UNIVERSIDAD DEL BÍO-BÍO

FACULTAD DE CIENCIAS EMPRESARIALES

DEPARTAMENTO DE SISTEMAS DE INFORMACIÓN

Predicción de la tasa de éxito en las asignaturas de primer año para los alumnos de la Universidad del Bío-Bío

Memoria para optar al Título de Ingeniero Civil en Informática

Autor

Gabriel Osorio Muñoz

Profesor Guía

Elizabeth Grandón Toledo

Concepción, Marzo 2018

Resumen

Hoy en día, tanto la deserción estudiantil como el rendimiento de los estudiantes son problemas que las instituciones de educación superior buscan solucionar. En la Universidad del Bío-Bío, según cifras de la Dirección General de Análisis Institucional (DGAI) cada año ingresan más de 2200 nuevos estudiantes a la institución. De los alumnos que ingresan, cerca de 1093 reprueban al menos una asignatura (Dirección de Admisión Registro y Control Académico). Actualmente la institución busca disminuir esta cantidad y una herramienta que permite predecir el número de estudiantes que reprueban asignaturas cada año es la minería de datos. Esta herramienta permite encontrar patrones repetitivos en grandes volúmenes de datos, por lo tanto, se podría descubrir las variables que influyen en el éxito académico de un estudiante.

El objetivo de este estudio es generar un modelo de predicción mediante el algoritmo R-CNR Tree para determinar las variables que influyen en el éxito académico de los recién ingresados a la universidad.

La población objetivo de este estudio consiste de 5082 estudiantes que se encuentran en estado de alumnos regulares y los cuales ingresaron a la universidad entre los años 2014-2016. De los 5082 registros, el 70% fue utilizado para entrenar los modelos mientras que el 30% fue usado para validarlo. Cada uno de estos registros contenía 34 variables de las cuales se seleccionaron 23. Dieciséis de ellas fueron elegidas basándose en estudios previos, y el resto fueron incluidas para analizar si realmente influyen en el éxito académico de un alumno. El éxito académico fue medido como el ratio de créditos aprobados durante el primer año sobre créditos que debería aprobar en el año por malla curricular.

Mediante el poder de predicción (KI) de cada variable, se logró obtener que de las 23 variables, 18 de ellas tenían la capacidad de predecir la variable de destino. Con estas 18 variables se construyó un conjunto base y uno de conjugación, compuesto por las primeras 7 variables y las 4 siguientes con mayor poder de predicción respectivamente.

Se generaron 4 modelos experimentales con 10 variables cada uno en la interfaz de Expert Analytics del software SAP Predictive Analytics. El resultado indica que el modelo 3 es el

mejor de los 4 modelos experimentales debido a que entrega una mayor precisión en los conjuntos de entrenamiento y validación, consiguió un menor número de falsos/negativos y falsos/positivos, y fue uno de los que más variables descartó para generar el árbol de decisión. Por otra parte, se obtuvo que las variables: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, son las que más influyen en el rendimiento académico de un estudiante.

Abstract

Today, both student dropout and student performance are problems that institutions of higher education seek to solve. According to figures of the General Direction of Institutional Analysis (DGAI) each year more than 2200 new students enroll at the Universidad del Bío-Bío. From these, about 1093 students fail at least one subject (Dirección de Admisión Registro y Control Académico). Currently the institution seeks to reduce this number and a tool that allows predicting the number of student who fail each year is the data mining. This tool allows finding repetitive patterns in large volumes of data; therefore, one could discover the variables that influence the academic success of a student.

The objective of this study is to generate a prediction model using the algorithm R-CNR Tree to determine the variables that influence the academic success of the newly admitted students to the university.

The target population of this study consists of 5082 students who are regular students and who enrolled the University between the years 2014-2016. Of the 5082 records, 70% was used to train the models while 30% was used to validate it. Each of these records contained 34 variables of which 23 were selected. Sixteen of them were chosen based on previous studies, and the rest were included to analyze whether they really influence the academic success of a student. Academic success was measured as the ratio of credits approved during the first year on credits that the student should approve in the year according to the curricular program.

By the prediction power (KI) of each variable, it was obtained that of the 23 variables, 18 of them had the ability to predict the dependent variable. With these 18 variables, it was created a base and a conjugation set, comprised by the first 7 variables and the following 4 with greater prediction power respectively.

Four experimental models were generated with 10 variables each in the Expert Analytics interface of the software SAP Predictive Analytics. The result was that model 3 is the best of the 4 experimental models because it delivers more precision in the training and validation sets, got a lower number of false/negative and false/positive, and it was the one that ruled out the most variables to generate the decision tree. On the other hand, it was obtained that the variables: Carrera, Nem, Ranking, Gender, Origin College, Commune Domicile, PSU Language and Age, are those that mostly influence the academic performance of a student.

Índice general

Índice Tablas

Índice Figuras

CAPÍTULO I

INTRODUCCIÓN

1.1. Introducción

Cada año más de doscientos mil nuevos estudiantes ingresan a la educación superior (CNED, 2017) y de esos cerca de 2300 lo hacen en la Universidad del Bío-Bío (DGAI, 2018). Debido al gran número de ingresos cada año, las universidades deben enfrentar algunos retos. Uno de ellos es la deserción estudiantil, que según Delen (2010) corresponde "al abandono de un programa de estudios antes de obtener el título o grado correspondiente, considerando un tiempo lo suficientemente largo como para descartar la posibilidad de reincorporación" (Citado en Miranda y Guzmán, 2017, p. 62). Sin embargo, la deserción no es el único problema que las instituciones de educación superior deben enfrentar, ya que el rendimiento académico de los estudiantes es un tema importante a tener en cuenta. Según Navarro (2003), el rendimiento académico es definido como "la expresión de las habilidades, actitudes y valores que son desarrollados por el alumno a través del proceso de enseñanza-aprendizaje, es decir son todas aquellas acciones dirigidas a la explicación e interpretación de lo aprendido y que se sintetizan en valores cuantitativos o cualitativos" (Citado en Avendaño *et al*., 2016, p.5).

Para poder hacer frente a estos problemas, las instituciones de educación superior han recurrido a una herramienta que en los últimos años se ha vuelto muy popular, la *minería de* datos (MD). Según Pal (2012), la minería de datos "es una tecnología utilizada para describir el descubrimiento de conocimiento y para buscar relaciones significativas, como patrones, asociaciones y cambios entre las variables en las bases de datos" (p.1). De la misma forma, Natek (2014) establece que esta herramienta "es utilizada para extraer y descubrir conocimiento valioso y significativo desde una gran cantidad de datos" (p.1). No obstante, para poder extraer dichos patrones la MD hace uso de diferentes algoritmos como de clasificación, asociación, regresión, entre otros.

Actualmente la Universidad del Bío-Bío cuenta con un modelo de deserción de sus alumnos, pero no así con un modelo para predecir el rendimiento académico de ellos. Por este motivo, el objetivo principal de este proyecto es construir un modelo predictivo de la tasa de éxito, el cual fue medido en base al ratio de créditos aprobados durante el primer año sobre créditos que

debería aprobar en el año por malla curricular, utilizando un algoritmo de árbol de decisión. Con este conocimiento, la Universidad podrá contar con mayor información para la toma de decisiones, desarrollando planes al comienzo del año académico con el objetivo de aumentar el rendimiento académico de los estudiantes y así evitar que el alumno abandone sus estudios superiores.

Es importante destacar que este proyecto cuenta con el patrocinio de la Dirección de Admisión, Registro y Control Académico (DARCA) de la Universidad del Bío-Bío.

1.2. Definición del proyecto

1.2.1. Justificación

El sector de la educación, en todo el mundo, ha sido testigo de un cambio radical en su funcionamiento. Hoy en día se reconoce como una industria y, como tal, se enfrenta a retos (Mishra *et al*., 2014). Uno de ellos es la deserción de los estudiantes de educación superior, la cual genera una serie de inconvenientes que afectan a los estudiantes y las universidades (Miranda & Guzmán, 2017).

La deserción no es el único problema que las instituciones de nivel superior deben enfrentar, ya que el rendimiento académico de los estudiantes es un tema importante a tener en cuenta. Según cifras de la Dirección de Admisión, Registro y Control Académico (DARCA) de la Universidad del Bío-Bío, en los años 2014, 2015 y 2016; más de un 49,7% de los estudiantes de primer año reprobó al menos una asignatura, por lo que la universidad, hoy en día, busca disminuir ese porcentaje. La Tabla 1 muestra el resumen del rendimiento académico de los estudiantes que ingresaron a la universidad entre los años 2014-2016.

Tabla 1: Rendimiento académico de los estudiantes entre los años 2014-2016

Año **Rendimiento**	**Alumnos de 1er año**	
	Aprueba todas sus asignaturas	**Reprueba al menos una asignatura**
2014	40,7%	59,3%
2015	50,3%	49,7%
2016	38,4%	61,4%

Fuente: Elaboración propia a partir de DARCA

Actualmente, la Universidad del Bío-Bío cuenta con una herramienta para analizar la deserción de sus alumnos pero no así un modelo para predecir el rendimiento académico de ellos.

El beneficio de desarrollar este proyecto será que la universidad cuente con un modelo predictivo de las tasas de éxito de los estudiantes de primer año. Al obtener esta información, la universidad podrá contar con mayor información para la toma de decisiones, desarrollando

planes al comienzo del año académico con el objetivo de aumentar el rendimiento académico de los estudiantes y así evitar que el alumno abandone sus estudios superiores.

1.2.2. Objetivo General

Predecir la tasa de éxito en las asignaturas de primer año para los alumnos de la Universidad del Bío-Bío, a través del ratio de créditos aprobados durante el primer año sobre créditos que debería aprobar en el año por malla curricular utilizando Minería de Datos.

1.2.3. Objetivos Específicos

- Recopilar información bibliográfica de estudios anteriormente desarrollados para definir el éxito de los estudiantes y las variables que podrían predecir.
- Estudiar modelos de Minería de Datos.
- Seleccionar el modelo más adecuado para el análisis.
- Generar y validar modelo de predicción.

1.2.4. Aporte y limitaciones

El aporte de realizar este proyecto será:

- Que la Universidad del Bío-Bío pueda contar con un modelo de predicción de éxito de las asignaturas de primer año.
- Que la universidad genere diversos planes para evitar que los alumnos tengan una baja tasa de éxito en las asignaturas.
- Que los estudiantes de primer año puedan tener un mejor rendimiento académico.
- Que el modelo de predicción generado pueda ser utilizado en otras instituciones de educación superior.

Por otra parte, este proyecto tiene algunas limitaciones las cuales son:

- Se hará uso de datos de estudiantes que ingresaron los años 2014,2015 y 2016, los que conforman en total una población de 5082.
- No se hará uso de variables motivacionales para generar el modelo.

CAPÍTULO II

MARCO TEÓRICO

2.1. Conceptualización de éxito académico

2.1.1. Concepto y definiciones de éxito académico

El rendimiento académico de los estudiantes es un problema que actualmente las universidades buscan solucionar. Lograr identificar a los alumnos que puedan tener un mal rendimiento hará que las instituciones generen planes para que estos puedan aumentar su rendimiento académico.

Navarro (2003) describe el rendimiento académico como "la expresión de las habilidades, actitudes y valores que son desarrollados por el alumno a través del proceso de enseñanza-aprendizaje, es decir son todas aquellas acciones dirigidas a la explicación e interpretación de lo aprendido y que se sintetizan en valores cuantitativos o cualitativos."(Citado en Avendaño *et al.*, 2016, p.5)

Por otra parte, Gutiérrez-Soto *et al.* (2008) establecen que "el éxito académico corresponde a la no reprobación de asignaturas en un semestre académico" (p.4).

De una forma similar, Barahona *et al.* (2016) definen que "el éxito académico corresponde al ratio de créditos aprobados sobre créditos inscritos en un periodo académico" (p. 27).

Como una forma de ser más preciso, para este estudio el éxito académico es definido como el ratio de créditos aprobados durante el primer año sobre créditos que debería aprobar en el año por malla curricular.

2.2. Clasificación de éxito académico

Según Tejedor (2003), el rendimiento académico, de forma esquemática, puede ser dividido en: Rendimiento inmediato y Rendimiento diferido.

2.2.1. Rendimiento Inmediato

Tejedor (2003) establece que el rendimiento inmediato es la obtención de calificaciones por parte de los alumnos durante sus estudios hasta lograr a la titulación de una carrera universitaria. El autor indica que este tipo de rendimiento puede ser dividido en rendimiento en sentido amplio, regularidad académica y rendimiento en sentido estricto.

1. **Rendimiento en sentido amplio:** Este rendimiento es dividido en 3 sub-áreas: éxito, retraso y abandono de estudios. El *éxito* se refiere a la conclusión y titulación de un alumno en una carrera universitaria durante los años previstos en el plan de estudios; el *retraso* es la conclusión de una carrera pero empleando más tiempo del establecido oficialmente, y; el *abandono de estudios* consiste en la no finalización de una carrera universitaria en particular.
2. **Regularidad académica:** Se refiere al porcentaje de no presentación a exámenes por parte de los estudiantes en un semestre o año.
3. **Rendimiento en sentido estricto:** Corresponde a las calificaciones que obtienen los estudiantes durante un periodo académico.

2.2.2. Rendimiento Diferido

Según Tejedor (2003), el rendimiento diferido se refiere a aplicar la formación recibida en la vida laboral y social. Este rendimiento es mucho más complejo de valorar debido a que entran en juego variables más personales y sociales de los sujetos, que son difíciles de cuantificar.

2.3. Variables que explican el rendimiento académico

Diversos autores han investigado sobre cuáles son las variables que influyen en el rendimiento académico de un alumno. Por ejemplo, González Tirados (1985) realizó una clasificación para determinar los factores que más influyen en el éxito o fracaso académico de los universitarios. En dicha clasificación, los factores se agrupaban en tres tipos: factores inherentes al alumno, al profesor y a la organización académica (como se cita en Tejedor, 2003).

Por otra parte, Tejedor (2003), basándose en investigaciones propias realizadas anteriormente, desarrolló una clasificación donde estableció cinco categorías de variables para analizar si influyen en un rendimiento académico. Estas categorías son: variables de identificación; variables psicológicas; variables académicas; variables pedagógicas, y; variables socio familiares.

A continuación se detallan cada una de estas variables.

2.3.1. Variables de identificación

En esta categoría se incluyen variables que identifican a un estudiante, como el género y la edad.

1. **Género:** Existen numerosas investigaciones sobre cómo esta variable influye en el rendimiento académico de un estudiante. Tejedor (2003) indica que aunque estas investigaciones generan resultados contradictorios, gran parte de los estudios concluyen que las mujeres tienen un mayor éxito académico. No se conoce con certeza por qué las mujeres tienen un mayor éxito que los hombres, sin embargo, "Salvador y García-Valcárcel; Goma *et al.*; Sánchez Gómez, y; Tejedor *et al.* (1998) establecen que "esto se debe a las distintas pautas de socialización y el esfuerzo de aptitudes diferenciales por sexos" (como se cita en Tejedor, 2003, p.7). No obstante, Guerrero *et al.* establecen que "no existen diferencias entre hombres y mujeres si el rendimiento académico es medido a través de la finalización o abandono de los estudios" (como se cita en Tejedor, 2003. p.7).
2. **Edad:** Esta variable puede ser considerada contradictoria debido a que en un curso, los estudiantes más jóvenes son los que obtienen mejores rendimiento y calificaciones, pero los alumnos de último año también tienen un excelente rendimiento y éstos tienen una mayor edad. Por este motivo, Tejedor (2003) establece que existe una clara relación entre un curso y la edad, y para poder relacionar esta última con el rendimiento académico se debe controlar la variable curso.

2.3.2. Variables psicológicas

Este tipo de variables no se puede examinar de manera externa el contexto sociofamiliar o entorno escolar, debido a que estas características surgen en el ámbito sociofamiliar y serán moduladas por las circunstancias del entorno escolar en que se desarrolla el alumno. En las variables psicológicas encontramos cuatro tipos de factores: inteligencia y aptitudes intelectuales, personalidad, motivación y, estilos cognitivos y estrategias de aprendizaje.

1. **Inteligencia y aptitudes intelectuales:** Según Tejedor (2003), estas variables son consideradas predictores muy deficientes del rendimiento académico. De hecho algunas opiniones establecen que ni los test de inteligencia ni los test de aptitudes permiten predecir si un alumno tendrá o no un buen rendimiento académico.
2. **Personalidad:** Según tejedor (2003), existen un gran número de investigaciones sobre cómo afecta la personalidad en el rendimiento académico, aunque todos entregan resultados contradictorios, pero en general, todos los rasgos de personalidad examinados tienen una baja contribución para predecir el éxito académico en la universidad y términos estadísticos, las correlaciones de esta variable apenas pasan de 0,3.
3. **Motivación:** Es considerada como una variable que facilita tener un buen rendimiento académico, aunque algunos estudios concluyan que no existe una relación entre ambas. Según Tejedor (2003) esto último se debe principalmente a que la motivación constituye un constructo multidimensional y a la baja confiabilidad de los instrumentos de medida utilizados.
4. **Estilos cognitivos y estrategias de aprendizaje:** En el ámbito universitario, González Tirados (1989) ha llegado a la conclusión que "un alumno que no posee un estilo de aprendizaje acorde con la carrera universitaria elegida tiene un mayor porcentaje de fracasar académicamente" (como se cita en Tejedor, 2003, p.9). Por otra parte, basándose en el Inventario de estilos de Aprendizaje (IEA) de Kole, donde se identifican cuatro estilos de aprendizaje: divergente, asimilador, convergente y de acomodación, se ha llegado a la conclusión que los estudiantes con estilos de aprendizaje convergente o asimilador, obtienen un mayor éxito académico.

2.3.3. Variables académicas

En esta categoría se incluyen dos factores: rendimiento académico previo y asistencia a clases.

1. **Rendimiento académico previo**: Diversas investigaciones han establecido que este factor es el mejor para poder predecir el rendimiento académico universitario, ya sea el

expediente de enseñanza secundaria, rendimiento en un curso o cursos anteriores de enseñanza universitaria y los resultados de la prueba de selectividad. "Según Apodaka *et al.*, el rendimiento académico previo en las enseñanzas medias es uno de los factores poderosos para predecir el rendimiento de un estudiante" (como se cita en Tejedor, 2003).

2. **Asistencia a clases:** Es un factor que permite obtener buenas calificaciones. Según López & López (1982), "los estudiantes universitarios con mayor grado de asistencia a clases obtienen calificaciones más altas "(como se cita en Tejedor, 2003, p.10).

2.3.4. Variables pedagógicas

Según Tejedor (2003), las metodologías seguidas en las aulas tienen una gran incidencia en el rendimiento de los estudiantes universitarios. Por este motivo, se han propuestos realizar cambios en las formas de enseñar y evaluar a los alumnos, sugiriendo la necesidad de una enseñanza que fomente la reflexión, la solución de problemas, la exposición de diversos puntos de vistas.

2.3.5. Variables socio familiares

1. **Estudios y situación laboral de los padres:** Los estudios de los padres ha sido una variable considerada en diversos estudios (Latiesa, 1983; Oroval, 1989; Salvador y Garcia-Vacárcel; 1989; Apodaka, 1991). Las conclusiones de estos estudios establecen que los estudios del padre tienen una incidencia prácticamente nula sobre el rendimiento de un estudiante. Sin embargo, según Apodaka *et al.* señalan que "los alumnos cuyas madres posean algún título universitario presentan una mayor regularidad académica del resto" (como se cita en Tejedor, 2003).
2. **Población de residencia:** En esta variable se incluyen el lugar de estudios del alumno, lugar de residencia del alumno durante el curso, coincidencia del lugar de estudio con el lugar de residencia, tipo de residencia del alumno durante el curso. Según Tejedor (2003), la coincidencia del hogar familiar con el lugar de estudio influye positivamente en el rendimiento académico de los estudiantes.

2.3.4. Clasificación de Shahiri, Husain & Abdul

Shahiri *et al.* (2015) realizaron una revisión sistemática de la literatura para identificar los atributos que son más importantes para predecir el rendimiento académico de los estudiantes.

Los autores encontraron que las variables que más se utilizan para la predicción del rendimiento son el promedio acumulado de notas (CGPA) y la evaluación interna. Un tercio de los artículos analizados utiliza la variable CGPA para realizar una predicción y esto se debe a que la mayoría de los investigadores consideró que esta variable posee un valor tangible para la futura movilidad educativa y profesional. La otra variable más utilizada, evaluación interna, la cual fue clasificada en notas de asignación, pruebas, trabajos de laboratorio, prueba de clase y asistencia.

Por otra parte, los otros atributos más utilizados son los datos demográficos de los estudiantes y la evaluación externa. En la Tabla 2 se puede observar las variables que incluyen cada atributo.

Tabla 2: Variables demográficas y de evaluación externa que influyen en el rendimiento académico.

Datos demográficos	**Evaluación externa**
Género	Nota obtenida en un examen final para un tema en particular
Edad	
Antecedentes familiares	
Discapacidad	

Fuente: Elaboración propia a partir de Shahiri *et al.* (2015)

La razón por la cual un gran número de investigadores utiliza los datos demográficos de los estudiantes, como el "género" se debe a que hombres y mujeres tienen distintos estilos de aprendizaje. Según Shahiri *et al.* (2015), la mayoría de las mujeres tienen diversos estilos de aprendizaje positivos en comparación con los hombres. Además, son mucho más disciplinadas en sus estudios y según Simsek & Balaban (2010) "tienen una estrategia de aprendizaje más efectiva en sus estudios" (como se cita en Shahiri *et al.*, 2015, p.417).

2.3.5. Otros estudios sobre rendimiento académico

Los estudios sobre que variables influyen en el rendimiento académico de un estudiante no son pocos, aunque gran parte de estos utilizan prácticamente las mismas variables (por ejemplo, edad, género, pruebas de selección, etc.) como base. A continuación se muestran otras variables que fueron analizadas por los investigadores y las cuales podrían incidir en el éxito académico de un alumno.

El estudio de Mayilvaganan & Kalpanadevi (2014) buscaba comparar diferentes técnicas de clasificación para predecir el rendimiento académico. Las técnicas que utilizaron fueron: algoritmo C4.5, AODE, Clasificador bayesiano, k-vecinos más cercanos; los cuales fueron aplicados a 19 variables entre las que destaca la ”especialidad” del estudiante. Sin embargo, esta última no fue considerada para el modelo final.

Barahona *et al.* (2016) buscaban determinar mediante, modelos de regresión lineal múltiple y regresión logística, los factores asociados al rendimiento académico de los alumnos y las variables determinantes en la deserción estudiantil. Se encontró que entre las variables de tipo socioeconómico, institucionales y académicas, el tipo de establecimiento si influye en el rendimiento de un estudiante mientras que variables como “Prioridad_postulacion_carrera” y “Ranking” no lo hacían.

En un estudio realizado por Salinas *et al.* (2017) encontraron que los alumnos con becas tienen un mayor porcentaje de aprobar el semestre académico que los estudiantes que no poseen beneficios estudiantiles.

Otras variables utilizadas fueron “trabajo” (Quadril & Kalyankar; 2010) y el “tipo de alojamiento durante sus estudios universitarios” (Rosas *et al.*, 2006) la primera fue descartada para generar el modelo final debido a su bajo aporte mientras que la segunda si fue incluida en el modelo.

2.4. Proceso de Descubrimiento del Conocimiento en Base de Datos

2.4.1. Concepto y definiciones del Proceso KDD

El proceso KDD (Knowledge Discovery in Databases - Descubrimiento del Conocimiento en Base de Datos) abarca campos de investigación muy dispares, y engloban desde la supercomputación, la estadística, bases de datos hasta el reconocimiento de patrones, y su objetivo principal es "extraer, almacenar y acceder a conocimiento en grandes volúmenes de datos" (González-Ruiz *et al.*, 2015, p.32).

Se define como "proceso de descubrir conocimiento útil desde una colección de datos" (Gamarra *et al.*, 2016, p.616).

Bernstein *et al.* (2005) señalan que "el proceso KDD es el resultado de un proceso exploratorio que involucra la aplicación de varios procedimientos algorítmicos para manipular datos, construir modelos a partir de estos datos, y manipular los modelos" (p.503).

Matheus *et al.* (1993) indican que el proceso KDD es un "procedimiento automático de grandes cantidades de datos sin procesar, para identificar patrones significativos y presentarlos como conocimiento apropiado para lograr los objetivos del usuario" (p.903).

Fayyad *et al.* (1996) define el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" (p.40).

2.4.2. Propiedades del conocimiento extraído

Basándose en la definición del KDD de Fayyad *et al*. (1996), se pueden conocer propiedades deseables que debería tener un conocimiento extraído.

- **Válido:** Los patrones que se encuentran deben seguir siendo precisos para datos nuevos, y no sólo para aquellos que hayan sido utilizados para su obtención.
- **Novedoso:** Que genere un aporte desconocido tanto para el sistema como para el usuario.

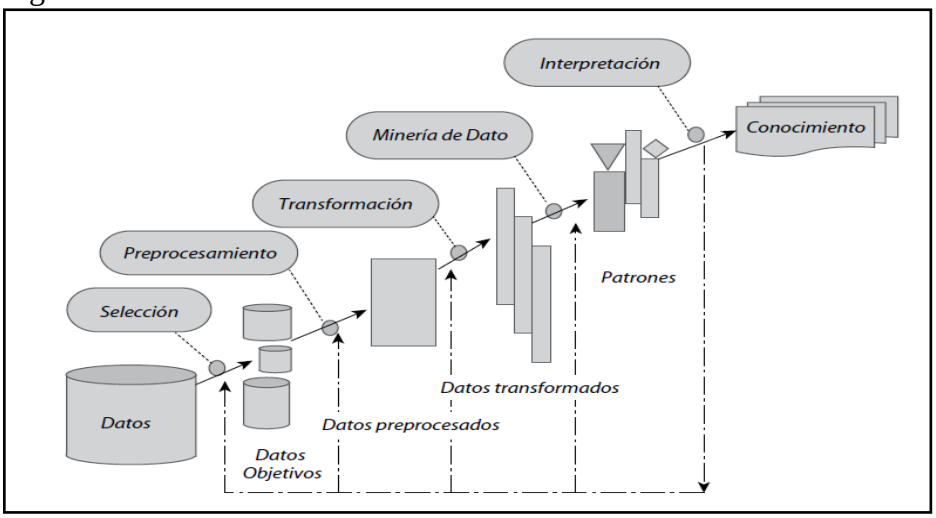
- **Potencialmente útil:** La información debe conducir a acciones que generen beneficios para el usuario.
- **Comprensible:** La extracción de patrones no comprensibles genera dificultades en la interpretación, revisión, validación y uso en la toma de decisiones.

2.4.3. Etapas del proceso KDD

2.4.3.1. Proceso típico de KDD

El proceso típico de KDD consta de cinco etapas: selección, pre-procesamiento, transformación, minería de datos e interpretación/evaluación. Las etapas de este proceso se muestran en la figura 1.

Figura 1: Proceso KDD



Fuente: Timarán-Pereira *et al.* (2016)

1. **Selección:** Definida la problemática y los objetivos, corresponde seleccionar los datos. En esta fase se selecciona todo o una muestra representativa del conjunto de datos objetivo, con el propósito de realizar un análisis para resolver el problema definido. Además, es importante unificar los datos debido a que pueden proceder de distintas fuentes.
2. **Pre-procesamiento:** En este paso se verifica que los datos a analizar sean de calidad para garantizar que el conocimiento a descubrir sea de un alto grado. Esta etapa incluye

tareas como remoción de datos ruidosos (noisy data), filtrado de datos atípicos (outliers), manejo de datos nulos, desconocidos y duplicados.

3. **Transformación:** Fase que modifica la estructura de los datos, a una estructura con características útiles para facilitar el análisis y la meta del proceso. Esto incluye transformar esquemas de datos a otros esquemas, métodos de reducción de dimensiones o de transformación para reducir la cantidad efectiva de variables bajo consideración o técnicas como clustering.
4. **Minería de datos:** Es la etapa de "descubrimiento" del proceso KDD y busca encontrar patrones insospechados dentro de un conjunto de datos mediante el uso de algoritmos de clasificación, agrupación, asociación, regresión, entre otras.
5. **Interpretación/evaluación:** Última fase del proceso KDD donde los usuarios interpretan los patrones descubiertos en base a tres criterios: precisión, claridad, e interés. Además, esta etapa puede incluir la eliminación de patrones redundantes y la traducción de los patrones en términos entendibles para el usuario.

2.4.3.2. Etapas KDD según Fayyad, Piatetsky-Shapiro & Smyth

Fayyad *et al*., (1996), señalan que el proceso KDD tiene nueve etapas: comprensión del dominio de la aplicación, creación del conjunto de datos, limpieza y pre-procesamiento de datos, reducción y proyección de datos, determinar la tarea de minería de datos, determinar el algoritmo de minería, minería de datos, interpretación y utilización del nuevo conocimiento. A continuación se describe brevemente cada etapa del proceso.

1. **Comprensión del dominio de la aplicación:** Fase donde todo el conocimiento disponible y relevante sobre el dominio de la aplicación es recolectado y donde además, se debería identificar los objetivos del proceso desde el punto de vista del usuario.
2. **Creación del conjunto de datos:** Etapa donde se selecciona un conjunto de datos o un subconjunto de variables en los cuales se realizará el descubrimiento. Además, podría ser necesario integrar los datos ya que podrían proceder de distintas fuentes.

3. **Limpieza y pre-procesamiento de datos:** En este paso se llevan a cabo operaciones básicas como la eliminación datos ruidosos (noisy data) y datos atípicos (outlier) y tratamiento de datos faltantes o perdidos (missing values).
4. **Reducción y proyección de datos:** Fase en la cual se buscan características útiles de representación de los datos dependiendo del objetivo. Además, incluye utilizar métodos de reducción o transformación de la dimensionalidad para reducir el número de variables consideradas o para descubrir representaciones invariantes en los datos.
5. **Determinar la tarea de minería de datos:** Etapa en la cual se debe determinar la tarea de minería de datos (por ejemplo, clasificación, regresión, agrupación, reglas de asociación, o análisis correlacional) para lograr los objetivos del estudio.
6. **Determinar el algoritmo de minería:** Etapa en la cual dependiendo de la regla de minería de datos seleccionada, se define el algoritmo (o los algoritmos) a utilizar para encontrar patrones en el conjunto de datos.
7. **Minería de datos:** Fase donde se aplica el algoritmo (o los algoritmos) a los datos para encontrar patrones de interés.
8. **Interpretación:** En este paso el usuario interpreta los patrones descubiertos y, posiblemente, regresa a pasos anteriores para nuevas iteraciones. Además, es posible visualizar los patrones encontrados, eliminar patrones redundantes o irrelevantes y traducir los patrones en términos comprensibles por el usuario.
9. **Utilización del nuevo conocimiento:** Última etapa del proceso en la cual se incorpora el conocimiento descubierto en el sistema y se toman acciones basadas en el conocimiento. Además, incluye la verificación y resolución de conflictos con el conocimiento extraído previamente.

2.5. Minería de datos

2.5.1. Concepto y definición de Minería de datos

Durante muchos años, grandes volúmenes de datos sólo eran almacenados en los repositorios debido a que se tenía un total desconocimiento de los beneficios que contenía toda esta

información. Teniendo en cuenta esto, se ha desarrollado una herramienta que permite convertir todos los datos en información valiosa, esta técnica es la Minería de Datos.

Pal (2012) señala que la Minería de Datos es "una tecnología utilizada para describir el descubrimiento de conocimiento y para buscar relaciones significativas, como patrones, asociaciones y cambios entre las variables en las bases de datos" (p.1).

Fayyad *et al.* (1996) establecen que la Minería de Datos es "la aplicación de algoritmos específicos para extraer patrones desde los datos" (p. 39).

Natek (2014) define que la Minería de Datos es "utilizada para extraer y descubrir conocimiento valioso y significativo desde una gran cantidad de datos" (p.1).

Ngai *et al.*, (2009) determinan que la Minería de Datos es un "proceso que utiliza técnicas estadísticas, matemáticas, de inteligencia artificial y de aprendizaje automático para extraer e identificar información útil y, posteriormente, obtener conocimiento de grandes bases de datos" (p. 2593).

En la educación cada vez es más común el uso de la Minería de Datos para mejorar el rendimiento académico de los estudiantes o identificar posibles desertores. Cuando la Minería de Datos es aplicada sobre este campo, se le conoce como Minería de Datos Educativa (EDM, por sus siglas en inglés). Badr *et al.*, (2016) define a la EDM como "el proceso utilizado para transformar datos sin procesar en información útil, para que profesores puedan tomar medidas correctivas y responder preguntas de investigación" (p. 81).

2.5.2. Objetivos de la Minería de Datos

La Minería de Datos posee tres objetivos importantes, los cuales son:

1. Obtener nuevo conocimiento útil a partir de la construcción de un modelo generado mediante un conjunto de datos.
2. Descubrir modelos inteligibles en los conjuntos de datos recopilados de diferentes repositorios de datos.

3. Tomar decisiones mucho más seguras y beneficiosas para la organización con los patrones descubiertos.

2.5.2.1. Objetivos Minería de Datos Educativa

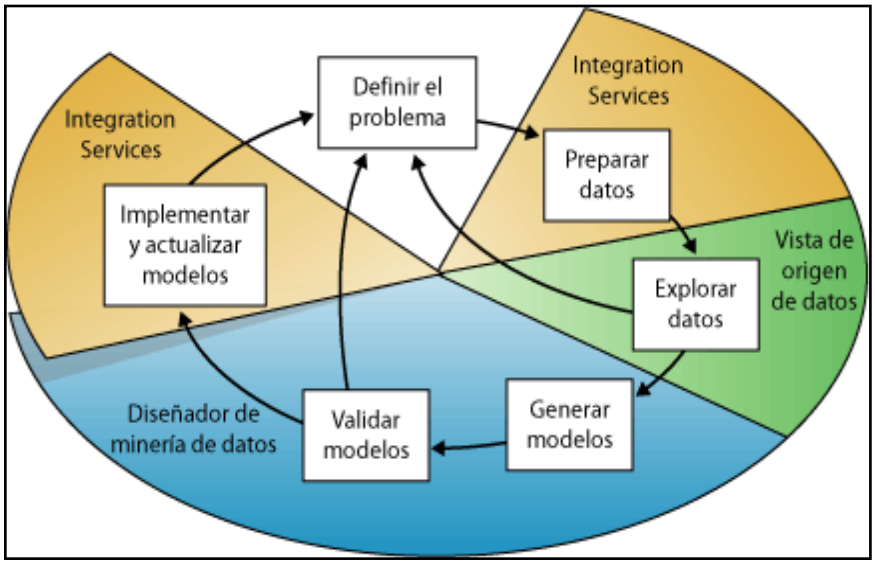
Según Baker & Yacef (2009) la minería de datos educativa posee cuatro objetivos:

- Desarrollar modelos para predecir el comportamiento de los estudiantes en el futuro.
- Descubrir o mejorar la estructura de conocimiento de dominio de un modelo.
- Estudiar el apoyo pedagógico para encontrar los tipos (de apoyo) que son más efectivos para diferentes grupos de estudiantes.
- Buscar evidencia empírica para ampliar las teorías educativas sobre cuáles son los factores claves que afectan el aprendizaje de un estudiante.

2.5.3. Etapas de la Minería de Datos

La Minería de Datos cuenta con seis etapas (Microsoft, 2017). Las etapas de esta técnica se muestran en la figura 2.

Figura 2: Etapas de la Minería de Datos



Fuente: Microsoft (2017)

A continuación se describe brevemente cada etapa de la Minería de Datos.

1. **Definir el problema:** Primera etapa del proceso donde se define de manera clara el problema y la forma en que se deben utilizar los datos para lograr resolver dicho problema.
2. **Preparar datos:** Fase en la cual se consolidan y limpian los datos identificados en la etapa anterior. Por lo general, los datos se encuentran dispersos por toda la empresa y es necesario unificarlos para poder realizar el análisis. Además, algunos de ellos no son válidos por lo que se eliminan para no generar problemas en el análisis.
3. **Explorar datos:** Los datos seleccionados para realizar el análisis se deben conocer perfectamente para tomar decisiones apropiadas en la creación de los modelos de minería de datos.
4. **Generar modelos:** Etapa en la cual se crea el modelo de minería de datos mediante los conocimientos adquiridos en la etapa de exploración de datos que ayuda a definir y crear el modelo.
5. **Validar modelos:** En este paso se exploran los modelos generados para comprobar su eficacia. Por lo general, cuando se genera un modelo, son creados con diversas configuraciones y se deben probar todos para ver cuál entrega mejores resultados para el problema.
6. **Implementar y actualizar modelos:** Última etapa del proceso y es donde se implementa el modelo que mejor resultados genera en el entorno de producción.

2.6. Modelos de Minería de Datos

2.6.1. Modelo predictivo

De acuerdo a Hernández (2007), los modelos predictivos intentan estimar valores futuros de ciertas variables que comúnmente se les denomina dependientes, mediante la utilización de diferentes campos de una base de datos, a la cuales se les conoce como variables independientes, es decir, los modelos predictivos hacen uso de datos ya conocidos para predecir de manera explícita valores futuros. Un ejemplo de este modelo sería determinar la posibilidad de que una entidad financiera realice un fraude bancario.

En los modelos predictivos identificamos los siguientes tipos: la clasificación y la regresión.

2.6.1.1. Clasificación

Según Hernández (2007), la clasificación debe ser la tarea con mayor nivel de utilización. En este tipo, se identifican instancias las cuales, mediante el valor del atributo, son relacionadas con una clase. El atributo anteriormente mencionado puede obtener diversos valores discretos y solamente, pertenecen a una clase en particular. Mientras que los otros atributos de la instancia son utilizados para predecir la clase.

La clasificación tiene como objetivo predecir clases con nuevas instancias que se desconocen en las clases, en otras palabras, busca aumentar la razón de precisión de la clasificación de las nuevas instancias que se van agregando. Esto se logra mediante el cálculo del cociente entre las predicciones correctas y el número total de predicciones.

En la tarea de la clasificación, existen diversas variantes como el aprendizaje de preferencias, aprendizaje de "rankings", aprendizaje de estimadores de probabilidad, entre otros.

2.6.1.2. Regresión

Consiste en aprender una función real que asigna a cada instancia un valor numérico. Esta es la principal diferencia respecto a la clasificación, ya que el valor a predecir es de tipo numérico (real). El objetivo principal de esta tarea es disminuir el error entre el valor predicho y el valor real, mediante el uso de la función generada.

Como la única diferencia con la clasificación es que la regresión predice valores numéricos, un modelo de regresión fácilmente podría convertirse en un modelo de clasificación.

2.6.2. Modelo descriptivo

Los modelos descriptivos buscan explorar e identificar patrones dentro de las propiedades de cada uno de los datos examinados, es decir, específica los patrones que existen en los datos pero no predicen nuevos valores. Un ejemplo de este modelo sería analizar la canasta de compra y la creación de grupos diferenciadores de empleados.

En los modelos descriptivos se identifican los siguientes tipos: agrupamiento, análisis correlacional y reglas de asociación.

2.6.2.1. Agrupamiento

El agrupamiento es considerado como una de los mejores dentro de los modelos descriptivos (Hernández, 2007), consiste en generar grupos naturales desde una colección de datos. En el agrupamiento, a diferencia de la clasificación, no busca analizar datos etiquetados en una clase, sino que realiza un análisis para generar esta etiqueta.

El agrupamiento para generar los grupos naturales se basa en el principio de maximizar la semejanza entre todos los elementos de un grupo pero al mismo tiempo minimizar la semejanza entre los diversos grupos.

2.6.2.2. Análisis correlacional

El análisis correlacional es utilizado para comprobar el grado de semejanza entre los valores de dos variables de tipo numéricas. Esta tarea tiene una fórmula estándar llamada "coeficiente de correlación r", es usada para la medición de la correlación lineal y entrega un valor real entre -1 y 1. Cuando r es 1 todas las variables están correlacionadas de manera perfecta, si r es -1 las variables están correlacionadas de manera perfecta pero negativamente, mientras que si r es 0 no existe correlación. Dicho de otra forma, cuando el r es positivo las variables se comportan de forma similar todas crecen o decrecen simultáneamente. Si r es negativo cuando una variable aumenta la otra disminuye.

2.6.2.3. Reglas de asociación

Las reglas de asociación son muy similares al análisis de correlaciones, su objetivo es la identificación de relaciones no explicitas entre los atributos categóricos. Esta tarea tiene muchos tipos de formulación, aunque la más común es "si un atributo X captura el valor d entonces el atributo Y obtiene el valor b".

Las reglas de asociación no involucran una relación causa-efecto, o sea, no necesariamente debe existir una causa para que haya una asociación en los datos.

2.7. Árboles de decisión

2.7.1. Definición

Los árboles de decisión son unos de los clasificadores más utilizados en la minería de datos. Consiste en un conjunto de condiciones que son organizadas en una estructura jerárquica, permitiendo tomar una decisión siguiendo las condiciones que se cumplen desde el inicio (raíz) del árbol hasta llegar a algunas de sus hojas.

Estos clasificadores son considerados como una forma de aprendizaje de reglas, debido a que cada rama del árbol se puede interpretar como una de esta, donde los nodos que existen entre la raíz y las hojas definen los términos de conjunción el cual constituye el antecedente de la regla, y la clase que se asigna a la hoja es el consecuente. Por lo general, estas interpretaciones son conocidas como "reglas de inducción", y tiene la siguiente forma:

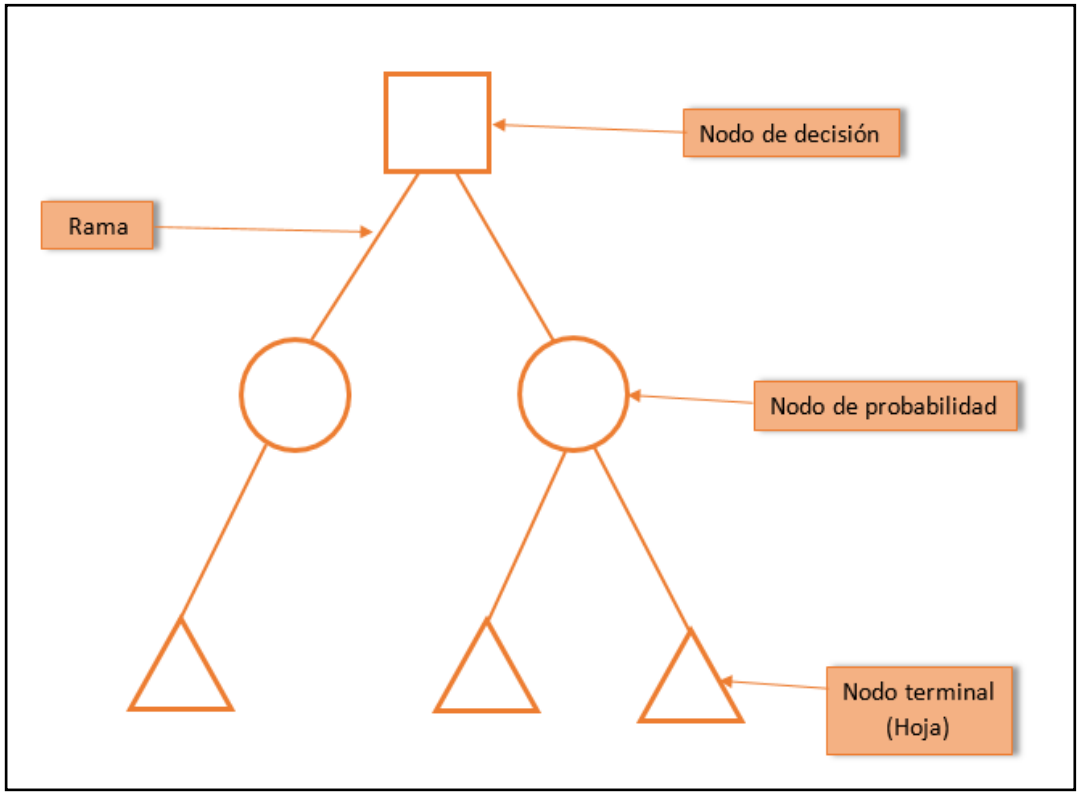
$$\boldsymbol{SI}\ cond_1\ \boldsymbol{Y}\ cond_2\ \boldsymbol{Y}\ \ldots \boldsymbol{Y}\ cond_n\ \boldsymbol{ENTONCES}\ pred$$

Un árbol de decisión es construido basándose en el algoritmo "divide y vencerás" (Hernández, 2007), el cual mediante un algoritmo de aprendizaje supervisado una serie de divisiones del espacio multivariable para incrementar la distancia entre los grupos en cada una de las divisiones.

2.7.2. Elementos de un árbol de decisión

Según Berlanga, Rubio & Vilà (2013) un árbol de decisión está compuesto por cuatro elementos: Nodo de decisión, nodo de probabilidad, nodo terminal y ramas, tal cual se muestra en la siguiente figura.

Figura 3: Elementos de un árbol de decisión



Fuente: Elaboración propia

- **Nodo de decisión:** Indica que en ese lugar del proceso se necesita tomar una decisión. Se representa por un cuadrado.
- **Rama:** Muestra los diferentes caminos que se pueden abordar cuando se toma una decisión.
- **Nodo de probabilidad:** Indica que en ese lugar del proceso sucede un acontecimiento aleatorio. Se representa por un círculo.
- **Nodo terminal (Hoja):** Nodo homogéneo el cual no requiere ninguna división adicional debido a que se encuentra en estado "puro". En otras palabras, indica un resultado.

2.7.3. Ventajas de los árboles de decisión

Rokach & Maimon (2015) establece las siguientes ventajas de los árboles de decisión:

- Los árboles de decisión son muy fáciles de interpretar
- Permiten manejar atributos nominales y numéricos
- Pueden manejar conjuntos de datos con errores
- Los árboles de decisión permiten el manejo de conjuntos de datos con valores faltantes
- Son considerados como un método no paramétrico, esto es, que las decisiones no incluyen supuestos sobre la distribución del espacio ni sobre la estructura del clasificador
- Disminuye la cantidad de variables independientes
- Facilita el entendimiento del conocimiento utilizado en la toma de decisiones

2.7.4. Desventajas de los árboles de decisión

Rokach & Maimon (2015) establece las siguientes desventajas de los árboles de decisión:

- Algunos algoritmos requieren que su atributo objetivo tenga solamente valores discretos
- Los árboles de decisión al utilizar el método "divide y vencerás" no funcionan correctamente cuando existen interacciones complejas
- Utilizar atributos irrelevantes y/o con ruido generan árboles inestables
- Requiere de gran esfuerzo para tratar datos faltantes o perdidos

2.7.5. Como construir un árbol de decisión

Existen diversos tipos de algoritmos de clasificación para árboles de decisión, pero todos ellos tienen la misma idea básica:

- Cada nodo no terminal se encuentra etiquetado con un atributo.
- Cada rama que sale de un nodo es etiquetada con un valor de ese atributo.
- Cada nodo terminal es etiquetado por un conjunto de casos, los que satisfacen todos los valores de atributos que se etiquetan en el camino desde ese nodo al nodo inicial.

Cuando se aplica un atributo A como criterio de selección se clasifican los casos en diferentes conjuntos. El objetivo principal es construir un árbol más simple que sea coherente con el conjunto de entrenamiento T. Para lograrlo se deben ordenar los atributos más relevantes, desde

el nodo raíz a los nodos terminales, de mayor a menor poder de clasificación. El poder de clasificación se refiere a la capacidad que tiene un atributo A para generar particiones en el conjunto de entrenamiento que se ajusten en un grado dado a las diversas clases posibles; de esta manera se introduce un orden o desorden (ruido) en el conjunto, el cual se puede medir. El poder de clasificación de un atributo se mide en base a su capacidad para disminuir la incertidumbre o entropía, y esta métrica es conocida como "ganancia de información". El atributo con mayor ganancia de información es elegido como un atributo que forme un nodo del árbol.

El árbol de decisión se construye de la siguiente manera:

- Calcular la entropía de cada atributo.
- Ordenar los atributos de mayor a menor capacidad de disminución de la entropía.
- Construir el árbol de decisión siguiendo el orden establecido en el punto anterior.

Según Timarán-Pereira *et al.* (2016), la ganancia de información que se obtiene al particionar el conjunto T, de acuerdo con un atributo A es definida como:

$$Gain(T, A) = I(T) - E(A)$$

Donde *I(T)* es la entropía del conjunto compuesto de T ejemplos y m diferentes clases Ci (i=1, m) y es calculada de la siguiente forma:

$$I(T) = -\sum p_i log_2 (p_i)$$

Donde, $p_i = S_i/S$ corresponde a la probabilidad de los posibles valores.

E(A) es la entropía del conjunto T cuando es particionado por los n diferentes valores del atributo A en n subconjuntos, $\{S_i, S_2, \ldots S_n\}$, donde S_j incluye esos ejemplos de T que poseen el valor a_j en A y S_{ij} la cantidad de ejemplos de la clase C_i en el subconjunto S_j.

E(A) se calcula de la siguiente manera:

$$E(A) = \sum S_{ij}/S * I(S_{ij})$$

Donde, S_{ij} es la cantidad de ejemplos de la clase C_i en el subconjunto S_j

$$I(S_{ij}) = -\sum p_{ij}\, log_2\, (p_{ij})$$

Donde $p_{ij} = S_{ij}/|S_i|$ corresponde a la probabilidad que un ejemplo de S_j pertenezca a la clase C_i.

2.7.5.1. Otra métrica: Impureza de Gini

El índice de Gini desarrollado por el estadístico italiano Corrado Gini es utilizado para medir la desigualdad de ingresos o para cualquier forma de distribución desigual. Existe una forma modificada de este índice llamada "Impureza de Gini", la cual mide las proporciones de las clases en un conjunto de datos y es comúnmente utilizado cuando la variable dependiente es una variable categórica. Además, su valor mínimo es cero y el máximo es 1 - (1/k), donde k es el número de categorías en la variable objetivo.

El índice Gini de un nodo t, $GINI(t)$, es definido como:

$$GINI(t) = \sum_{j \neq i} p(j/t)\, p(i/t)$$

Donde i, j son categorías de la variable dependiente. La ecuación del índice de Gini también se puede escribir como:

$$g(t) = 1 - \sum_{j} p^2\, (j/t)$$

Donde $p\ (j/t)$ es la proporción de la categoría objetivo j presente en el nodo t.

Cuando los casos de un nodo son distribuidos de manera uniforme entre todas las categorías, el índice de Gini toma su valor máximo 1 – (1/k), donde k es el número de categorías en la variable objetivo. Sin embargo, cuando los casos de un nodo pertenecen a la misma categoría, el índice de Gini es cero.

Según Soman *et al.* (2006), la función del criterio de Gini para dividir s en el nodo t se define de la siguiente manera:

$$GINI_{split}(s,t) = GINI\ (t) - p_L * GINI\ (t_L) - p_R * GINI\ (t_R)$$

Donde p_L es la proporción de casos en el nodo t enviados al nodo hijo izquierdo y p_R es la proporción de casos en el nodo t enviados al nodo hijo derecho.

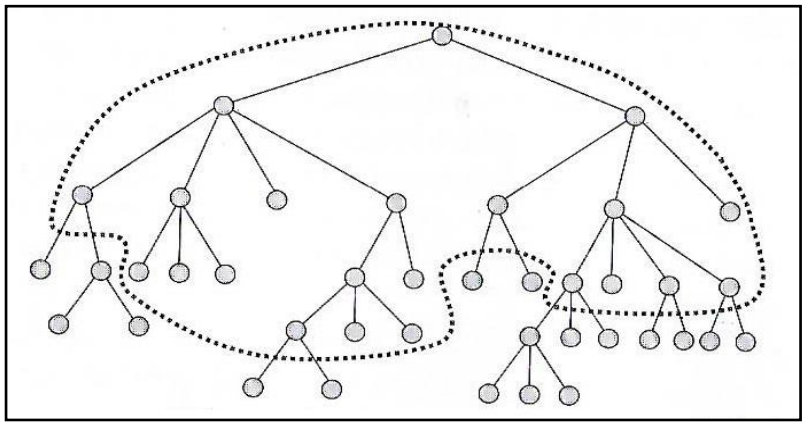
La división s es escogida para maximizar el valor de $GINI_{split}(s,t)$, y este valor informa la mejora del árbol.

2.7.6. Poda en los árboles de decisión

Según Hernández (2007), los árboles de decisión generan modelos muy completos y consistentes con respecto a la evidencia. A simple vista pueden parecer óptimos, pero en realidad tienen algunos problemas. En primer lugar, ajustarlos de manera excesiva a la evidencia genera que el modelo se comporte mal para nuevos ejemplos. En segundo lugar, la evidencia puede incluir errores en los atributos o en las clases, por lo que el modelo se ajustará a los errores y en consecuencia perjudicara el comportamiento global del modelo.

Para limitar este problema es necesario modificar los algoritmos de aprendizaje para la generación de modelos más generales, esto último se refiere a eliminar condiciones en las ramas. Dicho procedimiento, en los árboles de decisión se le conoce como “poda”, tal como se muestra en la siguiente figura.

Figura 4: Ejemplo de poda



Fuente: Hernández (2007)

En la figura 4 se puede observar que todos los nodos que se encuentran debajo del límite de la poda serán eliminados, debido a que son considerados demasiado específicos.

2.7.6.1. Métodos de poda

Según Hernández (2007), en la poda de un árbol se pueden distinguir dos métodos: prepoda y pospoda.

Prepoda: Procedimiento que se realiza mientras el árbol de decisión es construido. Este método determina el criterio de detención a la hora de seguir especializando una regla o rama. Los criterios de prepoda suelen basarse en los números de ejemplos por nodo, número de excepciones respecto a la clase mayoritaria o técnicas más sofisticadas, como el criterio MDL.

Pospoda: Procedimiento que se realiza después que el árbol de decisión es construido. Este método busca eliminar nodos de abajo hacia arriba hasta llegar a un límite. Los criterios de pospoda se basan en las mismas medidas que la prepoda, aunque este criterio suele obtener mejores resultados al realizarse con una visión mucho más completa del modelo. Sin embargo, la pospoda no genera nada que luego deba eliminarse, lo cual hace que este criterio sea menos eficiente que la prepoda.

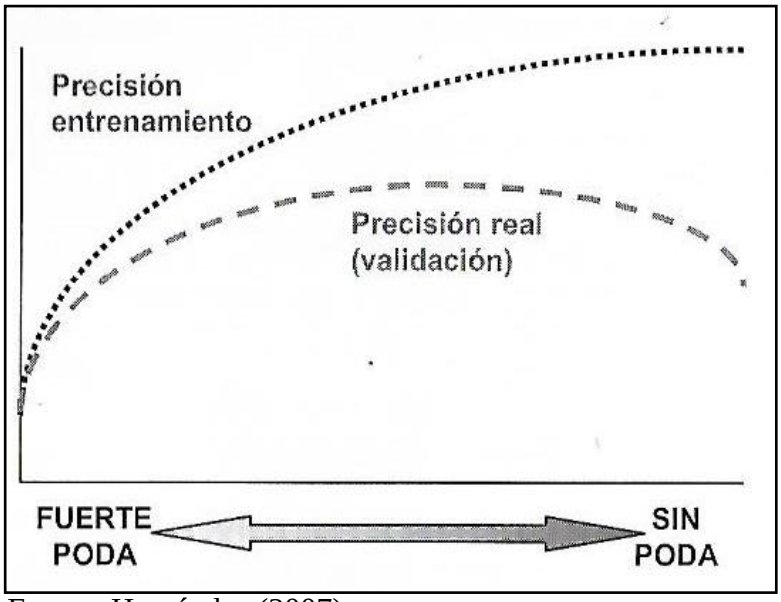
Estas dos técnicas pueden ser combinadas. Un ejemplo de ello es el algoritmo C4.5, del cual hablará en la siguiente sección, en donde utiliza una prepoda por cardinalidad y una pospoda basada en un criterio más sofisticado. El uso de prepoda o pospoda (o los dos) provoca que los

nodos hojas ya no sean puros, o sea, probablemente posean ejemplos de varias clases. Generalmente la clase con más ejemplos será elegida para etiquetar al nodo hoja y generar la predicción.

Por otra parte, es muy importante determinar el nivel de poda óptimo. Según Hernández (2007), diversos métodos de poda poseen al menos un parámetro que les permite decidir el grado de poda, aunque es prácticamente imposible encontrar un grado que funcione correctamente para todos los problemas debido a que muchos poseen más ruido, más ejemplos, o características especiales.

Una forma de ajustar el nivel de poda a un determinado problema es analizando el comportamiento que tiene este con respecto a los datos de validación. La siguiente figura muestra cómo cambia la precisión de un árbol con respecto a los datos de entrenamiento y los datos de validación.

Figura 5: Efecto beneficio de la poda



Fuente: Hernández (2007)

En la figura 5 se puede observar que cuando se aplica una poda a los datos de entrenamiento, esta no genera beneficios pero cuando se aplica una poda a los datos de validación el

comportamiento es esclarecedor; en cada problema existe un grado óptimo de poda y se puede estimar utilizando datos de validación.

2.7.7. Algoritmos de árboles de decisión

A continuación se describe brevemente algunos de los algoritmos de árboles de decisión:

2.7.7.1. ID3

El algoritmo ID3 considerado como un árbol de decisión muy simple fue introducido en 1986 por Quinlan Ross y se basa en el algoritmo de Hunts. Este algoritmo utiliza el concepto de "ganancia de información" como criterio de división y deja de crecer cuando en un valor de la característica objetivo tiene asignadas todas las instancias o cuando el mejor criterio de división (ganancia de información) no es mayor a cero. Sin embargo, ID3 no permite aplicar ningún tipo de poda, ni manejar atributos que sean numéricos o valores perdidos (Rokach & Maimon, 2015).

Rokach & Maimon (2015) establecen que una de las grandes ventajas de este algoritmo es su simplicidad, lo que permite que sea utilizado con mayor frecuencia. No obstante, ID3 tiene algunas desventajas como:

- ID3 utiliza una estrategia muy ambiciosa la cual no garantiza una solución óptima.
- ID3 solamente permite atributos nominales, entonces para poder usar datos continuos estos deben ser convertidos.

2.7.7.2. C4.5

El algoritmo C4.5 es una evolución del ID3 y también se basa en el algoritmo de Hunt. C4.5 utiliza la "razón de ganancia" como criterio de división y se detiene cuando la cantidad de instancias a dividir se encuentra por debajo de cierto límite. Además, permite utilizar atributos numéricos y poda, esta última solamente puede realizarse después de la fase de crecimiento (Rokach & Maimon, 2015).

Rokach & Maimon (2015) indican que el algoritmo C4.5 incluye varias mejoras con respecto al algoritmo ID3, las más importantes son:

- C4.5 permite utilizar poda la cual elimina las ramas que no ayudan a la precisión y las reemplaza con nodos hojas.
- C4.5 permite manejar valores perdidos de los atributos.

2.7.7.3. CART

El algoritmo CART (Classification And Regression Trees) fue desarrollado en 1984 por Breiman y, al igual que los algoritmos ID3 y C4.5, se basa en el algoritmo de Hunt. CART permite manejar atributos categóricos y continuos para construir el árbol (Rokach & Maimon, 2015).

El algoritmo hace uso del índice de Gini para seleccionar atributos para la construcción del árbol de decisión. A diferencia de los algoritmos ID3 y C4.5, CART produce divisiones binarias por lo que genera árboles binarios. Sin embargo, CART también puede generar árboles de regresión, cuyas hojas predicen un número real y no una clase (Rokach & Maimon, 2015).

2.7.7.4. CHAID

El algoritmo CHAID (CHi-square Automatic Interaction Detector) fue desarrollado a principio de los años setenta por investigadores de estadística aplicada. Consiste en un algoritmo de árbol estadístico y multidireccional que inspecciona todos los datos de manera rápida y eficaz, y crea perfiles o porciones en relación al resultado deseado. Además, permite detectar automáticamente interacciones mediante Chi-cuadrado (Berlanga *et al.*, 2013).

CHAID selecciona la variable independiente que muestra una interacción mayor con respecto a la variable dependiente. Cuando una categoría de un predictor no es significativamente distinta a la variable dependiente, estas se funden (Berlanga *et al.*, 2013).

2.8. SAP Predictive Analytics

SAP Predictive Analytics es un software de inteligencia empresarial desarrollado por la empresa SAP SE, el cual entrega soluciones de análisis estadístico y minería de datos para la elaboración de modelos predictivos que permitan descubrir conocimientos y relaciones ocultas

en los datos (https://help.sap.com/viewer/product/SAP_PREDICTIVE_ANALYTICS/3.0/es-ES, el 15 de Febrero de 2018).

SAP Predictive Analytics combina características de SAP InfiniteInsight y SAP Predictive Analysis en un solo software. Además, incluye dos interfaces de usuario, Automated Analytics y Expert Analytics.

2.8.1. SAP Predictive Analytics – Automated Analytics

SAP Predictive Analytics – Automated Analytics proporciona una solución de Minería de Datos para modelar datos con la mayor facilidad y rapidez posible, manteniendo resultados relevantes y fácilmente interpretables (https://help.sap.com/viewer/bc031e667eea409c9e08e7ab8b1e4c70/3.3/en-US/b474299e88ed46e0820c6d9d471e48a3.html, el 17 de Febrero de 2018). Esta interfaz incluye los siguientes módulos: Modeler, Social y Recommendation.

2.8.1.1. Modeler

Permite crear diferentes tipos de modelos y exportarlos para ser utilizados en su entorno de producción. Modeler proporciona las siguientes funciones:

- Crear un modelo de clasificación/regresión
- Crear un modelo de agrupación en clústeres
- Crear un análisis de serie temporal
- Crear reglas de asociación
- Cargar un modelo

2.8.1.2. Social

Permite extraer y utilizar información estructural que se encuentra almacenada en diversos conjuntos de datos, permitiendo mejorar las capacidades de decisión y predicciones de los modelos. Social proporciona las siguientes funciones:

- Crear un análisis de redes sociales
- Crear un análisis de colocación

- Crear un análisis de vía de acceso frecuente
- Cargar un modelo de análisis de redes sociales

2.8.1.3. Recommendation

Genera recomendaciones de productos basándose en los análisis de las redes sociales. Recommendation proporciona las siguientes funciones:

- Crear un recomendador nuevo
- Cargar un recomendador

2.8.2. SAP Predictive Analytics – Expert Analytics

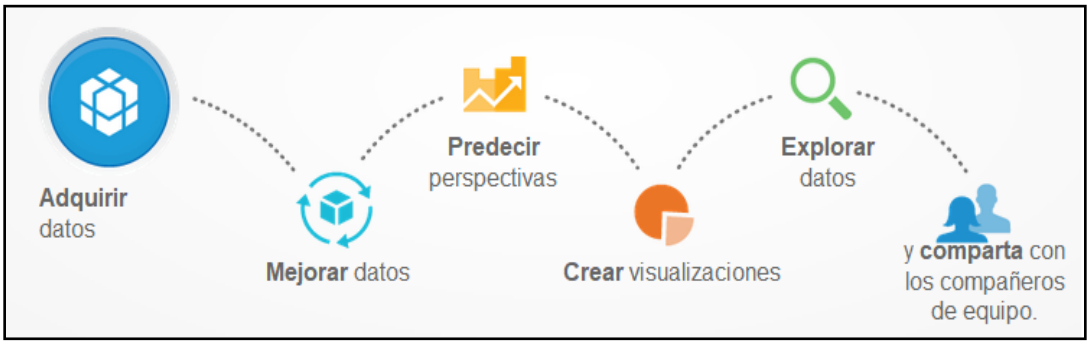
SAP Predictive Analytics – Expert Analytics es una interfaz con menor grado de automatización pero que ofrece una mayor potencia y flexibilidad (https://sapexperts.wispubs.com/bi/articles/sap-predictive-analytics-part-2-an-overview-of-the-expert-analytics-tool?id=9ea49d471cf44be29ea7283a669d7be9#.Wt5EacgvzIU , el 17 de Febrero de 2018). Expert Analytics permite hacer lo siguiente:

- Realizar análisis de datos, esto incluye la previsión de series de tiempo, detección de datos atípicos, análisis de tendencias, análisis de clasificación, análisis de segmentación y análisis de afinidad.
- Realizar análisis de datos a través de diversas técnicas de visualización como árboles de decisión, gráficos de matriz de dispersión, gráficos de clúster y coordenadas paralelas.
- Utilizar diversos algoritmos para realizar análisis predictivo y análisis estadístico de código abierto R.

2.8.2.1. Etapas Expert Analytics

El modo experto de SAP Predictive Analytics permite realizar un análisis predictivo siguiendo tres etapas: adquirir datos, mejorar datos y predecir perspectivas. Además, incluye tres paneles opcionales como crear visualizaciones, explorar datos y compartir. A continuación se muestra las etapas de la interfaz Expert Analytics.

Figura 6: Etapas de SAP Predictive Analytics- Expert Analytics



Fuente: Software SAP Predictive Analytics- Expert Analytics

1. **Adquirir datos:** Consiste en obtener datos desde una fuente para poder generar una predicción. Expert Analytics permite adquirir datos desde diversas fuentes de datos como hoja de trabajo Excel, archivos de texto como .csv o .txt, entre otros. Además, incluye una vista previa de estos, donde analiza los datos de manera sistemática y examina las columnas para identificar su tipo de dato.
2. **Mejorar datos:** Etapa en la cual los datos se mejoran debido a que generalmente estos se formatean de forma inconsistente la primera vez que se adquieren, lo que provoca que los usuarios no puedan interpretarlos de manera fácil.
3. **Predecir perspectivas:** Etapa en la cual se realiza el análisis de predicción, previa configuración de los algoritmos a utilizar. Cuando el análisis ha finalizado, los resultados se representan mediante diversos gráficos de visualización. Uno de ellos es la ***matriz de confusión*** la cual contiene información sobre la clasificación real y predictiva ejecutada por el algoritmo. Consiste en una matriz de n*n que determina el número de sucesos para todos los valores previstos teniendo en cuenta el valor real. Este tipo de gráfico se encuentra disponible cuando se ha seleccionado el método de salida clasificación y tendencia.
4. **Crear visualizaciones:** Permite crear dimensiones, indicadores y gráficos de datos mediante un panel.
5. **Explorar datos:** Permite la generación de diversos resúmenes e infográficos. Estos últimos son una composición visual de formas, imágenes, pictogramas y colores utilizados para expresar mensajes de forma mejorada.

6. **Compartir:** Expert Analytics permite exportar los modelos generados como vistas y procedimientos a SAP HANA (plataforma de computación in-memory para acelerar procesos de negocios) y a una carpeta local de un equipo.

CAPÍTULO III

METODOLOGÍA

En este proyecto no se utilizará una metodología de desarrollo de software propiamente tal. Sin embargo, se podría ser considerar como evolutivo en el sentido de que se estará interactuando con el usuario de manera constante para que pueda visualizar los resultados parciales que se han generado y mostrar cómo se va mejorando la predicción del modelo si se cambian los algoritmos y las variables.

Para generar el modelo se hará uso del proceso KDD el cual es un "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" (Fayyad *et al*, 1996). Este proceso consta de 5 etapas, las cuales se resumen a continuación:

1. **Selección de datos:** Etapa en la cual se selecciona el conjunto de datos objetivo, con el propósito de realizar un análisis para resolver el problema definido.
2. **Pre-procesamiento de datos:** Fase que verifica que los datos del conjunto sean de calidad para garantizar que el conocimiento a descubrir sea de un alto grado. Esta etapa incluye tareas como la eliminación de datos ruidosos, filtrado de datos atípicos, manejo de datos nulos, desconocidos y duplicados.
3. **Transformación de los datos:** Etapa en la cual se modifica la estructura de los datos, a una estructura con características útiles para facilitar el análisis y la meta del proceso.
4. **Minería de datos:** Es la fase donde se aplican los algoritmos para descubrir patrones insospechados dentro de un conjunto de datos.
5. **Interpretación/Evaluación de resultados:** Última etapa del proceso KDD donde se interpretan los patrones descubiertos en base a tres criterios: precisión, claridad, e interés.

A continuación se describe la herramienta utilizada y los algoritmos para generar el modelo.

3.1. Herramienta elegida

Existen diversas herramientas que permiten realizar un análisis predictivo. Para este estudio se hará uso del software SAP Predictive Analytics desarrollado por la empresa SAP SE. La elección de esta herramienta radica principalmente en que como la Universidad del Bío-Bío se

encuentra dentro del programa "SAP University Alliances" lo cual permite integrar tecnologías SAP a la enseñanza, entonces se desea conocer y tener un primer acercamiento, sobre todo en el área de la educación, de la herramienta que ha desarrollado esta empresa para el análisis predictivo.

En SAP Predictive Analytics se han identificado algunas ventajas, las cuales son:

- El software se encuentra completamente en español
- Permite utilizar 27 diferentes algoritmos
- Entrega tres tipos de informes: información general del modelo, informe ejecutivo y resumen de parámetros de modelación; los cuales se pueden exportar a pdf
- Posee un sitio oficial de ayuda para conocer funcionalidades del software
- Existen video-tutoriales para conocer cómo funcionan los algoritmos

3.2. Tipo de modelo y algoritmos a utilizar

Para el presente estudio se utilizara el modelo predictivo de clasificación basado en árboles de decisión. Se utilizó el algoritmo R-CNR Tree (CART) para generar el árbol de decisión y el algoritmo Inter Quartile Range para encontrar valores atípicos en algunas variables del conjunto de datos. La configuración utilizada para generar el árbol de decisión usando R-CNR Tree es la siguiente:

Tabla 3: Configuración del algoritmo R-CNR Tree.

	Propiedad	**Descripción**
Información de salida	Modo de salida	Tendencia
Manejo de datos de entrada	Valores perdidos	Rpart
Poda de árboles	División mínima	10
	Parámetro de complejidad	0.001
	Profundidad máxima	6
Comportamiento	Dividir criterios	Gini

Fuente: Elaboración propia.

El modo de salida será “Tendencia”, el cual predice valores para la variable dependiente y agrega una columna adicional con valores previstos en la salida. Para manejar los valores perdidos se utilizará el algoritmo “Rpart”, el cual elimina observaciones que le faltan a la variable dependiente pero retiene estas observaciones en las variables independientes. En cuanto a la “división mínima” esta consiste en un número mínimo de observaciones que son necesarios para dividir un nodo, en este caso se utilizó el valor predeterminado 10. La “profundidad máxima” significa el nivel de nodo máximo que existirá en el árbol final, por otra parte el “parámetro de complejidad” busca controlar la expansión del árbol y evita divisiones que no mejoran el ajuste. Finalmente se hará uso de Gini (Impureza de Gini) como criterio de división del nodo.

CAPÍTULO IV

RESULTADOS

4.1. Selección de datos

Etapa en la cual se procede a obtener los datos de los alumnos de primer año. Estos datos son obtenidos a partir de la Dirección y Admisión, Registro y Control Académico (DARCA). Los datos se recibieron en diferentes archivos por lo que se tuvo que realizar un trabajo de unificación de datos para poder observar mejor lo que se va a analizar.

4.2. Pre-procesamiento de datos

En esta etapa se verifica que los datos del conjunto sean de calidad para garantizar que el conocimiento a descubrir sea de un alto nivel.

El conjunto de datos está compuesto por 35 variables: 34 variables independientes y una variable dependiente. A continuación se definen cada una de estas variables.

- **Variables independientes**

  En la Tabla 4 muestra las 34 variables independientes del conjunto de datos.

  Tabla 4: Variables independientes

Variable	Descripción
Año	Variable numérica que indica el año que ingresó el alumno a la universidad.
Rut	Variable numérica que identifica a un estudiante en particular.
Genero	Variable que indica el sexo del estudiante.
Edad	Variable numérica que indica la edad del estudiante.
Tipo_colegio	Variable que indica el tipo de colegio del cual proviene un alumno, ya sea municipal, particular o subvencionado.
Tipo_enseñanza	Variable que indica el tipo de enseñanza que recibió el estudiante en la enseñanza media.

Variable	Descripción
PSU_Lenguaje	Variable que indica el puntaje que obtuvo el alumno en la PSU de Lenguaje
PSU_Matematica	Variable que indica el puntaje que obtuve el alumno en la PSU de matemática.
PSU_Ciencias	Variable que indica el puntaje que obtuve el alumno en la PSU de ciencias.
PSU_Historia	Variable que indica el puntaje que obtuve el alumno en la PSU de matemática.
Nem	Variable que indica la nota de enseñanza media que obtuvo el alumno.
Ranking	Variable que indica el puntaje ranking de notas que obtuvo el alumno.
BEA	Variable que indica si el estudiante posee la beca excelencia académica.
Orden_postulacion	Variable de tipo numérica que indica el orden de preferencia que el alumno le da a su carrera.
Años_acreditacion	Variable que indica la cantidad de años que está acreditada la carrera en que un alumno se ha matriculado.
Becas_internas	Variable que indica si el alumno posee beneficios estudiantiles internos
Duración_carrera	Variable que indica la duración de la carrera.
Sede	Variable que indica la sede a la cual pertenece la carrera del alumno.

Variable	Descripción
Gratuidad	Variable que indica si el estudiante posee gratuidad.
Discapacidad	Variable numérica que indica si el alumno posee alguna discapacidad.
Vivía_enseñanza_media	Variable numérica que indica con quien vivía el alumno en la enseñanza media.
Vive_año_actual	Variable numérica que indica con quien vivirá el alumno en la enseñanza superior.
Hijos	Variable numérica que indica la cantidad de hijos que tiene un alumno.
Trabajado_alguna_vez	Variable que indica si el alumno ha trabajado alguna vez antes de ingresar a la enseñanza superior.
Piensa_trabajar	Variable que indica si el alumno piensa trabajar durante la enseñanza superior.
Motivación_trabajar	Variable numérica que indica la motivación que tiene el estudiante para trabajar.
Tipo_ingreso	Variable que indica el proceso por el cual el alumno ingresó a la universidad
Carrera	Variable numérica que indica la carrera en la cual se ha matriculado el estudiante.
Provincia_domicilio	Variable que indica la provincia de domicilio del alumno.
Procedencia_colegio	Variable que indica la provincia del colegio de la cual viene un estudiante.

Variable	Descripción
Créditos_totales_aprobados	Variable numérica que indica el total de créditos que aprobó el alumno durante su primer año en la universidad
Créditos_totales_malla	Variable numérica que indica el total de créditos que debería aprobar el alumno según su malla curricular.
Porcentaje_éxito	Variable que indica los créditos que aprobó el alumno sobre los créditos que debería aprobar según su malla.

Fuente: Elaboración propia.

- **Variable dependiente**

  La variable dependiente para este estudio se ha denominado **Éxito_académico**. Esta variable nos indica si un alumno tuvo o no éxito académico. En la siguiente tabla se pueden observar las opciones que existen para esta variable.

  Tabla 5: Variable dependiente

Éxito_académico	**Si**: El alumno de primer año aprobó todas las asignaturas de su malla curricular.
	No: El alumno de primer año no aprobó todas las asignaturas de su malla curricular.

  Fuente: Elaboración propia.

  Para asignar un "Si" o un "No" a un estudiante en particular esto depende principalmente de la variable independiente "**Porcentaje_éxito**". Cuando esta variable es mayor o igual a 1, significa que el estudiante si aprobó todas las asignaturas de primer año por lo que se le asignará un "Si" en la variable dependiente. En caso contrario, cuando la variable independiente "**Porcentaje_éxito**" es menor a 1, el estudiante no

aprobó todas las asignaturas de primer año por lo que se le asignará un "No" en la variable dependiente.

Se realiza una limpieza en el conjunto de datos en donde son eliminados los datos que cumplan algunos de los siguientes puntos:

- No se tiene información sobre cuantos créditos aprobó durante el año un alumno en particular.
- El alumno es un desertor, ya sea temporal o definitivo.
- El alumno no realizó una inscripción de asignaturas.
- El alumno posee valores nulos o sin información (SinInf). Sin embargo, esto sólo se aplica a las variables que podrían ser incluidas en los modelos, mientras que las variables que serán excluidas sí podrían contener valores nulos o sin información (SinInf). Esto último solo ocurre para las variables PSU de Ciencias y PSU de Historia.

Por otra parte, la limpieza para los valores atípicos (outliers) es diferente. Para filtrar estos datos se hace uso del algoritmo "Inter Quartile Range", el cual se encuentra disponible en el software SAP Predictive Analytics en la interfaz Expert Analytics.

El algoritmo "Inter Quartile Range" es aplicado a tres variables: Edad, Ranking y Nem. La configuración del algoritmo se puede apreciar en la Tabla 6.

Tabla 6: Configuración del algoritmo Inter Quartile Range

Inter Quartile Range	
Modo de salida	Eliminar valores atípicos
Coeficiente Fence	1.5

Fuente: Elaboración propia

En la Tabla 6 se puede observar que al encontrar valores atípicos estos serán eliminados, además, el coeficiente Fence es la desviación permitida para los valores desde el rango intercuartil, 1.5 es el valor predeterminado. Los resultados de aplicar el algoritmo a las variables antes descritas se pueden observar en la Tabla 7.

Tabla 7: resultados del algoritmo Inter Quartile Range

Variable	**Número de valores atípicos detectados**
Edad	669
Ranking	0
NEM	0

Fuente: Elaboración propia.

En la Tabla 7 se observa que tanto la variable NEM como la variable Ranking no se encontraron datos atípicos mientras que para la variable Edad se encontraron 669 valores atípicos, los cuales fueron eliminados por el algoritmo.

Luego de haber realizado la limpieza de datos, se obtiene un archivo Excel con datos de calidad. Basado en los estudios de Tejedor (2003), Shahiri *et al.* (2015), Mayilvaganan & Kalpanadevi (2014), Barahona *et al.* (2016), Salinas *et al.* (2017), Quadril & Kalyankar (2010) y Rosas *et al.* (2006), se han seleccionado las siguientes variables que podrían ser ingresadas al modelo final. Estas variables se pueden observar en la Tabla 8.

Tabla 8: Variables para modelo según diversos autores

Variables
Género
Edad
Tipo_enseñanza
Nem
PSU_Lenguaje
PSU_Matemática
Provincia_domicilio
Procedencia_colegio
Discapacidad
Tipo_colegio
BEA

Orden_postulacion
Becas_internas
Vive_año_actual
Piensa_trabajar
Ranking

Fuente: Elaboración propia.

Además, se incorporó 7 nuevas variables para analizar si realmente influyen en el éxito académico de un alumno. Estas variables se pueden observar en la Tabla 9.

Tabla 9: Variables a analizar si influyen en el rendimiento académico

Variables
Carrera
Renovación_curricular
Tipo_ingreso
Hijos
Vivía_enseñanza_media
Trabajado_alguna_vez
Gratuidad

Fuente: Elaboración propia.

4.3. Transformación de datos

En esta etapa se modificó la estructura de los datos, a una estructura con características que permita facilitar el análisis para cumplir con el objetivo del proceso. Algunas de estas modificaciones son: cambiar algunas variables de tipo carácter a tipo numérico y aplicar algoritmos de agrupación a otras. Se seleccionó el algoritmo "Auto Clustering" para agrupar datos semejantes de manera de que al ejecutar el algoritmo de clasificación sea más fácil su interpretación. En el software Predictive Analytics, este algoritmo no posee una configuración fija, por lo que se estableció un número mínimo de clústeres a generar (6). El número máximo

de clústeres es definido de acuerdo al número donde los datos se solapan menos, aunque la idea principal es que estos no superen los 10 clústeres.

A continuación, se muestra cada una de las transformaciones realizadas a cada variable para predecir el rendimiento académico. En la columna izquierda aparece la etiqueta por la cual fue reemplazada la información aunque en la variable "Carrera", estas etiquetas se muestran en columnas hacia la derecha.

- **Variable Género**

  Las dos etiquetas para esta variable son:

  Tabla 10: Caracterización variable "Genero"

Genero	
F	Femenino
M	Masculino

  Fuente: Elaboración propia.

- **Variable Tipo_colegio**

  Las tres etiquetas para esta variable son:

  Tabla 11: Caracterización variable "Tipo_colegio"

Tipo_colegio	
1	Particular
2	Subvencionado
3	Municipal

  Fuente: Elaboración propia.

- **Variable Tipo_enseñanza**

  Las dos etiquetas para esta variable son:

  Tabla 12: Caracterización variable "Tipo_enseñanza"

Tipo_enseñanza	
HC	Humanista científico
TP	Técnico profesional

  Fuente: Elaboración propia.

- **Variable BEA**

  Las dos etiquetas para esta variable son:

  Tabla 13: Caracterización variable "BEA"

Variable BEA	
Si	Alumno con beca excelencia académica
No	Alumno sin beca excelencia académica

  Fuente: Elaboración propia.

- **Variable Orden_postulación**

  Las cinco etiquetas para esta variable son:

  Tabla 14: Caracterización variable "Orden_postulación"

Variable Orden_postulación	
1	Preferencia 1
2	Preferencia 2
3	Preferencia 3
4	Preferencia 4, 5, 6 y 7

0	Preferencia 8, 9, 10 y no postulo a ninguna carrera

Fuente: Elaboración propia.

- **Variable Becas_internas**

Las dos etiquetas para esta variable son:

Tabla 15: Caracterización variable "Becas_internas"

Variable Becas_internas	
Si	Alumnos con beca interna
No	Alumnos sin beca interna

Fuente: Elaboración propia.

- **Variable Gratuidad**

Las dos etiquetas para esta variable son:

Tabla 16: Caracterización variable "Gratuidad"

Variable Gratuidad	
Si	Alumnos con gratuidad
No	Alumnos sin gratuidad

Fuente: Elaboración propia.

- **Variable Discapacidad**

  Las seis etiquetas de esta variable son:

  Tabla 17: Caracterización variable "Discapacidad"

Variable Discapacidad	
0	Ninguna
1	Dificultad física
2	Dificultad del habla
3	Dificultades psiquiátricas mentales intelectuales
4	Dificultad auditiva
5	Dificultad en la visión

  Fuente: Elaboración propia.

- **Variable Vivía_enseñanza_media**

  Las seis etiquetas de esta variable son:

  Tabla 18: Caracterización variable "Vivía_enseñanza_media"

Variable Vivía_enseñanza_media	
0	Ambos padres
1	Solo madre
2	Solo padre
3	Familiares y otros
4	Internado

  Fuente: elaboración Propia.

- **Variable Vive_año_actual**

  Las seis etiquetas de esta variable son:

  Tabla 19: Caracterización variable “Vive_año_actual”

Variable Vive_año_actual	
0	Solo
1	Padres (uno o ambos)
2	Amistades
3	Aún no definido u otro
4	Familiares
5	Pareja

  Fuente: Elaboración propia

- **Variable Hijos**

  Las cinco etiquetas de esta variable son:

  Tabla 20: Caracterización variable "Hijos"

Variable Hijos	
0	Ninguno
1	1 hijo
2	2 hijos
3	3 hijos
4	4 hijos

  Fuente: Elaboración propia

- **Variable Trabajado_alguna_vez**

  Las dos etiquetas de esta variable son:

  Tabla 21: Caracterización variable "Trabajado_alguna_vez"

Variable Trabajado_alguna_vez	
Si	El alumno si ha trabajo alguna vez antes de ingresar a la universidad
No	El alumno no ha trabajo alguna vez antes de ingresar a la universidad

  Fuente: Elaboración propia

- **Variable Piensa_trabajar**

  Las dos etiquetas de esta variable son:

  Tabla 22: Caracterización variable “Piensa_trabajar”

Variable Piensa_trabajar	
Si	El alumno tiene pensado trabajar durante la educación superior
No	El alumno no tiene pensado trabajar durante la educación superior.

  Fuente: Elaboración propia

- **Variable Tipo_ingreso**

  Las cinco etiquetas de esta variable son:

  Tabla 22: Caracterización variable “Tipo_ingreso”

Variable “Tipo_ingreso”	
ART	Alumnos que ingresaron mediante ART
BEA	Alumnos que ingresaron mediante BEA
DEP	Alumnos que ingresaron mediante DEP
ETNIA	Alumnos que ingresaron mediante ETNIA
HF	Alumnos que ingresaron mediante HF
PSU	Alumnos que ingresaron mediante PSU
PSU_ESPECIAL	Alumnos que ingresaron mediante PSU_ESPECIAL
RUR	Alumnos que ingresaron mediante RUR

Fuente: Elaboración propia

- **Provincia_domicilio y Procedencia_colegio**

  Las etiquetas de estas variables se muestran a continuación:

Tabla 23: Caracterización variables "Provincia_domicilio" y "Procedencia_colegio"

Provincia	**Comunas**			
Antofagasta	Antofagasta			
Arauco	Arauco	Cañete	Contulmo	Curanilahue
	Lebu	Los Álamos	Tirúa	
Arica	Arica			
Aysén	Aysén			
Biobío	Alto Biobío	Antuco	Cabrero	Laja
	Los Ángeles	Mulchen	Nacimiento	Negrete
	Quilaco	Quilleco	San Rosendo	Santa Bárbara
	Tucapel	Yumbel		

Provincia	**Comunas**			
Cachapoal	Coltauco	Doñihue	Graneros	Las Cabras
	Machalí	Malloa	Mostazal	Peumo
	Pichidegua	Quinta de Tilcoco	Rancagua	Rengo
	San Vicente			
Cardenal Caro	Pichilemu			
Cauquenes	Cauquenes	Chanco	Pelluhue	
Cautín	Curarrehue	Freire	Loncoche	Padre Las Casas
	Pitrufquén	Temuco	Villarrica	Gorbea
	Nueva imperial	Pucón		
Chacabuco	Colina	Lampa		
Chiloé	Ancud	Castro	Dalcahue	Puqueldón
	Quellón	Quinchao		
Colchagua	Chimbarongo	Palmilla	Peralillo	San Fernando
	Santa Cruz			
Concepción	Chiguayante	Concepción	Coronel	Florida
	Hualpén	Hualqui	Lota	Penco
	San Pedro de la Paz	Santa Juana	Talcahuano	Tomé
Copiapó	Copiapó			
Cordillera	Puente alto			
Coyhaique	Coyhaique			
Curicó	Curicó	Hualañé	Molina	Romeral
	Teno			
El Loa	Calama			
Elqui	Coquimbo	La Serena	Vicuña	

Provincia	Comunas			
General Carrera	Chile Chico			
Iquique	Alto hospicio	Iquique		
Limarí	Punitaqui			
Linares	Colbún	Linares	Longaví	Parral
	San Javier	Retiro	Villa Alegre	Yerbas Buenas
Llanquihue	Calbuco	Puerto Montt	Puerto varas	
Los Andes	Los Andes			
Magallanes	Punta Arenas			
Maipo	Buin	Paine	San Bernardo	
Malleco	Angol	Curacautín	Los Sauces	Purén
	Renaico	Traiguén	Collipulli	Victoria
Marga Marga	Limache	Quilpué		
Ñuble	Bulnes	Chillan	Chillan Viejo	Cobquecura
	Colemu	Coihueco	El Carmen	Huape
	Ninhue	Ñiquén	Pemuco	Pinto
Ñuble	Portezuelo	Quillón	Quirihue	Ránquil
	San Carlos	San Fabián	San Ignacio	San Nicolás
	Trehuaco	Yungay	El Carmen	
Osorno	Purranque			
Quillota	Quillota			
Ranco	La unión			
San Felipe de Aconcagua	San Felipe			
Santiago	Conchalí	El Bosque	Estación Central	Huechuraba
	La Cisterna	La Florida	La Granja	La Reina

Provincia	**Comunas**			
	Las Condes	Lo Prado	Macul	Maipú
	Pedro Aguirre Cerda	Peñalolén	Pudahuel	Quilicura
	Quinta normal	Renca	San Joaquín	Santiago
	Cerrillos	Independencia	La Pintana	Lo Barnechea
	Ñuñoa	Providencia	San Miguel	San Ramón
	Vitacura			
Talagante	Peñaflor	Talagante	El Monte	
Talca	Constitución	Curepto	Maule	San Clemente
	Talca			
Tamarugal	Pozo Almonte			
Tierra del Fuego	Porvenir			
Valdivia	Lanco	Valdivia		
Valparaíso	Viña del Mar			

Fuente: Elaboración Propia

A continuación se detalla la transformación efectuada utilizando el algoritmo "Auto clustering" con el fin de agrupar datos semejantes y simplificar las variables del modelo:

- **Variable PSU_Lenguaje**

  Se generaron seis etiquetas, las cuales se pueden observar a continuación:

Tabla 24: Caracterización variable "PSU_Lenguaje"

PSU_Lenguaje	
1	Estudiantes con PSU Lenguaje entre 362 y 501
2	Estudiantes con PSU Lenguaje entre 502 y 520
3	Estudiantes con PSU Lenguaje entre 521 y 548
4	Estudiantes con PSU Lenguaje entre 549 y 597
5	Estudiantes con PSU Lenguaje entre 598 y 624
6	Estudiantes con PSU Lenguaje entre 625 y 811

Fuente: Elaboración propia.

- **Variable PSU_Matemática**

Se generaron seis etiquetas, las cuales se pueden observar a continuación:

Tabla 25: Caracterización variable "PSU_Matematica"

PSU_Matemática	
1	Estudiantes con PSU Matemática entre 330 y 535
2	Estudiantes con PSU Matemática entre 543 y 553
3	Estudiantes con PSU Matemática entre 554 y 560
4	Estudiantes con PSU Matemática entre 561 y 604
5	Estudiantes con PSU Matemática entre 605 y 624 o entre 536 y 542
6	Estudiantes con PSU Matemática entre 625 y 817

Fuente: Elaboración propia.

- **Variable Nem**

  Se generaron nueve etiquetas, las cuales se pueden observar a continuación:

  Tabla 26: Caracterización variable "Nem"

Variable Nem	
1	Estudiante con NEM entre 4.89 y 5.44
2	Estudiante con NEM entre 5.45 y 5.56
3	Estudiante con NEM entre 5.57 y 5.71
4	Estudiante con NEM entre 5.72 y 5.79
5	Estudiante con NEM entre 5.80 y 5.91
6	Estudiante con NEM entre 5.92 y 6.11
7	Estudiante con NEM entre 6.12 y 6.29
8	Estudiante con NEM entre 6.30 y 6.33
9	Estudiante con NEM entre 6.34 y 7.00

Fuente: Elaboración propia.

- **Variable Ranking**

  Se generaron seis etiquetas, las cuales se pueden observar a continuación:

Tabla 27: Caracterización variable "Ranking"

Variable Ranking	
1	Estudiantes con puntaje ranking entre 404 y 504
2	Estudiantes con puntaje ranking entre 505 y 568
3	Estudiantes con puntaje ranking entre 569 y 632
4	Estudiantes con puntaje ranking entre 633 y 681
5	Estudiantes con puntaje ranking entre 682 y 726
6	Estudiantes con puntaje ranking entre 727 y 850

Fuente: Elaboración propia.

- **Variable Carrera**

  Se generaron nueve etiquetas, las cuales se pueden observar a continuación:

Tabla 28: Caracterización variable “Carrera"

Agrupación	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Carreras	2924	2952	2972	2901	2927	2954	2959	2966	2921
	2926		2973	2904	2928	2955	2963	2967	
	2930		2976	2905	2929	2957	2964	2968	
	2932		2978	2910				2969	
	2935		2980	2915				2971	
	2937		2982	2918					
	2945		2986	2919					
	2949			2920					
	2951								

Fuente: Elaboración Propia

Los códigos descritos en la Tabla 28 corresponden a las siguientes carreras:

Tabla 29: Códigos por carrera

Código	**Carrera**	**Clúster**
2924	Ingeniería Civil en Industrias de la Madera	1
2926	Ingeniería Civil Mecánica	1
2930	Ingeniería de Ejecución en Electricidad	1
2932	Ingeniería de Ejecución Mecánica	1
2935	Ingeniería de Ejecución en Electrónica	1
2937	Ingeniería de Ejecución en Computación e Informática	1
2945	Contador Público y Auditor (Concepción)	1
2949	Ingeniería Comercial (Concepción)	1
2951	Ingeniería en Alimentos	1
2952	Enfermería	2

Código	Carrera	Clúster
2972	Pedagogía en Castellano y Comunicación	3
2973	Pedagogía en Ciencias Naturales con mención Biología, Física o Química	3
2976	Pedagogía en Inglés	3
2978	Pedagogía en Historia y Geografía	3
2980	Pedagogía en Educación Parvularia	3
2982	Pedagogía en Educación General Básica	3
2986	Pedagogía en Educación Matemática	3
2901	Arquitectura	4
2904	Diseño Industrial	4
2905	Ingeniería en Construcción	4
2910	Trabajo Social (Concepción)	4
2915	Bachillerato en Ciencias (Concepción)	4
2918	Ingeniería Estadística	4
2919	Ingeniería Civil Química	4
2920	Ingeniería Civil Industrial	4
2927	Ingeniería Civil en Informática (Concepción)	5
2928	Ingeniería Civil en Automatización	5
2929	Ingeniería Civil Eléctrica	5
2954	Nutrición y Dietética	6
2955	Fonoaudiología	6
2957	Ingeniería Civil en Informática (Chillán)	6
2959	Ingeniería Comercial (Chillán)	7
2963	Contador Público y Auditor (Chillán)	7
2964	Diseño Gráfico	7
2966	Trabajo Social (Chillán)	8
2967	Psicología	8

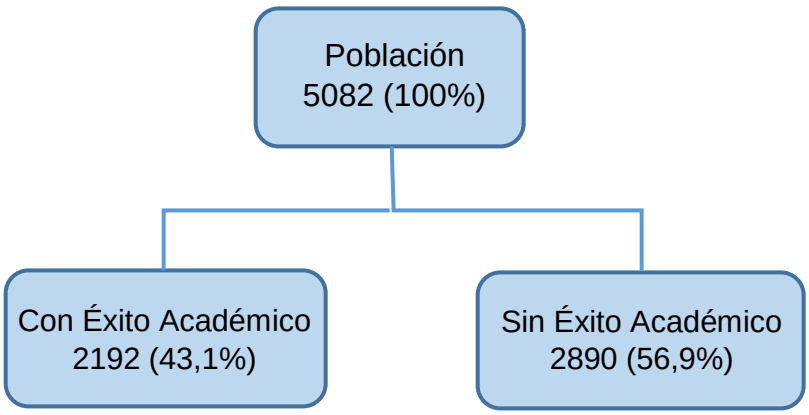
Código	Carrera	Clúster
2968	Bachillerato en Ciencias (Chillán)	8
2969	Ingeniería en Recursos Naturales	8
2971	Pedagogía en Educación Física	8
2921	Ingeniería Civil	9

Fuente: Elaboración propia.

4.4. Minería de datos

Para el presente estudio, y luego de finalizar las etapas de pre-procesamiento y tratamiento de datos, se cuenta con la información presentada en la figura 7:

Figura 7: Población total del conjunto de datos y su desempeño

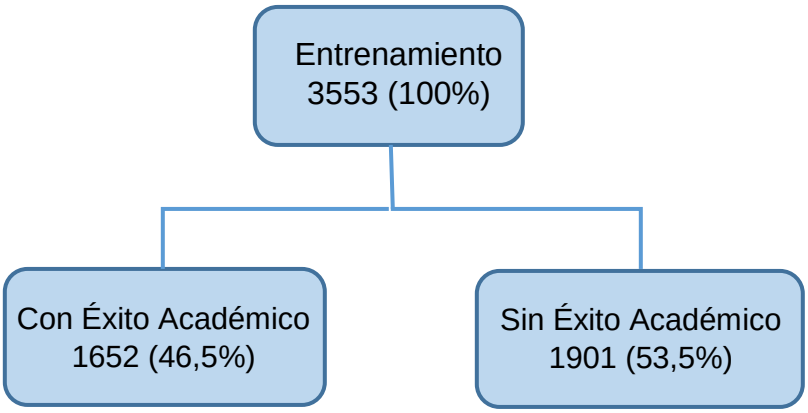


Fuente: Elaboración propia a partir de datos de DARCA

En la Figura 7 se muestra la población de estudiantes con la cual se trabajará en el estudio finalizadas las etapas de limpieza y transformación. Además se muestra la cantidad de alumnos que sí tuvieron éxito académico y los que no lo tuvieron.

Por otra parte, la población de estudiantes fue dividida en dos sub-conjuntos: el primero llamado "entrenamiento, el cual abarca el 70% de la población mientras que el segundo llamado "validación" abarca el 30% restante. Lo anteriormente descrito se puede observar en la Figura 8 y 9, respectivamente.

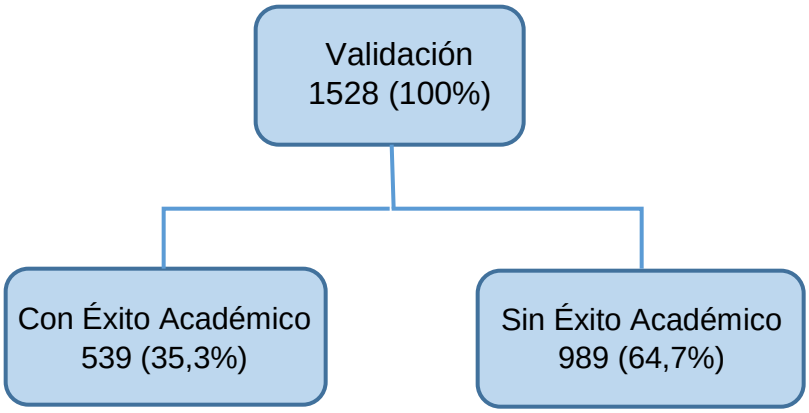
Figura 8: Población del entrenamiento y su desempeño



Fuente: Elaboración propia.

En la Figura 8 se muestra la cantidad de datos que fueron seleccionados para el entrenamiento del modelo, además se muestra la cantidad de alumnos que sí tuvieron éxito académico y los que no lo tuvieron.

Figura 9: Población de la validación y su desempeño



Fuente: Elaboración propia.

En la Figura 9 se muestra la cantidad de datos que fueron seleccionados para la validación del modelo, además se muestra la cantidad de alumnos que sí tuvieron éxito académico y los que no lo tuvieron.

4.4.1. Selección de variables

Para seleccionar las variables que serán incluidas en los diferentes modelos a generar, se utilizó el poder de predicción (KI) de cada una de estas. El poder de predicción (KI) representa la

capacidad de que una variable sola pueda predecir la variable de destino. Cabe mencionar que esta información se obtiene desde la opción "informes estadísticos" y generada automáticamente por el software SAP Predictive Analytics en la interfaz de Automated Analytics.

El poder de predicción de cada variable se muestra a continuación:

Tabla 30: Poder predictivo (KI) de cada variable

Variable	**KI**
Carrera	0,3506
Nem	0,2052
Ranking	0,1891
Género	0,1818
Procedencia_colegio	0,1721
Provincia_domicilio	0,1550
PSU_Lenguaje	0,1165
Edad	0,0858
Renovación_curricular	0,0615
Tipo_colegio	0,0603
Piensa_trabajar	0,0375
Vivía_enseñanza_media	0,0350
Gratuidad	0,0273
Becas_internas	0,0269
PSU_Matemática	0,0238
Tipo_enseñanza	0,0213
Orden_postulación	0
BEA	0
Discapacidad	0
Vive_año_actual	0

Hijos	0
Trabajado_alguna_vez	0
Tipo_ingreso	0

Fuente: Elaboración propia.

Basándose en la información de la tabla anterior, se puede observar que las primeras siete variables contienen el mayor poder de predicción (KI), las cuales se consideraron para obtener un conjunto fijo de variables que todos los modelos deberán tener como base.

Este conjunto fijo está compuesto por las siguientes variables:

- Carrera
- Nem
- Ranking
- Género
- Procedencia_colegio
- Provincia_domicilio
- PSU_Lenguaje

Además, se consideraron las próximas cuatro variables con mayor poder de predicción (KI), las cuales se usaron para conjugar con las variables base:

- Edad
- Renovación_curricular
- Tipo_colegio
- Piensa_trabajar

De esta forma se generó 4 modelos experimentales compuestos por 10 variables.

Modelo 1: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, Renovación_curricular, Tipo_colegio.

Modelo 2: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, Renovación_curricular, Piensa_trabajar.

Modelo 3: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, Tipo_colegio, Piensa_trabajar.

Modelo 4: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Renovación_curricular, Piensa_trabajar, Tipo_colegio.

4.5. Interpretación/Evaluación de resultados

En esta sección se muestra la manera en que el algoritmo crea el árbol de decisión para cada uno de los 10 modelos experimentales y algunas de las reglas más interesantes de cada modelo. Al final de esta sección se muestran algunas observaciones encontradas al interpretar los resultados.

Cabe mencionar que para crear el árbol de decisión, el algoritmo sigue la siguiente estructura:

Número Nodo, variable, condición, [número de variables que entraron al nodo], (probabilidad de "No" probabilidad de "Si").

4.5.1. Modelo 1

El primer modelo contempla las variables: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, Renovación_curricular y Tipo_colegio. La representación de este modelo mediante el algoritmo R-CNR Tree se muestra a continuación:

```
1) Raíz [3051] No (0.52802360 0.47197640)
  2) Carrera < 1.5 [703]  No (0.71692745 0.28307255)
    4) Nem < 7.5 [618]  No (0.74595469 0.25404531)
      8) Provincia_domicilio = Arauco, Biobío, Cachapoal, Cauquenes, Chiloé, Concepción,
                               Curicó, Elqui, Iquique, Limarí, Linares, Malleco, Ñuble, Santiago,
                               Talca [613]  No (0.75203915 0.24796085)
       16) Procedencia_colegio = Arauco, Arica, Biobío, Cachapoal, Cauquenes, Chiloé, Elqui,
                                 Limarí, Malleco, Santiago, Talca [122]
                                 No (0.86885246 0.13114754) *
       17) Procedencia_colegio = Concepción, Curicó, Iquique, Linares, Ñuble [491]
                                 No (0.72301426 0.27698574)
         34) Provincia_domicilio = Concepción, Curicó, Iquique, Linares, Ñuble, Santiago [475]
```

No (0.74105263 0.25894737) *
35) Provincia_domicilio = Biobío [16] Si (0.18750000 0.81250000) *
9) Provincia_domicilio = Llanquihue, Maipo, Osorno, Quillota [5] Si (0.0000000 1.0000000) *
5) Nem >= 7.5 [85] No (0.50588235 0.49411765)
10) Procedencia_colegio = Biobío, Colchagua, Malleco 15 No (0.93333333 0.06666667) *
11) Procedencia_colegio = Arauco, Concepción, Ñuble, Viña del Mar [70]
Si (0.41428571 0.58571429)
22) Renovacion_curricular = "Con" 56 Si (0.48214286 0.51785714)
44) Edad >= 19.5 [4] No (1.00000000 0.00000000) *
45) Edad < 19.5 [52] Si (0.44230769 0.55769231)
90) PSU_Lenguaje < 3.5 [26] No (0.57692308 0.42307692) *
91) PSU_Lenguaje >=3.5 [26] Si (0.30769231 0.69230769) *
23) Renovacion_curricular = "Sin" [14] Si (0.14285714 0.85714286) *
3) Carrera >=1.5 [2348] Si (0.47146508 0.52853492)
6) Carrera >=3.5 [1867] No (0.51312266 0.48687734)
12) Procedencia_colegio = Biobío, Colchagua, Los Andes, San Felipe de Aconcagua,
Talagante, Tierra del Fuego, Valparaíso [90]
No (0.85555556 0.14444444) *
13) Procedencia_colegio = Antofagasta, Arauco, Cachapoal, Cardenal Caro, Cauquenes,
Cautín, Chacabuco, Chiloé, Coyhaique, Concepción, Copiapó,
Curicó, Elqui, General Carrera, Linares, Llanquihue, Magallanes,
Malleco, Ñuble, Quillota, Santiago, Talca, Valdivia [1777]
Si (0.49577940 0.50422060)
26) Provincia_domicilio = Arauco, Cachapoal, Cauquenes, Cautín, Chacabuco,
Concepción, Copiapó, Curicó, Linares, Llanquihue, Magallanes,
Malleco, Ñuble, Talca [1701] No (0.51205173 0.48794827)
52) Nem < 5.5 [602] No (0.59800664 0.40199336)
104) Carrera < 7.5 [448] No (0.64508929 0.35491071) *
105) Carrera >= 7.5 [154] Si (0.46103896 0.53896104) *
53) Nem >= 5.5 [1099] Si (0.46496815 0.53503185)
106) Edad >= 18.5 [419] No (0.55131265 0.44868735) *
107) Edad < 18.5 [680] Si (0.41176471 0.58823529) *
27) Provincia_domicilio = Antofagasta, Biobío, Cardenal Caro, Chiloé, Elqui, General
Carrera, Marga Marga, Santiago, Valdivia [76]
Si (0.13157895 0.86842105) *
7) Carrera < 3.5 [481] Si (0.30977131 0.69022869)
14) Edad >=18.5 [213] Si (0.44600939 0.55399061)
28) Provincia_domicilio = Biobío, Cachapoal, Colchagua, Curicó, El Loa, Ñuble, Santiago
[188] Si (0.46808511 0.53191489)

```
      56) Genero = “M” [63] No (0.58730159 0.41269841)
        112) PSU_Lenguaje >= 1.5 [57] No (0.63157895 0.36842105) *
        113) PSU_Lenguaje < 1.5 [6] Si (0.16666667 0.83333333) *
      57) Genero = “F” [125] Si (0.40800000 0.59200000)
        114) Ranking >= 5.5 [32] No (0.56250000 0.43750000) *
        115) Ranking < 5.5 [93] Si (0.35483871 0.64516129) *
    29) Provincia_domicilio = Cauquenes, Concepción, Linares, Talca [25]
                                Si (0.28000000 0.72000000) *
  15) Edad < 18.5 [268] Si (0.20149254 0.79850746)
    30) Nem < 5.5 [114] Si (0.32456140 0.67543860)
      60) Ranking < 1.5 [11] No (0.63636364 0.36363636) *
      61) Ranking >= 1.5 [103] Si (0.29126214 0.70873786) *
    31) Nem >=5.5 [154] Si (0.11038961 0.88961039) *
```

En la representación anterior, se puede observar que el algoritmo descartó la variable “Tipo_colegio” debido a que no la consideraba importante para generar el árbol de decisión. Por otra parte, en la representación se puede interpretar las siguientes reglas de inducción:

- ***SI*** *Carrera < 1.5* ***Y*** *Nem <7.5* ***Y*** *(Provincia_domicilio = Maipo* ***OR*** *Provincia_domicilio = Llanquihue* ***OR*** *Provincia_domicilio = Osorno* ***OR*** *Provincia_domicilio = Quillota)* ***ENTONCES*** *Éxito Académico*
- ***SI*** *Carrera < 1.5* ***Y*** *Nem >= 7.5* ***Y*** *(Procedencia_colegio = Arauco* ***OR*** *Procedencia_colegio = Concepción* ***OR*** *Procedencia_colegio = Ñuble* ***OR*** *Procedencia_colegio = Viña del Mar)* ***Y*** *Renovacion_curricular = Con* ***Y*** *Edad >= 19.5* ***ENTONCES*** *Sin Éxito Académico*

4.5.2. Modelo 2

El segundo modelo consideraba las variables: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, Renovación_curricular, Piensa_trabajar. La representación de este modelo mediante el algoritmo R-CNR Tree se muestra en el anexo A. En dicha representación, se puede observar que el algoritmo descartó la variable “Piensa_trabajar” debido a que no la consideraba importante para generar el árbol de

decisión. Por otra parte, en la representación se puede interpretar las siguientes reglas de inducción:

- ***SI*** *Carrera < 1.5* ***Y*** *Nem <7.5* ***Y*** *(Provincia_domicilio = Maipo* ***OR*** *Provincia_domicilio = Llanquihue* ***OR*** *Provincia_domicilio = Osorno* ***OR*** *Provincia_domicilio = Quillota)* ***ENTONCES*** *Éxito Académico*
- ***SI*** *Carrera < 1.5* ***Y*** *Nem >= 7.5* ***Y*** *(Procedencia_colegio = Arauco* ***OR*** *Procedencia_colegio = Concepción* ***OR*** *Procedencia_colegio = Ñuble* ***OR*** *Procedencia_colegio = Viña del Mar)* ***Y*** *Renovacion_curricular = Con* ***Y*** *Edad >= 19.5* ***ENTONCES*** *Sin Éxito Académico*

4.5.3. Modelo 3

El Modelo 3 consideraba las variables: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Edad, Tipo_colegio y Piensa_trabajar. La representación de este modelo mediante el algoritmo R-CNR Tree se muestra en el anexo B. En dicha representación, se puede observar que el algoritmo descartó las variables "Tipo_colegio" y "Piensa_trabajar" debido a que no las consideraba importantes para generar el árbol de decisión. Por otra parte, en la representación se puede interpretar las siguientes reglas de inducción:

- ***SI*** *Carrera < 1.5* ***Y*** *Nem <7.5* ***Y*** *(Provincia_domicilio = Maipo* ***OR*** *Provincia_domicilio = Llanquihue* ***OR*** *Provincia_domicilio = Osorno* ***OR*** *Provincia_domicilio = Quillota)* ***ENTONCES*** *Éxito Académico*
- ***SI*** *Carrera < 1.5* ***Y*** *Nem >= 7.5* ***Y*** *(Procedencia_colegio = Arauco* ***OR*** *Procedencia_colegio = Concepción* ***OR*** *Procedencia_colegio = Ñuble* ***OR*** *Procedencia_colegio = Viña del Mar)* ***Y*** *(Provincia_domicilio = Arauco* ***OR*** *Provincia_domicilio = Concepción* ***OR*** *Provincia_domicilio = Ñuble)* ***Y*** *PSU_Lenguaje >= 3.5* ***Y*** *Edad >=19.5* ***ENTONCES*** *Sin Éxito Académico*

4.5.4. Modelo 4

El Modelo 4 consideraba las variables: Carrera, Nem, Ranking, Género, Procedencia_colegio, Provincia_domicilio, PSU_Lenguaje, Renovación_curricular, Piensa_trabajar y Tipo_colegio. La representación de este modelo mediante el algoritmo R-CNR Tree se muestra en el anexo C. En dicha representación, se puede observar que el algoritmo descartó las variables "PSU_Lenguaje", "Género", "Piensa_trabajar" y "Renovación_curricular" debido a que no las consideraba importantes para generar el árbol de decisión. Por otra parte, en la representación se puede interpretar las siguientes reglas de inducción:

- ***SI*** *Carrera >= 1.5* ***Y*** *Carrera < 3.5* ***Y*** *Ranking >= 1.5* ***Y*** *Tipo_colegio < 2.5* ***Y*** *(Procedencia_colegio = Biobío* ***OR*** *Procedencia_colegio = Cachapoal* ***OR*** *Procedencia_colegio = Colchagua* ***OR*** *Procedencia_colegio = Curicó* ***OR*** *Procedencia_colegio = Santiago)* ***Y*** *Ranking >= 3.5* ***ENTONCES*** *Éxito Académico*
- ***SI*** *Carrera < 1.5* ***Y*** *Nem <7.5* ***Y*** *(Provincia_domicilio = Maipo* ***OR*** *Provincia_domicilio = Llanquihue* ***OR*** *Provincia_domicilio = Osorno* ***OR*** *Provincia_domicilio = Quillota)* ***ENTONCES*** *Éxito Académico*

A continuación, se describe las variables utilizadas en los 4 modelos experimentales.

Tabla 31: Descripción de las variables de los modelos experimentales

Variable	**Tipo**	**Media**	**Desviación Estándar**
Edad	Numérico	18,9811	2,09319
Tipo_colegio	Numérico	2,3674	0,54136
PSU_Lenguaje	Numérico	3,4630	1,66606
Carrera	Numérico	4,2485	2,49830
Ranking	Numérico	3,7393	1,65824
Nem	Numérico	5,4370	2,58058
		N	**%**
Género	Nominal		
• F		2222	43.7
• M		2860	53.3
Piensa_trabajar	Nominal		
• Si		1647	32.4
• No		3435	67.6
Renovación_curricular	Nominal		
• Con		3907	76.9
• Sin		1175	23.1
Provincia_ domicilio	Nominal		
• Antofagasta		1	0.0
• Arauco		206	4.1
• Biobío		378	7.4
• Cachapoal		32	0.6
• Cardenal Caro		1	0.0
• Cauquenes		43	0.8
• Cautín		12	0.2

• Chacabuco		4	0.1
• Chiloé		11	0.2
• Colchagua		12	0.2
• Concepción		1782	35.1
• Copiapó		2	0.0
• Cordillera		7	0.1
• Coyhaique		3	0.1
• Curicó		32	0.6
• El Loa		1	0.0
• Elqui		4	0.1
• General Carrera		1	0.0
• Iquique		3	0.1
• Limarí		1	0.0
• Linares		207	4.1
• Llanquihue		19	0.4
• Los Andes		1	0.0
• Magallanes		4	0.1
• Maipo		2	0.0
• Malleco		35	0.7
• Marga Marga		2	0.0
• Ñuble		2206	43.4
• Osorno		1	0.0
• Quillota		1	0.0
• Ranco		1	0.0
• San Felipe de Aconcagua		1	0.0
• Santiago		32	0.6

• Talagante		2	0.0
• Talca		27	0.5
• Valdivia		2	0.0
• Valparaíso		3	0.1
Procedencia_colegio	Nominal		
• Antofagasta		2	0.0
• Arauco		206	4.1
• Arica		1	0.0
• Biobío		286	5.6
• Cachapoal		34	0.7
• Cardenal Caro		1	0.0
• Cauquenes		42	0.8
• Cautín		16	0.3
• Chacabuco		3	0.1
• Chiloé		12	0.2
• Colchagua		12	0.2
• Concepción		1869	36.8
• Copiapó		3	0.1
• Cordillera		4	0.1
• Coyhaique		4	0.1
• Curicó		32	0.6
• El Loa		1	0.0
• Elqui		5	0.1
• General Carrera		1	0.0
• Iquique		4	0.1
• Limarí		1	0.1
• Linares		208	4.1

• Llanquihue		20	0.4
• Los Andes		1	0.0
• Magallanes		1	0.0
• Maipo		3	0.1
• Malleco		42	0.8
• Ñuble		2185	43.0
• Osorno		1	0.0
• Quillota		2	0.0
• Ranco		1	0.0
• San Felipe de Aconcagua		2	0.0
• Santiago		40	0.8
• Talagante		3	0.1
• Talca		26	0.5
• Tierra del Fuego		1	0.0
• Última Esperanza		1	0.0
• Valdivia		2	0.0
• Valparaíso		3	0.1
• Viña del Mar		1	0.0
Éxito_academico	Nominal		
• Si		2192	43.1
• No		2890	56.9

Fuente: Elaboración propia.

A continuación se muestran los resultados que se obtuvieron al aplicar el algoritmo R-CNR Tree a cada uno de los 4 modelos experimentales descritos anteriormente.

4.6. Entrenamiento

La primera parte del estudio consiste en entrenar los 4 modelos experimentales con 3553 datos, correspondientes al 70% de la población total (5082). Cada modelo cuenta con 10 variables. Los resultados del entrenamiento se muestran en la Tabla 32.

Tabla 32: Resultados entrenamiento

	Precisión	**Falsos/negativos**	**Falsos/positivos**	**Variables Descartadas**
Modelo 1	67 %	550	462	1
Modelo 2	67 %	550	462	1
Modelo 3	67 %	549	460	2
Modelo 4	66 %	561	488	4

Fuente: Elaboración propia

La Tabla 32 está compuesta por cuatro columnas: precisión, falsos/negativos, falsos/positivos y variables descartadas. La **precisión** de un modelo se refiere a la capacidad de obtener los mismos resultados en otros datos; los **falsos/negativos** indican que un alumno tuvo éxito académico, pero la herramienta predijo que no lo tuvo; **falsos/positivos** se refiere a que el alumno no tuvo éxito académico, pero la herramienta predijo que si lo tuvo; **variables descartadas corresponde** al número de variables que el algoritmo excluyó para generar el árbol de decisión.

Se puede observar que la precisión de los 4 modelos oscila entre el 66% y 67% siendo los modelos 1, 2 y 3 los que lograron un mayor porcentaje de precisión. En cuanto a los falsos/negativos el modelo 3 es el que obtuvo un menor número de asignaciones, lo siguió los modelos 1 y 2 con 550, y finalmente el modelo 4, el cual es el que posee un mayor número de

asignaciones (561). En los falsos/positivos se encontró que el modelo 3 es el que consiguió un menor número de asignaciones, lo siguió el modelo 1 y 2 con 462, y finalmente el modelo 4 con 488. Por último, en la columna "variables descartadas" se encontró que los modelos 1 y 2 descartaron una variable para generar el árbol de decisión, el modelo 3 descartó dos mientras que el modelo 4 excluyó a dos variables.

4.7. Validación

La segunda parte del estudio consiste en validar los resultados obtenidos por los 4 modelos experimentales. En la validación, se utiliza el 30% (1528) de la población total (5082). Los resultados de la validación se muestran en la Tabla 33.

Tabla 33: Resultados validación

	Precisión	**Falsos/negativos**	**Falsos/positivos**	**Variables Descartadas**
Modelo 1	71 %	272	117	2
Modelo 2	71 %	272	117	2
Modelo 3	71 %	282	114	3
Modelo 4	71 %	266	133	2

Fuente: Elaboración propia.

En la Tabla 33 se puede observar que todos los modelos experimentales obtuvieron la misma precisión, 71%. En los falsos/negativos el modelo 4 es el que logró un menor número de asignaciones (266), lo siguió los modelos 1 y 2 con 272, y finalmente el modelo 3, el cual es el que obtuvo un mayor número de asignaciones (282). En cuanto a los falsos/positivos se encontró que el modelo 3 fue el que logró un menor número de asignaciones, lo siguió los modelos 1 y 2 con 117, y finalmente el modelo 4 con 133. Finalmente, en la columna "variables descartadas" se encontró que los modelos 1, 2 y 4 descartaron 2 variables para generar el árbol de decisión mientras que el modelo 3 excluyó a 3.

CAPÍTULO V

DISCUSIÓN

De los 4 modelos generados anteriormente se puede observar lo siguiente:

- Tres de los 4 modelos experimentales conservó las 7 variables base para generar el árbol de decisión
- Ninguno de los 4 modelos experimentales utilizó sus 10 variables para generar el árbol de decisión.
- El "Género" y "PSU_Lenguaje" fueron las únicas variables base descartada por el algoritmo R-CNR Tree, ambas variables fueron excluidas en el modelo 4.
- Entre las variables a conjugar, "Piensa_trabajar" es la que más veces fue descartada por el algoritmo.
- La "Edad" fue la única variable a conjugar que no fue descartada por el algoritmo.
- El modelo 4 fue el que más variables descartó para generar el árbol de decisión.

Si bien, todos los modelos entregaron resultados muy similares, se consideró al modelo 3 como el mejor de todos debido a los siguientes puntos:

- Obtuvo una de las mayores precisiones tanto en el entrenamiento como en la validación lo que significa que puede identificar de manera correcta si un estudiante tendrá o no éxito académico en otros conjuntos de datos.
- En el conjunto de entrenamiento obtuvo menor número de falsos/negativos y falsos/positivos, y en la validación, logró un mejor número de falsos/positivos. Esto permite que el modelo tenga un menor porcentaje de error para clasificar incorrectamente a un estudiante.
- Fue uno de los modelos que más descartó variables para generar el árbol de decisión, lo cual permite que sea más simple de interpretar y en caso de ser implementado en algún sistema permitirá ahorrar tiempo y recursos.

5.1. Interpretación del árbol de decisión

A continuación, se muestra el árbol de decisión generado por el modelo 3 (Figura 10). Cabe mencionar que cada nodo del árbol tiene tres segmentos. En la parte superior se encuentra el "nombre" del nodo y a su izquierda la etiqueta de la variable que predomina en el nodo. En el centro encontramos dos gráficos, el de la izquierda corresponde al porcentaje que no tuvo éxito académico mientras que en el de la derecha muestra el porcentaje que si lo tuvo. En la parte inferior, se muestra el porcentaje de la población total que cumple la condición del nodo.

Más adelante se describe la forma de interpretar el árbol de decisión.

Figura 10: Árbol de decisión del modelo 3

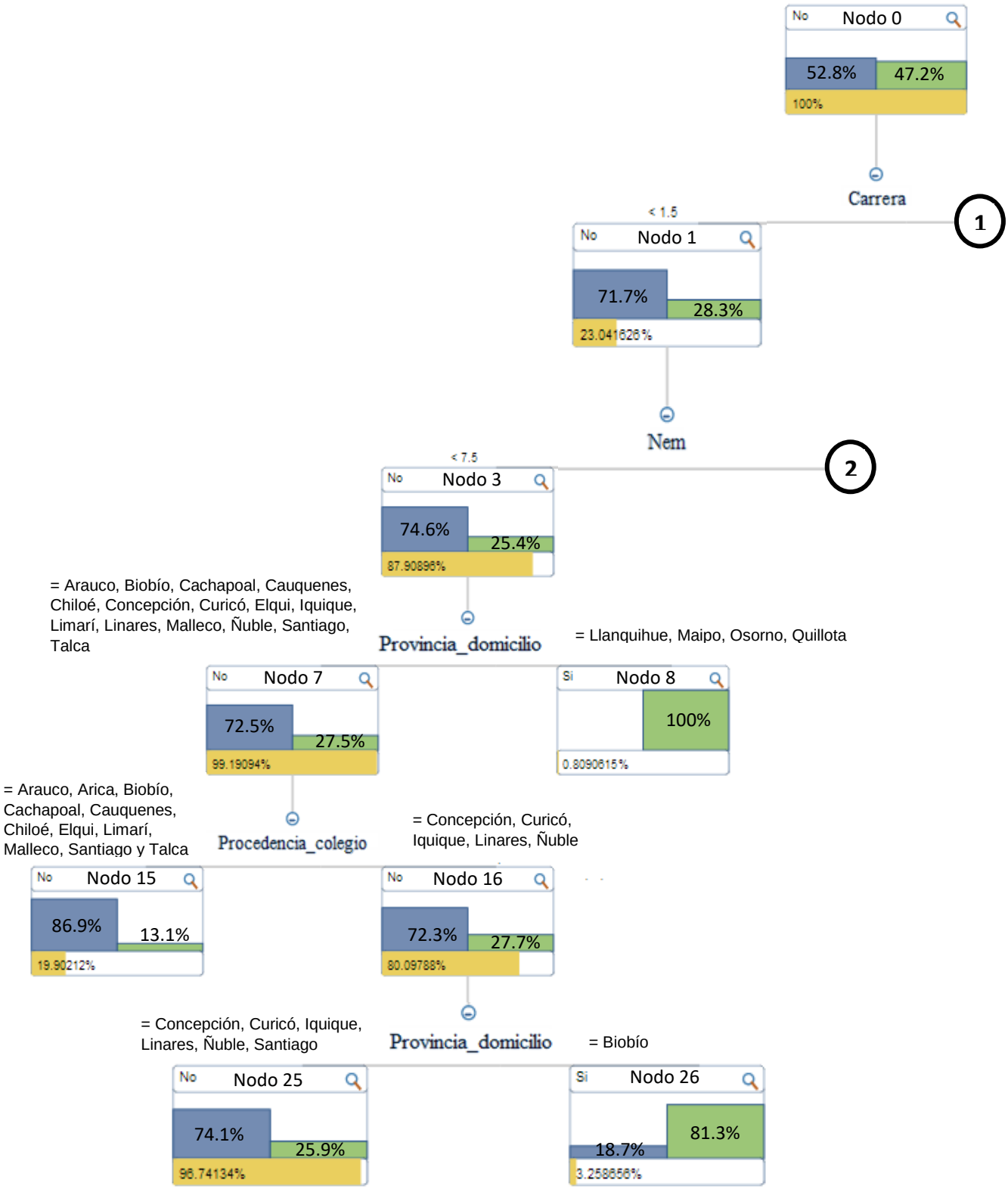
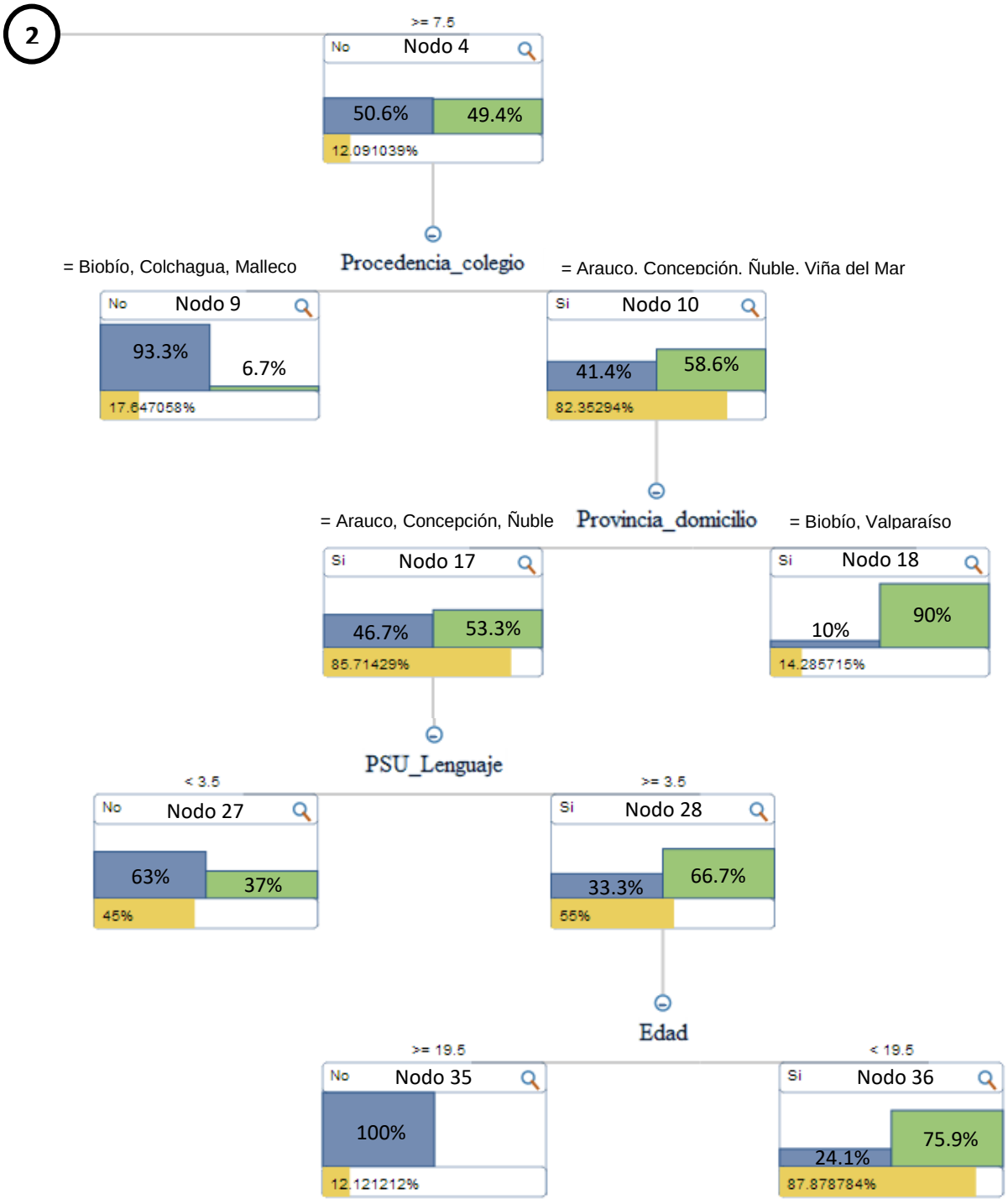
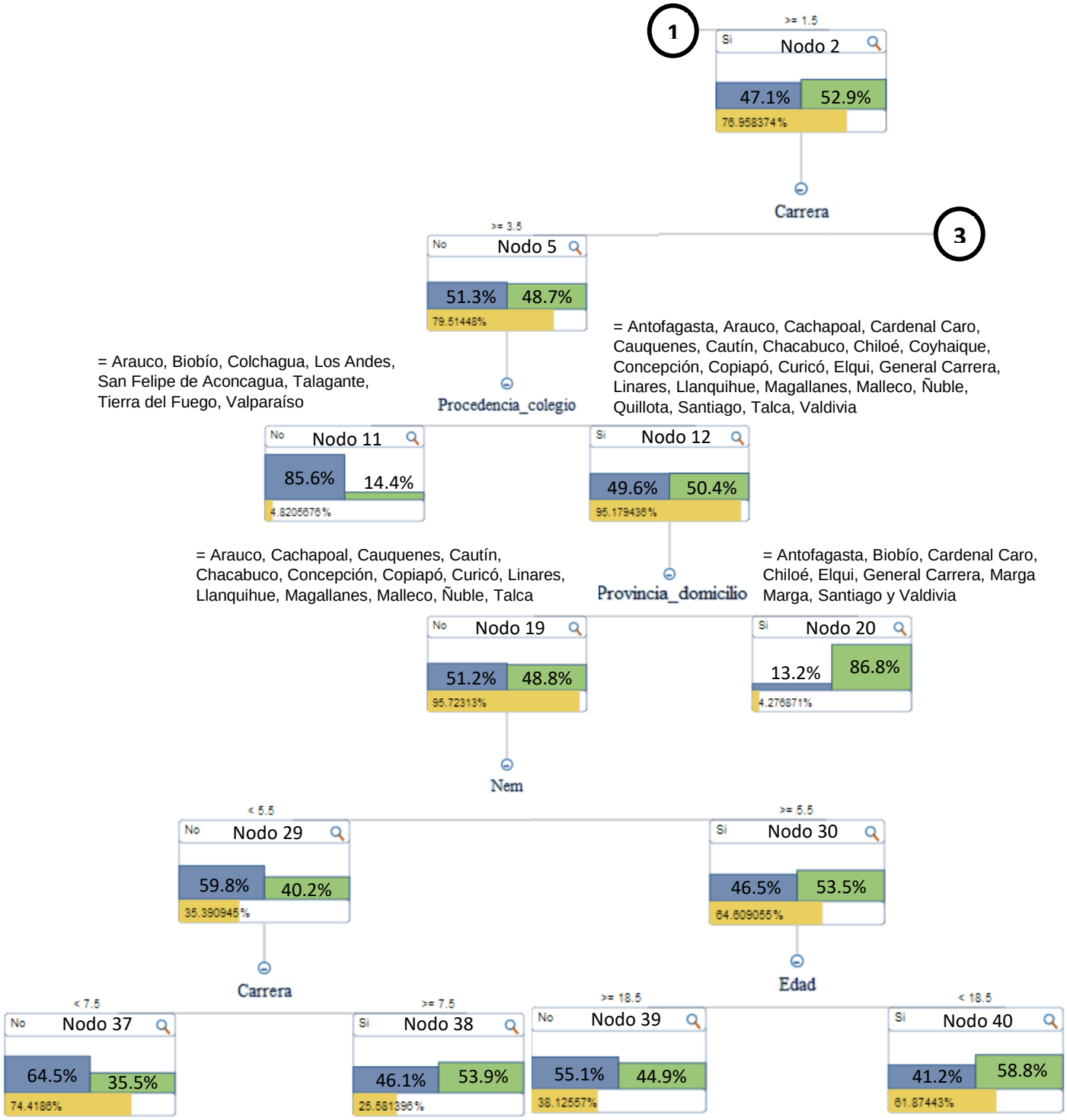


Figura 10: Árbol de decisión del modelo 3 (continuación)



Fuente: Elaboración propia.

Figura 10: Árbol de decisión del modelo 3 (continuación)



Fuente: Elaboración propia.

Figura 10: Árbol de decisión del modelo 3 (continuación)

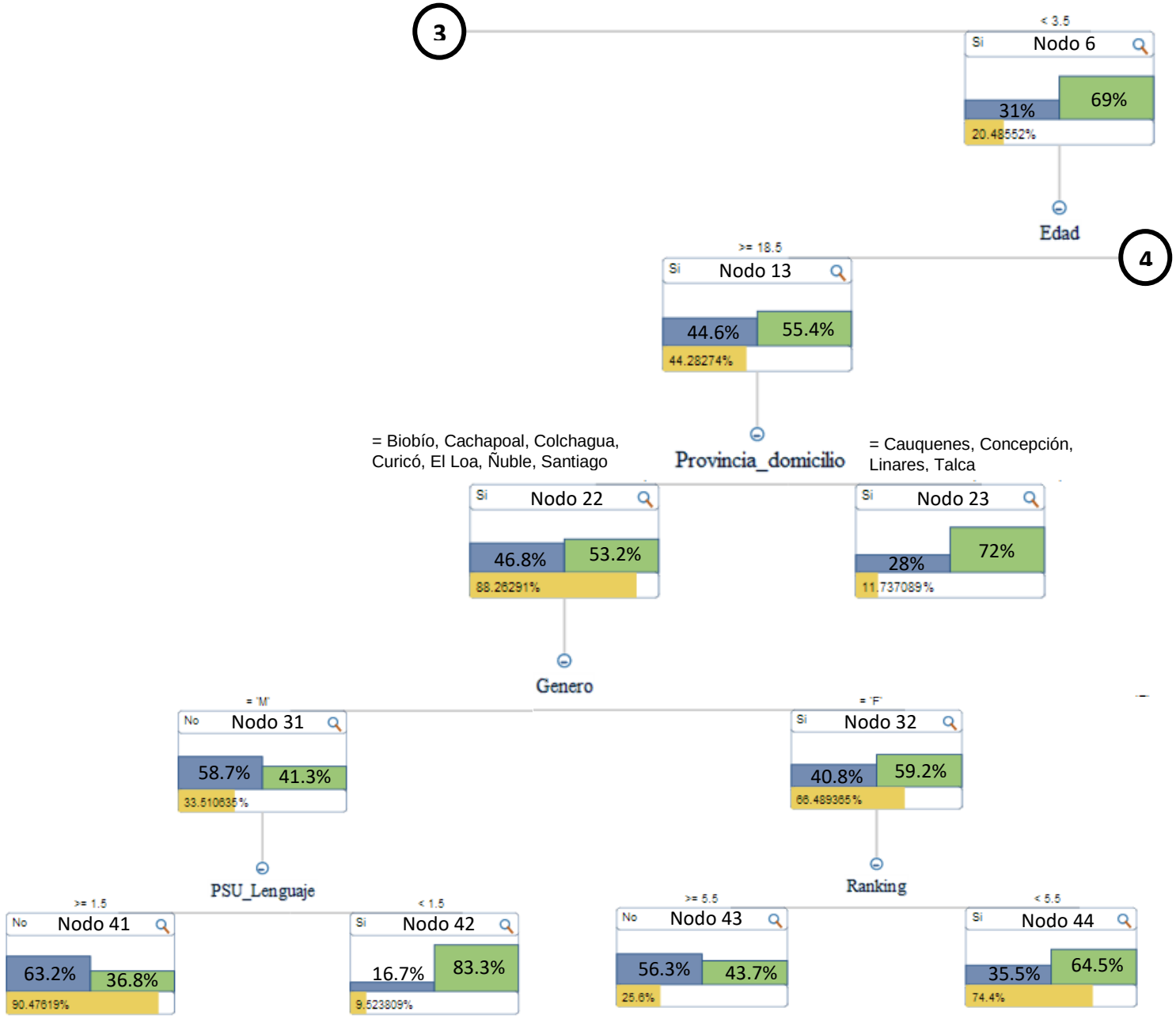
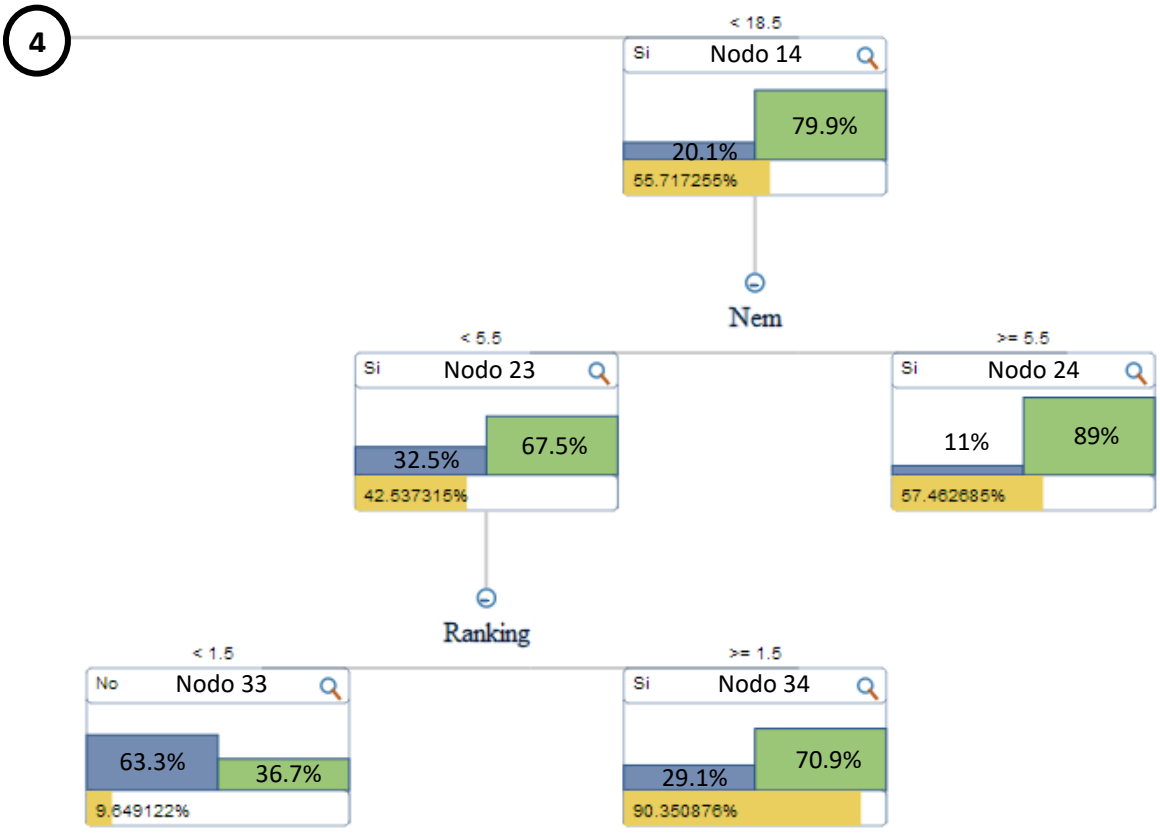


Figura 10: Árbol de decisión del modelo 3 (continuación)



Fuente: Elaboración propia.

A continuación se muestra la forma de interpretar el árbol de decisión:

1. En primer lugar, tenemos el nodo 0 (cero) el cual describe la variable dependiente: porcentaje de estudiantes que tienen éxito académico (1440 = 47.2%) y de los que no tienen éxito académico (1611 = 52.8%).
2. Seguidamente se observa que la variable dependiente se ramifica en dos nodos: Nodo 1 y 2 que pertenecen a la variable "Carrera", lo que indica que esta es la variable principal predictora.
3. El nodo 1 nos describe que el 71.7% de los estudiantes que se matriculan en las carreras que se encuentran en la agrupación 1, no tiene éxito académico.
4. El nodo 1 se ramifica en los nodos 3 y 4 pertenecientes a la variable "Nem". Se observa en el nodo 3 que los alumnos cuya nota Nem se encuentra dentro de los primeras siete agrupaciones, el 74.6% no tiene éxito académico mientras que en el nodo 4, el cual contiene el resto de agrupaciones, el 50.6% de los alumnos no tiene éxito académico.
5. El nodo 3 se ramifica en los nodos 7 y 8, los cuales pertenecen a la variable "Provincia_domicilio". En el nodo 7 encontramos a los estudiantes cuyo domicilio se encuentra en las provincias de: Arauco, Biobío, Cachapoal, Cauquenes, Chiloé, Concepción, Curicó, Elqui, Iquique, Limarí, Linares, Malleco, Ñuble, Santiago y Talca, y el 75.2% de estos no tiene éxito académico. Por otra parte, si la provincia de domicilio de los estudiantes es: Llanquihue, Maipo, Osorno o Quillota (nodo 8), el 100% tiene éxito académico.
6. El nodo 7 se vuelve a ramificar en los nodos 15 y 16 pertenecientes a la variable "Procedencia_colegio". En el nodo 15 se observa que los estudiantes que provienen de colegios de: Arauco, Arica, Biobío, Cachapoal, Cauquenes, Chiloé, Elqui, Limarí, Malleco, Santiago y Talca, el 86.9% no tiene éxito académico. Por otra parte, si los estudiantes provienen de colegios de: Concepción, Curicó, Iquique, Linares y Ñuble (nodo 16), el 72.3% tiene éxito académico
7. El nodo 16 se ramifica en los nodos 25 y 26 que pertenecen a la variable "Provincia_domicilio". En el nodo 25 encontramos a los estudiantes cuyo domicilio se encuentra en las provincias de: Concepción, Curicó, Iquique, Linares, Ñuble y Santiago,

y el 74.1% de estos no tiene éxito académico. En el nodo 26, se encuentra la provincia del Biobío (nodo 26), y donde el 81.3% de os estudiantes tiene éxito académico.

8. El nodo 4 se ramifica en los nodos 9 y 10, los cuales pertenecen a la variable "Procedencia_colegio". El nodo 9 contiene las provincias de: Biobío, Colchagua y Malleco, y el 93.3% de los estudiantes no tiene éxito académico. En el nodo 10 se contiene a las provincias de: Arauco, Concepción, Ñuble y Viña del Mar, y los estudiantes que provienen de estas provincias el 58.6% tiene éxito académico.
9. El nodo 10 se ramifica en los nodos 17 y 18 pertenecientes a la variable "Provincia_domicilio". En el nodo 17 encontramos a los estudiantes cuyo domicilio se encuentra en las provincias de: Arauco, Concepción y Ñuble, y el 53.3% de estos tiene éxito académico. Por otra parte, si la provincia de domicilio de los estudiantes es Biobío o Valparaíso (nodo 18), el 90% tiene éxito académico.
10. El nodo 17 se vuelve a ramificar en los nodos 27 y 28 que pertenecen a la variable "PSU_Lenguaje". En el nodo 27, el 63% de los estudiantes no tiene éxito académico si su puntaje se encuentra dentro de las primeras tres agrupaciones mientras que en el nodo 28, el cual contiene el resto de agrupaciones, el 66.7% de los estudiantes tiene éxito académico.
11. El nodo 28 se ramifica en los nodos 35 y 36 pertenecientes a la variable "Edad". Cuando la edad es mayor o igual a 19.5 años (nodo 25), el 100% de los estudiantes no tiene éxito. En cambio, cuando la edad es menor a 19.5 años, el 75.9% tiene éxito académico.
12. El nodo 2 nos indica que el 52.9% de los estudiantes que se matriculan en las carreras que no pertenecen a la agrupación 1, si tiene éxito académico.
13. El nodo 2 se ramifica en los nodos 5 y 6 pertenecientes a la variable "Carrera". En el nodo 5 se encuentran las carreras que no pertenecen a las primeras tres agrupaciones, y el 51.3% de los estudiantes no tiene éxito académico mientras que el nodo 6, el cual contiene las agrupaciones 4, 5, 6, 7, 8 y 9, el 69% de los estudiantes si tuvo éxito académico.
14. El nodo 5 se ramifica en los nodos 11 y 12 que pertenecen a la variable "Procedencia_colegio". En el nodo 11 se encuentran las provincias de: Biobío,

Colchagua, Los Andes, San Felipe de Aconcagua, Talagante, Tierra del Fuego y Valparaíso, y el 85.6% de los estudiantes que provienen de estas provincias no tiene éxito académico. Por otra parte, en el nodo 12 se encuentran las provincias de: Antofagasta, Arauco, Cachapoal, Cardenal Caro, Cauquenes, Cautín, Chacabuco, Chiloé, Coyhaique, Concepción, Copiapó, Curicó, Elqui, General Carrera, Linares, Llanquihue, Magallanes, Malleco, Ñuble, Quillota, Santiago, Talca y Valdivia, y el 50.4% de los estudiantes si tienen éxito académico.

15. El nodo 12 se ramifica en los nodos 19 y 20 pertenecientes a la variable "Provincia_domicilio". El nodo 19 se encuentran las provincias de: Arauco, Cachapoal, Cauquenes, Cautín, Chacabuco, Concepción, Copiapó, Curicó, Linares, Llanquihue, Magallanes, Malleco, Ñuble y Talca, y el 51.2% de los estudiantes que provienen de estas provincias no tiene éxito académico. Por otro lado, en el nodo 20 se encuentran las provincias de: Antofagasta, Biobío, Cardenal Caro, Chiloé, Elqui, General Carrera, Marga Marga, Santiago y Valdivia, y el 86.8% de los estudiantes que provienen de estas provincias si tienen éxito académico.
16. El nodo 19 se ramifica en los nodos 29 y 30, los cuales pertenecen a la variable "Nem". En el nodo 29 se encuentran las notas que pertenecen a las primeras 5 agrupaciones, ahí el 59.8% de los estudiantes no tiene éxito académico mientras que en el nodo 30 donde se encuentra el resto de las agrupaciones, el 53.5% de los estudiantes si tiene éxito académico.
17. El nodo 29 se vuelve a ramificar en los nodos 37 y 38 pertenecientes a la variable "Carrera". Los estudiantes que se matriculan en carreras que se encuentran dentro de las primeras siete agrupaciones (nodo 37), un 64.5% no tiene éxito académico mientras que el nodo 38 contiene el resto de agrupaciones de las carreras, ahí el 53.9% de los estudiantes si tiene éxito académico.
18. El nodo 30 también se vuelve a ramificar en los nodos 39 y 40 que pertenecen a la variable "Edad". Cuando la edad es mayor o igual a 18.5 años (nodo 39), el 55.1% de los estudiantes no tiene éxito. En cambio, cuando la edad es menor a 18.5 años, el 58.8% tiene éxito académico.

19. El nodo 6 se ramifica en los nodos 13 y 14 pertenecientes a la variable "Edad". Si la edad es mayor o igual a 18.5 años (nodo 13), un 55.4% tiene éxito académico. En cambio, si la edad es menor a 18.5 años (nodo 14), el 79.9% tiene éxito.
20. El nodo 13 se ramifica en los nodos 21 y 22, los cuales pertenecen a la variable "Provincia_domicilio". Los estudiantes que provienen de las provincias de: Biobío, Cachapoal, Colchagua, Curicó, El Loa, Ñuble o Santiago (Nodo 21), el 53.2% tiene éxito académico, en cambio, los estudiantes que provienen de las provincias de: Cauquenes, Concepción, Linares o Talca, el 72% si tiene éxito.
21. El nodo 21 se vuelve a ramificar en los nodos 31 y 32 pertenecientes a la variable "Género". Si el género es masculino, un 58.7% no tiene éxito, en cambio cuando el género es femenino, el 59.3% tiene éxito académico.
22. El nodo 31 se ramifica en los nodos 41 y 42 pertenecientes a la variable "PSU_Lenguaje". Si el puntaje se encuentra dentro de la primera agrupación (nodo 42), el 83.3% tiene éxito académico, en caso contrario, si el puntaje no se encuentra dentro de esta agrupación, el 63.2% de los estudiantes no tiene éxito académico.
23. El nodo 32 se ramifica en los nodos 43 y 44 pertenecientes a la variable "Ranking". Si el puntaje ranking se encuentra dentro de las primeras cinco agrupaciones, el 64.5% tiene éxito. En cambio, si el puntaje no se encuentra dentro de estas agrupaciones, el 56.3% no tiene éxito.
24. El nodo 14 se vuelve a ramificar en los nodos 23 y 24 pertenecientes a la variable "Nem". Si la nota nem se encuentra dentro de las primera cinco agrupaciones (nodo 23), un 67.5% tiene éxito. Por otra parte, si la nota se encuentra en el resto de agrupaciones el 89% tiene éxito.
25. Finalmente el nodo 23 se ramifica en los nodos 33 y 34 pertenecientes a la variable "Ranking". El 63.6% de los estudiantes cuyo puntaje se encuentra dentro de la primera agrupación (nodo 33), no tiene éxito académico. Por otra parte, si el puntaje no se encuentra dentro de esta agrupación, el 70.9% de los estudiantes tiene éxito académico.

Similar a lo propuesto por Tejedor (2003) y Shahiri *et al*. (2015), el modelo de predicción obtenido en esta investigación arrojó que las variables Nem, Género, Procedencia colegio, Provincia domicilio, Puntaje PSU Lenguaje y Edad son importantes para la predicción del rendimiento académico de los estudiantes.

Por otro lado, en este estudio surgieron como importantes variables la Carrera y el Ranking. La primera, y de acuerdo a los estudios previos analizados, no se consideró como una variable que permite predecir el rendimiento académico de los estudiantes. La segunda, a diferencia de Barahona (2016), no fue encontrada significativa.

Las Universidades podrían enfocarse en recolectar datos asociados a los estudiantes (Carrera, Ranking, Nem, Género, Procedencia colegio, Provincia domicilio, Puntaje PSU Lenguaje y Edad) para predecir el éxito académico de primer año y crear planes e intervenciones para aumentar el rendimiento para aquellos estudiantes que tengan una baja tasa de éxito.

CAPITULO VI: CONCLUSIONES

Para este estudio se utilizaron 5082 datos de estudiantes regulares que ingresaron a la universidad entre los años 2014-2016. De los 5082 datos, el 70% fue utilizado para entrenar los modelos mientras que el 30% fue usado para validarlo. Cada uno estos datos contenían 34 variables de las cuales se seleccionaron 23; 16 de ellas fueron elegidas basándose en estudios previos y el resto (descartando la variable de destino) fueron incluidas para analizar si realmente influyen en el éxito académico de un alumno, estas variables son: Carrera, Renovación curricular, Tipo de ingreso, Hijos, Vivía enseñanza media, Trabajado_alguna_vez y Gratuidad.

Mediante el poder de predicción (KI) de cada variable, el cual se puede obtener en la interfaz Automated Analytics del software SAP Predictive Analytics, se logró obtener que de las 23 variables, 18 de ellas tenían la capacidad de predecir la variable de destino. Con estas 18 variables se construyó un conjunto base y uno de conjugación para generar distintos modelos. El primer conjunto abarcó las primeras 7 variables mientras que el segundo las 4 siguientes con mayor poder de precisión.

Se generaron 4 modelos experimentales con 10 variables cada uno, aunque ninguno de ellos obtuvo una gran precisión en el conjunto de entrenamiento y validación. Finalmente se seleccionó el modelo 3 como el mejor debido a que obtuvo una mayor precisión en el entrenamiento y validación, logró mejores números de falsos/negativos y falsos/positivos, y fue uno de los que más variables descarto para generar el árbol de decisión, lo cual permite que el modelo sea más fácil de interpretar.

A pesar de que el modelo 3 arrojó una precisión de 71%, se puede obtener información interesante a través del árbol de decisión generado por el algoritmo:

- La carrera en la que se matricula un estudiante es la principal variable predictora del éxito académico.

- Los estudiantes cuya nota Nem se encuentra dentro de las agrupaciones 8 y 9 (nota de 6,3 a 7) tienen un mayor porcentaje de lograr el éxito académico. Por otro lado, los estudiantes cuya nota Nem se encuentra dentro de las agrupaciones 1 al 7 (nota de 4,6 a 6.29) tienen un menor porcentaje de aprobar todas las asignaturas.
- El 100% de los estudiantes tuvo éxito académico cuando su provincia de domicilio fue: Llanquihue, Maipo, Osorno o Quillota.
- Entre un 64% a 70% de los estudiantes tuvo éxito cuando su puntaje de Ranking era superior a 504 (dentro de las agrupaciones 2, 3, 4 o 5).
- Los estudiantes más jóvenes tienen un mayor porcentaje de aprobar todas las asignaturas de primer año.
- Los hombres tienen un mayor porcentaje (58.7%) de no tener éxito académico que las mujeres (40.8%).
- En cuanto al puntaje PSU Lenguaje, se encontró que los estudiantes con mejor puntaje tienen mayor probabilidad de no tener éxito académico (agrupaciones 1, 2 y 3).
- Los estudiantes cuyos colegios se encontraba en las provincias de: Arauco, Concepción, Ñuble o Viña del Mar son los que tuvieron mayor porcentaje de éxito académico.

Para futuras investigaciones se podrían crear los mismos modelos, pero utilizando distintos algoritmos disponibles en la herramienta y comparar los resultados obtenidos. Por otra parte, se podría incluir variables emocionales y de asistencia a clases las cuales no fueron contempladas en este estudio, y así averiguar si estas influyen en el rendimiento académico de un estudiante. Además, se podría utilizar un mayor conjunto de datos y otro software predictivo (por ejemplo, WEKA) para analizar si entrega resultados similares a los obtenidos en este estudio.

Bibliografía

Avendaño, C., Gutiérrez, K., Salgado, C., & Dos-Santos, M. (2016). Rendimiento Académico en Estudiantes de Ingeniería Comercial: Modelo por Competencias y Factores de Influencia. *Formación Universitaria*, *9*(3), 03-10. http://dx.doi.org/10.4067/s0718-50062016000300002

Ayuda automática para Automated Analytics. SAP Help Portal: Extraído de https://help.sap.com/viewer/769493aa825848758ab80f7754ea34a0/3.0/es-ES/ el 18 de Febrero de 2018

Ayuda online de Expert Analytics. SAP Help Portal: Extraído de https://help.sap.com/viewer/94dbf2ba9d4047618880187451c3b253/3.3/es-ES el 18 de Febrero de 2018

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. *Procedia Computer Science*, *82*, 80-89. http://dx.doi.org/10.1016/j.procs.2016.04.012

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining,* 1(1), 3-16

Barahona Urbina, Planck, Veres Ferrer, Ernesto, & Aliaga Prieto, Verónica. (2016). Deserción académica de la Universidad de Atacama, Chile. *Comuni@cción*, *7*(2), 27-37. Recuperado en 20 de febrero de 2018, de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2219-71682016000200003&lng=es&tlng=es.

Berlanga Silvente, V., Rubio Hurtado, M. J., Vilà Baños, R. (2013). Cómo aplicar árboles de decisión en SPSS. [En línea] REIRE, Revista d'Innovació i Recerca en Educació, 6 (1), 65-79. Accesible en: http://www.ub.edu/ice/reire.htm

Bernstein, A., Provost, F., & Hill, S. (2005). Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 17(4), 503-518

Conceptos de minería de datos. (2017). Docs.microsoft.com: Extraído de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts el 10 de Febrero de 2018

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17, 37-54.

Gamarra, C., Guerrero, J., & Montero, E. (2016). A knowledge discovery in databases approach for industrial microgrid planning. *Renewable And Sustainable Energy Reviews, 60,* 615-630. http://dx.doi.org/10.1016/j.rser.2016.01.091

González-Ruiz, S.L., Gómez-Gallego, I., Pastrana-Brincones, J.L., & Hernández-Mendo, A.. (2015). Algoritmos de clasificación y redes neuronales en la observación automatizada de registros. *Cuadernos de Psicología del Deporte*, *15*(1), 31-40. https://dx.doi.org/10.4321/S1578-84232015000100003

Gutiérrez-Soto, C., Oliva P., & Paredes A. (2008). Una aplicación de Minería de Datos en la educación superior. *JCCC'08.*

Hernández Orallo, J., Ramírez Quintana, M., & Ferri Ramírez, C. (2007). *Introducción a la minería de datos*. Madrid: Pearson.

Matheus, C., Chan, P., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 5(6), 903-913

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of classification techniques for predicting the performance of students academic environment. *2014 International Conference On Communication And Network Technologies*. http://dx.doi.org/10.1109/cnt.2014.7062736

Miranda, M., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación Universitaria*, *10*(3), 61-68. http://dx.doi.org/10.4067/s0718-50062017000300007

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining Students' Data for Prediction Performance. *2014 Fourth International Conference On Advanced Computing & Communication Technologies*. http://dx.doi.org/10.1109/acct.2014.105

Natek, S., & Zwilling, M. (2018). *Student data mining solution–knowledge management system related to higher education institutions*. Expert Systems With Applications, 1-8. http://dx.doi.org/10.1016/j.eswa.2014.04.024

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems With Applications*, *36*(2), 2592-2602. http://dx.doi.org/10.1016/j.eswa.2008.02.021

Overview of Automated Analytics SAP Help Portal: Extraído de https://help.sap.com/viewer/bc031e667eea409c9e08e7ab8b1e4c70/3.3/en-US/b474299e88ed46e0820c6d9d471e48a3.html el 17 de Febrero de 2018

Pal, S. (2012). Mining Educational Data to Reduce Dropout Rates of Engineering Students. *International Journal Of Information Engineering And Electronic Business*, *4*(2), 1-7. http://dx.doi.org/10.5815/ijieeb.2012.02.01

Quadril, M., & Kalyankar, N. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 10(2), 2-5

Rokach, L., & Maimon, O. (2015). Data Mining with Decision Trees: Theory and 2nd ed. World Scientific.

Rosas, M, Chacín, F, García, J, Ascanio, M, & Cobo, M. (2006). Modelos de regresión lineal múltiple en presencia de variables cuantitativas y cualitativas para predecir el rendimiento estudiantil. *Revista de la Facultad de Agronomía*, *23*(2), 197-214. Recuperado en 01 de marzo de 2018, de http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0378-78182006000200007&lng=es&tlng=es

Salinas Oviedo, D., Hernández Castillo de Tejeda, A., & Barboza-Palomino, M. (2018). Condición de becario y rendimiento académico en estudiantes de una universidad peruana. *Revista Electrónica De Investigación Educativa*, *19*(4), 124. http://dx.doi.org/10.24320/redie.2017.19.4.1348

SAP Predictive Analytics. Extraído de https://help.sap.com/viewer/product/SAP_PREDICTIVE_ANALYTICS/3.0/es-ES el 15 de Febrero de 2018

SAP Predictive Analytics Part 2: An Overview of the Expert Analytics Tool. Extraído de https://sapexperts.wispubs.com/bi/articles/sap-predictive-analytics-part-2-an-overview-of-the-expert-analytics-tool?id=9ea49d471cf44be29ea7283a669d7be9#.Wt5RY8gvzIW el 17 de Febrero de 2018

Shahiri, A., Husain, W., & Rashid, N. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414-422. http://dx.doi.org/10.1016/j.procs.2015.12.157

Soman, K.P., Diwakar, S., & Ajay, V. (2006). *Data Mining: Theory and practice*. PHI Learning. Recuperado de: https://books.google.cl/books?id=TmvWI-b77AIC

Tejedor, F. (2003). Poder explicativo de algunos determinantes del rendimiento en los estudios universitarios. *Revista Española de Pedagogía*, 61(224), 5-32

Timarán Pereira, S., Hernández Arteaga, I., Caicedo Zambrano, S., Hidalgo Troya, A., & Alvarado Pérez, J. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. http://dx.doi.org/10.16925/9789587600490

Anexos

Anexo A: Representación modelo 2 mediante el algoritmo R-CNR Tree

```
1) Raíz [3051] No (0.52802360 0.47197640)
  2) Carrera < 1.5 [703] No (0.71692745 0.28307255)
    4) Nem < 7.5 [618] No (0.74595469 0.25404531)
      8) Provincia_domicilio = Arauco, Biobío, Cachapoal, Cauquenes, Chiloé, Concepción,
                               Curicó, Elqui, Iquique, Limarí, Linares, Malleco, Ñuble, Santiago,
                               Talca [613] No (0.75203915 0.24796085)
       16) Procedencia_colegio = Arauco, Arica, Biobío, Cachapoal, Cauquenes, Chiloé, Elqui,
                                 Limarí, Malleco, Santiago, Talca [122]
                                 No (0.86885246 0.13114754) *
       17) Procedencia_colegio = Concepción, Curicó, Iquique, Linares, Ñuble [491]
                                 No (0.72301426 0.27698574)
         34) Provincia_domicilio = Concepción, Curicó, Iquique, Linares, Ñuble, Santiago [475]
                                   No (0.74105263 0.25894737) *
         35) Provincia_domicilio = Biobío [16] Si (0.18750000 0.81250000) *
      9) Provincia_domicilio = Llanquihue, Maipo, Osorno, Quillota [5] Si (0.0000000 1.0000000) *
    5) Nem >= 7.5 [85] No (0.50588235 0.49411765)
      10) Procedencia_colegio = Biobío, Colchagua, Malleco [15] No (0.93333333 0.06666667) *
      11) Procedencia_colegio = Arauco, Concepción, Ñuble, Viña del Mar [70]
                                Si (0.41428571 0.58571429)
        22) Renovacion_curricular = “Con” 56 [Si] (0.48214286 0.51785714)
          44) Edad >= 19.5 [4] No (1.00000000 0.00000000) *
          45) Edad < 19.5 [52] Si (0.44230769 0.55769231)
            90) PSU_Lenguaje < 3.5 [26] No (0.57692308 0.42307692) *
            91) PSU_Lenguaje >= 3.5 [26] Si (0.30769231 0.69230769) *
        23) Renovacion_curricular = “Sin” [14] Si (0.14285714 0.85714286) *
  3) Carrera >= 1.5 [2348] Si (0.47146508 0.52853492)
    6) Carrera >= 3.5 [1867] No (0.51312266 0.48687734)
      12) Procedencia_colegio = Biobío, Colchagua, Los Andes, San Felipe de Aconcagua,
                                Talagante, Tierra del Fuego, Valparaíso [90]
                                No (0.85555556 0.14444444) *
      13) Procedencia_colegio = Antofagasta, Arauco, Cachapoal, Cardenal Caro, Cauquenes,
                                Cautín, Chacabuco, Chiloé, Coyhaique, Concepción, Copiapó,
                                Curicó, Elqui, General Carrera, Linares, Llanquihue, Magallanes,
                                Malleco, Ñuble, Quillota, Santiago, Talca, Valdivia [1777]
```

Si (0.49577940 0.50422060)

26) Provincia_domicilio = Arauco, Cachapoal, Cauquenes, Cautín, Chacabuco, Concepción, Copiapó, Curicó, Linares, Llanquihue, Magallanes, Malleco, Ñuble, Talca [1701] No (0.51205173 0.48794827)

52) Nem < 5.5 [602] No (0.59800664 0.40199336)

104) Carrera < 7.5 [448] No (0.64508929 0.35491071) *

105) Carrera >= 7.5 [154] Si (0.46103896 0.53896104) *

53) Nem >= 5.5 [1099] Si (0.46496815 0.53503185)

106) Edad >= 18.5 [419] No (0.55131265 0.44868735) *

107) Edad < 18.5 [680] Si (0.41176471 0.58823529) *

27) Provincia_domicilio = Antofagasta, Biobío, Cardenal Caro, Chiloé, Elqui, General Carrera, Marga Marga, Santiago, Valdivia [76] Si (0.13157895 0.86842105) *

7) Carrera < 3.5 [481] Si (0.30977131 0.69022869)

14) Edad >= 18.5 [213] Si (0.44600939 0.55399061)

28) Provincia_domicilio = Biobío, Cachapoal, Colchagua, Curicó, El Loa, Ñuble, Santiago [188] Si (0.46808511 0.53191489)

56) Genero = “M” [63] No (0.58730159 0.41269841)

112) PSU_Lenguaje >= 1.5 [57] No (0.63157895 0.36842105) *

113) PSU_Lenguaje < 1.5 [6] Si (0.16666667 0.83333333) *

57) Genero = “F” [125] Si (0.40800000 0.59200000)

114) Ranking >= 5.5 [32] No (0.56250000 0.43750000) *

115) Ranking < 5.5 [93] Si (0.35483871 0.64516129) *

29) Provincia_domicilio = Cauquenes, Concepción, Linares, Talca [25] Si (0.28000000 0.72000000) *

15) Edad < 18.5 [268] Si (0.20149254 0.79850746)

30) Nem < 5.5 [114] Si (0.32456140 0.67543860)

60) Ranking < 1.5 [11] No (0.63636364 0.36363636) *

61) Ranking >= 1.5 [103] Si (0.29126214 0.70873786) *

31) Nem >= 5.5 [154] Si (0.11038961 0.88961039) *

Anexo B: Representación modelo 3 mediante el algoritmo R-CNR Tree

```
 1) Raíz [3051] No (0.52802360 0.47197640)
 2) Carrera < 1.5 [703] No (0.71692745 0.28307255)
  4) Nem < 7.5 [618] No (0.74595469 0.25404531)
    8) Provincia_domicilio = Arauco, Biobío, Cachapoal, Cauquenes, Chiloé, Concepción,
                             Curicó, Elqui, Iquique, Limarí, Linares, Malleco, Ñuble, Santiago,
                             Talca [613] No (0.75203915 0.24796085)
     16) Procedencia_colegio = Arauco, Arica, Biobío, Cachapoal, Cauquenes, Chiloé, Elqui,
                               Limarí, Malleco, Santiago, Talca [122]
                               No (0.86885246 0.13114754) *
     17) Procedencia_colegio = Concepción, Curicó, Iquique, Linares, Ñuble [491]
                               No (0.72301426 0.27698574)
       34) Provincia_domicilio = Concepción, Curicó, Iquique, Linares, Ñuble, Santiago [475]
                                 No (0.74105263 0.25894737) *
       35) Provincia_domicilio = Biobío [16] Si (0.18750000 0.81250000) *
    9) Provincia_domicilio = Llanquihue, Maipo, Osorno, Quillota [5] Si (0.0000000 1.0000000) *
  5) Nem >= 7.5 [85] No (0.50588235 0.49411765)
   10) Procedencia_colegio = Biobío, Colchagua, Malleco [15] No (0.93333333 0.06666667) *
   11) Procedencia_colegio = Arauco, Concepción, Ñuble, Viña del Mar [70]
                             Si (0.41428571 0.58571429)
     22) Provincia_domicilio = Arauco, Concepción, Ñuble [60] Si (0.46666667 0.53333333)
       44) PSU_Lenguaje < 3.5 [27] No (0.62962963 0.37037037) *
       45) PSU_Lenguaje >= 3.5 [33] Si (0.33333333 0.66666667)
         90) Edad >= 19.5 [4] No (1.00000000 0.00000000) *
         91) Edad < 19.5 [29] Si (0.24137931 0.75862069) *
     23) Provincia_domicilio = Biobío, Valparaíso [10] Si (0.10000000 0.90000000) *
 3) Carrera >= 1.5 [2348] Si (0.47146508 0.52853492)
  6) Carrera >= 3.5 [1867] No (0.51312266 0.48687734)
   12) Procedencia_colegio = Biobío, Colchagua, Los Andes, San Felipe de Aconcagua,
                             Talagante, Tierra del Fuego, Valparaíso [90]
                             No (0.85555556 0.14444444) *
   13) Procedencia_colegio = Antofagasta, Arauco, Cachapoal, Cardenal Caro, Cauquenes,
                             Cautín, Chacabuco, Chiloé, Coyhaique, Concepción, Copiapó,
                             Curicó, Elqui, General Carrera, Linares, Llanquihue, Magallanes,
                             Malleco, Ñuble, Quillota, Santiago, Talca, Valdivia [1777]
                             Si (0.49577940 0.50422060)
     26) Provincia_domicilio = Arauco, Cachapoal, Cauquenes, Cautín, Chacabuco, Concepción,
```

```
                        Copiapó, Curicó, Linares, Llanquihue, Magallanes, Malleco,
                        Ñuble, Talca [1701] No (0.51205173 0.48794827)
    52) Nem < 5.5 [602] No (0.59800664 0.40199336)
     104) Carrera < 7.5 [448] No (0.64508929 0.35491071) *
     105) Carrera >= 7.5 [154] Si (0.46103896 0.53896104) *
    53) Nem >= 5.5 [1099] Si (0.46496815 0.53503185)
     106) Edad >= 18.5 [419] No (0.55131265 0.44868735) *
     107) Edad < 18.5 [680] Si (0.41176471 0.58823529) *
   27) Provincia_domicilio = Antofagasta, Biobío, Cardenal Caro, Chiloé, Elqui, General
                        Carrera, Marga Marga, Santiago, Valdivia [76]
                        Si (0.13157895 0.86842105) *
 7) Carrera < 3.5 [481] Si (0.30977131 0.69022869)
  14) Edad >= 18.5 [213] Si (0.44600939 0.55399061)
   28) Provincia_domicilio = Biobío, Cachapoal, Colchagua, Curicó, El Loa, Ñuble, Santiago
                        [188] Si (0.46808511 0.53191489)
    56) Genero = "M" [63] No (0.58730159 0.41269841)
     112) PSU_Lenguaje >= 1.5 [57] No (0.63157895 0.36842105) *
     113) PSU_Lenguaje < 1.5 [6] Si (0.16666667 0.83333333) *
    57) Genero = "F" [125] Si (0.40800000 0.59200000)
     114) Ranking >= 5.5 [32] No (0.56250000 0.43750000) *
     115) Ranking < 5.5 [93] Si (0.35483871 0.64516129) *
   29) Provincia_domicilio = Cauquenes, Concepción, Linares, Talca [25]
                        Si (0.28000000 0.72000000) *
  15) Edad < 18.5 [268] Si (0.20149254 0.79850746)
   30) Nem < 5.5 [114] Si (0.32456140 0.67543860)
    60) Ranking < 1.5 [11] No (0.63636364 0.36363636) *
    61) Ranking >= 1.5 [103] Si (0.29126214 0.70873786) *
   31) Nem >= 5.5 [154] Si (0.11038961 0.88961039) *
```

Anexo C: Representación modelo 4 mediante el algoritmo R-CNR Tree

1) Raíz [3051] No (0.52802360 0.47197640)
2) Carrera < 1.5 [703] No (0.71692745 0.28307255)
4) Nem < 7.5 [618] No (0.74595469 0.25404531)
8) Provincia_domicilio = Arauco, Biobío, Cachapoal, Cauquenes, Chiloé, Concepción, Curicó, Elqui, Iquique, Limarí, Linares, Malleco, Ñuble, Santiago, Talca [613] No (0.75203915 0.24796085)
16) Procedencia_colegio = Arauco, Arica, Biobío, Cachapoal, Cauquenes, Chiloé, Elqui, Limarí, Malleco, Santiago, Talca [122] No (0.86885246 0.13114754) *
17) Procedencia_colegio = Concepción, Curicó, Iquique, Linares, Ñuble [491] No (0.72301426 0.27698574)
34) Provincia_domicilio = Concepción, Curicó, Iquique, Linares, Ñuble, Santiago [475] No (0.74105263 0.25894737) *
35) Provincia_domicilio = Biobío [16] Si (0.18750000 0.81250000) *
9) Provincia_domicilio = Llanquihue, Maipo, Osorno, Quillota [5] Si (0.0000000 1.0000000) *
5) Nem >= 7.5 [85] No (0.50588235 0.49411765)
10) Procedencia_colegio = Biobío, Colchagua, Malleco [15] No (0.93333333 0.06666667) *
11) Procedencia_colegio = Arauco, Concepción, Ñuble, Viña del Mar [70] Si (0.41428571 0.58571429) *
3) Carrera >= 1.5 [2348] Si (0.47146508 0.52853492)
6) Carrera >= 3.5 [1867] No (0.51312266 0.48687734)
12) Procedencia_colegio = Biobío, Colchagua, Los Andes, San Felipe de Aconcagua, Talagante, Tierra del Fuego, Valparaíso [90] No (0.85555556 0.14444444) *
13) Procedencia_colegio = Antofagasta, Arauco, Cachapoal, Cardenal Caro, Cauquenes, Cautín, Chacabuco, Chiloé, Coyhaique, Concepción, Copiapó, Curicó, Elqui, General Carrera, Linares, Llanquihue, Magallanes, Malleco, Ñuble, Quillota, Santiago, Talca, Valdivia [1777] Si (0.49577940 0.50422060)
26) Provincia_domicilio = Arauco, Cachapoal, Cauquenes, Cautín, Chacabuco, Concepción, Copiapó, Curicó, Linares, Llanquihue, Magallanes, Malleco, Ñuble, Talca [1701] No (0.51205173 0.48794827)
52) Nem < 5.5 [602] No (0.59800664 0.40199336)
104) Carrera < 7.5 [448] No (0.64508929 0.35491071) *
105) Carrera >= 7.5 [154] Si (0.46103896 0.53896104) *
53) Nem >= 5.5 [1099] Si (0.46496815 0.53503185)
106) Carrera < 5.5 [483] 1 No (0.52173913 0.47826087) *
107) Carrera >= 5.5 [616] Si (0.42045455 0.57954545) *

```
   27) Provincia_domicilio = Antofagasta, Biobío, Cardenal Caro, Chiloé, Elqui, General Carrera,
                             Marga Marga, Santiago, Valdivia [76] Si (0.1315789 0.86842105) *
 7) Carrera < 3.5 [481] Si (0.30977131 0.69022869)
  14) Ranking < 1.5 [49] No (0.57142857 0.42857143)
    28) Provincia_domicilio = Biobío, Linares, Ñuble, Santiago [44] No (0.6136363 0.3863636) *
    29) Provincia_domicilio = Concepción, Talca [5] Si (0.20000000 0.80000000) *
  15) Ranking >= 1.5 [432] Si (0.28009259 0.71990741)
    30) Tipo_colegio >= 2.5 [167] Si (0.35329341 0.64670659) *
    31) Tipo_colegio < 2.5 [265] Si (0.23396226 0.76603774)
      62) Procedencia_colegio = Biobío, Cachapoal, Colchagua, Curicó, Santiago [11]
                                No (0.54545455 0.45454545)
       124) Ranking < 3.5 [7] No (0.85714286 0.14285714) *
       125) Ranking >= 3.5 [4] Si (0.00000000 1.00000000) *
      63) Procedencia_colegio = Cautín, Concepción, Linares, Llanquihue, Malleco, Ñuble,
                                Talagante [254] Si (0.22047244 0.77952756) *
```