riesgo 1 | Alta probabilidad de ser retenidos de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 1, poseen un riesgo medio-bajo. |
| Bajo riesgo 2 | Alta probabilidad de ser retenidos de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 2, poseen un riesgo medio-alto. |
| Bajo riesgo 3 | Alta probabilidad de ser retenidos de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 3, poseen un riesgo alto. |
| Alto riesgo 4 | Alta probabilidad de abandono de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 4, poseen un riesgo bajo. |
| Alto riesgo 1 | Alta probabilidad de abandono de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 1, poseen un riesgo medio-bajo. |
| Alto riesgo 2 | Alta probabilidad de abandono de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 2, poseen un riesgo medio-alto. |
| Alto riesgo 3 | Alta probabilidad de abandono de acuerdo con asistencia y aprobación de asignaturas. Además, al pertenecer a la clase 3, poseen un riesgo alto. |

Fonte: Elaboración propia

Con las cuatro clases definidas, las trece variables de entrada y el programa 4eMka2, se obtienen las reglas de asociación por grupo mediante conjuntos aproximados, cuyos resultados se muestran en la Tabela 8.

Debido a la baja solidez (strength) de las 12 reglas obtenidas, y del mismo modo que en la etapa del clustering, se decidió profundizar el análisis obteniendo de manera independiente las reglas de decisión para estudiantes que abandonan y los que no (Tabela 9).

Del análisis conjunto de las reglas de asociación con mayor consistencia (strength > 80 %), obtenidas para los alumnos de primer año que abandonan, se obtiene que las variables que presentan

Tabela 8. Resultado de la aplicación de la técnica de conjuntos aproximados a los cuatro grupos y 13 variables.

| Clase | Regla | Estado | Condiciones | Strength | Support | Coverage |
|-------|-------|----------|---|----------|---------|----------|
| 1 | 1-1 | Abandona | Promedio Puntaje Leng. y Mat. ≤ 510 Asignaturas inscritas ≤ 7 | 10,53 % | 14 | 133 |
| | 1-2 | Retenido | Asignaturas inscritas ≥ 15 Asistencia ≥ 90 % | 11,70 % | 31 | 265 |
| | 1-3 | Retenido | Puntaje de Cs. o Hist. ≥ 598 Asignaturas inscritas ≥ 9 | 10,94 % | 29 | 265 |
| 2 | 2-1 | Abandona | Riesgo Ranking $\leq 1,0$ Promedio Puntaje Leng. y Mat. ≤ 425 Asignaturas aprobadas ≤ 38 % | 16,24 % | 57 | 351 |
| | 2-2 | Abandona | NEM $\leq 4,90$ Asignaturas inscritas ≤ 7 Asistencia ≥ 81 % | 13,11 % | 46 | 351 |
| | 2-3 | Retenido | NEM $\geq 5,30$ Promedio Puntaje Leng. y Mat. ≥ 429 Asignaturas aprobadas ≥ 93 % | 9,82 % | 56 | 570 |
| 3 | 3-1 | Abandona | NEM $\leq 5,47$ Puntaje de Cs. o Hist. ≤ 407 Asignaturas aprobadas ≤ 0 % | 24,63 % | 101 | 410 |
| | 3-2 | Abandona | NEM $\leq 5,10$ Asignaturas inscritas ≤ 8 Asistencia ≤ 71 % | 20,00 % | 82 | 410 |
| | 3-3 | Retenido | Riesgo Ranking $\geq 2,0$ Puntaje de Cs. o Hist. ≥ 384 Asignaturas inscritas ≥ 14 Asistencia ≥ 88 % | 14,47 % | 67 | 463 |
| 4 | 4-1 | Abandona | Asignaturas inscritas ≤ 7 Asistencia ≤ 75 % | 40,63 % | 26 | 64 |
| | 4-2 | Abandona | Asignaturas inscritas ≤ 7 Asignaturas aprobadas ≤ 0 % | 35,94 % | 23 | 64 |
| | 4-3 | Retenido | Riesgo Ranking ≥ 598 Asignaturas inscritas ≥ 15 | 35,14 % | 65 | 185 |

Fonte: Elaboración propia

Tabela 9. Resultado de la aplicación de la técnica de conjuntos aproximados a los cuatro grupos y 13 variables, por condición de deserción.

| Clase | Subclase | Regla | Condiciones | Strength | Support | Coverage |
|-------|-----------|-------|--|----------|---------|----------|
| 1 | Abandonan | 1-1 | Puntaje de Cs. o Hist. ≤ 585 | 83,86 % | 256 | 305 |
| | | 1-2 | Asistencia ≥ 36 % | 100,00 % | 715 | 715 |
| | Retenidos | 1-3 | Asignaturas aprobadas ≥ 40 % | 96,78 % | 692 | 715 |
| | | 1-4 | Asignaturas inscritas ≥ 7 | 94,97 % | 679 | 715 |
| 2 | Abandonan | 2-1 | Promedio Puntaje Leng. y Mat. ≤ 498 | 81,07 % | 882 | 1088 |
| | | 2-2 | Asignaturas inscritas ≤ 10 | 66,08 % | 719 | 1088 |
| | Retenidos | 2-3 | Asistencia ≥ 41 % | 99,69 % | 1949 | 1955 |
| | | 2-4 | Asignaturas aprobadas ≥ 27 % | 97,80 % | 1912 | 1955 |
| | | 2-5 | Asignaturas inscritas ≥ 7 % | 96,06 % | 1878 | 1955 |
| 3 | Abandonan | 3-1 | Puntaje de Cs. o Hist. ≤ 477 | 97,94 % | 667 | 681 |
| | | 3-2 | Asignaturas inscritas ≤ 10 | 66,67 % | 453 | 679 |
| | Retenidos | 3-3 | Asistencia ≥ 49 % | 99,66 % | 875 | 878 |
| | | 3-4 | Asignaturas inscritas ≥ 7 % | 96,01 % | 843 | 878 |
| | | 3-5 | Asignaturas aprobadas ≥ 36 % | 95,90 % | 842 | 878 |
| 4 | Abandonan | 4-1 | Promedio Puntaje Leng. y Mat. ≤ 720 | 96,60 % | 142 | 147 |
| | | 4-2 | Puntaje de Cs. o Hist. ≤ 714 | 93,20 % | 137 | 147 |
| | | 4-3 | Asignaturas aprobadas ≤ 93 % | 65,31 % | 96 | 147 |
| | Retenidos | 4-4 | Asignaturas inscritas ≥ 8 | 97,05 % | 427 | 440 |
| | | 4-5 | Asistencia ≥ 77 % | 96,36 % | 424 | 440 |

Fonte: Elaboración propia

una mayor incidencia en esta decisión están ligadas a factores previos al ingreso (Promedio Puntaje Lenguaje y Matemática, Puntaje de Ciencias o Historia). Este mismo análisis en los estudiantes que son retenidos muestra que las variables más influyentes están relacionadas con factores de posingreso a estudios superiores (Asistencia, Asignaturas aprobadas, Asignaturas inscritas).

Así, al no encontrar variables cuyas reglas permitan explicar de forma transversal el hecho que un estudiante decida abandonar o no, independientemente de la clase a la que pertenece, se realizó un último análisis. Se buscan reglas de asociación tomando como variable decisional la clase a la que pertenece el estudiante (Tabela 10) y no la variable Abandono. De esta forma se puede verificar la complementariedad existente entre los distintos grupos.

Tabela 10. Resultado de la aplicación de la técnica de conjuntos aproximados a los cuatro grupos y 13 variables, por grupo.

| Clase | Regla | Condiciones | Strength | Support | Coverage |
|-------|-------|--|----------|---------|----------|
| 1 | 1-1 | Asignaturas inscritas ≤ 10 | 77,48 % | 320 | 413 |
| 2 | 2-1 | Asistencia ≥ 51 % | 93,52 | 2885 | 3085 |
| | 2-2 | Asignaturas aprobadas ≥ 42 % | 79,48 % | 2452 | 3085 |
| 3 | 3-1 | Asignaturas inscritas ≥ 6 | 93,65 | 1446 | 1544 |
| | 3-2 | Asistencia ≥ 38 % | 96,05 % | 1483 | 1544 |
| | 3-3 | Asignaturas aprobadas ≥ 43 % | 71,05 % | 1097 | 1544 |
| 4 | 4-1 | Promedio Puntaje Leng. y Mat. ≤ 572 | 97,61 | 573 | 587 |
| | 4-2 | Asignaturas inscritas ≥ 6 | 96,93 % | 569 | 587 |
| | 4-3 | Puntaje de Cs. o Hist. ≥ 588 | 79,22 % | 465 | 587 |
| | 4-4 | Asignaturas aprobadas ≥ 45 % | 93,02 % | 546 | 587 |
| | 4-5 | Asistencia ≥ 50 % | 91,82 % | 539 | 587 |

Fonte: Elaboración propia

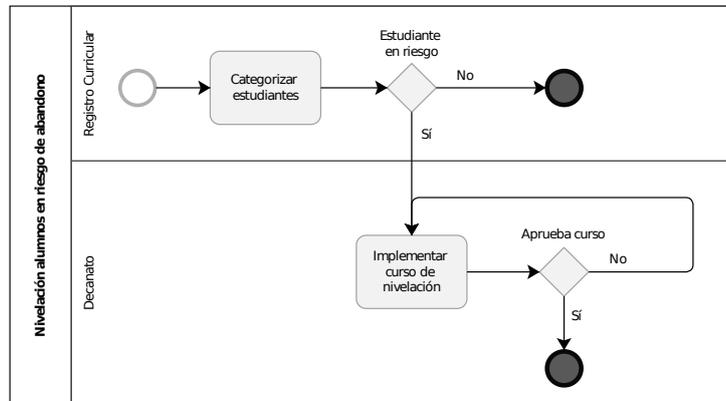
Se observa que existe cierta complementariedad entre los factores que definen cada grupo, los cuales se repiten entre clases. La complementariedad se define por los factores relacionados al desarrollo del estudiante posingreso a su primer año de ES. Sin embargo, los valores obtenidos en estas reglas no permiten definir una clara demarcación entre los valores que definen a cada clase, para validar con la información disponible no es posible desarrollar un modelo que pueda explicar la deserción.

Con esto, se valida la importancia que Asistencia y Asignaturas aprobadas tienen en el proceso de gestión estudiantil, y por consecuencia, en el manejo del abandono, sin embargo, no poseen la incidencia suficiente como para explicar el fenómeno por sí solas.

Finalmente, al considerar los perfiles de estudiantes 1, 2, 3 y 4 (Tabela 5), que tienen una dificultad de retención baja, alta, muy alta, y muy baja, respectivamente, y las reglas de asociación obtenidas para cada grupo (Tabelas 9 e 10), se propone un modelo de gestión estudiantil basado en el trabajo colaborativo de distintas unidades académicas en dos ámbitos: mitigación del abandono (fig. 9) y fomento de la retención (fig. 10). Para mitigar el abandono, se proponen cursos de nivelación obligatorios para los estudiantes con déficit en los resultados de sus pruebas de selección (Tabela 9), donde se desarrollen competencias académicas, valores y habilidades sociales.

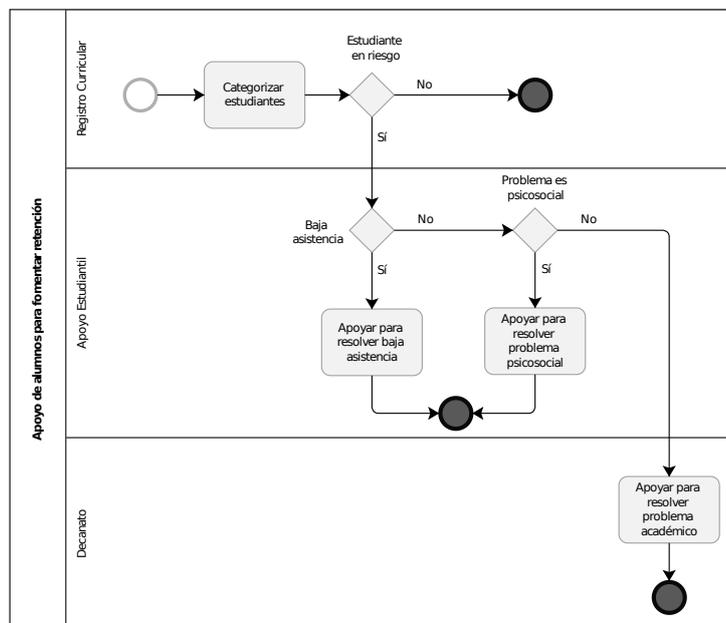
Para fomentar la retención, según factores ligados al posingreso (Tabela 9), Registro Curricular monitoreará la asistencia, inscripción y aprobación de asignaturas, y derivará a los alumnos a distintas unidades académicas, según sea el problema. A los estudiantes con baja asistencia los conducirá a una unidad de Apoyo Estudiantil, que indagará las causas y brindará el apoyo, ya sea psicológico o financiero. A quienes presenten baja inscripción de asignaturas se derivan a Apoyo Estudiantil o al Decanato (Dirección de Carrera), según su necesidad es psicosocial o académica, respectivamente. Y a los estudiantes, según porcentaje de aprobación, los derivará a Apoyo Estudiantil o al Decanato.

Figura 9. Modelo de gestión estudiantil para disminuir el abandono



Fonte: Elaboración propia.

Figura 10. Modelo de gestión estudiantil para fomentar la retención



Fonte: Elaboración propia

4 Conclusiones

La presente investigación valida una metodología basada en CRISP-DM, correlación y aprendizaje no supervisado, para obtener perfiles de estudiantes en riesgo de abandono. Así, permite proponer mejoras en la gestión estudiantil, que propenden a disminuir este riesgo.

Por otro lado, no se pudo generar un modelo explicativo de la deserción dado que, al utilizar registros históricos, la información se restringió solo a los datos disponibles. Esto representa una oportunidad de mejora, y demuestra la importancia de implementar sistemas de información institucionales que contengan los datos socioeconómicos, demográficos, psicosociales y académicos necesarios de los estudiantes.

La variedad de enfoques y heterogeneidad de los factores involucrados en el fenómeno del abandono estudiantil hacen de su análisis un trabajo muy dependiente del contexto en el cual se lleva a cabo. El estudio se realizó con estudiantes de primer año de pregrado en jornada diurna de una universidad privada independiente chilena y no adscrita al Sistema de acceso universitario. Así, la comparación de los resultados obtenidos con los de otros estudios es pertinente y válida en tanto se trate del mismo tipo de institución, área de conocimientos, tipo de estudiante, entorno social, entre otros. En base a lo anterior, se puede afirmar que en el contexto estudiado las variables que presentan una mayor incidencia en la decisión de abandonar están ligadas a factores previos al ingreso (puntajes en pruebas de selección y NEM) y académicas posingreso (Asignaturas aprobadas y Asistencia). Esto se correlaciona con lo encontrado por (BARRIOS RUBIO, 2011; LARROUCAU, 2015; CARVAJAL; GONZÁLEZ, 2018; VELÁSQUEZ; HITPASS, 2014), y especialmente con (MIRANDA; GUZMÁN, 2017), que es la investigación más afín.

El presente estudio innova en su contexto en la incorporación de la variable Aprobación de asignaturas, y en el uso del clustering jerárquico aglomerativo y de los conjuntos aproximados para encontrar perfiles de estudiantes en riesgo de abandono y los respectivos factores predominantes.

Por último, los resultados de esta investigación contribuyen al desarrollo de la teoría del abandono estudiantil en las universidades chilenas, a su gestión, y establece un fundamento metodológico para futuros estudios empíricos.

Referências

ADDINSOFT. *XLSTAT | Software estadístico Excel*. [S.l.: s.n.]. Disponível em: <<https://www.xlstat.com/es/>>. Acesso em: 9 ago. 2021.

ANTONENKO, P. D.; TOY, S.; NIEDERHAUSER, D. S. Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, v. 60, n. 3, p. 383–398, 2012. DOI: 10.1007/s11423-012-9235-8.

ARRIAGA, J.; VELÁSQUEZ, M. *Proyecto ALFA-III "Gestión Universitaria Integral del Abandono": Construcción colectiva del concepto de abandono en la educación superior para su medición y análisis*. [S.l.], 2013.

BARRIOS RUBIO, A. Deserción Universitaria en Chile. Incidencia del financiamiento y otros factores asociados. *Revista CIS*, n. 14, p. 59–72, 2011. Disponível em: <<http://www.techo.org/wp-content/uploads/2013/02/barrios.pdf>>.

BEAN, J. P. Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome. *American Educational Research Journal*, v. 22, n. 1, p. 35–64, 1985. DOI: 10.3102/00028312022001035.

BEAN, J. P.; METZNER, B. S. A Conceptual Model of Nontraditional Undergraduate Student Attrition. *Review of Educational Research*, v. 55, n. 4, p. 485–540, 1985. DOI: 10.3102/00346543055004485.

BECKER, G. Investment in human capital: A theoretical analysis. *The Journal of Political Economy*, v. 70, n. 5, p. 9–49, 1962.

BEHR, A. et al. Motives for dropping out from higher education—An analysis of bachelor's degree students in Germany. *European Journal of Education*, v. 56, n. 2, p. 325–343, 2021. DOI: 10.1111/ejed.12433.

BOUZAYANE, S.; SAAD, I. Weekly predicting the at-risk MOOC learners using dominance-based rough set approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10254 LNCS, p. 160–169, 2017. DOI: 10.1007/978-3-319-59044-8_18.

- BRAXTON, J. M.; JOHNSON, R. M.; SHAW-SULLIVAN, A. Appraising Tinto's theory of college student departure. In: AGATHON PRESS (Ed.). *Higher Education Handbook of theory and research*, Vol. 12. NY, EE.UU.: [s.n.], 1997. DOI: 10.1353/csd.2014.0061.
- BRAXTON, J. M.; MCCLENDON, S. A. The Fostering of Social Integration and Retention through Institutional Practice. *Journal of College Student Retention: Research, Theory & Practice*, v. 3, n. 1, p. 57–71, 2002. DOI: 10.2190/rgxj-u08c-06vb-jk7d.
- BURITICÁ, Nicolas Clavijo et al. Selection of supplier management policies using clustering and fuzzy-AHP in the retail sector. *International Journal of Logistics Systems and Management*, v. 34, n. 3, p. 352–374, 2019. ISSN 17427975. DOI: 10.1504/IJLSM.2019.103089.
- CABRERA, L.; BETHENCOURT, J. T. et al. El problema del abandono de los estudios universitarios. *RELIEVE: Revista Electrónica de Investigación y Evaluación Educativa*, v. 12, n. 2, p. 171–203, 2006.
- CABRERA, L.; BETHENCOURT BENÍTEZ, J. T. et al. Un estudio transversal retrospectivo sobre prolongación y abandono de estudios universitarios. *Revista ELección de Investigación y EValuación Educativa*, v. 12, n. 1, p. 105–127, 2006.
- CALDERÓN, Dora Inés et al. *Fenómeno deserción en Cultiva*. [S.l.], 2017. p. 83.
- CALVACHE F., L. C. et al. Aplicación de técnicas de minería de datos para la identificación de patrones de deserción estudiantil como apoyo a las estrategias de SARA (Sistema de acompañamiento para el rendimiento académico). In: OCTAVA Conferencia Latinoamericana sobre el Abandono en la Educación Superior. [S.l.: s.n.], 2018. p. 1177–1185.
- CARVAJAL, C. M; GONZÁLEZ, J. A. Variables Sociodemográficas y Académicas Explicativas de la Deserción de Estudiantes en la Facultad de Ciencias Naturales de la Universidad de Playa Ancha (Chile). v. 11, n. 2, p. 3–12, 2018.
- CENTRO INTERUNIVERSITARIO DE DESARROLLO (CINDA). *Repitencia y Deserción Universitaria en América Latina*. Edição: Colección Gestión Universitaria. Santiago de Chile: Colección Gestión Universitaria, 2006. p. 380.
- CHAPMAN, P. et al. The CRISP-DM Process Model. [S.l.], 1999.
- COMISIÓN NACIONAL DE ACREDITACIÓN (CNA). *Criterios y estándares para la acreditación de universidades*. Santiago de Chile, 2020.
- DELEN, D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, Elsevier B.V., v. 49, n. 4, p. 498–506, 2010. DOI: 10.1016/j.dss.2010.06.003. Disponible em: <<http://dx.doi.org/10.1016/j.dss.2010.06.003>>.
- DÍAZ PERALTA, C. Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios Pedagógicos*, v. XXXIV, n. 2, p. 65–86, 2008.
- DONOSO, S.; SCHIEFELBEIN, E. Analisis de los modelos explicativos de retencion de estudiantes en la universidad: Una vision desde la desigualdad social. *Estudios Pedagogicos*, v. 33, n. 1, p. 7–27, 2007. DOI: 10.4067/s0718-07052007000100001.
- ETHINGTON, C. A. A psychological model of student persistence. *Research in Higher Education*, v. 31, n. 3, p. 279–293, 1990. DOI: 10.1007/BF00992313.
- FÉNYES, H.; MOHÁCSI, M.; PALLAY, K. Career consciousness and commitment to graduation among higher education students in Central and Eastern Europe. *Economics and Sociology*, v. 14, n. 1, p. 61–75, 2021. DOI: 10.14254/2071-789X.2021/14-1/4.
- FERREYRA, M. M. et al. *At a Crossroads: Higher Education in Latin America and the Caribbean*. v. 23. Washington, 2017. p. 272. DOI: 10.2753/RSL1061-1975230163.
- FRANCISCO, J.; MUÑOZ, R.; SALDAÑA, I. H. ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. *Revista de la Educación Superior*, XL (4), n. 160, p. 29–49, 2011.
- GONZÁLEZ, L. E.; URIBE, D. Estimaciones sobre la "repitencia" y deserción en la Educación Superior Chilena. Consideraciones sobre sus implicaciones. *Calidad en la Educación*, n. 17, p. 75–90, 2018. DOI: 10.31619/caledu.n17.408.
- GONZÁLEZ F, L. E.; URIBE JORQUERA, D.; GONZÁLEZ VIDAL, S. *Estudio sobre la repitencia y deserción en la Educación Superior Chilena*. Santiago, Chile, 2005.

- GRECO, S.; MATARAZZO, B.; SLOWINSKI, R. Multicriteria classification by dominance-based rough set approach. *Handbook of data mining and knowledge discovery*, p. 1–14, 2002. Disponível em: <http://idss.cs.put.poznan.pl/site/fileadmin/projects-images/4emka%7B%5C_%7Dmethodology.pdf>.
- GRECO, S.; MATARAZZO, B.; SLOWINSKI, R. The Use of Rough Sets and Fuzzy Sets in MCDM. In: GAL, T.; HANNE, T.; STEWART, T. (Ed.). *Advances in Multiple Criteria Decision Making*. Boston, MA: Kluwer Academic Publishers, 1999. p. 397–455. DOI: 10.1007/978-1-4615-5025-9_14.
- HENRÍQUEZ, N.; ESCOBAR-RIFF, D. Construcción de un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción de la Universidad Metropolitana de Ciencias de la Educación. *Revista Mexicana de Investigación Educativa RMIE*, v. 21, p. 14056666, 2016.
- HERNÁNDEZ, J. et al. Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. *Archivos Venezolanos de Farmacología y Terapéutica*, v. 37, n. 5, p. 587–595, 2018.
- HIMMEL, E. Modelo de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación*, n. 17, p. 91, 2002. DOI: 10.31619/caledu.n17.409.
- HINOJOSA V., Mauricio. Adaptation of the Balanced Scorecard to Latin American Higher Education Institutions in the Context of Strategic Management: A Systematic Review with Meta-analysis. In: INTERNATIONAL Conference of Production Research-Americas. Santiago de Chile: Springer, 2021. 1408 CCIS, p. 2176–2190. ISBN 9783030763091. DOI: 10.1007/978-3-030-76310-7_10.
- HOFFLINGER, A. *Diseño y construcción de un modelo de detección temprana de deserción para las carreras de pregrado FID ULagos*. Osorno, Chile, 2020. Disponível em: <http://pedi.ulagos.cl/wp-content/uploads/2021/08/AT%7B%5C_%7Dinforme-final-2.pdf>.
- HOSSLER, D.; BEAN, J. P. *The strategic management of college enrollments*. Edição: Jossey-Bass. 1st ed. New York: [s.n.], 1990. p. 330.
- HUESCA RAMÍREZ, M. G.; CASTAÑO CORVO, M. B. Causas de deserción de alumnos de primeros semestres de una universidad privada. *Revista Mexicana de Orientación Educativa*, v. 5, n. 12, p. 34–39, 2007.
- ISHITANI, T. T.; DESJARDINS, S. L. A Longitudinal Investigation of Dropout From College in the United States* an Overview of Student Departure Theory. *J. College Student Retention*, v. 4, n. 2, p. 173–201, 2002.
- KEHM, B. M.; LARSEN, M. R.; SOMMERSEL, H. B. Student dropout from universities in Europe: A review of empirical literature. *Hungarian Educational Research Journal*, v. 9, n. 2, p. 147–164, 2020. DOI: 10.1556/063.9.2019.1.18.
- KIRTON, M. Transitional factors influencing the academic persistence of firstsemester undergraduate freshmen. In: 2-A. DISSERTATION Abstracts International Section A: Humanities & Social Sciences. [S.l.: s.n.], 2000. v. 61, p. 522.
- KUH, G. D. et al. *Student Success in College: Creating Conditions That Matter*. Edição: Jossey-Bass; 1st ed. [S.l.]: Jossey-Bass, 2010. p. 416.
- LABORATORY OF INTELLIGENT DECISION SUPPORT SYSTEMS IDSS. *Software 4eMka2*. [S.l.: s.n.]. Disponível em: <<http://idss.cs.put.poznan.pl/site/4emka.html>>. Acesso em: 23 ago. 2021.
- LARROUCAU, T. Estudio de los factores determinantes de la deserción en el sistema universitario chileno. *Estudio de políticas públicas*, v. 1, n. 1, p. 1–23, 2015. DOI: 10.5354/0719-6296.2015.38351.
- LEMAITRE, M. J. *Aseguramiento de la Calidad en América Latina*. Edição: IESALC-UNESCO. [S.l.]: IESALC-UNESCO, 2017. p. 33–4.
- LUJAN, J. R.; RESENDIZ, A. N. Hacia la construcción de un modelo causal en el análisis de la deserción. *Universidad Autónoma Metropolitana Iztapalapa*, 1981.
- MALDONADO, S. et al. Redefining profit metrics for boosting student retention in higher education. *Decision Support Systems*, v. 143, August 2020, 2021. DOI: 10.1016/j.dss.2021.113493.
- MEDRANO, L. A. et al. Creencias irracionales, rendimiento y deserción académica en ingresantes universitarios. *Liberabit*, v. 16, n. 2, p. 183–191, 2010.
- MINISTERIO DE EDUCACIÓN. *Ley 20.129: Establece un Sistema Nacional de Aseguramiento de la Calidad de la Educación Superior*. [S.l.: s.n.], 2006.

- MINISTERIO DE EDUCACIÓN. *Ley 21.091 Sobre Educación Superior*. [S.l.: s.n.], 2018.
- MIRANDA, M. A.; GUZMÁN, J. Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formacion Universitaria*, v. 10, n. 3, p. 61–68, 2017. DOI: 10.4067/S0718-50062017000300007.
- MUNIZAGA, F.; CIFUENTES, M. B. Retención y Abandono Estudiantil en la Educación Superior Universitaria en América Latina y el Caribe : Una Revisión Sistemática. *archivos analíticos de políticas educativas*, v. 26, n. 61, 2018. DOI: 10.14507/epaa.26.3348.
- NORA, A. The Depiction of Significant others in Tinto's "Rites of Passage": A Reconceptualization of the Influence of Family and Community in the Persistence Process. *Journal of College Student Retention: Research, Theory & Practice*, v. 3, n. 1, p. 41–56, 2002. DOI: 10.2190/byt5-9f05-7f6m-5ycm.
- NORA, A.; RENDON, L. I. Determinants of predisposition to transfer among community college students: A structural model. *Research in Higher Education*, v. 31, n. 3, p. 235–255, 1990. DOI: 10.1007/BF00992310.
- OECD. *Education at a Glance 2021: OECD indicators*. Paris, 2021. DOI: 10.4135/9781529714395.n163. Disponible em: <<https://doi.org/10.1787/b35a14e5-en>>.
- OECD. How many students drop out of tertiary education? In: HIGHLIGHTS from Education at a Glance, 2008. Paris: OECD Publishing, 2009. p. 24–26. Disponible em: <<http://dx.doi.org/10.1787/401536355051>>.
- ROSÁRIO, P. et al. Autoeficacia y utilidad percibida como condiciones necesarias para un aprendizaje académico autorregulado. *Anales de Psicología*, v. 28, n. 1, p. 37–44, 2012.
- RYAN, M. P.; GLENN, P. A. Increasing One-Year Retention Rates by Focusing on Academic Competence: An Empirical Odyssey. *Journal of College Student Retention: Research, Theory & Practice*, v. 4, n. 3, p. 297–324, 2003. DOI: 10.2190/kunn-a2ww-rfqt-py3h.
- SCHREIBER, B.; LUDEMAN, R. B. *Student Affairs and Services in Higher Education: Global Foundations , Issues , and Best Practices*. Third ed. [S.l.: s.n.], 2020.
- SCHULTZ, T. W. *Investment in Human Capital*. v. 51. [S.l.: s.n.], 1962. p. 1–20.
- SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES). *Deserción de primer año y Reingreso a la Educación Superior en Chile: Análisis de la cohorte 2015*. Santiago de Chile, 2019. p. 3–26.
- SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES). *Informe 2020 Retención de 1er Año de pregrado. Cohortes 2015 - 2019*. Santiago de Chile, 2020. p. 1–16.
- SERVICIO DE INFORMACIÓN DE EDUCACIÓN SUPERIOR (SIES). *Panorama de la educación superior en Chile 2014*. Santiago de Chile, 2014. p. 57. Disponible em: <http://www.mifuturo.cl/images/Estudios/Estudios%7B%5C_%7DSIES%7B%5C_%7DDIVESUP/panorama%7B%5C_%7Dde%7B%5C_%7Dla%7B%5C_%7Deducacion%7B%5C_%7Dsuperior%7B%5C_%7D2014%7B%5C_%7Dsies.pdf>.
- SPADY, W. G. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, v. 1, n. 1, p. 64–85, 1970. DOI: 10.1007/BF02214313.
- ST. JOHN, E. P. et al. Economic Influences on Persistence Reconsidered: How Can Finance Research Inform the Reconceptualization of Persistence Models? *Reworking the Student Departure Puzzle*, v. 1, p. 29–47, 2000. DOI: 10.2307/j.ctv176kvf4.5.
- THOMAS, L. Student retention in higher education: The role of institutional habitus. *Journal of Education Policy*, v. 17, n. 4, p. 423–442, 2002. DOI: 10.1080/02680930210140257.
- THURLOW, L. C. The Political Economy of Income Redistribution Policies. *The ANNALS of the American Academy of Political and Social Science*, v. 409, n. 1, p. 146–155, 1973. DOI: 10.4337/9781784712105.00017.
- TINTO, V. Colleges as Communities: Taking Research on Student Persistence Seriously. *The Review of Higher Education*, v. 21, n. 2, p. 169–177, 1998.
- TINTO, V. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, v. 45, n. 1, p. 89–125, 1975.
- TINTO, V. *Leaving college: rethinking the causes and cures of student attrition*. Edição: University of Chicago Press. 2nd ed. Chicago: University of Chicago Press, 1993. p. 312.
- TINTO, V. Moving from theory to action. In: ROWMAN & LITTLEFIELD PUBLISHERS (Ed.). *College Student Retention: Formula for Student Success*. Westport, Connecticut.: [s.n.], 2005. Epilogue, p. 317–333.

- TSENG, S. F. et al. Who will pass? Analyzing learner behaviors in MOOCs. *Research and Practice in Technology Enhanced Learning*, Research e Practice in Technology Enhanced Learning, v. 11, n. 1, 2016. DOI: 10.1186/s41039-016-0033-5. Disponível em: <<http://dx.doi.org/10.1186/s41039-016-0033-5>>.
- UNESCO-UIS. *International Standard Classification of Education: ISCED 2011*. [S.l.: s.n.], 2012. p. 84. Disponível em: <<http://uis.unesco.org/en/isced-mappings>>.
- UNESCO-UIS / OECD / EUROSTAT. *UOE data collection on education systems - Volume1 - Manual: Concepts, definitions and classifications*. v. 1. Paris, 2005.
- VARGAS, Ma. et al. CDIO project approach to design Polynesian canoes by first-year engineering students. *International Journal of Engineering Education*, v. 35, n. 5, p. 1336–1342, 2019.
- VELÁSQUEZ, L.; HITPASS, B. El nivel de Actividad en el Proceso Educativo como Indicador de Riesgo de Deserción Estudiantil medido en tiempo real con apoyo de tecnología BAM. *JCC Workshop on Business Process Management*, November, 2014. DOI: 10.13140/2.1.3217.8880.
- VIJAYA, V.; SHARMA, S.; BATRA, N. Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects, COMITCon 2019, IEEE*, p. 568–573, 2019. DOI: 10.1109/COMITCon.2019.8862232.
- WASSERMAN, K. N. Psychological and developmental differences between students who withdraw from college for personal-psychological reasons and continuing students. In: DISSERTATION Abstracts International Section A: Humanities & Social Sciences. [S.l.: s.n.], 2001. p. 915.
- WIRTH, R. CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, n. 24959, p. 29–39, 2000.
- YIP, M. C.W. The linkage among academic performance, learning strategies and self-efficacy of Japanese university students: a mixed-method approach. *Studies in Higher Education*, Taylor & Francis, v. 46, n. 8, p. 1565–1577, 2019. DOI: 10.1080/03075079.2019.1695111. Disponível em: <<https://doi.org/10.1080/03075079.2019.1695111>>.

Contribuições dos autores

Mauricio Hinojosa V.: Aquisição de financiamento, Programas, Validação, Escrita – rascunho original, Escrita – revisão e edição; **Iván Derpich:** Conceituação, Supervisão; **Miguel Alfaro:** Administração de projetos, Supervisão; **David Ruete:** Curadoria de dados, Análise Formal, Aquisição de financiamento, Metodologia, Recursos, Validação; **Alejandro Caroca:** Conceituação, Recursos, Validação; **Gustavo Gatica:** Curadoria de dados, Análise Formal, Metodologia, Administração de projetos, Programas, Supervisão, Escrita – revisão e edição .