UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MODELO PARA LA AUTOMATIZACIÓN DEL PROCESO DE DETERMINACIÓN DE RIESGO DE DESERCIÓN EN ESTUDIANTES UNIVERSITARIOS

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

ERWIN SERGIO FISCHER ANGULO

SANTIAGO DE CHILE
2 0 1 2

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MODELO PARA LA AUTOMATIZACIÓN DEL PROCESO DE DETERMINACIÓN DE RIESGO DE DESERCIÓN EN ALUMNOS UNIVERSITARIOS

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

ERWIN SERGIO FISCHER ANGULO

PROFESOR GUÍA
NELSON ANTRANIG BALOIAN TATARYAN

MIEMBROS DE LA COMISIÓN
PABLO BARCELO BAEZA
SEBASTIAN RIOS PEREZ
M[a.] ANGELICA PINNINGHOFF JUNEMANN

SANTIAGO DE CHILE
MAYO 2012

Resumen

A pesar de los esfuerzos en políticas públicas para brindar acceso a la universidad, la deserción universitaria se ha convertido en un problema prioritario a ser investigado y tratado. La tasa de deserción ha llegado a constituir uno de los principales indicadores de eficiencia interna dentro de cualquier institución de educación superior. Invertir más tiempo en diagnósticos de las causas de la deserción con metodologías adecuadas que permitan predecir ésta con mayor efectividad, contribuye a mejorar la relación efectividad-costo en la gestión de la unidad académica.

El objetivo del presente proyecto consiste en investigar y proponer una metodología que permita identificar en forma automática a los estudiantes con mayor riesgo de deserción de las carreras de Ingeniería de la Universidad de Las Américas.

Para la implementación de este proyecto se adoptó la metodología CRISP-DM que estructura el proceso de minería de datos en seis fases, que interactúan entre ellas de forma iterativa. Se aplicaron los modelos de Redes Neuronales, Árboles de decisión y Cluster K-medianas para analizar el comportamiento de los estudiantes, evaluando factores como el puntaje promedio obtenido en la Prueba de Selección Universitaria (PSU), el promedio de notas obtenido en la enseñanza media, La edad a la fecha de Ingreso a la institución y el género de los estudiantes.

La exactitud de los modelos es calculada a partir del conjunto de datos de pruebas, los cuales indican que ningún modelo predictivo arrojó resultados positivos, debido a esto se analizó el proceso y se llegó a la conclusión que es muy probable que los datos de entrada no eran suficientemente confiables. Dado que dentro de los límites de este trabajo era imposible conseguir datos fidedignos y completos, esta tesis propone una metodología para enfrentar estudios de minería de datos educativa donde se eviten los problemas descritos

Como trabajo futuro se propone implementar un proceso de captura de variables relevantes para la investigación, al momento del ingreso del estudiante a la institución superior, de tal manera de facilitar la generación de un almacén de datos para ayudar a la toma de decisiones.

TABLA DE CONTENIDOS

Índice de Figuras

Índice de Tablas

1. Introducción

Universidad de Las Américas contribuye a la movilidad social de las personas, posibilitando en ellas el desarrollo de las competencias necesarias para que asuman, con mejor preparación, los desafíos de la sociedad contemporánea. Además reconoce que el único camino posible para aumentar la fuerza laboral capacitada es la educación. También reconoce que cualquier joven debe tener la oportunidad de cursar estudios superiores, independiente de su origen socioeconómico y su recorrido escolar y, que los actuales trabajadores también deben tener la posibilidad de acceder a la educación superior y/o al perfeccionamiento profesional.

Uno de los problemas a enfrentar en la universidad es la deserción en los primeros años de las carreras. La deserción universitaria afecta tanto en los ámbitos personales como en los institucionales, sociales y económicos. En lo personal, implica una condición de fracaso que afecta emocionalmente por la discrepancia con las aspiraciones personales e incide en la trayectoria ocupacional de los individuos. En lo institucional, implica una disminución del rendimiento académico de la universidad. En lo social, la deserción contribuye a generar inequidad y desequilibrios sociales y desvirtúa los objetivos que la sociedad le ha entregado a la educación superior. En lo económico, el costo que esto implica para los sistemas es considerable.[1]

Universidad de Las Américas, consciente de los efectos que provocan la interrupción o abandono de los estudios universitarios por parte de los alumnos, ha realizado un gran esfuerzo para aumentar la retención de los mismos en sus carreras.

Un estudio [2] realizado recientemente ha demostrado que la tasa de retención anual[1] total de Universidad de Las Américas, para el año 2006 es de 77%. La Universidad ha desarrollado campañas de apoyo a los estudiantes, asignando profesores tutores al ingreso a la universidad. En el ámbito docente, se ha implementado un programa de nivelación de competencias en el área de Comunicación, Psicológica y Cultural, además de incorporar un programa de nivelación (Propedéutico) en Matemática para los alumnos de primer año del área de la Ingeniería.

[1] La retención anual considera a los estudiantes que permanecen en la universidad de un año para otro, la cual es calculada como la razón en porcentaje, del número de estudiantes matriculados en el año en curso que estaban matriculados el año anterior.

Con el objeto de mejorar la tasa de retención, se hace necesario tener un mecanismo que permita determinar el riesgo de deserción de los alumnos, para aplicar medidas en forma temprana.

Desde un punto de vista de las tecnologías de la información, se define Inteligencia de Negocios (Business Intelligence) como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada para su análisis y conversión en conocimiento, así dando soporte a la toma de decisiones sobre el negocio.[3]

El rendimiento académico de los estudiantes se basa en diversos factores como las variables ambientales, personales, sociales, psicológicas entre otras. La minería de datos, como un área dentro de los sistemas DSS, es una colección de métodos de las estadísticas, las ciencias de la computación, la ingeniería y la inteligencia artificial para identificar patrones interesantes[4]. En la minería de datos se hace hincapié en la identificación de patrones en los grandes volúmenes de datos (descubrimiento), por lo que se considera esta técnica con el objeto de determinar patrones de comportamiento de las variables que tienen un impacto en la retención estudiantil universitaria.

Una de las técnicas más importantes en minería de datos educativa es la clasificación. La clasificación es una técnica de minería de datos predictiva, realiza la predicción a partir resultados conocidos que se encuentran en diferentes datos [5]. Los modelos de predicción tienen el objetivo específico que nos permite predecir los valores desconocidos de las variables de interés dado los valores conocidos de otras variables. Los modelos de predicción pueden ser considerado como el aprendizaje a partir de un mapeo de un conjunto de mediciones de vectores de entrada a una salida escalar [6]. La Clasificación mapea un conjunto de datos en grupos predefinidos de clases. Se conoce como aprendizaje supervisado, porque las clases se determinan antes de examinar los datos.

Los modelos de predicción que incluyen todos los datos personales, sociales, psicológicos y ambientales son necesarios para la predicción eficaz de la deserción de los estudiantes. Lograr una predicción con una alta precisión es beneficioso para identificar inicialmente a los estudiantes con riesgo de deserción. Se requiere que los estudiantes identificados puedan ser asistidos tempranamente para que su desempeño mejore.

2. Objetivos y alcances

Los objetivos de la presente investigación se orientan a mejorar los índices de retención en la educación superior, los que se presentan a continuación.

2.1. Objetivo general

Desarrollar un indicador que permita clasificar en forma automática a los alumnos con mayor riesgo de deserción de las carreras de Ingeniería de la Universidad de Las Américas, para trabajar un programa docente y de tutorías, focalizado con el objeto de mejorar los índices de retención.

2.2. Objetivos específicos

- Conocer diferentes formas para desarrollar modelos de predicción.
- Identificar los diferentes factores que afectan la permanencia del estudiante en su carrera académica.
- Proponer un modelo predictor de riesgo de deserción universitaria en la Facultad de Ciencias de la Ingeniería de Universidad de Las Américas.
- Validar el modelo de predicción desarrollado.

2.3. Alcance del proyecto

El presente proyecto está circunscrito en la Facultad de Ciencias de la Ingeniería de Universidad De Las Américas, específicamente en la Escuela de Tecnologías de la Información, abarcando las sedes Providencia, Santiago Centro y La Florida.

Se aplicarán técnicas de minería de datos con modelos en Redes Neuronales, Árboles de decisión y Clustering con datos históricos de admisión comprendiendo el período de ingreso 2001 al 2008.

Se utilizará el módulo de Analysis Services de SQL Server 2008, que contiene las características y herramientas necesarias de minería de datos, con los algoritmos requeridos para esta investigación.

2.4. Método de trabajo

El método de trabajo utilizado para este proyecto corresponde a una serie de actividades ordenadas en etapas orientadas a obtener el resultado requerido:

Etapa 1 – Definición del problema

El primer paso en esta investigación, es establecer una definición del problema con el objeto que comprender lo que se pretende resolver.

Etapa 2 – Revisión bibliográfica

En la segunda etapa, se desarrolla la revisión bibliográfica, que abarca los aspectos conceptuales relacionados con el problema de la deserción universitaria y los métodos que generalmente se han utilizado para estudiar las variables asociadas a este fenómeno con el objeto comprender los avances en esta área.

Etapa 3 – Diseño de la solución

Enseguida en la tercera etapa se realiza un diseño de la solución a través de la aplicación de una metodología de minería de datos.

Etapa 4 – Construcción de un prototipo

Más tarde en la cuarta etapa se construye un prototipo de la solución, identificando las variables relevantes, y planteando un modelo en un software de minería de datos.

Etapa 5 – Análisis y evaluación

Posteriormente se debe realizar el análisis y evaluación, de los resultados obtenidos, comparando los distintos modelos aplicados.

Etapa 6 – Conclusiones y trabajo futuro

Finalmente se presentan las conclusiones de este trabajo, como también se proponen líneas de investigación para futuros estudios, considerando nuevas variables y atributos.

3. Definición del problema

Durante mucho tiempo, la deserción universitaria fue considerada como un fenómeno normal e, incluso, como una muestra de la exigencia de la universidad y de la carrera particular; sin embargo, hoy se ve como un signo de ineficiencia y como un gran costo para el país, los estudiantes y para las instituciones de educación superior, lo que pasó a convertirse en un problema que hay que entender para poder combatirlo.

En la Tabla 1 Se presenta la tasa de retención de los alumnos nuevos de la Facultad de Ciencias de la Ingeniería (FACI) de Universidad de Las Américas, para el período 2001 al 2006, considerando, las escuelas de Ingeniería Industrial, de Construcción Civil y de Tecnologías de la información. [2].

Esc	2001	2002	2003	2004	2005	2006
Industrias	81%	58%	76%	73%	76%	63%
Construccion	71%	69%	68%	73%	68%	55%
Tec. Info	76%	75%	73%	73%	74%	66%
FACI	77%	68%	74%	73%	75%	64%

Tabla 1 - Tasa de retención alumnos nuevos

Según se puede observar en la grafica de la Figura 1, se aprecia una tendencia a la baja en los índices de retención, por lo que se hace imprescindible conocer los determinantes de este indicador, que permita detectar en forma temprana los posibles desertores para trabajar un programa docente y de tutorías, focalizado con el objeto de mejorar los índices de retención.

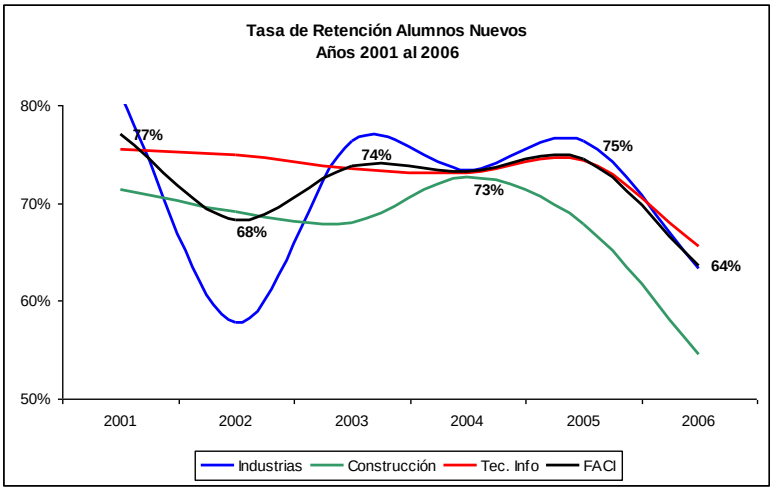

Figura 1 - Tasa de retención alumnos nuevos

Se espera determinar que el problema de la deserción es originado por ciertas variables asociadas a los ámbitos académico socioeconómico, institucionales y personales, tales como: los apoyos financieros o académicos que recibe, el género del estudiante, su edad, el número

de hermanos que tiene, el puesto que ocupa en la familia, el nivel educacional de los padres, entre otros factores, sin embargo a la fecha es discutible o desconocido el impacto o contribución que tiene cada uno de ellos en la deserción.

Las tres causas más determinantes en la deserción de estudiantes en primer año universitario son: problemas vocacionales, situación económica de sus familias, y rendimiento académico[7]

Este proyecto está orientado a caracterizar el problema y generar un conjunto de indicadores de deserción universitaria que relacione estas variables.

4. Revisión bibliográfica

Existen dos modelos principales desde los cuales se puede estudiar el problema de los determinantes de la deserción: el paradigma funcionalista y el paradigma dialéctico. La perspectiva funcionalista tiene un enfoque individualista de la educación, donde lo más importante son los talentos, habilidades, dones y esfuerzo individuales. Por otro lado, el paradigma dialéctico enfoca la deserción dentro de todo el sistema educativo, donde no solo importa el estudiante como individuo, sino como parte de todo el sistema educativo.

En el informe “Estudio de la deserción estudiantil en la educación superior en Colombia” [8] se presentan distintos factores asociados que permiten predecir la deserción, como el académico, el socioeconómico, los institucionales e individuales y como parte de estos tenemos el programa en el que está inscrito, los apoyos financieros o académicos que recibe, el género del estudiante, su edad, el número de hermanos que tiene y el puesto que ocupa en la familia y el nivel educacional de los padres, entre otros factores, los que se presentan en la Tabla 2 - Factores determinantes de la deserción universitaria.

Individuales	Académicos	Institucionales	Socioeconómicos
• Edad, género, estado civil • Entorno familiar • Calamidad y problemas de salud • Integración social • Incompatibilidad horaria en actividades extra académicas	• Orientación profesional • Rendimiento académico • Calidad del programa • Métodos de estudio • Calificación en examen admisión • Insatisfacción en el programa u otros factores académicos • Numero de materias	• Normalidad académica • Tipo de colegio • Becas y forma de financiamiento • Recursos universitarios • Orden público • Entorno político • Relaciones con los profesores y otros estudiantes	• Estrato • Trabajo del estudiante • Situación laboral de los padres • Dependencia económica • Personas a cargo • Nivel educativo de los padres • Entorno macroeconómico del país

Tabla 2 - Factores determinantes de la deserción universitaria.[8]

4.1. Conceptualización y metodología sobre deserción estudiantil

En los estudios sobre deserción asociados a los distintos segmentos de enseñanza, abarcando enseñanza básica, media y universitaria, se puede apreciar una falta de consenso respecto de un concepto unificado que permita una recolección de datos con una metodología igualmente unificada. En las investigaciones se han utilizado definiciones complementarias y diferentes. En este capítulo se muestra un panorama de la diversidad de definiciones que se han manejado en los diferentes estudios. [9], [10].

El fenómeno comprende a quienes no siguieron la trayectoria normal de la carrera, bien sea por cancelar su matrícula o por no matricularse. Cuantitativamente el fenómeno puede expresarse como el número de estudiantes que abandonan la universidad en un período determinado, antes de haber obtenido el título correspondiente, relativo al total de estudiantes asociado a la cohorte correspondiente.

La deserción también sería consecuencia de interacciones insuficientes con otros (estudiantes, profesores y personal administrativo) en la escuela y congruencia insuficiente con los modelos de valores predominantes en la colectividad escolar [9]

El fenómeno se puede observar desde tres ópticas diferentes [10]. En primer lugar, la individual, que se refiere al hecho de que la persona llega a la universidad buscando obtener un título que lo acredite ante la sociedad como alguien que tiene la idoneidad intelectual. En consecuencia, quien no logra esta meta individual es llamado desertor. En segundo lugar, se encuentra la óptica institucional, que se relaciona con el choque del estudiante contra los preceptos institucionales que lo repelen, llevándolo lentamente a comprender que debe retirarse, unas veces conscientemente, otras de manera irracional y dolorosa. En tercer y último lugar, se encuentra la deserción, que en la óptica estatal está en la base de la organización educativa del país.

Quienes han establecido tipos de deserción, definen el fenómeno como el hecho de que el alumno no registre actividad académica por un período académico de dos años. En tal caso, el desertor inicial sería aquel que no registra inscripción al año siguiente, y el desertor avanzado

sería el individuo que habiendo aprobado más de la mitad de las materias del plan de estudios, no registra inscripción durante dos años [11].

Tal vez el intento más cercano de aclaración del concepto de deserción universitaria es el plasmado en “Deserción Estudiantil Universitaria. Conceptualización” [12], ensayo que apunta a conceptualizar sobre este fenómeno, de tal forma que sea posible para la comunidad académica plantear estrategias y políticas educativas que conduzcan a detectar posibles desertores y promuevan la prevención del fenómeno en la educación superior. Pretenden, además, diferenciar la deserción de otros fenómenos tales como la mortalidad estudiantil, el ausentismo y el retiro forzoso.

Los autores parten de una concepción de deserción estudiantil entendida no sólo como el abandono definitivo de las aulas, sino como el abandono de la formación académica; es una decisión personal del sujeto y no obedece a un retiro académico forzoso (por la falta de éxito del estudiante en el rendimiento académico, como es el caso de expulsión por bajo promedio académico) o a un retiro por asuntos disciplinarios. Los autores de esta investigación sostienen que es preciso diferenciar entre deserción (y variables asociadas), y mortalidad estudiantil, dado que la primera es intrasujeto y la segunda es extrasujeto.

Por otro lado, se afirma que la deserción es todo un proceso, a veces lento, que va creciendo y reforzándose en el interior del sujeto, quien lo manifiesta en la decisión definitiva, para bien o para mal de él mismo y de su entorno. [8]

En el mismo estudio [8], se mencionan las siguientes clases de deserción en educación, no excluyentes entre sí:

- **Deserción total**: abandono definitivo de la formación académica individual.
- **Deserción discriminada por causas**: según la causa de la decisión.
- **Deserción por facultad** (escuela o departamento): cambio facultad - facultad.
- **Deserción por programa**: cambio de programa en una misma facultad.
- **Deserción a primer semestre de carrera**: por inadecuada adaptación a la vida universitaria.
- **Deserción acumulada**: sumatoria de deserciones en una institución.

Adicionalmente, se involucran en el fenómeno de la deserción como actores relevantes no sólo a los desertores, sino también a padres de familia de desertores, excompañeros de estudio, profesores, directivas y administradores académicos.

En el documento “La Educación Superior en Colombia – Década de Los 90 [13] se define la deserción estudiantil como la cantidad de estudiantes que abandonan el sistema de educación superior entre uno y otro período académico (semestre) de un año, calculada a partir del balance entre el estado del primer semestre, disminuido en los egresados del mismo periodo y adicionado con los alumnos nuevos del siguiente período, lo cual genera el nuevo estado ideal de alumnos matriculados sin deserción. Al hacer la diferencia entre este último dato y el real reportado por las instituciones, se obtiene la deserción correspondiente al período en mención.

En cuanto a indicadores y metodología para calcular la deserción, los autores Osorio y Jaramillo [14] han tenido en cuenta los siguientes:

- **Índices de deserción semestral**: relación entre el número total de alumnos desertores del programa (i) en el período (t) y el número total de estudiantes matriculados en dicho programa para el mismo período.
- **Índices de deserción por cohorte**: diferencia entre el número de estudiantes que ingresan a la cohorte (c) en el período (t) y la cantidad de ellos que se matriculan en el período (t +1).
- **Índices de deserción promedio por nivel**: promedio simple de los índices de deserción por semestre calendario, calculados como el número total de desertores de cada nivel sobre el total de matriculados en dicho nivel del programa (i).
- **Tasa ponderada de deserción por nivel**: muestra la expectativa de deserción para el programa (i). Se calcula ponderando la tasa de deserción con el promedio de la distribución de la población matriculada, en los semestres de duración de la carrera.

De las anteriores definiciones se desprende una primera diferenciación para los casos de deserción, consistente en que el abandono puede ser transitorio o definitivo. Además, podemos encontrar en ellas la diferenciación entre lo que se conoce como movilidad interna o externa del sistema.
La deserción académica implica un análisis de todo el sistema universitario y del “equipaje del alumno”, como su orientación vocacional, la familia, entre otros aspectos.

También la movilidad o cambio de carrera o institución hace que se modifique el concepto de deserción. Usualmente habrá que considerar entonces si la deserción es de programa, de institución o del sistema universitario.

Definitivamente debe distinguirse entre la deserción (no académica) o intra-sujeto, y la mortalidad (o deserción académica) o extra-sujeto. La deserción académica podrá ser entonces por razones disciplinarias o por rendimiento y la deserción no académica, por retiro “voluntario”.

Se distingue el abandono voluntario del no voluntario, encuadrando la deserción académica como la no voluntaria y la no académica, como la voluntaria. Sin embargo, la deserción no académica no siempre será tan “voluntaria”.

En el estudio del Instituto Tecnológico de Parral se intenta una comprensión del fenómeno que vaya más allá de la medición. En este último, citando a Vicent Tinto [15], se señala que el éxito en la educación superior está moldeado por las mismas fuerzas que moldean el éxito en general. Si logramos identificar quiénes podrían ser vulnerables a factores de deserción y retención, podremos determinar qué políticas pueden adelantarse por la universidad como determinantes de retención.

En el estudio del Instituto Tecnológico de Parral, se critica que se entienda que en los casos de deserción es el individuo quien no tiene capacidades. Para ellos la deserción no es simplemente un fracaso del individuo, sino que se debe revisar el sistema mismo, en especial el sistema de acreditación o certificación; así la deserción se mira también como un fracaso del sistema, donde los exámenes o evaluaciones se utilizan para marginar de la educación a los reprobados. En este estudio también se traen a colación investigaciones que intentan explicar el problema de la deserción a partir de las variables académicas, y se resalta cómo parece que siempre hay unos niveles de deserción, y que cuando esos niveles sobrepasan ciertos límites, el fenómeno se vuelve preocupante.

Algunos de los factores asociados a la deserción que se plantean en los estudios tomados para el desarrollo del concepto de deserción son:

- Ambientes educativos
- Ambiente familiar
- Trayectoria educativa y acompañamiento al estudiante en su formación

- Edad
- Adaptación social del estudiante con sus pares u homólogos
- Modelos pedagógicos diferentes a los del bachillerato.
- Programas curriculares rígidos, de alta intensidad temática y tiempos reducidos.
- Evaluaciones extenuantes y avasalladoras
- Cursos no asociados ni aplicables al desarrollo profesional.
- Factores económicos.
- Cantidad de oferentes (el mercado de la educación).
- La orientación profesional y vocacional.
- Hijos de padres a los que no les interesa la educación
- Bajo aprovechamiento de oportunidades educativas
- Problemas de disciplina
- Problemas con la justicia
- Falta de interés
- Nivel socioeconómico bajo
- Ausentismo
- Salud psicosomática
- Relaciones interpersonales
- Procedencia de entornos violentos
- Baja empatía por el trabajo de sus pares
- Desmotivación hacia la carrera o la Universidad
- Resistencia a desarrollar actividades formantes
- Inapetencia por el conocimiento

La deserción no es un problema del individuo; desde luego que el desertor es aquel en donde todo se concentra, pero ello no es suficiente para declararlo culpable. De cualquier forma sí se debe mirar directamente al desertor, que es donde converge todo el proceso de la deserción.

Métodos de Predicción de Deserción

Ahora que hemos analizado los conceptos y metodologías asociados a la deserción estudiantil, es necesario explorar algunos métodos que permitan predecir este comportamiento.

Para lograr los objetivos propuestos, existen varios caminos, a saber: Modelos estadísticos, y modelos de Inteligencia de negocios, que comprende a los sistemas de soporte a la toma de decisión, entre otros.

Con respecto a los modelos de duración, en la investigación *"Determinantes de la deserción y graduación universitaria: Una aplicación utilizando modelos de duración*" [16], se plantea que esta técnica posee varias **ventajas** con respecto a técnicas clásicas como la estimación de modelos logit "clásicos", de regresión o análisis discriminante. Estas últimas son de naturaleza estática, mientras que el análisis de duración capta la temporalidad y la variación de las circunstancias a lo largo del tiempo, siendo un enfoque más dinámico. El concepto central de un modelo de duración no es la probabilidad de que un evento ocurra (ej. probabilidad de que un individuo abandone sus estudios el tercer año), sino más bien la probabilidad *condicional* de que esto ocurra, dado un conjunto de variables.

Estas técnicas estadísticas se han utilizado de forma sistemática para abordar distintos problemas de modelación y predicción a partir de distintas variables. Sin embargo, el gran volumen de datos del que se dispone hoy día hace que estas técnicas tarden mucho en numerosos problemas de interés. La necesidad de métodos eficientes y automáticos para explorar bases de datos ha motivado un rápido avance de disciplinas conocidas hoy como Inteligencia de negocios (BI) en la que los sistemas de Soporte a la toma de decisiones a través de minería de datos (*data mining)*, permiten desarrollar métodos que operen de forma automática a partir de un conjunto de datos para capturar distintos patrones de comportamiento que sean apropiados para resolver un problema.

Las redes probabilísticas: Son modelos apropiados para el tratamiento de problemas con incertidumbre, y utilizan técnicas estadísticas modernas de inferencia y estimación para ajustar los parámetros a los datos y obtener conclusiones en base a los modelos resultantes. Por otra parte, los sistemas de minería de datos, permiten encontrar patrones de comportamiento, en un conjunto de datos que representan la realidad codificada a partir de los sistemas transaccionales.

Finalmente, se puede buscar una solución al problema de la predicción de la deserción estudiantil a través de un sistema de soporte a la toma de decisiones (**DSS** por sus siglas en inglés *Decision Support System*).

4.2. Sistemas de apoyo a la toma de decisiones

Un DSS es un sistema informático utilizado para servir de apoyo en el proceso de toma de decisiones. Esto significa ayudar a las personas a generar alternativas basadas en el ***uso de datos y modelos*** en la solución de problemas no estructurados.

Aunque no hay un consenso tan claro de lo que se entiende por DSS[17], podemos considerar la siguiente definición[18] "Un sistema computacional que apoya a personas que toman decisiones enfrentados a problemas no estructurados". Finalmente podemos clasificar los sistemas DSS como una herramienta de la Inteligencia de negocios (BI) enfocada al análisis de los datos de una organización.

Dentro de los sistemas de soporte a la toma de decisiones se pueden observar aplicaciones basadas en minería de datos predictiva a través de la utilización de diferentes técnicas de clasificación que realiza la predicción a partir de resultados conocidos que se encuentran en diferentes datos[5].

4.3. Proceso de extracción de conocimientos

El conocimiento es una combinación de valores, información contextualizada y experiencias que proporcionan un marco para evaluar e incorporar nuevas experiencias e información. Desde el punto de vista de las Organizaciones, se puede definir el conocimiento como aquella información que permite generar acciones asociadas a satisfacer las demandas del mercado, y apoyar las nuevas oportunidades a través de la explotación de las competencias centrales de la Organización.[19].

El proceso de extracción de conocimiento a partir de datos, ha sido definido de diferentes maneras por diversos autores. Sin embargo, aunque con diferentes palabras, todos refieren las mismas ideas. Por citar una de las definiciones; se puede entender como: "el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en bases de datos, Data-Warehouses o cualquier otro medio de almacenamiento de información." [20]. Es el proceso de descubrimiento del conocimiento, señalado como *Knowledge Discovery in Databases* (KDD) [21], que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

KDD es un proceso de extracción no trivial de información potencialmente útil a partir de un gran volumen de datos en el cual la información está implícita y que no se conoce previamente. Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones. En la Figura 2 - El proceso KDD, se presentan las actividades relacionadas con este proceso [21].

Las investigaciones en esta área incluyen análisis estadístico de datos, técnicas de representación del conocimiento, razonamiento basado en casos [CBR: Case Based Reasoning], razonamiento aproximado, redes neuronales y visualización de datos. Las tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y *clustering*, el reconocimiento de patrones, el modelado predictivo y la detección de dependencias, entre otros.

Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos, novedosos para el sistema (para el usuario siempre que sea posible) y potencialmente útiles.

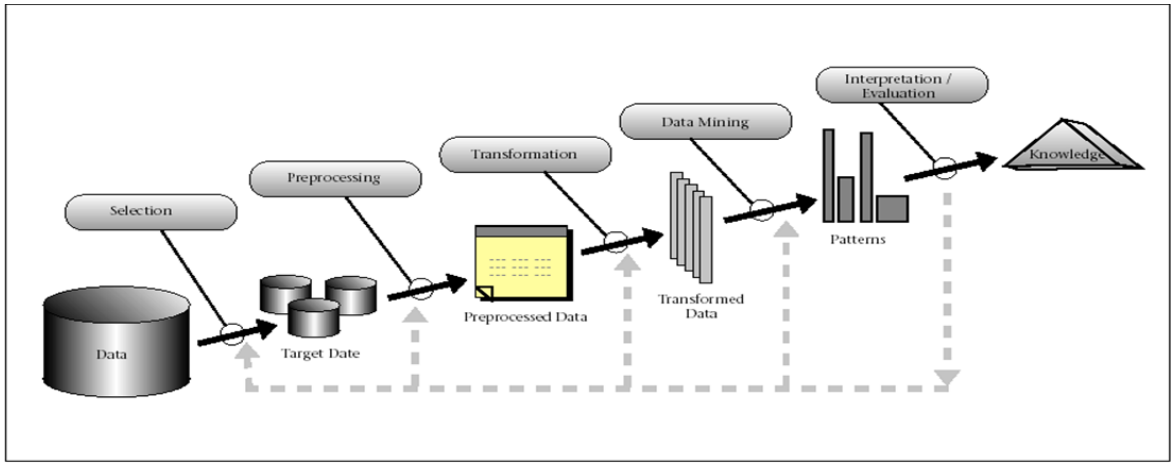

Figura 2 - El proceso KDD

Como muestra la figura anterior, las etapas del proceso KDD se dividen en 5 fases y son:

1. **Selección de datos**. En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

2. **Preprocesamiento**. Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.
3. **Transformación**. Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.
4. **Data Mining**. Es la fase de modelamiento propiamente tal, en donde son aplicados métodos inteligentes con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.
5. **Interpretación y Evaluación**. Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos

El conocimiento detectado fuera del proceso de KDD se utiliza generalmente para apoyar las decisiones de gestión. Por lo tanto, desemboca en un Sistema de Soporte a la Decisión (DSS) o en la automatización de algún proceso tal como marketing para la comercialización directa.

4.3.1. Metodologías para el proceso de extracción de conocimiento

Teniendo en cuenta que el proceso de extracción de conocimiento KDD, es un proceso no trivial, ha surgido la necesidad de una aproximación sistemática para la realización de los proyectos de Data Mining, por lo que diversas empresas y consorcios han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión de pasos que le dirijan a obtener buenos resultados. Así la empresa SAS propone la utilización de la metodología SEMMA. En 1999 un importante consorcio de empresas europeas, NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para el desarrollo de la metodología de libre distribución CRISP-DM (Cross- Industry Standard Process for Data Mining). Siendo estas las principales metodologías utilizadas para la realización de proyectos de Data Mining [22]. Estas metodologías comparten la misma esencia estructurando el proyecto de Data Mining en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Data Mining en un proceso iterativo e interactivo.

4.3.1.1. Metodología SEMMA

El Instituto SAS desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso (Sample, Explore, Modify, Model, Assess).

El proceso se inicia con la **extracción** de la muestra sobre la que se va a aplicar el análisis. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles. La forma más común de obtener una muestra es la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Este método de muestreo se denomina muestreo aleatorio simple.

La metodología SEMMA establece que para cada muestra considerada para el análisis del proceso se debe asociar el nivel de confianza de la muestra. Una vez determinada una muestra o conjunto de muestras representativas de la población en estudio, la metodología SEMMA indica que se debe proceder a una **exploración** de la información disponible para simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. Para lograr este objetivo se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.

La tercera fase de la metodología consiste en la **manipulación** de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo. Una vez que se han definido las entradas del modelo, con el formato adecuado para la aplicación de la técnica de modelado, se procede al análisis y **modelado** de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como técnicas basadas en datos tales como redes neuronales, técnicas adaptativas, lógica fuzzy, árboles de decisión, reglas de asociación y computación evolutiva. Finalmente, la última fase del proceso consiste en la **valoración** de los

resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas muestras.

4.3.1.2. Metodología CRISP-DM

La metodología CRISP-DM [23] consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos.

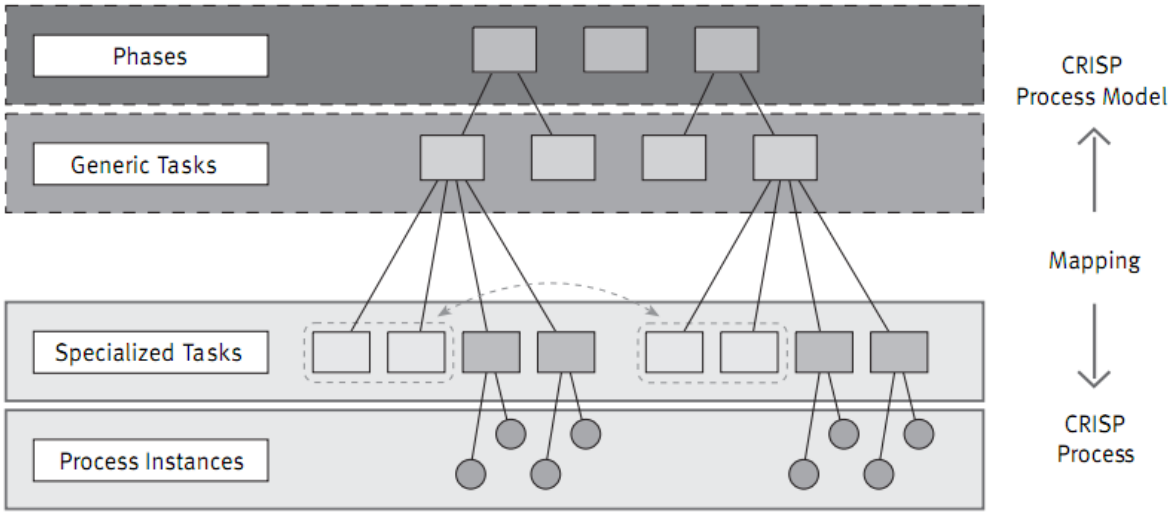

Figura 3 - Cuatro niveles de detalle de la metodología CRISP-DM

A nivel más general, el proceso está organizado en fases, estando cada fase a su vez estructurada en varias tareas genéricas de segundo nivel. Las tareas genéricas se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea genérica “limpieza de datos”, en el tercer nivel se indican las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, “limpieza de datos numéricos”, o “limpieza de datos categóricos”. El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Data Mining específico.

La metodología CRISP-DM proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de Data Mining: el modelo de referencia y la guía del usuario. El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de Data Mining en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de Data Mining

específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

Fase de análisis del problema: incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

Fase de análisis de datos: comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis. Una vez realizado el análisis de datos, la metodología establece que se proceda a la preparación de los datos, de tal forma que puedan ser tratados por las técnicas de modelado.

Fase de preparación de datos: incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. La fase de preparación de los datos, se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto, las fases de preparación y modelado interactúan de forma sistemática.

Fase de modelado: se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas que serán utilizadas en esta fase se seleccionan en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requerimientos del problema.
- Tiempo necesario para obtener un modelo.
- Conocimiento de la técnica.

Antes de proceder al modelado de los datos, se debe establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Una vez

realizadas estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.

Fase de evaluación: se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.

Fase de explotación: normalmente los proyectos de Data Mining no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. Además, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados (Fayyad 1996). En la Figura 4 - El proceso CRISP DM se puede ver un esquema de las diferentes fases de la metodología y las tareas generales que se deben desarrollar en cada fase.

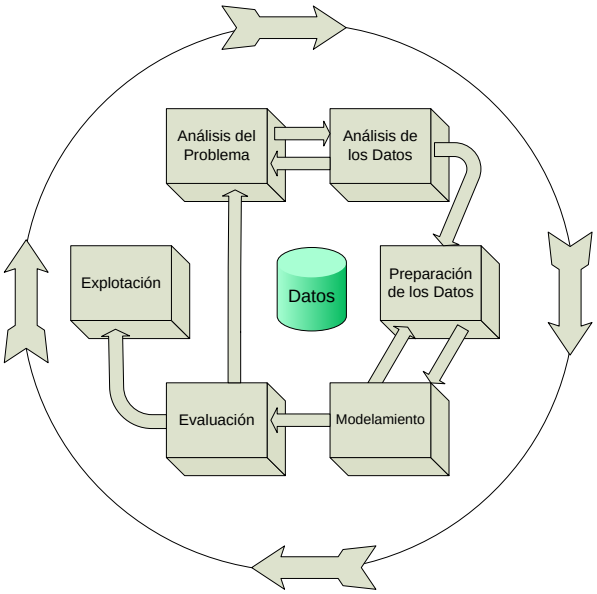

Figura 4 - El proceso CRISP DM

4.3.1.3. Comparación de Metodologías

La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Data Mining donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza efectuando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista más global, se puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Otra diferencia significativa entre la metodología SEMMA y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales, ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte, la metodología CRISP-DM si bien está fuertemente ligada a SPSS adquirida por IBM, ha sido diseñada como una metodología neutra, respecto a la herramienta que se utilice para el desarrollo del proyecto de Minería de datos, siendo su distribución libre y gratuita.

Bajo este último criterio, es que se ha decidido utilizar la metodología CRISP-DM como guía para el desarrollo de este proyecto.

4.4. Técnicas de Minería de Datos

Las técnicas de minería de datos se están volviendo más ampliamente usadas cada día, debido a que permiten descubrir patrones y tendencias desde las bases de datos existentes en las empresas [24].

La minería de datos (Data Mining) es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos [25]. Los algoritmos de Data Mining se enmarcan en el proceso completo de extracción de conocimiento a partir de datos según se indica en la sección anterior.

La Minería de Datos se apoya en la aplicación de métodos matemáticos de análisis, utilizando diferentes algoritmos y técnicas de Clasificación, tales como Clustering, Regresión, Inteligencia

artificial, Redes neuronales, Reglas de asociación, Árboles de decisión, Algoritmos genéticos, entre otras, que son de gran utilidad para llevar a cabo el análisis inteligente de grandes volúmenes de información digital. La minería de datos relacionada con la educación, se denomina “Minería de datos educativa” [26]

La clasificación es la técnica más aplicada de minería de datos, que emplea un conjunto pre-clasificado de ejemplos para desarrollar un modelo que puede clasificar a una población de grandes volúmenes de registros. Este enfoque emplea con frecuencia métodos de Árbol de decisión o Redes neuronales. El proceso de clasificación de datos implica el aprendizaje y la clasificación. En el aprendizaje los datos de entrenamiento son analizados por el algoritmo de clasificación. La clasificación de los datos de prueba se utiliza para estimar la precisión de las reglas de clasificación. Si la precisión es aceptable las reglas se pueden aplicar a las tuplas de datos. El algoritmo de aprendizaje de clasificación utiliza estos ejemplos pre-clasificados para determinar el conjunto de parámetros necesarios para una discriminación adecuada. El algoritmo codifica estos parámetros en un modelo llamado clasificador.

4.4.1. Redes neuronales artificiales

El cerebro humano en su biología está formado por miles de millones de neuronas que se conectan entre sí, transmitiendo información entre ellas, luego de que esta procesa, se genera una respuesta en función del estímulo recibido

Las redes neuronales forman parte fundamental del cerebro, en el cual se generan todos los procesos necesarios asociados al aprendizaje como respuesta a un estímulo generado en el ambiente. Es por esto que como una alternativa para obtener sistemas artificiales de aprendizaje surgen las Redes Neuronales Artificiales.

Como parte de esta investigación, se considera importante comenzar describiendo cómo funcionan las redes neuronales, para luego describir el comportamiento de las RNA.

El cerebro humano continuamente recibe señales de entrada de muchas fuentes y las procesa a manera de crear una apropiada respuesta de salida. Nuestros cerebros cuentan con millones de neuronas que se interconectan para elaborar "Redes Neuronales". Estas redes ejecutan los millones de instrucciones necesarias para mantener una vida normal.

Las neuronas son las células que forman la corteza cerebral de los seres vivos, cada una está formada por elementos llamados cuerpo, axón y dendritas, como se muestra en la Figura 5 - Diagrama básico de una neurona

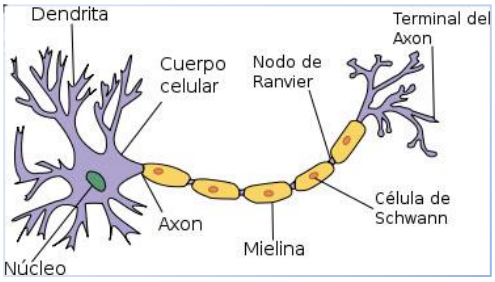

Figura 5 - Diagrama básico de una neurona

Las dendritas forman una estructura de filamentos muy fina que rodean el cuerpo de la neurona. El axón es un tubo largo y delgado que se ramifica en su extremo en pequeños bulbos finales que casi tocan las dendritas de las células vecinas. La pequeña separación entre los bulbos finales y las dendritas se denomina *sinapsis*, en las cuales se produce una transformación de los impulsos eléctricos en un mensaje neuroquímico, mediante la liberación de unas sustancias llamadas *neurotransmisores*.

Una red neuronal se define como una población de neuronas físicamente interconectadas o un grupo de neuronas aisladas que reciben señales que procesan a la manera de un circuito reconocible. La comunicación entre neuronas, que implica un proceso electroquímico, implica que, una vez que una neurona es excitada a partir de cierto umbral, ésta se despolariza transmitiendo a través de su axón una señal que excita a neuronas aledañas, y así sucesivamente. El sustento de la capacidad del sistema nervioso, por tanto, radica en dichas conexiones.

En la literatura podemos encontrar algunas definiciones:

[27] “Una red neuronal es un procesador masivamente paralelo distribuido que es propenso por naturaleza a almacenar conocimiento experimental y hacerlo disponible para su uso”. Este mecanismo se parece al cerebro en dos aspectos:

- El conocimiento es adquirido por la red a través de un proceso que se denomina aprendizaje.
- El conocimiento se almacena mediante la modificación de la fuerza o peso sináptico de las distintas uniones entre neuronas."

[28]"Una red neuronal es un modelo computacional con un conjunto de propiedades específicas, como son la habilidad de adaptarse o aprender, generalizar u organizar la información, todo ello basado en un procesamiento eminentemente paralelo".

Las redes neuronales artificiales engloban diferentes estructuras de procesamiento paralelo distribuido, algunas de ellas biológicamente inspiradas, en las que un número de elementos simples de proceso no lineal se interconectan en forma más o menos densa, las cuales, dependiendo del tipo de arquitectura neuronal, pueden tener diferentes aplicaciones. Pueden utilizarse para el reconocimiento de patrones, la compresión de información y la reducción de la dimensionalidad, el agrupamiento, la clasificación, la visualización, entre otros aspectos.

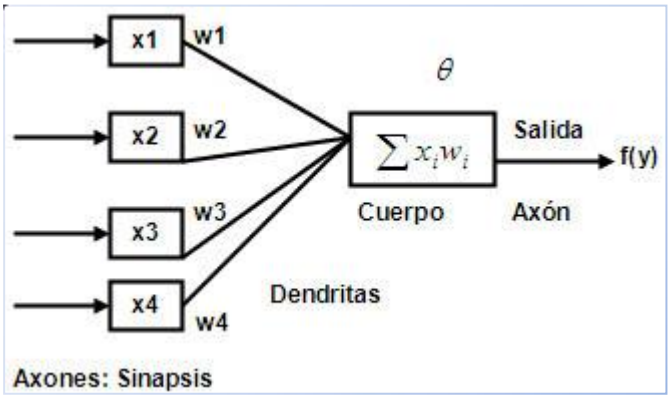

Figura 6 - Modelo de una neurona artificial

La Figura 6 - Modelo de una neurona artificial, ilustra el modelo habitualmente utilizado para una neurona artificial. Esencialmente, la operación de la neurona involucra el cálculo de una función de entrada, a partir de las señales que ingresan a la misma, y la posterior aplicación de una función de activación, en general no lineal. En algunos casos, existe una entrada de umbral (offset), que puede ser siempre asociada a una entrada adicional de valor 1 y peso igual al umbral. La salida de cada neurona dependerá, entonces, de sus señales de entrada, del peso

asociado a cada entrada, y de las características de las funciones de entrada y activación. En la mayoría de los casos, los elementos plásticos de la red son los pesos de las conexiones, y el mecanismo utilizado para adaptar esos pesos se conoce como algoritmo de entrenamiento o aprendizaje. Los elementos citados permiten clasificar los diferentes tipos de redes encontradas en la literatura según los siguientes aspectos:

4.4.1.1. Estructura de una red neuronal

La Figura 7 - Red neuronal artificial perceptrón simple, representa un red neuronal artificial denominada *perceptrón simple* con *n* neuronas de entrada, *m* neuronas en su capa oculta y una neurona de salida, la **Capa Entrada**: almacena la información bruta suministrada a la red o realiza un sencillo pre-proceso de la misma, la **Capa Oculta**: se encargan de recibir, procesar y memorizar la información y la **Capa Salida**: almacena la respuesta de la red para que pueda ser leída, dando un resultado final.

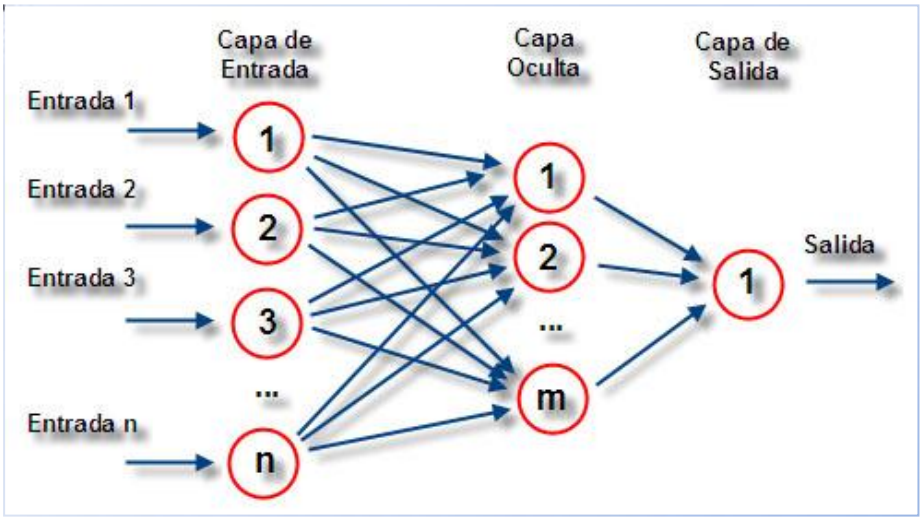

Figura 7 - Red neuronal artificial perceptrón simple

4.4.1.2. Funciones de una red neuronal artificial

Según [22], las funciones de una Red Neuronal Artificial son:

Función de propagación o de red: Esta calcula el valor de base o entrada total a la unidad, lo hace mediante la suma ponderada de todas las entradas recibidas, es decir, de las entradas multiplicadas por el peso o valor de las conexiones.
La función de propagación proporciona el valor del potencial postsináptico de la neurona *i* en función de sus pesos y entradas, en un tiempo *t*. Equivale a la combinación de las señales excitatorias e inhibitorias de las neuronas biológicas.

Función de Activación: Es quizás la característica principal o definitoria de las neuronas, es decir, la que mejor define el comportamiento de la misma. Se usan diferentes tipos de funciones, desde simples funciones de umbral a funciones no lineales. Con ésta se pretende calcular el nivel o estado de activación de la neurona en función de la entrada total.

De esta forma, la función de activación o transferencia será aquel elemento de la topología de una red que permite calcular el estado de actividad de una neurona, transformando la entrada neta en un valor cuyo rango normalmente va de (0 a 1) o de (–1 a 1), dependiendo de su estado de actividad o inactividad. Así, una neurona podrá estar totalmente inactiva (0) –1) o totalmente activa (1) [4].

Las funciones de activación más comunes se detallan en el Anexo – Redes neuronales artificiales

4.4.1.3. Características de una red neuronal artificial

Según [29] las redes neuronales poseen ciertas características. Estas son:

Paralela: Las RNA cuentan con una gran cantidad de neuronas, cada una de ellas trabajando simultáneamente con una parte del problema mayor.

Distribuida: Las RNA cuentan con una gran cantidad de neuronas, a través de las cuales distribuyen su memoria. Esto los diferencia de los sistemas computacionales tradicionales, los que sólo cuentan con un procesador y memoria fija.

Adaptativa: Las RNA tienen la capacidad de adaptarse al entorno modificando sus pesos sinápticos, aprendiendo de las experiencias, consiguiendo generalizar conceptos

a partir de casos particulares, permitiéndole encontrar una solución aceptable al problema.

Existen múltiples tipos de clasificación de RNA. Algunos autores las clasifican según el aprendizaje o según las capas. Otros según conexión; existen, además, distintos tipos de RNA.

Clasificación de RNA según aprendizaje

Para [4], las RNA se pueden clasificar según su aprendizaje, en:

- Supervisado
- No supervisado

El primero menciona que el entrenamiento es controlado por un agente externo, quien determina la respuesta que debería generar la red a partir de una entrada determinada. En el caso contrario está el no supervisado, en el cual no existe un agente externo que determine la respuesta deseada, generando que la red sólo reconozca regularidades en el conjunto de entradas. Un aprendizaje que mezcla estos dos anteriores es el Híbrido, en el cual, algunas capas tienen aprendizaje supervisado y otras no supervisado.

Similar al aprendizaje supervisado, está el reforzado, en el cual se le indica a la red el error que comete (error global), es decir, se indica el éxito o fracaso del resultado obtenido.

Según sus capas, las redes las podemos dividir en: monocapas y multicapas. Según la conexión, se pueden dividir las RNA en: Feedforward, Feedback

En las RNA Feedforward, las neuronas de cada nivel sólo están conectadas con las neuronas de los niveles posteriores por lo tanto la información se propaga hacia delante. Mientras que en las segundas, las neuronas pueden estar conectadas con neuronas de niveles previos, posteriores, o de su mismo nivel.

Las topologías clásicas de RNA son tratadas en el Anexo – Redes neuronales artificiales

4.4.1.4. Ventajas de las redes neuronales artificiales

Debido a su constitución y a sus fundamentos, las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro. Por ejemplo, son capaces de

aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. Esto hace que tengan numerosas ventajas, según [29] éstas son:

Aprendizaje adaptativo: capacidad de aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.

Auto-organización: una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.

Tolerancia a fallas: la destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

Operación en tiempo real: los cómputos neuronales pueden ser realizados en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.

Fácil inserción dentro de la tecnología existente: se pueden obtener circuitos integrados especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Esto facilitará la integración modular en los sistemas existentes.

4.4.1.5. Desventajas de las redes neuronales artificiales

Según [4] las redes neuronales a pesar de todas las ventajas que tiene, también presenta desventajas; éstas son:

Las RNA constituyen un método de resolución de problemas creativo, es decir, que dada las especificaciones de un problema, se desconoce la topología (características de la RNA) con la que se va a solucionar del modo más eficiente

Una vez entrenada una red neuronal, se hace difícil interpretar su funcionamiento, aún más, no es fácil asegurar con qué grado de acierto responderá ante casos nunca vistos.

Los modelos neuronales necesitan una herramienta de procesamiento poderosa. Esto se manifiesta principalmente en el proceso de aprendizaje, pero esto se puede contrarrestar con la facilidad de implementación en dispositivos de hardware específicos.

4.4.2. Arboles de decisión

Los arboles de decisión son una técnica de minería de datos, que establece un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.[4].

El autor plantea que la tarea de aprendizaje para la cual los árboles de decisión se adecuan mejor es la clasificación. Clasificar es determinar de entre varias clases, a qué clase pertenece un objeto; la estructura de condición y ramificación de un árbol de decisión es idónea para este problema. La característica más importante del problema de la clasificación es que se las clases son disjuntas, es decir, una instancia es de la clase a o de la clase b, pero no puede ser al mismo tiempo de las clases a y b.

Los modelos de árboles de decisiones son comúnmente usados en la minería de datos para examinar los datos e inducir las reglas para realizar predicciones. Se pueden utilizar diferentes algoritmos para construir árboles de decisión tales como Detección Automática de Interacciones (CHAID Chi Square Automatic Interaction Detection), Clasificación y Árboles de Regresión (CART Classification and Regression Trees), Quest y C5.0.

Los árboles de decisión crecen a través de una división iterativa de grupos discretos, donde la meta es maximizar la “distancia” entre grupos por cada división. Una de las distinciones entre los diferentes métodos de “división” es como miden esta distancia. Se puede pensar que cada división de los datos en nuevos grupos debe ser diferente uno de otro tanto como sea posible. Esto también es llamado “purificación” de grupos.

Los árboles de decisión usados para predecir variables categóricas son llamados árboles de clasificación, y los árboles usados para predecir variables continuas son llamados árboles de regresión. Los árboles de decisión manejan datos no numéricos muy bien. La habilidad para

aceptar datos categóricos minimiza la cantidad de transformaciones en los datos y la explosión de variables de predicción inherentes en las redes neuronales.

4.4.3. Agrupamiento o Clustering

Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores que utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

La técnica consiste en agrupar un conjunto de datos, sin tener clases predefinidas, basándose en la similitud de los valores de los atributos de los distintos datos. Esta agrupación, a diferencia de la clasificación, se realiza de forma no supervisada, ya que no se conoce de antemano las clases del conjunto de datos de entrenamiento. El clustering identifica grupos, o regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional. El clustering se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters . K-Means [6] es un método particional de clustering donde se construye una partición de una base de datos D de n objetos en un conjunto de k grupos, buscando optimizar el criterio de particionamiento elegido. En K-Means cada grupo está representado por su centro. K-Means intenta formar k grupos, con k predeterminado antes del inicio del proceso. Asume que los atributos de los objetos forman un vector espacial. El objetivo que se intenta alcanzar es minimizar la varianza total intra-grupo o la función de error cuadrático.

Desde un punto de vista práctico, el clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, y muchas otras.

En los últimos años han surgido una gran variedad de algoritmos de clustering, algunos de los cuales se detallan a continuación.

EM

EM pertenece a una familia de modelos que se conocen como *Finite Mixture Models*, los cuales se pueden utilizar para segmentar conjuntos de datos. Es un método de clustering probabilístico. Se trata de obtener la FDP (Función de Densidad de Probabilidad) desconocida a la que pertenecen el conjunto completo de datos. Esta FDP se puede aproximar mediante una combinación lineal de NC componentes, definidas a falta de una serie de parámetros $\{\vartheta\} = \cup\{\vartheta_j \; \forall j = 1..NC\}$, que son los que hay que averiguar,

$$P(x) = \sum_{j=1}^{NC} \pi_j \; p(x;\theta_j), \qquad \sum_{j=1}^{NC} \pi_j = 1.$$

Donde π_j son las probabilidades a priori de cada cluster cuya suma debe ser 1, que también forman parte de la solución buscada, P(x) denota la FDP arbitraria y $p(x; \vartheta_j)$ la función de densidad del componente j. Cada cluster se corresponde con las respectivas muestras de datos que pertenecen a cada una de las densidades que se mezclan. Se pueden estimar FDP de formas arbitrarias, utilizándose FDP normales n-dimensionales, t-Student, Bernoulli, Poisson, y log-normales.

El ajuste de los parámetros del modelo requiere alguna medida de su bondad, es decir, cómo de bien encajan los datos sobre la distribución que los representa. Este valor de bondad se conoce como el likelihood de los datos. Se trataría entonces de estimar los parámetros buscados θ, maximizando este likelihood (este criterio se conoce como ML-Maximun Likelihood). Normalmente, lo que se calcula es el logaritmo de este likelihood, conocido como log-likelihood ya que es más fácil de calcular de forma analítica. La solución obtenida es la misma, gracias a la propiedad de monotonicidad del logaritmo. La forma de esta función log-likelihood es:

$$L(\theta,\pi) = \log \prod_{n=1}^{NI} P(x_n)$$

Donde NI es el número de instancias, que se suponen independientes entre si. El algoritmo EM, procede en dos pasos que se repiten de forma iterativa:

- Expectation: Utiliza los valores de los parámetros, iniciales o proporcionados por el paso Maximization de la iteración anterior, obteniendo diferentes formas de la FDP buscada.
- Maximization: Obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior.

Después de una serie de iteraciones, el algoritmo EM tiende a un máximo local de la función L. Finalmente se obtendrá un conjunto de clusters. Cada uno de estos cluster estará definido por los parámetros de una distribución normal.

K-medianas

El algoritmo K-medianas (o K-Means) es probablemente el algoritmo de agrupamiento más conocido. Es un método de agrupamiento heurístico con número de clases conocido (K). El algoritmo está basado en la minimización de la distancia interna (la suma de las distancias de los patrones asignados a un agrupamiento al centroide de dicho agrupamiento). De hecho, este algoritmo minimiza la suma de las distancias al cuadrado de cada patrón al centroide de su agrupamiento.

El algoritmo es sencillo y eficiente. Además, procesa los patrones secuencialmente (por lo que requiere un almacenamiento mínimo). Sin embargo, está sesgado por el orden de presentación de los patrones (los primeros patrones determinan la configuración inicial de los agrupamientos) y su comportamiento depende enormemente del parámetro K.

5. Diseño de la solución

Con el objeto de buscar un indicador de riesgo de deserción de alumnos, en este trabajo se aplican técnicas de minería de datos utilizando la metodología CRISP-DM que estructura el ciclo de vida de un proyecto de Minería de datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto:

La investigación consiste en implementar un proceso predictor de deserción aplicando minería de datos basado en Redes Neuronales, Árboles de decisión y Clustering, con el objeto de verificar cual algoritmo tiene mejor comportamiento en la solución al problema.

La principal fuente de datos para llevar a cabo esta investigación la constituyen los registros históricos de alumnos con información académica y financiera, registrada en Universidad de Las Américas durante el período académico correspondiente al año 2001 hasta 2006.

Las variables a estudiar por alumno son: Edad, Puntaje PSU, Puntaje ingreso UDLA, Notas de Enseñanza Media, Ingreso Familiar, Asignaturas reprobadas, asistencia a clases, y Colegio origen.

Como principal herramienta, para este proyecto se utiliza el módulo *Analysis Services* de minería de datos que provee SQL Server 2008.

5.1. Análisis de datos

La fase de análisis de datos ha comprendido la recolección de los datos de los estudiantes de la facultad de Ingeniería de Universidad de las Américas, obtenidos desde diferentes sistemas de información de la Universidad.

Inicialmente se tienen los datos de los estudiantes de las seis sedes con que cuenta la universidad, los que se detallan en la Tabla 3 - Alumnos nuevos por sede, que entrega el total de alumnos nuevos ingresados en las distintas sedes de la universidad, para la carrera de Ingeniería de Ejecución Informática en la modalidad de entrega diurnos.

Alumnos Nuevos por Año										
Campus	Carrera	tipo entrega	2001	2002	2003	2004	2005	2006	2007	2008
PR	INGENIERÍA DE EJECUCIÓN EN INFORMÁTICA	D	117	77	78	54	20	19	10	3
LF	INGENIERÍA DE EJECUCIÓN EN INFORMÁTICA	D	31	51	52	33	16	14	16	2
MP	INGENIERÍA DE EJECUCIÓN EN INFORMÁTICA	D		41	32	20	8	15	15	1
SC	INGENIERÍA DE EJECUCIÓN EN INFORMÁTICA	D		52	59	58	38	34	25	39
CO	INGENIERÍA DE EJECUCIÓN EN INFORMÁTICA	D			26	16	12	11	11	10
VL	INGENIERÍA DE EJECUCIÓN EN INFORMÁTICA	D				20	23	19	17	18
Total			148	221	247	201	117	112	94	73

Tabla 3 - Alumnos nuevos por sede

La selección de las variables a utilizar para este estudio se basa principalmente atendiendo los factores descritos en Tabla 2 - Factores determinantes de la deserción universitaria, detallados en el apartado 4.1 Conceptualización y metodología sobre deserción estudiantil.

Inicialmente se rescataron más de 17 atributos, sin embargo, manualmente se descartaron algunos de ellos ya que son considerados como irrelevantes para el estudio. Los atributos con sus descripciones y posibles valores son presentados en la Tabla 4 - Atributos relativos a los estudiantes. El atributo clase es el indicador de deserción de los estudiantes de Ingeniería de ejecución en computación y es denominado “Estado”.

Tipo	Atributo	Descripción	Posibles valores
Individuales	RUT	Identificador de la tupla	Los numeros de rut
	Nombre	Nombre y apellidos del estudiante	Todos los nombre posibles
	FeIngreso	Fecha Ingreso a la Universidad	rango 2000 a 2004
	FecNac	Fecha de nacimiento	rango de fechas posibles
	Sexo	Género	M (masculino), F (femenino)
	Comuna	Comuna de residencia del estudiante	Las comunas de la region Metropolitana
Académicos	Prom_PSU	Promedio PSU	rango de puntages 200 a 900
	Prom_NEM	Promedio NEM	Rango de notas 4 a 7
Institucionales	Sede	Sede donde estudia el alumno	LF, SC, PR, MP, CO, VL
	Colegio	Colegio Origen	Nombre de los colegios
	ComC	Comuna Colegio	comunas de Chile
Socio económicos	Financiamiento	Nivel socioeconómico	Interno, Externo, Contado
	NeduMadre	Nivel educativo de la Madre	Primarios o Sin estudios, Secundarios, Universitarios
	TrabE	Trabajo del estudiante	Si, No
otro	FecEgre	Fecha de egreso	rango 2004 a 2009
	FecTit	Fecha de titulación	rango 2004 a 2009
	Estado	El estado a predecir (La clase)	Titulado, Desertor

Tabla 4 - Atributos relativos a los estudiantes

5.2. Proceso de selección de variables y preparación de datos

En esta etapa, los datos recolectados fueron preparados en un formato adecuado para el proceso de minería de datos, a utilizar con Análisis Services de SQL server 2008. En el proceso de preparación de datos, se limpiaron los datos removiendo los valores inconsistentes usando los mismos valores estándar para todos los datos. El proceso de depuración incluyo completar los valores faltantes utilizando el enfoque de reemplazo por valores que preserven la Media o la Varianza para los atributos numéricos o por la Moda para aquellos atributos nominales.

Hay que considerar que respecto a las notas de enseñanza media hay algunos estudios realizados, como es el caso del trabajo desarrollado por Ronald Fischer y Andrea Repeto [30], que ponen de manifiesto que las notas de enseñanza media tienen una capacidad predictiva importante.

Posteriormente, se realiza un proceso de selección de las variables de entrada para el modelo. Así se determina qué variables, de todas las obtenidas en el análisis de datos realizado, presentan una mayor relevancia para este estudio; o cuales aportaban una información redundante o secundaria. De este modo, las variables que finalmente son consideradas para este estudio son las mostradas en la Tabla 5 – BD81Total, que contiene los registros de los estudiantes de la carrera ingeniería de ejecución informática (81) de las sedes Providencia (PR), Santiago Centro (SC) y La Florida (LF)

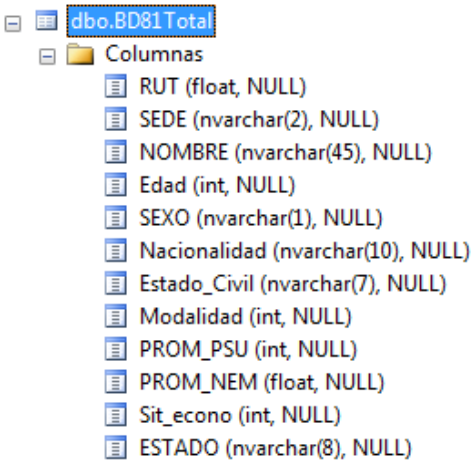

Tabla 5 – BD81Total

La distribución de los datos según el atributo sexo, se muestra en la Tabla 6

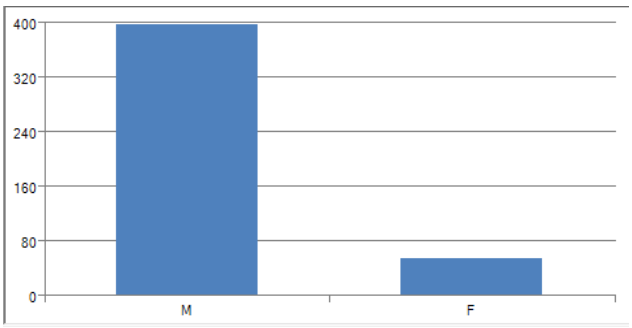

Tabla 6 - Distribución de estudiantes, según sexo

La distribución de los estudiantes según el puntaje PSU, se aprecia en la Tabla 7

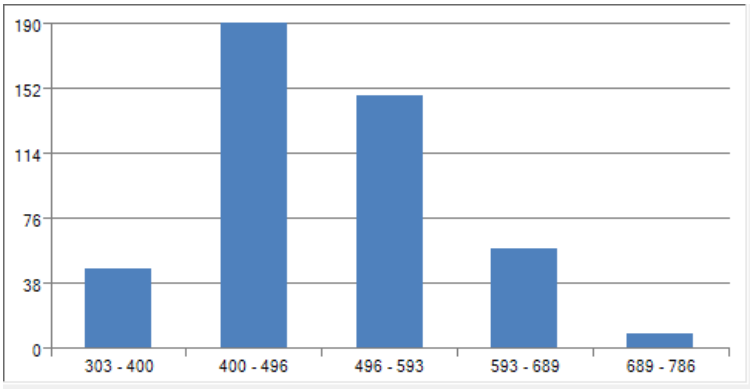

Tabla 7 - Distribución de estudiantes según PSU

La Tabla 8 muestra La distribución de alumnos, según promedio notas de enseñanza media.

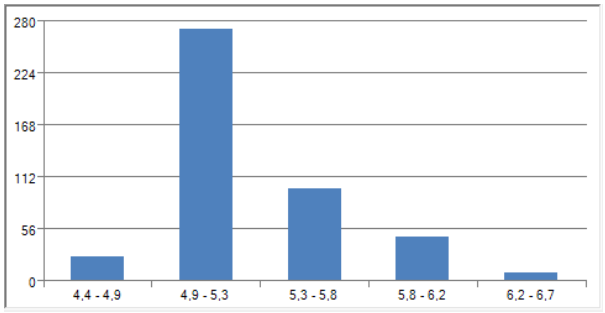

Tabla 8 - Distribución de estudiantes según Notas Enseñanza Media

El atributo financiamiento presentado en la Tabla 9, permite estimar el nivel socioeconómico familiar, de acuerdo al siguiente criterio: Las familias de más altos ingresos pagan al contado, las de ingreso intermedio solicitan crédito externo con instituciones financieras y las de más bajos recursos, recurren al crédito interno de la universidad.

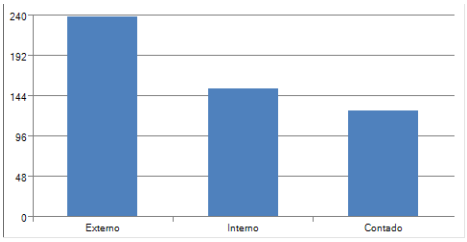

Tabla 9 - Distribución de estudiantes por Financiamiento

5.3. La construcción de los modelos

El siguiente paso consiste en la construcción de los modelos de minería de datos, usando Redes Neuronales, como principal objetivo, además de Árboles de decisión y Cluster, con el objeto de tener un parámetro de comparación respecto a qué modelo presenta una mejor respuesta frente a la estimación de la variable Estado (La clase).

Para la construcción de los modelos, finalmente se dispone de 452 registros con información de los alumnos de las carreras de Ingeniería de ejecución informática de las cohortes 2001 al 2004, según se muestra en la Tabla 10 - Total estudiantes por sede

Sede	2001	2002	2003	2004	
LF	26	34	32	21	113
PR	75	57	52	40	224
SC		40	39	36	115
	101	131	123	97	452

Tabla 10 - Total estudiantes por sede

Del total de 452 registros se seleccionaron en forma aleatoria 317 registros para el modelo de Minería de datos, dejando un total de 135 registros para “validar” el Modelo.

Sede	2001	2002	2003	2004	
LF	18	24	24	13	79
PR	54	45	33	25	157
SC		27	30	24	81
	72	96	87	62	317

Tabla 11 - Estudiantes por sede para el modelo

5.4. Estructura de Minería de datos

La estructura de minería de datos es una estructura de datos que define el dominio de datos a partir del cual se generan los modelos de minería de datos. Una única estructura de minería de

datos puede contener varios modelos de minería de datos que comparten el mismo dominio. Las unidades de creación de la estructura de minería de datos son las columnas, que describen los datos que contiene el origen de datos. Estas columnas contienen información respecto al tipo de datos, el tipo de contenido y el modo en que se distribuyen los datos.

Con los datos y la segmentación propuesta en el párrafo anterior, se ha preparado una estructura de minería de datos, considerando los siguientes atributos descritos en la Tabla 12, con el objeto de aplicar diferentes modelos de minería de datos, y evaluar el comportamiento de cada modelo.

Atributo	Uso	Tipo de datos	Tipo de contenido	Valores
Edad	Entrada	Long	Continuo	17 - 34
ESTADO	Sólo predicción	Text	Discreta	Desertor Titulado
Estado_Civil	Entrada	Text	Discreta	CASADO SOLTERO
Modalidad	Entrada	Long	Discreta	1 2 3
Nacionalidad	Entrada	Text	Discreta	CHILENO EXTRANJERO
PROM_NEM	Entrada	Double	Continuo	4.6 - 6.7
PROM_PSU	Entrada	Long	Continuo	303 - 786
RUT	Entrada	Long	Clave	
SEXO	Entrada	Text	Discreta	F M
Sit_econo	Entrada	Long	Discreta	1 2 3

Tabla 12 - Atributos para la estructura de minería de datos

En la Tabla 13 se presenta la estructura de minería de datos, y los parámetros asociados, para la base de datos de alumnos:

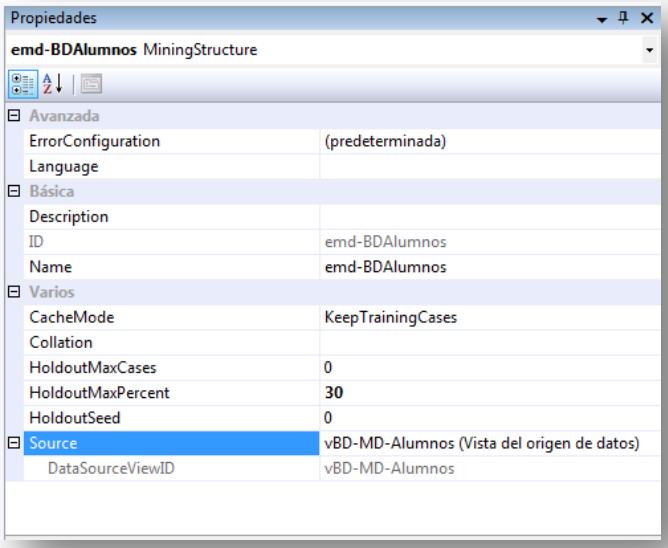

Tabla 13 - Estructura de minería de datos BDAlumnos

Donde:

- ***HoldoutMaxCases*** = 0: Especifica el número máximo de casos en el origen de datos que se van a utilizar en la partición de exclusión que contiene el conjunto de pruebas para la estructura de minería de datos emd-BDAlumnos. Los casos restantes en el conjunto de datos se usan para el entrenamiento. Un valor 0 indica que no hay ningún límite con respecto al número de casos que se pueden considerar como el conjunto de pruebas.
- **HoldoutMaxPercent** = 30: Especifica el porcentaje máximo de casos en el origen de datos que se van a usar en la partición de exclusión que contiene el conjunto de pruebas para la estructura de minería de datos emd-BDAlumnos. Los casos restantes se usan para aprendizaje. Un valor 0 indica que no hay ningún límite con respecto al número de casos que se pueden considerar como el conjunto de pruebas.
- Si especifica los valores de HoldoutMaxPercent y HoldoutMaxCases el algoritmo limita el conjunto de pruebas al menor de los dos valores.
- Si HoldoutMaxCases está establecido en el valor predeterminado de 0 y no se ha establecido un valor para HoldoutMaxPercent, el algoritmo utiliza el conjunto de datos completo para entrenamiento.

Una vez definida la estructura de minería de datos (EMD-Tesis), se procede a definir los modelos a desarrollar en esta investigación, esto es: Árbol de decisión, Cluster K-medianas y

Redes Neuronal, con los atributos indicados en la Tabla 12. A continuación se presentan los modelos propuestos con sus respectivas características:

5.4.1. Modelo de Red neuronal Artificial

En este proyecto se utiliza el algoritmo de red neuronal de Microsoft que combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente, usa estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción (Clase) basándose en los atributos de entrada.

Los modelos de minería de datos construidos con el algoritmo de red neuronal de Microsoft, pueden contener varias redes, en función del número de columnas que se utilizan para la entrada y la predicción, o sólo para la predicción. El número de redes que contiene un único modelo de minería de datos depende del número de estados que contienen las columnas de entrada y las columnas de predicción que utiliza el modelo.

Nombre del modelo	Red neuronal
Descripción del modelo	
Algoritmo	Microsoft_Neural_Network
Procesado por última vez	10-10-2011 20:05
Número de nodos de entrada	15
Número de nodos de salida	2
Número de nodos ocultos	21

Tabla 14 - Información sobre el modelo de Red Neuronal

El algoritmo evalúa y extrae los datos de entrenamiento del origen de datos. Un porcentaje de los datos de entrenamiento, denominado *datos de exclusión*, se reserva para evaluar la precisión de la red. Durante el proceso de entrenamiento, la red se evalúa de forma inmediata después de cada iteración mediante los datos de entrenamiento. Cuando la precisión deja de aumentar, el proceso de entrenamiento se detiene. La tabla Tabla 14, presenta la información establecida para el modelo de Red Neuronal.

El algoritmo de red neuronal de Microsoft crea modelos de minería de datos de regresión y de clasificación mediante la generación de una red de perceptrón multicapa de neuronas.

La tabla Tabla 15, muestra los valores de los parámetros utilizados para el algoritmo de Red Neuronal, para mas detalles sobre estos parámetros ver en Anexo - Referencia técnica del algoritmo de red neuronal.

Nombre	Valor
HIDDEN_NODE_RATIO	4
HOLDOUT_PERCENTAGE	30
HOLDOUT_SEED	0
MAXIMUM_INPUT_ATTRIBUTES	255
MAXIMUM_OUTPUT_ATTRIBUTES	255
MAXIMUM_STATES	100
SAMPLE_SIZE	10000

Tabla 15 - Parámetros del algoritmo Red Neuronal

5.4.2. Modelo de Árbol de decisión

El algoritmo de árboles de decisión de Microsoft genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

Nombre del modelo	Árboles de Decisión
Descripción del modelo	
Algoritmo	Microsoft_Decision_Trees
Procesado por última vez	10-10-2011 20:03
Número de árboles	1
Número de nodos en un árbol 'ESTADO'	3

Tabla 16 – Información del modelo Árbol de decisión

El algoritmo de árboles de decisión de Microsoft utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos de Analysis Services utilizan la selección de características para mejorar el rendimiento y la calidad del análisis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Entre los métodos que se usan para determinar si hay que dividir el árbol figuran métricas estándar del sector para la entropía y las redes Bayesianas. La Tabla 17 - Parámetros del algoritmo Árbol de decisión presenta los valores utilizados para el modelo Árbol de decisión., para mas detalles sobre los parámetros del algoritmo, ver Anexo - Referencia técnica algoritmo de árboles de decisión.

Parámetros de algoritmo	
Nombre	Valor
COMPLEXITY_PENALTY	0,5
FORCE_REGRESSOR	
MAXIMUM_INPUT_ATTRIBUTES	255
MAXIMUM_OUTPUT_ATTRIBUTES	255
MINIMUM_SUPPORT	10
SCORE_METHOD	4
SPLIT_METHOD	3

Tabla 17 - Parámetros del algoritmo Árbol de decisión

5.4.3. Modelo de Cluster

El algoritmo de clústeres de Microsoft es un algoritmo de segmentación suministrado por Analysis Services. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. La información del modelo de Cluster se presenta en la Tabla 18.

Nombre del modelo	Cluster K-Mediana
Descripción del modelo	
Algoritmo	Microsoft_Clustering
Procesado por última vez	10-10-2011 20:07
Número de clústeres	2
Compatibilidad con el clúster 'Cluster 1'	178
Compatibilidad con el clúster 'Cluster 2'	44
Promedio (probabilidad de filas)	0,511924386

Tabla 18 - Información del modelo Cluster

El algoritmo de clústeres de Microsoft proporciona dos métodos para crear clústeres y asignar puntos de datos a dichos clústeres. El primero, el algoritmo K-medianas, es un método de agrupación en clústeres duro. Esto significa que un punto de datos puede pertenecer a un solo clúster, y que únicamente se calcula una probabilidad de pertenencia de cada punto de datos de ese clúster. El segundo, el método Expectation Maximization (EM), es un método de agrupación en clústeres blando. Esto significa que un punto de datos siempre pertenece a varios clústeres, y que se calcula una probabilidad para cada combinación de punto de datos y clúster.

Parámetros de algoritmo	
Nombre	**Valor**
CLUSTER_COUNT	2
CLUSTER_SEED	0
CLUSTERING_METHOD	3
MAXIMUM_INPUT_ATTRIBUTES	255
MAXIMUM_STATES	100
MINIMUM_SUPPORT	1
MODELLING_CARDINALITY	10
SAMPLE_SIZE	50000
STOPPING_TOLERANCE	10

Tabla 19 - Parámetros del algoritmo Cluster

El algoritmo K-medianas proporciona dos métodos para realizar un muestreo en el conjunto de datos: K-medianas no escalable, que carga el conjunto de datos completo y realiza una pasada de agrupación en clústeres, y K-medianas escalable, donde el algoritmo usa los primeros 50.000 casos y lee más casos únicamente si necesita más datos para lograr un buen ajuste del modelo a los datos. Los valores para los parámetros del algoritmo de Cluster K-medianas, se muestran en la Tabla 19, para más detalles sobre los parámetros del algoritmo ver Anexo - Referencia técnica del algoritmo de clústeres.

5.5. Fase de evaluación

La validación es el proceso de evaluar cuál sería el rendimiento de los modelos de minería de datos con datos reales. Es muy importante validar los modelos de minería entendiendo su calidad y sus características antes de implementarlos en un entorno de producción.

En general se utiliza la exactitud de la clasificación o la tasa de error para medir el desempeño de un modelo de clasificación en el conjunto de pruebas, según se describe en [6]. La exactitud de la clasificación se calcula a partir del conjunto de pruebas en el que también se puede utilizar para comparar el rendimiento relativo de los clasificadores diferentes en el mismo dominio. Sin embargo, con el fin de hacerlo, las etiquetas de clase de los registros de prueba deben ser conocidas. Por otra parte una metodología de evaluación es necesaria para evaluar el modelo de clasificación y calcular la precisión de la clasificación.

La validación cruzada es un método establecido para evaluar la exactitud de los modelos de minería de datos. La validación cruzada, divide sucesivamente los datos de la estructura de minería en subconjuntos, genera modelos en los subconjuntos y, a continuación, mide la exactitud del modelo para cada partición. Revisando las estadísticas devueltas, puede determinar el grado de confiabilidad del modelo de minería de datos y comparar más fácilmente los modelos que se basan en la misma estructura.

5.5.1. Evaluación Red Neuronal

Para ver cómo el modelo pone en correlación las variables de entrada con la variable de predicción ESTADO, se utiliza el Visor de redes neuronales seleccionando los estados concretos de atributos de entrada. Considerando todas las variables de entrada, y como se aprecia en Tabla 20, se tiene que la modalidad 3 (executive) y modalidad (2) vespertinos tiene un impacto favorable a “Titulado”, sobre la modalidad de estudios diurna que favorece “desertor”

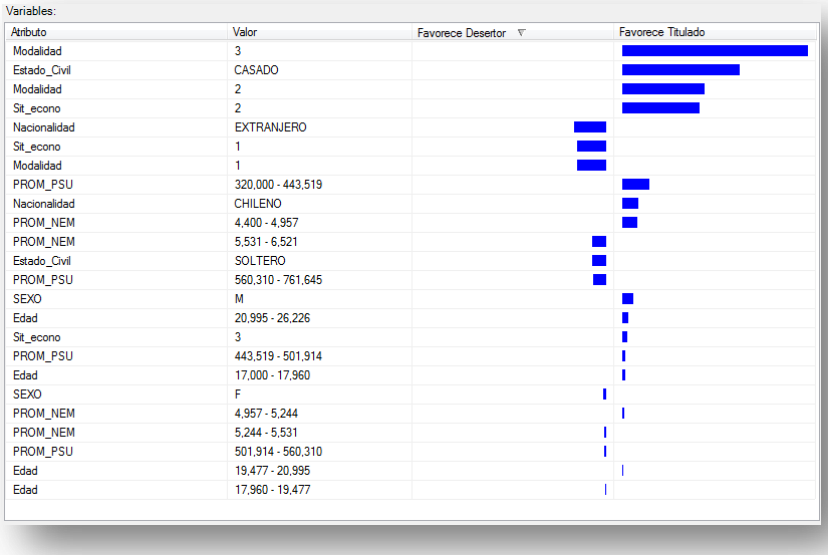

Tabla 20 - Visor de Redes Neuronales

El modelo de red neuronal, presenta una estimación para el atributo *Estado*, con una predicción sin errores de un 64%, (ver Tabla 21), sin embargo, tiene un valor negativo para logaritmo y la mejora del modelo predictivo, que lo hace un modelo con una predicción peor que la predicción aleatoria.

Información de la columna de resultados para 'ESTADO'	
Tipo de contenido	Discreta
Sin errores	64 %
Error	36 %
Logaritmo	-0,649887621
Mejora respecto al modelo predictivo	-0,026489755

Tabla 21 - Resultados Red neuronal

5.5.2. Evaluación Árbol de decisión

Cuando se crea un modelo de árbol de decisión ver Tabla 22, se genera un árbol independiente por cada atributo de predicción. Un árbol de decisión se compone de una serie de divisiones, con la división más importante, determinada por el algoritmo, a la izquierda del visor en el nodo. *Todos*. Las divisiones adicionales se muestran a la derecha. La división del nodo Todos es la más importante porque contiene la condición más determinante de división del conjunto de datos y, por tanto, la que ocasiona la primera división asociada a la variable sexo, influyendo en la variable de resultado *Estado*.

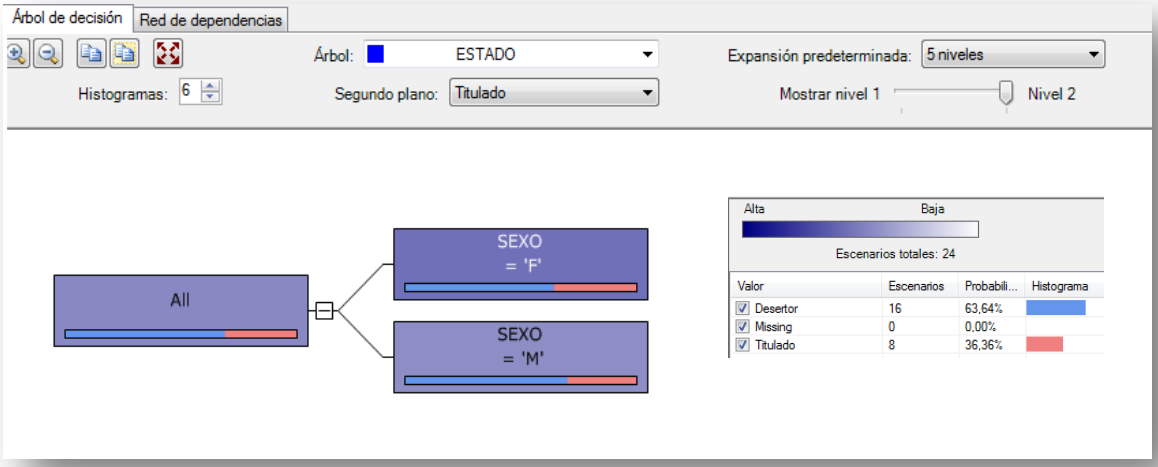

Tabla 22 - Visor de Árboles de decisión

En la Tabla 23, se presenta la estimación para el atributo *Estado* del modelo Árbol de decisión con una predicción sin errores de un 68%, manteniendo un valor negativo para logaritmo y medida para la mejora del modelo predictivo, lo hace un modelo con una predicción mejor que el modelo Red Neuronal.

Información de la columna de resultados para 'ESTADO'	
Tipo de contenido	Discreta
Divisiones importantes	SEXO
Sin errores	68 %
Error	32 %
Logaritmo	-0,623163402
Mejora respecto al modelo predictivo	0,000234453

Tabla 23 - Información de la columna Estado para árbol de decisión

5.5.3. Evaluación Cluster

El Visor de clústeres muestra los modelos de minería de datos que se generan con el algoritmo de agrupación en clústeres de Microsoft. Este es un algoritmo de segmentación que se utiliza para explorar datos con el fin de identificar anomalías en ellos y crear predicciones.

La Tabla 24, presenta una vista general de los clústeres que crea el modelo. Esta vista muestra cada atributo, junto con la distribución del atributo en cada clúster. Un recuadro informativo por cada celda muestra las estadísticas de la distribución y otro por cada encabezado de columna muestra el llenado del clúster.

Los atributos discretos se muestran como barras de color y los atributos continuos se muestran como un gráfico en forma de rombo que representa la media y la desviación estándar de cada cluster.

Tabla 24 - Vista general de Cluster

La Tabla 25 - Información para la columna Estado para Cluster, presenta las estimaciones para el atributo *Estado* del modelo Cluster K-mediana con una predicción sin errores de un 68%, con valores similares al modelo Árbol de decisión.

Información de la columna de resultados para 'ESTADO'	
Tipo de contenido	Discreta
Sin errores	68 %
Error	32 %
Logaritmo	-0,622316539
Mejora respecto al modelo predictivo	0,00108132

Tabla 25 - Información para la columna Estado para Cluster

5.5.4. Evaluación comparada de los algoritmos

En esta sección se presentan las estrategias para la validación de los modelos aplicados en esta investigación. No hay ninguna regla completa que pueda indicarle si un modelo es suficientemente bueno, o si cuenta con suficientes datos. En general las medidas de minería de datos pertenecen a las categorías de precisión, confiabilidad y utilidad.

La ***precisión*** es una medida que indica hasta qué punto el modelo pone en correlación un resultado con los atributos de los datos que se han proporcionado. Existen varias medidas de precisión, pero todas ellas dependen de los datos que se utilicen. En realidad, podrían faltar valores o estos ser aproximados, o incluso diferentes procesos podrían cambiar los datos. En particular, en la fase de exploración y desarrollo, podría decidir aceptar una cierta cantidad de errores en los datos, sobre todo si éstos son suficientemente uniformes en sus características. Por tanto, es necesario equilibrar las mediciones de precisión mediante las valoraciones de confiabilidad.

La ***confiabilidad*** evalúa la manera en la que se comporta un modelo de minería de datos en conjuntos de datos diferentes. Un modelo de minería de datos es confiable si genera el mismo tipo de predicciones o encuentra los mismos tipos generales de patrones independientemente de los datos de prueba que se proporcionen.

La ***utilidad*** incluye diferentes métricas que le indican si el modelo proporciona información útil. También podría descubrir que un modelo que, de hecho parece correcto, no tiene sentido porque está basado en correlaciones cruzadas de los datos.

Se separaran los datos en conjuntos de datos de entrenamiento y pruebas, para evaluar con precisión el rendimiento de todos los modelos con los mismos datos. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de prueba para comprobar la precisión del modelo mediante la creación de consultas de predicción

De los 317 registros seleccionados aleatoriamente para el desarrollo de los modelos, se reservaron para el proceso de prueba un 30% es decir 95 registros, utilizando el resto de los datos (222 registros), para el entrenamiento. Una vez completado el modelo, éste se utiliza para realizar las predicciones en función del conjunto de prueba. Dado que los datos del conjunto de entrenamiento se seleccionan de forma aleatoria a partir de los mismos datos utilizados para el

entrenamiento, es poco probable que las métricas de precisión que se derivan de la prueba se vean afectadas por discrepancias en los datos, y por tanto, reflejarán mejor las características del modelo.

5.5.4.1. Validación cruzada

Permite particionar un conjunto de datos en muchas secciones transversales de menor tamaño y crear varios modelos en dichas secciones para probar la validez del conjunto de datos completo.

Los datos se dividen en particiones, cada una se utiliza a su vez como datos de pruebas, mientras que los datos restantes se utilizan para entrenar un nuevo modelo. A continuación en la *Tabla 26 – Estimaciones de validación Cruzada de los modelos* en estudio, se presenta el resumen de las medidas de precisión detalladas para cada partición para los modelos Árbol de decisión, Red Neuronal y Cluster K-medianas. Al comparar las medidas de los modelos generados para cada sección transversal, puede hacerse una idea del grado de confiabilidad del modelo de minería con respecto a todo el conjunto de datos

Índice de partición	Tamaño de partición	Prueba	Medida	Árbol de Decisión	Red Neuronal	Cluster K-Mediana
1	14	Classification	Pass	10	9	7
2	15	Classification	Pass	10	10	5
3	16	Classification	Pass	10	7	9
4	16	Classification	Pass	11	9	9
5	15	Classification	Pass	7	9	8
			Promedio	9,6184	8,7763	7,6447
			Desviación estándar	1,3559	0,9947	1,5018
1	14	Classification	Fail	4	5	7
2	15	Classification	Fail	5	5	10
3	16	Classification	Fail	6	9	7
4	16	Classification	Fail	5	7	7
5	15	Classification	Fail	8	6	7
			Promedio	5,6184	6,4605	7,5921
			Desviación estándar	1,3374	1,5082	1,194
1	14	Likelihood	Log Score	-0,5949	-0,6903	-0,6907
2	15	Likelihood	Log Score	-0,5936	-0,9564	-0,7503
3	16	Likelihood	Log Score	-0,6496	-0,8421	-0,691
4	16	Likelihood	Log Score	-0,6085	-0,8999	-0,6858
5	15	Likelihood	Log Score	-0,7564	-0,7042	-0,6931
			Promedio	-0,6409	-0,8216	-0,702
			Desviación estándar	0,0608	0,1042	0,0241
1	14	Likelihood	Lift	0,088	-0,0074	-0,0078
2	15	Likelihood	Lift	0,0973	-0,2655	-0,0593
3	16	Likelihood	Lift	0,0357	-0,1568	-0,0057
4	16	Likelihood	Lift	0,0768	-0,2146	-0,0004
5	15	Likelihood	Lift	-0,0655	-0,0132	-0,0022
			Promedio	0,0462	-0,1346	-0,0149
			Desviación estándar	0,0593	0,1035	0,0222
1	14	Likelihood	Root Mean Square Error	0,3758	0,3051	0,4444
2	15	Likelihood	Root Mean Square Error	0,4092	0,1533	0,4285
3	16	Likelihood	Root Mean Square Error	0,3588	0,3168	0,4386
4	16	Likelihood	Root Mean Square Error	0,3794	0,2877	0,4402
5	15	Likelihood	Root Mean Square Error	0,3912	0,368	0,429
			Promedio	0,3826	0,2864	0,4361
			Desviación estándar	0,0169	0,0712	0,0062

Tabla 26 – Estimaciones de validación Cruzada de los modelos en estudio

En la estimación de validación cruzada de la Tabla 26, se presentan las 5 particiones y los resultados de las distintas métricas (Clasificación sin errores, Clasificación errónea, y las probabilidades puntuación de registro (Log score), elevación y error cuadrático medio). Para cada una de ellas, se entrega en el resumen de cada partición los siguientes indicadores: el promedio y la desviación estándar, calculados con la siguiente fórmula:

$$\sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

Clasificación sin errores

Esta métrica representa el recuento de casos clasificados correctamente y de acuerdo a los resultados obtenidos indica que el modelo *Árbol de decisión*, presenta en promedio una mejor probabilidad de predicción con un 63%, sin embargo, la desviación estándar (1,36) es más alta que la obtenida con el modelo *Red Neuronal* (0,99), que presenta una probabilidad de 58%.

Clasificación errónea

Respecto a la clasificación errónea, el modelo Árbol de decisión presenta una mejor desviación estándar (1,34) que el modelo *Red Neuronal* (1,51), el modelo *Cluster K-mediana*, con una desviación estándar de 1,19, presenta una probabilidad 50%.

Probabilidad

Los indicadores asociados a la probabilidad, se indican a continuación:

Puntuación del registro, Esta medida representa la proporción entre dos probabilidades, convertido a una escala logarítmica. Los tres modelos estudiados, presentan valores negativos para esta métrica, lo que significa que la predicción es peor que la predicción aleatoria. Según se puede apreciar que el modelo *Árbol de decisión* (-0,64), presenta en promedio una estimación más cercana a la predicción aleatoria, que los modelos *Cluster K-mediana (-0,82) y Red Neuronal (-0,70)*

Elevación, Este indicador representa la proporción entre la probabilidad de predicción real y la probabilidad marginal en los casos de prueba y muestra hasta qué punto mejora la probabilidad cuando se utiliza el modelo. En esta medida, se tiene que el modelo *Árbol de decisión (0,05)*, presenta en promedio una mejor estimación entre la probabilidad de predicción real y la probabilidad marginal en los casos de prueba, respecto a los modelos *Cluster K-mediana (-0,13) y Red Neuronal (-0,01)*

Error cuadrático medio, Este indicador corresponde a la Raíz cuadrada del error promedio para todos los casos de partición, dividido por el número de casos en la partición. Según los valores presentados en la Tabla 26, tenemos que el modelo *Red Neuronal (0,29),* tiene un indicador mejor que *Cluster K-mediana (0,38) y Árbol de decisión (0,44),* sin embargo los modelos, *Cluster K-mediana (0,01) y Árbol de decisión (0,02),* tienen una menor desviación estándar para este indicador que *Red Neuronal (0,07).*

5.5.4.2. Gráfico de elevación

Un ***gráfico de elevación*** es un método para visualizar la mejora que se obtiene al utilizar un modelo de minería de datos, si lo compara con una estimación aleatoria.

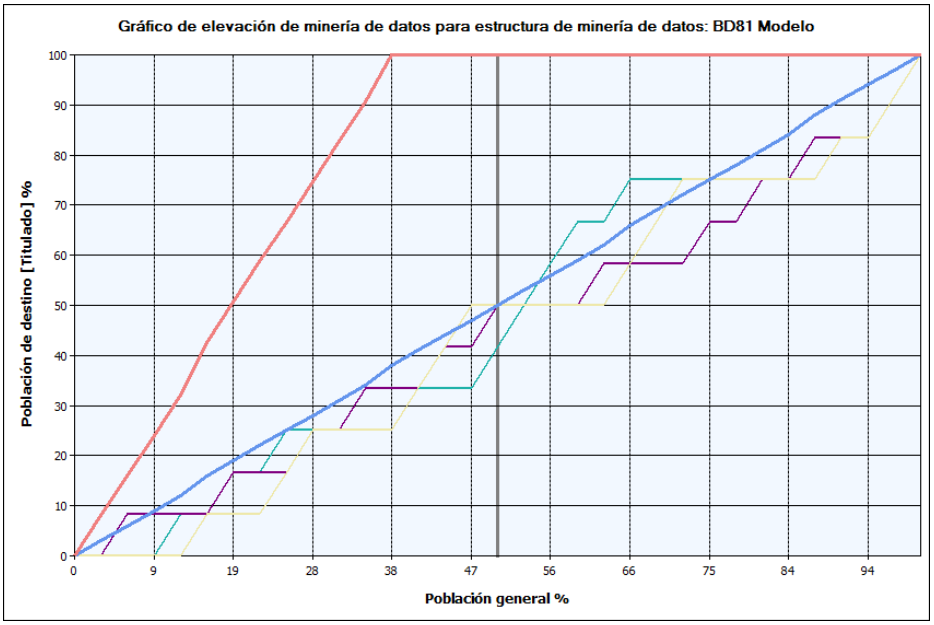

Figura 8 - Gráfico de elevación para los modelos en estudio Estado = Titulado

En la Figura 8 - Gráfico de elevación para los modelos en estudio Estado = Titulado , el atributo de destino es [Estado] y el valor de destino es Titulado, lo que representa que el estudiante es probable que se Titule. El gráfico de elevación muestra así la mejora que el modelo proporciona al identificar a los estudiantes que es probable que se Titulen.

Leyenda de minería de datos

Porcentaje de población: 50,00%

Serie, Modelo	Puntuación	Población de destino	Probabilidad de predicción
Árbol de Decisión	0,56	41,67%	43,94%
Red Neuronal	0,55	50,00%	45,95%
Cluster K-Mediana	0,54	50,00%	58,33%
Modelo de estima...		50,00%	
Modelo ideal para...		100,00%	

Tabla 27 - Leyenda grafico de elevación para ESTADO = Titulado

El eje X del gráfico representa el porcentaje del conjunto de datos de prueba que se usa para comparar las predicciones. El eje Y del gráfico representa el porcentaje de valores que se predicen con el Estado = “Titulado”. En el gráfico, la línea azul representa la línea aleatoria y la roja el modelo ideal.

En la *Tabla 27 - Leyenda grafico de elevación para ESTADO = Titulado*, el valor de **Probabilidad de predicción** representa el umbral necesario para incluir un estudiante entre los

casos "con probabilidad de Titularse". Para cada caso, el modelo calcula la exactitud de cada predicción y almacena ese valor, que puede utilizar para filtrar o elegir estudiantes.

El valor de **Puntuación** ayuda a comparar los modelos calculando la efectividad del modelo a través de una población normalizada. La mayor puntuación, representando el mejor modelo la obtiene *Árbol de decisión*, con un puntaje de 0,56, siguiendo Red Neuronal (0,55) y Cluster K-mediana (0,54) .

5.5.4.3. Matriz de clasificación

Una ***matriz de clasificación*** es un método para ordenar las estimaciones buenas y malas en una tabla, para analizar con qué precisión predice el modelo el valor de destino.

'Árboles de Decisión'				
	Desertor(Real)	Titulado(Real)	Desertor(Real)	Titulado(Real)
Desertor	**100,00 %**	100,00 %	**62**	33
Titulado	0,00 %	**0,00 %**	0	**0**
Correcta	**100,00 %**	**0,00 %**	**62**	**0**
Incorrecta	**0,00 %**	**100,00 %**	**0**	**33**

Cluster K-Mediana'				
	Desertor(Real)	Titulado(Real)	Desertor(Real)	Titulado(Real)
Desertor	**100,00 %**	100,00 %	**62**	33
Titulado	0,00 %	**0,00 %**	0	**0**
Correcta	**100,00 %**	**0,00 %**	**62**	**0**
Incorrecta	**0,00 %**	**100,00 %**	**0**	**33**

Red neuronal'				
	Desertor(Real)	Titulado(Real)	Desertor(Real)	Titulado(Real)
Desertor	**85,48 %**	90,91 %	**53**	30
Titulado	14,52 %	**9,09 %**	9	**3**
Correcta	**85,48 %**	**9,09 %**	**53**	**3**
Incorrecta	**14,52 %**	**90,91 %**	**9**	**30**

Tabla 28 – Matriz de Clasificación para los modelos en Estudio

Para generar una matriz de clasificación, se cuenta el número de predicciones buenas y erróneas, utilizando los valores reales existentes en el conjunto de datos de prueba. La matriz es una herramienta valiosa porque no sólo muestra la frecuencia con que el modelo predice un valor correctamente, sino que también muestra qué valores predice incorrectamente. Una matriz de clasificación muestra el recuento real de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada atributo de predicción, la Tabla 28, presenta los resultados para la matriz de clasificación

Del total de 95 registros que se seleccionaron para realizar el test a los modelos, se aprecia que el modelo *Árbol de decisión y Cluster K-mediana* clasifican correctamente al 100% de los estudiantes *Desertores*, sin embargo fallan en el 100% de los estudiantes *Titulados*, el modelo *Red Neuronal*, logra 85,48% de clasificación correcta para los *Desertores*, y un 9,09% de clasificación correcta para los *Titulados*.

5.5.4.4. Precisión y Alcance

La Tabla 29, representa las medidas de precisión y alcance (precision and recall) para los modelos en estudio. Estas medidas están basadas en la matriz de clasificación (matriz de confusión). La precisión mide la probabilidad que si el modelo clasifica a un estudiante en una categoría, el estudiante realmente pertenezca a dicha categoría. Por otro lado el alcance mide la probabilidad de que si un estudiante pertenece a una categoría, el modelo lo asigne a la categoría. Sin embargo para comparar modelos, es común utilizar F-Measure correspondiente a la media armónica de la precisión y el alcance.

	Árbol Decisión	**K-Mediana**	**Red Neuronal**
Precision	65%	65%	64%
Recall	100%	100%	85%
Accuracy	65%	65%	59%
F-Measure	**79%**	**79%**	**73%**

Tabla 29 - F-Measure para los modelos en estudio

De acuerdo a los resultados obtenidos, claramente los modelos **Árbol Decisión** y **K-Mediana**, no logran discriminar adecuadamente (todo es desertor), y el modelo **Red Neuronal**, al menos detecta titulados.

6. Discusión.

Del análisis anterior podemos deducir que los datos facilitados no nos permiten obtener una predicción satisfactoria. Más aún al analizar la influencia de cada una de las variables en forma independiente como se muestra en la Tabla 30, en donde se observa que no existe una correlación entre aquellas variables respecto al atributo Estado (la clase).

Attributes	SEDE	NOMBRE	Edad	SEXO	Nacionalidad	Estado_Civil	Modalidad	PROM_PSU	PROM_NEM	Sit_econo	ESTADO
SEDE	1	0.409	0.116	-0.011	-0.213	-0.037	-0.028	0.049	-0.046	0.072	0.077
NOMBRE	0.409	1	-0.050	0.008	-0.093	0.001	-0.021	0.026	0.150	-0.081	-0.031
Edad	0.116	-0.050	1	0.033	-0.091	0.119	0.095	0.079	-0.272	-0.055	0.055
SEXO	-0.011	0.008	0.033	1	0.010	-0.025	0.028	0.043	-0.143	0.064	-0.081
Nacionalidad	-0.213	-0.093	-0.091	0.010	1	0.002	0.105	0.048	0.027	-0.047	-0.030
Estado_Civil	-0.037	0.001	0.119	-0.025	0.002	1	-0.015	0.048	-0.033	-0.010	0.019
Modalidad	-0.028	-0.021	0.095	0.028	0.105	-0.015	1	0.056	-0.060	-0.024	0.058
PROM_PSU	0.049	0.026	0.079	0.043	0.048	0.048	0.056	1	0.244	0.025	0.010
PROM_NEM	-0.046	0.150	-0.272	-0.143	0.027	-0.033	-0.060	0.244	1	0.025	0.041
Sit_econo	0.072	-0.081	-0.055	0.064	-0.047	-0.010	-0.024	0.025	0.025	1	-0.027
ESTADO	0.077	-0.031	0.055	-0.081	-0.030	0.019	0.058	0.010	0.041	-0.027	1

Tabla 30 - Matriz de Correlación

Sin embargo, es conocido la influencia del atributo NEM [30] en el desempeño académico de los estudiantes luego de la educación media. Esto nos hace sospechar fuertemente que estos datos podrían no haber sido recabados con la debida rigurosidad. Lamentablemente, en el marco de este trabajo no fue posible comprobar la veracidad de los datos, dado que estos fueron entregados por el área de sistemas de la universidad sin tener la posibilidad de corroborarlos.

Esta investigación ha permitido descubrir que el enfoque de minería de datos guiado por el dominio (Domain Driven Data Mining (D3M)), ha tomado relevancia en estos últimos años [31], y bajo este contexto, la ***minería de datos educativos*** es un campo de investigación que está ganando popularidad según se puede apreciar en los diferentes trabajos publicados recientemente sobre este tema [5] [26][32], por lo que hay modelos y resultados en esta área que pueden ser aplicados para futuros trabajos sobre el desempeño académico de los estudiantes.

La minería de datos educativos (Educational Data Mining, EDM) está orientada al desarrollo de métodos para explorar los tipos únicos de los datos que provienen de los centros educativos, tales como datos administrativos de matriculas en colegios, institutos o universidades, expedientes académicos informatizados de los estudiantes, registro de actividad en portales educativos (plataformas de e-learning), sistemas de aprendizaje colaborativo guiados por

computador, entre otras fuentes de información y el uso de los métodos para transformar dichos datos en información para entender mejor el proceso de aprendizaje de los estudiantes buscando la mejora de la calidad y la rentabilidad del sistema educativo.

7. Recomendaciones para minería de datos educativos.

La minería de datos educativa tiene necesidades específicas que no se presentan en otros dominios. La que se centra en la identificación, extracción y evaluación de variables relacionadas con el proceso de aprendizaje de los estudiantes según lo descrito por Alaa el-Halees [26].

Como se estableció desde un comienzo el presente proyecto ha tenido como objetivo desarrollar un indicador que permita clasificar en forma automática a los estudiantes con mayor riesgo de deserción, al momento que se matricula en la institución. Sin embargo, de acuerdo a la experiencia adquirida a través del trabajo de investigación con los datos de los estudiantes pre grado de la facultad de ingeniería de la Universidad de las Américas, y los estudiantes de post-grado de otra institución, se puede apreciar que los sistemas de información actuales de estas instituciones, no están orientados a capturar las variables relevantes para la realización de estudios de minería de datos educativos asociados a la deserción.

Dado lo anterior, en este trabajo proponemos un conjunto de atributos asociados al problema de deserción universitaria [8] agrupados en factores individuales, académicos, institucionales y socioeconómico, que se recomienda sean capturados al momento que el estudiante se matricula en la institución, con el objeto poder crear un repositorio de registro histórico de las condiciones de entrada, completándolo con el resultado del desempeño de los estudiantes en su carrera, indicando si es desertor o no. La Tabla 31 resume los atributos que a nuestro juicio deberían captar estas instituciones de manera de tener la oportunidad de hacer un estudio serio de las variables que identifican a los alumnos que desertan de la carrea. Esta tabla entre otros factores, está inspirada en el trabajo desarrollado por Martha Artunduaga sobre variables que influyen en el rendimiento académico en la universidad [33], En este estudio se destacan entre otros, los factores socioeconómico familiar y nivel educativo de los padres:

Tipo	Atributo	Descripción	Posibles valores
Individuales	RUT	Identificador de la tupla	Los numeros de rut
	Nombre	Nombre y apellidos del estudiante	Todos los nombre posibles
	FecNac	Fecha de nacimiento del estudiante	rango de fechas posibles
	Sexo	Género del estudiante	M (masculino), F (femenino)
	Comuna	Comuna de residencia del estudiante	Las comunas de la region Metropolitana
	LugarDeVivienda	indicador si el estudiante vive con la famila, solo, o con otros	ConFamilia, con Amigos, solo
	PracticaDeporte	Indicador si el estudiante practica deporte con Regularidad	Si, No
	HabitosConsumo	Habitos de Consumo	Fuma, Bebe, ambos, no aplica
Académicos	FeIngreso	Fecha Ingreso a la Universidad	Fecha valida en el rango de estudio
	Prom_PSU	Promedio PSU	rango de puntages 200 a 900
	Prom_NEM	Promedio NEM	Rango de notas 4 a 7
Institucionales	SedeUniversidad	Sede donde estudia el alumno	LF, SC, PR, MP, CO, VL
	ComunaUniversidad	Comuna de la Sede de la universidad	Las comunas de la region Metropolitana
	Colegio	Colegio Origen	Nombre de los colegios
	ComunaColegio	Comuna Colegio	comunas de Chile
	TipoColegio	Tipo de administración del Colegio	Municipal, Particular Subencionado, Particular
	TipoEnseñanza	Tipo de enseñanza del Colegio	Científico-Humanista (EMCH), Técnico-
	PuntajeSIMCEColegio	Promedio SIMCE	resultados prueba SIMCE Colegio
Socio económicos	GrupoSocioeconomicoFamiliar	Nivel socioeconómico	Bajo, Medio Bajo, Medio, Medio alto, Alto
	NeduMadre	Nivel educativo de la Madre	Primarios o Sin estudios, Secundarios,
	NeduPadre	Nivel educativo de la Padre	Primarios o Sin estudios, Secundarios,
	TrabE	Trabajo del estudiante	Si, No
	Estado	El estado a predecir (La clase)	Titulado, Desertor

Tabla 31 - Variables a considerar en estudios futuros

Grupo socioeconómico familiar. El rendimiento académico y el porcentaje de culminación de estudios universitarios, está relacionado con el origen sociocultural de la familia. Se afirma que las tasas de éxito para los estudiantes de medios favorecidos, es superior a los de origen modesto, siendo estos últimos los que presentan mayores índices de abandono de sus estudios. Con ello, se afirma que vivir en entornos pobres es un factor de riesgo de fracaso escolar.

En el mismo estudio se indica que el **nivel educativo de los padres** influye en el rendimiento académico de los hijos. Las investigaciones han demostrado que cuando la madre ha realizado estudios universitarios, los estudiantes alcanzan mejores resultados académicos.

Respecto a las variables cognoscitivas, que tienen que ver con el rendimiento académico previo del alumno tenemos dos indicadores: las Notas de Enseñanza Media (**NEM**) y resultados de la Prueba de Selección Universitaria (**PSU**). El trabajo desarrollado por Fischer y Repeto [30], ponen de manifiesto que las notas de enseñanza media tienen una capacidad predictiva importante, sin embargo, sería bueno, establecer un mecanismo de “normalización” de este indicador, de acuerdo al resultado SIMCE del colegio que egresa el estudiante, con alguna ponderación sobre los años de estudio en dicha institución

En el trabajo de [5], se incorpora el atributo **lugar de residencia,** con el dominio (Familia, Amigos, Solo), entre otros atributos para construir un sistema basado en reglas, que permite predecir el resultado de nota final en un curso bajo estudio.

La metodología propuesta en este trabajo, basada en la metodología CRISP-DM sugiere dar especial atención a las actividades indicadas en la Tabla 32, asociadas a los pasos propuestos para el desarrollo de proyectos de minería de datos educativos.

Análisis de datos	Preparación de datos	Modelado	Evaluación
Recolección inicial de datos *Capturar al Momento de Inscripción del estudiante en la institución* **Verificar la calidad de los datos** *Establecer un mecanismo de acceso a otras bases de datos para validar la calidad de los datos ingresados, tales como PSU y NEM*	**Integrar datos** *Establecer una integración con los resultados SIMCE de los colegios de egreso de los postulantes, con el objeto de ponderar el atributo NEM.*	**Selección de la técnica de modelado** *Árbol de decisión* *Redes Neuronales* *Cluster k-Medianas*	**Evaluar en Modelo** *Considerar que el atributo NEM, tiene una capacidad predictiva sobre el rendimiento de los estudiantes*

Tabla 32 - Metodología propuesta

Las recomendaciones antes detalladas, están orientadas a mejorar el proceso de toma de datos al momento que el estudiante se matricula, y poder validar aquella información con organismos externos tales como el Departamento de Evaluación, Medición y Registro Educacional, DEMRE y el Sistema de Medición de Calidad de la Educación SIMCE con el objeto de mejorar las predicciones en los futuros estudios.

8. Conclusiones

A continuación se resumen las principales ideas generadas durante el desarrollo del presente trabajo. A partir de ellas se podrán reconocer con qué nivel fueron alcanzados los objetivos propuestos. Adicionalmente se presentan las propuestas para las mejoras en el desarrollo de proyectos de minería de datos educativa.

Esta investigación es un punto de partida en la aplicación sistemas de soporte a las decisiones basados en modelos de minería de datos, para analizar y evaluar los factores que influyen en la deserción universitaria. Los administradores de los establecimientos universitarios podrán usar la metodología propuesta por este trabajo para identificar y establecer procedimientos que permitan capturar en forma temprana la información de las variables relevantes para establecer los modelos de predicción para trabajar programas focalizados con el objeto de mejorar los índices de retención.

A continuación, se describen las principales conclusiones de este proyecto:

1) Se estudió el estado del arte en diferentes técnicas para desarrollar modelos de predicción, basados en sistemas de soporte a las decisiones, utilizando minería de datos, tales como Árboles de Decisión, Redes Neuronales y técnicas de clasificación, como Cluster K-mediana.
2) Se seleccionó un subconjunto de aquellos modelos que han presentado un mejor desempeño en esta área, para llevarlo al computador con el objeto de aplicarles las herramientas pertinentes, probando modelos para clasificar en forma automática a los alumnos con mayor riesgo de deserción y ha permitido tener una aproximación acerca del modelo más efectivo para esta labor, determinando que el modelo *Red Neuronal*, tiene un mejor comportamiento respecto a los modelos de *Árbol de decisión* y *Cluster K-mediana*, dado por los distintos indicadores, presentados en sección 5.5.4 Evaluación comparada de los algoritmos.
3) Sin embargo, los valores obtenidos, no permiten validar positivamente dichos modelos, dado que su capacidad de predicción es menor que la estimación aleatoria según se aprecia en la Figura 8 - Gráfico de elevación para los modelos en estudio Estado = Titulado, como en las estimaciones buenas y malas presentadas en la pruebas realizadas con los valores reales existentes en el conjunto de datos de prueba y presentados en la *Tabla 28 – Matriz de Clasificación para los modelos en Estudio*

4) En esta investigación se ha descubierto que en general las instituciones no recolectan la suficiente información referente a la caracterización del estudiante al momento de ingresar a estudiar, que permitan establecer modelos de predicción de retención.
5) Por lo anterior en este trabajo se presenta una metodología basada en recomendaciones para el desarrollo de proyectos de minería de datos educativa, proponiendo actividades y un conjunto de atributos que deben ser contemplados y capturados en forma temprana, para estudios posteriores.
6) Este trabajo permitió apreciar la importancia que tiene el proceso de recopilación de datos, abarcando las fases de análisis y preparación de los datos descrito en el apartado Proceso de extracción de conocimientos asociado a la metodología CRISP-DM.

Como trabajo futuro se propone recoger un gran conjunto de datos reales incorporando nuevas variables a la base de datos de estudiantes universitarios y aplicar los modelos a estos datos. Además, de aplicar otros métodos de clasificación para poner a prueba el método más adecuado que se adapte a la estructura de los datos de los estudiantes y dar una mejor precisión en la clasificación.

El conjunto de datos existente para el estudio y los atributos contenidos en ellos son fundamentales para poder lograr los niveles de predicción necesarios para la validación positiva de los modelos, por lo que se propone, profundizar en la determinación de las variables relevantes para el problema de la deserción y establecer un procedimiento de captura de dichas variables, al momento que el alumno se matricule en la institución.

9. Bibliografía

[1] Luis Eduardo González and Daniel Uribe, "Estimaciones sobre la "Repitencia" y deserción en la educación superior Chilena. Consideraciones sobre sus implicaciones," *Revista Calidad en la educación*, pp. 75-90, 2002.

[2] Comisión Acreditación UDLA, "Documento para la acreditación institucional de la Universidad de Las Américas," Santiago, Estudio 2008.

[3] Zbigniew Michalewicz, Martin Schmidt, Matthew Michalewicz Michalewicz, and Constantin Chiriac, *Adaptive Business Intelligence*. New York: Springer, 2007.

[4] José Hernández, Maíia José Ramírez, and César Ferri, *Introducción a la Minería de Datos*. Madrid: Pearson Educación S. A., 2004.

[5] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar, "Mining Student Data Using Decision Trees," in *The 2006 International Arab Conference on Information Technology (ACIT'2006)*, Jordan, 2006.

[6] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*. Boston: Pearson Addison Wesley, 2006.

[7] Centro de Microdatos, "Estudio Sobre Causas De La Deserción Universitaria," Santiago, Informe Final 2008.

[8] Universidad Nacional ICFES, "Estudio de la deserción estudiantil en la educación superior en Colombia," Bogotá, Documento sobre estado del arte 2002.

[9] Rafael Flores, "La deserción en los programas tecnológicos del SED, 1983 – 1986," Medellín, 1987.

[10] Alexander Quintero and Elmer Castaño, "DESERCIÓN ESTUDIANTIL EN EL PROGRAMA DE AGRONOMÍA DE LA UNIVERSIDAD DE CALDAS (1998-2006)," *Agronomía*, pp. 23-43, 2006.

[11] Gloria Salazar, "Aproximación a un análisis sobre la deserción académica, 1994-1997," Bogotá, Monografía para optar al Título de Especialista en Gerencia Social de la Educación 1999.

[12] Gabriel Páramo and Carlos Correa, "Deserción Estudiantil Universitaria. Conceptualización," Medellín, Revista Abril - Mayo – Junio 1999.

[13] ICFES, "La Educación Superior en Colombia – Década de Los 90," Bogotá, 2003.

[14] Ana Rocio Osorio and Catalina Jaramillo, "Deserción Estudiantil en los Programas de

] Pregrado 1995-1998," Medellín, Estudio 1999.

[15] Vincent Tinto, *Una reconsideración de las teorías de deserción estudiantil, Handbook of theory and research*. New York: Agathon Press, 1986.

[16] Paula Inés Giovagnoli, "Determinantes de la deserción y graduación universitaria: Una aplicación utilizando modelos de duración," Tesis de la Maestría en Economía de la UNLP dirigida por el Dr. Alberto Porto. 2002.

[17] Daniel Power. (2007, Marzo) A Brief History of Decision Support Systems. [Online]. http://DSSResources.COM/history/dsshistory.html

[18] Peter G. W. Keen and Michael S. Scott Morton, *Decision Support Systems: An Organizational Perspective*. MA: Addison-Wesley, 1978.

[19] Guillermo Matos, Ricardo Chalmeta, and Oscar Coltell. (2006) Metodología para la Extracción del Conocimiento Empresarial a partir de los Datos. [Online]. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642006000200011&lng=es&nrm=iso

[20] López Pérez and Daniel Santín, *Mineríia de Datos - Técnicas y Herramientas*. Madrid: Thomson, 2007.

[21] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, pp. 37-54, 1991.

[22] José Emilio Gondar Nores, *Redes Neuronales Artificiales*.: Data Mining Institute, 2001.

[23] Pete Chapman et al. (1999) CRISP-DM 1.0 Step-by-step data mining. [Online]. ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[24] Daniel T Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New York: John. Wiley & Sons, 2004.

[25] Basilio Sierra Araujo, *Aprendizaje Automático: Conceptos básicos y avanzados*. Madrid: Pearson Educación S. A., 2006.

[26] Alaa el-Halees, "Mining Students Data To Analyze Learning Behavior: A Case Study," Gaza, 2009.

[27] Simon Haykin, *Neural Networks*. New York: Macmillan College (IEEE Press Book), 1994.

[28] Ben Krose and Patrick Van der Smagt, *An Introduction to Neural Networks*. Amsterdam:

] University of Amsterdam, 1993.

[29 Juan Rabunal and Julian Dorrado, *Neural networks: algorithms, applications, and*] *programming techniques*.: Idea Group Reference, 2005.

[30 Ronald Fischer and Andrea Repetto, *Método de Selección y Resultados Académicos:*] *Escuela de Ingeniería de la Universidad de Chile*.: Estudios Públicos, 2003.

[31 " Domain-Driven Data Mining: Challenges and Prospect," in *IEEE Transactions on*] *Knowledge and* , 2010, pp. 755-76.

[32 Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students'] Performance," in *International Journal of Advanced Computer Science and Applications*, India, 2011, pp. Vol. 2, No. 6.

[33 Martha Artunduaga Murillo, "VARIABLES QUE INFLUYEN EN EL RENDIMIENTO] ACADÉMICO EN LA UNIVERSIDAD," Universidad Complutense de Madrid, Madrid, Doctoranda Universidad Complutense de Madrid 2008.

[34 Rodrigo Rolando, Juan Salamanca, and Alfredo Lara, "Retención de Primer Año en el] Pregrado:Descripción y Análisis de la cohorte de ingreso 2007," Santiago, 2010.

[35 Baker Ryan and Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and] Future Visions," *JEDM - Journal of Educational Data Mining*, vol. 1, no. 1, Octubre 2009.

10. Anexo – Redes neuronales artificiales

Este anexo presenta las funciones de activación y las topologías clásicas encontradas en las redes neuronales.

10.1. Funciones de activación en redes neuronales

Las funciones de activación más comunes son:

Función Lineal: Esta función es útil en aquellos casos en que el *output* es una variable continua o en aquellos en que se desea que la red aprenda los eventos menos frecuentes. A diferencia de la función Sigmoide, la lineal no se hace menos sensible al alejarse de cero.

La expresión matemática es:

$$F(x) = x$$

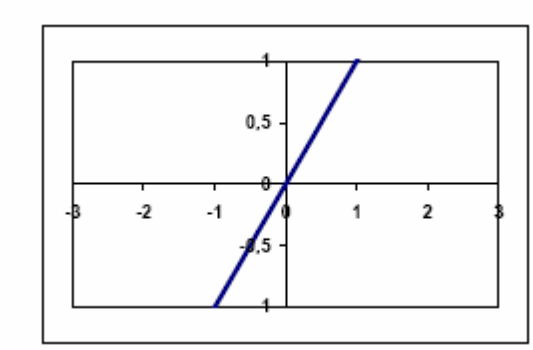

Figura 9 - Función lineal

Función Logística o Sigmoide: Esta es una de las funciones más comunes. Su rango es (0, 1). Se utiliza para concentrar el aprendizaje en valores no extremos en donde se deberían encontrar la mayor parte de los casos. Por ello es de especial utilidad cuando los outputs son categorías. En términos matemáticos es:

$$F(x) = \frac{1}{(1 + e^{-x})}$$

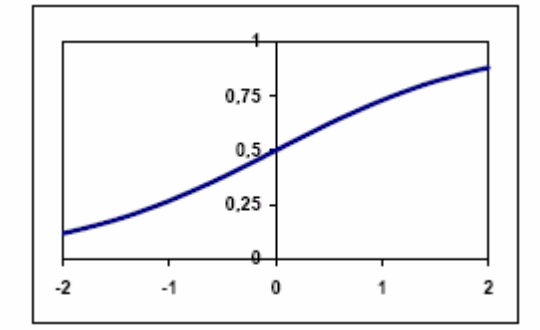

Figura 10 - Función logística

Función tangente hiperbólica: Esta función tiene las mismas propiedades que la logística, sin embargo, el rango de salida de esta función permite respuestas simétricas (-1, 1); manteniendo una intermedia en cero. Esta función suele converger antes que la función logística, sin embargo, no necesariamente generaliza igualmente bien. En términos matemáticos, queda expresada en:

$$F(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$$

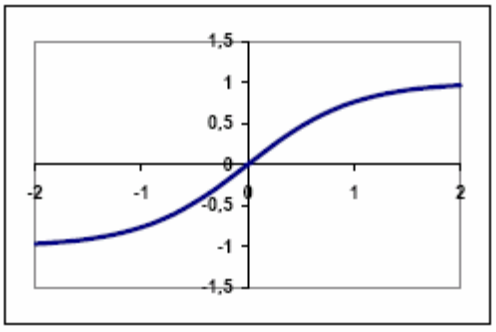

Figura 11 - Función tangente hiperbólica

10.2. Topologías clásicas de redes neuronales artificiales.

Las topologías clásicas de redes neuronales artificiales (RNA) son:

Perceptrón simple: Este tipo de RNA es el modelo más simple. Es una red unidireccional compuesta por dos capas de neuronas, una de entrada y la otra de salida; por lo tanto, en este modelo las neuronas de entrada únicamente envían la información a las neuronas de salida. El aprendizaje de este tipo de red es del tipo supervisado y se basa principalmente en la regla de corrección de error con respecto a la salida deseada.

Perceptrón multicapa: Este tipo de RNA es una ampliación del anterior, ya que incorpora uno o más niveles de unidades ocultas. Un perceptrón multicapa es una RNA con alimentación hacia delante y está compuesta de varias capas de neuronas entre la entrada y la salida, permitiendo de esta manera establecer regiones de decisión mucho más complejas en comparación con el perceptrón simple. Generalmente estas RNA se entrenan mediante el algoritmo backpropagation.

Redes Hebbianas: Este tipo de redes tiene un aprendizaje no supervisado. Se basa en la regla de Hebb, la cual dice que cuando una neurona activa a otra, la sinapsis queda reforzada. Este tipo de redes nos indica que si en el momento de la asociación entre las neuronas, dos o más de ellas se activan simultáneamente, éstas actuarán en conjunto incrementando o potencializando la sinapsis, ya que ahora la activación o desactivación de una de ellas influye en la otra; por consiguiente, se pueden activar varias neuronas en la salida.

Redes Kohonen: Este tipo de RNA posee la capacidad de formar mapas de características de manera similar al cerebro. El objetivo de este modelo es demostrar que un estímulo externo (información de entrada) por sí solo es suficiente para forzar la formación de estos mapas. Los que se constituyen de la información de entrada, la cual mediante la semejanza de sus datos, forma diferentes categorías. Esta red utiliza el aprendizaje no supervisado de tipo competitivo, es decir, las neuronas compiten por activarse y sólo una de ellas permanece activa ante una determinada información de entrada, provocando que los pesos de las conexiones se ajusten en función de la neurona vencedora.

Redes Hopfield: La red Hofield funciona como una memoria asociativa no lineal, que puede almacenar internamente patrones presentados de forma incompleta o con ruido. Esta red está formada por neuronas conectadas simétricamente (al existir una conexión desde la neurona N_i a la neurona N_j, también existe la conexión desde N_j a N_i y ambas con el mismo peso $W_{ij} = W_{ji}$) y el conjunto permitido de valores de entrada y salida es (0,1) pudiendo ser (-1,1) o sea, binario. Este modelo es similar al Perceptrón, pero muestra una característica adicional y es que las neuronas de la capa media, presentan conexiones de salida hacia otras neuronas de la capa media. Este hecho, hace que en esta capa se dé un *feedback* entre sus neuronas, de forma que al activarse una de las neuronas, las otras cambian su estado de activación, que a la vez hará cambiar el suyo. Por lo tanto, el patrón de activación se transmitirá sólo cuando se llegue a un equilibrio. Esta red no implica cálculo de pesos sinápticos, ya que éstos se mantienen constantes.

Según procesamiento, existen modelos estáticos y dinámicos de RNA. Los modelos estáticos configuran una red con un conjunto de datos determinados, la entrenan y una vez que la red alcanza buen desempeño se fija su configuración y sus pesos, luego con esa red fija se prueba un conjunto de datos (*testing*) para evaluar finalmente el desempeño de la red. Los modelos dinámicos incorporan los datos de pruebas (extramuestral) dentro de la red y estos nuevos datos van modificando la red, por lo que la hacen dinámica, ya que incorpora la nueva información. Dentro de los modelos dinámicos, tenemos:

Redes Recursivas: Este procedimiento consiste en ir agregando información a la que ya se tiene para realizar una nueva predicción. De esta forma, si el conjunto extramuestral posee *"m"* datos, al evaluar el funcionamiento de la red se considera sólo la predicción del primer valor. Posteriormente, el dato analizado sale del conjunto extramuestral (quedando ésta con *"m-1"* datos), pasando a formar parte del conjunto de entrenamiento, por lo que la muestra de *"n"* datos que contiene los valores de entrada se incrementa a "*n+1*". Luego, al realizar una nueva iteración, lo que implica reconstruir los pesos del modelo para cada una de las *"m-1"* proyecciones, la red logra aprender del error cometido en la predicción y ajustar nuevamente los pesos estimados. Este proceso se repite hasta que en el conjunto extramuestral queda con sólo un dato, por lo que la red recalcula los pesos de las diferentes capas hasta el momento en que la última observación (correspondiente al período "*t-1*") es utilizada para proyectar el valor que la variable de salida podría alcanzar en el momento "*t*", el cual representa el futuro inmediato.

Redes Rolling: Por otro lado, el procedimiento “Rolling” pretende capturar el proceso de adaptación de expectativas, de manera tal que se le otorgue una mayor importancia a aquella información recientemente agregada y descartando la más antigua, simulando el hecho de que los agentes ajustan sus expectativas dando una mayor ponderación a la información más reciente. Específicamente, el procedimiento mantiene constante el tamaño del conjunto muestral *“n”*, pero varía desde el punto de vista de su composición, pues para cada iteración, éste se va desplazando a través del conjunto total de datos, de forma de incluir la próxima observación, pero a la vez eliminando la más antigua (- *1+n+1=n).* Así, este procedimiento se repite *“m-1”* veces, permitiéndole a la red recalcular sus pesos de acuerdo a lo aprendido.

11. Referencia técnica del algoritmo de red neuronal.

El algoritmo de red neuronal de Microsoft provisto por el modulo analysis services de minería de datos, usa una red de tipo *perceptrón multinivel*, que también se denomina *red de tipo regla delta de propagación hacia atrás*, compuesta por tres niveles de neuronas o *perceptrones*. Estos niveles son: un nivel de entrada, un nivel oculto opcional y un nivel de salida.

Esta documentación no abarca una discusión detallada sobre redes neuronales de tipo perceptrón multinivel. En este tema, se explica la implementación básica del algoritmo, incluido el método usado para normalizar los valores de entrada y de salida, y los métodos de selección de características usados para reducir la cardinalidad de los atributos. En este tema, se describen los parámetros y otros valores que se pueden usar para personalizar el comportamiento del algoritmo; además, se proporcionan vínculos a información adicional sobre cómo consultar el modelo.

11.1. Implementación del algoritmo de red neuronal de Microsoft

En una red neuronal de tipo perceptrón multinivel, cada neurona recibe una o más entradas y genera una o más salidas idénticas. Cada salida es una función no lineal simple de la suma de las entradas a la neurona. Las entradas pasan de los nodos del nivel de entrada a los nodos del nivel oculto y, a continuación, al nivel de salida; no existe ninguna conexión entre neuronas del mismo nivel. Si no se incluye ningún nivel oculto, tal y como pasa en un modelo de regresión logística, las entradas pasan directamente desde los nodos del nivel de entrada a los nodos del nivel de salida.

Existen tres tipos de neuronas en una red neuronal creada con el algoritmo de red neuronal de Microsoft:

- **Input neurons**

Las neuronas de entrada proporcionan valores de atributo de entrada para el modelo de minería de datos. En el caso de los atributos de entrada discretos, las neuronas de entrada suelen representar un único estado del atributo de entrada. Esto incluye los valores ausentes, si los datos de entrenamiento contienen valores NULL para ese atributo. Un atributo de entrada discreto que tiene más de dos estados genera una neurona de entrada por cada estado y una neurona de entrada para un estado ausente, si existen valores NULL en los datos de

entrenamiento. Un atributo de entrada continuo genera dos neuronas de entrada: una neurona para un estado ausente y otra neurona para el valor del propio atributo continuo. Las neuronas de entrada proporcionan entradas para una o más neuronas ocultas.

- **Hidden neurons**

Las neuronas ocultas reciben entradas de las neuronas de entrada y proporcionan salidas a las neuronas de salida.

- **Output neurons**

Las neuronas de salida representan valores de atributo de predicción para el modelo de minería de datos. En el caso de los atributos de entrada discretos, una neurona de salida suele representar un único estado de predicción para un atributo de predicción, incluidos los valores que faltan. Por ejemplo, un atributo de predicción binario produce un nodo de salida que describe un estado ausente o existente, que indica si existe un valor para ese atributo. Una columna booleana utilizada como un atributo de predicción genera tres neuronas de salida: una neurona para un valor true, otra neurona para un valor false y una última neurona para un estado existente o ausente. Un atributo de predicción discreto que tiene más de dos estados genera una neurona de salida por cada estado y una neurona de salida para un estado ausente o existente. Las columnas de predicción continuas generan dos neuronas de salida: una neurona para un estado existente o ausente y otra neurona para el valor de la propia columna continua. Si se generan más de 500 neuronas de salida al revisar el conjunto de columnas de predicción, Analysis Services genera una red nueva en el modelo de minería de datos para representar las neuronas de salida adicionales.

Una neurona recibe la entrada de otras neuronas o de otros datos, dependiendo del nivel de la red en que se encuentra. Una neurona de entrada recibe entradas de los datos originales. Las neuronas ocultas y las neuronas de salida reciben entradas de la salida de otras neuronas de la red neuronal. Las entradas establecen relaciones entre neuronas; estas relaciones sirven como ruta de análisis para un conjunto específico de escenarios.

Cada entrada tiene un valor asignado denominado *peso*, que describe la relevancia o importancia de dicha entrada en la neurona oculta o en la neurona de salida. Cuanto mayor sea el peso asignado a una entrada, más importante o relevante será el valor de dicha entrada. Los

pesos pueden ser negativos, lo cual implica que la entrada puede desactivar, en lugar de activar, una neurona específica. El valor de cada entrada se multiplica por el peso para poner de relieve la importancia de la entrada de una neurona específica. En el caso de pesos negativos, el efecto de multiplicar el valor por el peso es una pérdida de importancia.

Cada neurona tiene una función no lineal sencilla asignada denominada *función de activación*, que describe la relevancia o importancia de una neurona determinada para ese nivel de una red neuronal. Las neuronas ocultas usan una función *tangente hiperbólica* (*tanh*) para su función de activación, mientras que las neuronas de salida usan una función *sigmoidea*. Ambas son funciones no lineales continuas que permiten que la red neuronal modele relaciones no lineales entre neuronas de entrada y salida.

11.1.1. Redes neuronales de entrenamiento

Existen varios pasos implicados en el entrenamiento de un modelo de minería de datos que utiliza el algoritmo de red neuronal de Microsoft. Estos pasos están muy influenciados por los valores que se especifican en los parámetros de algoritmo.

En primer lugar, el algoritmo evalúa y extrae los datos de entrenamiento del origen de datos. Un porcentaje de los datos de entrenamiento, denominado *datos de exclusión*, se reserva para evaluar la precisión de la red. Durante el proceso de entrenamiento, la red se evalúa de forma inmediata después de cada iteración mediante los datos de entrenamiento. Cuando la precisión deja de aumentar, el proceso de entrenamiento se detiene.

Los valores de los parámetros *SAMPLE_SIZE* y *HOLDOUT_PERCENTAGE* se usan para determinar el número de casos de muestra de los datos de aprendizaje y el número de casos que se apartan para los datos de exclusión. El valor del parámetro *HOLDOUT_SEED* se usa para determinar aleatoriamente los casos individuales que se apartan para los datos de exclusión.

Nota:

Estos parámetros de algoritmo son diferentes de las propiedades HOLDOUT_SEED y HOLDOUT_SIZE, que se aplican a una estructura de minería de datos para definir un conjunto de datos de prueba.

A continuación, el algoritmo determina el número y la complejidad de las redes que admite el modelo de minería de datos. Si el modelo contiene uno o más atributos que sólo se utilizan para la predicción, el algoritmo crea una única red que representa todos estos atributos. Si el modelo de minería de datos contiene uno o más atributos que se utilizan para la entrada y la predicción, el proveedor del algoritmo construye una red para cada atributo.

En el caso de los atributos de entrada y de predicción que tienen valores discretos, cada neurona de entrada o de salida representa respectivamente un único estado. En el caso de los atributos de entrada y de predicción que tienen valores continuos, cada neurona de entrada o de salida representa respectivamente el intervalo y la distribución de valores del atributo. El número máximo de estados admitidos en cada caso depende del valor del parámetro de algoritmo *MAXIMUM_STATES*. Si el número de estados para un atributo específico supera el valor del parámetro de algoritmo *MAXIMUM_STATES*, se eligen los estados más comunes o relevantes para dicho atributo, hasta alcanzar el máximo permitido; el resto de los estados se agrupa como valores ausentes para el análisis.

A continuación, el algoritmo utiliza el valor del parámetro *HIDDEN_NODE_RATIO* al determinar el número inicial de neuronas que se crearán para la capa oculta. Puede establecer *HIDDEN_NODE_RATIO* en 0 para evitar la creación de una capa oculta en las redes que genera el algoritmo para el modelo de minería de datos y tratar la red neuronal como una regresión logística.

El proveedor de algoritmos evalúa iterativamente el peso de todas las entradas de la red simultáneamente, tomando el conjunto de datos de entrenamiento reservado anteriormente y comparando el valor real conocido de cada escenario de los datos de exclusión con la predicción de la red, en un proceso conocido como *aprendizaje por lotes*. Una vez que el algoritmo ha evaluado el conjunto completo de los datos de entrenamiento, revisa el valor predicho y real de cada neurona. El algoritmo calcula el grado de error, si lo hay, y ajusta los pesos asociados con las entradas de esa neurona, trabajando hacia atrás desde las neuronas de salida a las de entrada en un proceso conocido como *propagación hacia atrás*. A continuación, el algoritmo repite el proceso en todo el conjunto de datos de entrenamiento. Dado que el algoritmo puede admitir múltiples pesos y neuronas de salida, el algoritmo de gradiente conjugado se utiliza para guiar el proceso de entrenamiento en la asignación y evaluación de los pesos de las entradas. Esta documentación no abarca una discusión sobre el algoritmo de gradiente conjugado.

11.1.2. Selección de características

Si el número de atributos de entrada es mayor que el valor del parámetro *MAXIMUM_INPUT_ATTRIBUTES*, o si el número de atributos de predicción es mayor que el valor del parámetro *MAXIMUM_OUTPUT_ATTRIBUTES*, se usa un algoritmo de selección de características para reducir la complejidad de las redes que se incluyen en el modelo de minería de datos. La selección de características reduce el número de atributos de entrada o de predicción a los más relevantes estadísticamente para el modelo.

Todos los algoritmos de minería de datos de Analysis Services usan automáticamente la selección de características para mejorar el análisis y reducir la carga de procesamiento. El método usado para la selección de características en los modelos de red neuronal depende del tipo de datos del atributo. Como referencia, en la tabla siguiente se muestran los métodos de selección de características usados para los modelos de red neuronal; además, se muestran los métodos de selección de características usados para el algoritmo de regresión logística, que está basado en el algoritmo de red neuronal.

Algoritmo	Método de análisis	Comentarios
Red neuronal	Puntuación interestingness Entropía de Shannon Bayesiano con prioridad K2 Dirichlet bayesiano con prioridad uniforme (predeterminado)	El algoritmo de red neuronal puede utilizar ambos métodos, siempre y cuando los datos contengan columnas continuas. Predeterminado.
Regresión logística	Puntuación interestingness Entropía de Shannon Bayesiano con prioridad K2 Dirichlet bayesiano con prioridad uniforme (predeterminado)	Dado que no se puede pasar un parámetro a este algoritmo para controlar el comportamiento de la selección de características, se utilizan los valores predeterminados. Por consiguiente, si todos los atributos son discretos o discretizados, el valor predeterminado es BDEU.

Los parámetros de algoritmo que controlan la selección de características para un modelo de red neuronal son MAXIMUM_INPUT_ATTRIBUTES, MAXIMUM_OUTPUT_ATTRIBUTES y MAXIMUM_STATES. También puede controlar el número de niveles ocultos mediante el establecimiento del parámetro HIDDEN_NODE_RATIO.

11.1.3. Métodos de puntuación

La *puntuación* es un tipo de normalización que, en el contexto del entrenamiento de un modelo de red neuronal, hace referencia al proceso de convertir un valor, como una etiqueta de texto discreta, en un valor que se pueda comparar con otros tipos de entradas y que se pueda pesar en la red. Por ejemplo, si un atributo de entrada es Sexo y los valores posibles son Hombre y Mujer, y otro atributo de entrada es Ingresos, con un intervalo de valores variable, los valores para cada atributo no son comparables directamente y, por consiguiente, deben estar codificados a una escala común para que se puedan calcular los pesos. Puntuar es el proceso de normalizar tales entradas para los valores numéricos, específicamente, para un intervalo de probabilidades. Las funciones usadas para la normalización también ayudan a distribuir más uniformemente los valores de entrada en una escala uniforme para que los valores extremos no distorsionen los resultados del análisis.

Las salidas de la red neuronal también están codificadas. Si hay un único destino para la salida (es decir, la predicción), o varios destinos que se usan solo para la predicción, no para la entrada, el modelo crea una red única y es posible que no sea necesario normalizar los valores. Sin embargo, si se usan varios atributos para la entrada y la predicción, el modelo debe crear varias redes; por tanto, se deben normalizar todos los valores y, al salir de la red, las salidas deberán estar codificadas.

La codificación de las entradas se basa en la suma de cada valor discreto de los casos de entrenamiento y en la multiplicación de ese valor por su peso. Esto se denomina *suma ponderada*, que se pasa a la función de activación del nivel oculto. Para la codificación, se usa la puntuación-z:

Valores discretos

$$\mu = p - \textit{la probabilidad anterior de un estado}$$

Valores continuos

$$\text{Valor presente} = 1 - \mu/\sigma$$

Una vez codificados los valores, se realiza una suma ponderada de las entradas, con los extremos de la red como pesos.

La codificación de las salidas usa la función sigmoidea, que tiene propiedades que la hacen muy útil para la predicción. Una de esas propiedades es que, sin tener en cuenta cómo se ajusta la escala de los valores originales, y sin tener en cuenta si los valores son negativos o positivos, la salida de esta función es siempre un valor entre 0 y 1, lo que resulta apropiado para la estimación de probabilidades. Otra propiedad útil es que la función sigmoidea tiene un efecto suavizador que hace que cuando los valores se alejan del punto de inflexión, la probabilidad del valor se aproxima lentamente a 0 o a 1.

11.2. Personalizar el algoritmo de red neuronal

El algoritmo de red neuronal de Microsoft admite varios parámetros que afectan al comportamiento, al rendimiento y a la precisión del modelo de minería de datos resultante. También puede modificar la forma en la que el modelo procesa los datos; para ello, puede establecer marcadores de modelado en las columnas o marcadores de distribución que especifiquen cómo se deben procesar los valores dentro de la columna.

11.2.1. Establecer los parámetros del algoritmo

A continuación, se describen los parámetros que se pueden usar con el algoritmo de red neuronal de Microsoft.

HIDDEN_NODE_RATIO

Especifica la proporción entre neuronas ocultas y neuronas de entrada y de salida. La siguiente fórmula determina el número inicial de neuronas de la capa oculta:

HIDDEN_NODE_RATIO * SQRT(Total input neurons * Total output neurons)

El valor predeterminado es 4,0.

HOLDOUT_PERCENTAGE

Especifica el porcentaje de escenarios de los datos de entrenamiento utilizados para calcular el error de exclusión, que se utiliza como parte de los criterios de detención durante el entrenamiento del modelo de minería de datos.

El valor predeterminado es 30.

HOLDOUT_SEED

Especifica un número que se utiliza para inicializar el generador pseudoaleatorio cuando el algoritmo determina aleatoriamente los datos de exclusión. Si este parámetro se establece en 0, el algoritmo genera la inicialización basada en el nombre del modelo de minería de datos, para garantizar que el contenido del modelo permanece intacto al volver a realizar el proceso.

El valor predeterminado es 0.

MAXIMUM_INPUT_ATTRIBUTES

Determina el número máximo de atributos de entrada que se pueden proporcionar al algoritmo antes de emplear la selección de características. La función de selección de atributos de entrada se deshabilita cuando este valor se establece en 0.

El valor predeterminado es 255.

MAXIMUM_OUTPUT_ATTRIBUTES

Determina el número máximo de atributos de salida que se pueden proporcionar al algoritmo antes de emplear la selección de características. La característica de selección de atributos de salida se deshabilita cuando este valor se establece en 0.

El valor predeterminado es 255.

MAXIMUM_STATES

Especifica el número máximo de estados discretos por atributo que admite el algoritmo. Si en número de estados de un atributo específico es mayor que el número especificado para este parámetro, el algoritmo utiliza los estados más frecuentes de este atributo y trata al resto como estados que faltan.

El valor predeterminado es 100.

SAMPLE_SIZE

Especifica el número de escenarios que se van a utilizar para realizar el entrenamiento del modelo. El algoritmo utiliza el valor menor entre este número o el porcentaje del total de escenarios que no están incluidos en los datos de exclusión, según se especifica en el parámetro HOLDOUT_PERCENTAGE.

En otras palabras, si HOLDOUT_PERCENTAGE se establece en 30, el algoritmo utilizará el valor de este parámetro o un valor igual al 70 por ciento del número total de casos, según cuál sea menor.

El valor predeterminado es 10.000.

11.2.2. Marcadores de modelado

El algoritmo de red neuronal de Microsoft admite los siguientes marcadores de modelado.

NOT NULL

Indica que la columna no puede contener un valor NULL. Se producirá un error si Analysis Services encuentra un valor NULL durante el entrenamiento del modelo.

Se aplica a las columnas de la estructura de minería de datos.

MODEL_EXISTENCE_ONLY

Indica que el modelo solo debe considerar si existe un valor para el atributo o si falta un valor. No importa el valor exacto.

Se aplica a las columnas del modelo de minería de datos.

11.2.3. Marcadores de distribución

El algoritmo de red neuronal de Microsoft admite los siguientes marcadores de distribución. Los marcadores solo se usan como sugerencias para el modelo; si el algoritmo detecta una distribución diferente, usará la distribución encontrada, no la proporcionada en la sugerencia.

Normal

Indica que los valores de la columna se deben tratar como si representasen la distribución normal o gaussiana.

Uniforme

Indica que los valores de la columna se deben tratar como si estuviesen distribuidos uniformemente; es decir, la probabilidad de cualquier valor es más o menos la misma y depende del número total de valores.

Logarítmica normal

Indica que los valores de la columna se deben tratar como si estuviesen distribuidos según la curva *logarítmica normal*; esto significa que el logaritmo de los valores se distribuye normalmente.

Requisitos

Un modelo de red neuronal debe contener por lo menos una columna de entrada y una columna de salida.

11.2.3.1. Columnas de entrada y de predicción

El algoritmo de red neuronal de Microsoft admite las columnas de entrada y de predicción específicas que se enumeran en la tabla siguiente.

Columna	Tipos de contenido
Atributo de entrada	Continuous, Cyclical, Discrete, Discretized, Key, Table y Ordered
Atributo de predicción	Continuous, Cyclical, Discrete, Discretized y Ordered

Nota:
Se admiten los tipos de contenido Cyclical y Ordered, pero el algoritmo los trata como valores discretos y no realiza un procesamiento especial.

12. Anexo - Referencia técnica algoritmo de árboles de decisión.

El algoritmo de árboles de decisión de Microsoft es un algoritmo híbrido que incorpora distintos métodos para crear un árbol, y admite varias tareas de análisis, incluyendo la regresión, la clasificación y la asociación. El algoritmo de árboles de decisión de Microsoft admite el modelado de los atributos discretos y continuos.

En este tema se explica la implementación del algoritmo, se describe cómo personalizar su comportamiento para distintas tareas y se proporcionan vínculos a información adicional sobre cómo consultar los modelos de árboles de decisión.

12.1. Implementación del algoritmo de árboles de decisión

El algoritmo de árboles de decisión de Microsoft aprende las redes bayesianas gracias a una combinación de experiencia previa y datos estadísticos. Una parte importante del algoritmo es la metodología para evaluar el valor de información de las *prioridades* necesarias para el aprendizaje. El enfoque está basado en la consideración de la *equivalencia de probabilidad*, que dice que los datos no deberían ayudar a discriminar estructuras de red que de otra forma representarían las mismas aserciones de independencia condicional.

Se supone que cada caso tiene una única red bayesiana anterior y una única medida de confianza para dicha red. Mediante estas redes anteriores, el algoritmo calcula las *probabilidades posteriores* relativas de las estructuras de red dados los datos de entrenamiento actuales, e identifica las estructuras de red con las probabilidades posteriores más altas.

El algoritmo de árboles de decisión de Microsoft usa distintos métodos para calcular el mejor árbol. El método usado dependerá de la tarea, que puede ser la regresión lineal, la clasificación o el análisis de la asociación. Un solo modelo puede contener varios árboles para distintos atributos de predicción. Es más, cada árbol puede contener varias bifurcaciones, dependiendo del número de atributos y de valores que contienen los datos. La forma y profundidad del árbol integrado en un modelo determinado depende del método de puntuación y del resto de parámetros usados. Los cambios en los parámetros también pueden afectar al lugar donde se dividen los nodos.

12.1.1. Generar el árbol

Cuando el algoritmo de árboles de decisión de Microsoft crea el conjunto de posibles valores de entrada, realiza una *feature selection* para identificar los atributos y los valores que ofrecen la mayor cantidad de información, y no tiene en cuenta los valores que son muy raros. El algoritmo también agrupa los valores en *bandejas* para crear agrupaciones de valores que se pueden procesar como una unidad para optimizar el rendimiento.

Un árbol se genera mediante la determinación de las correlaciones entre una entrada y el resultado deseado. Una vez correlacionados todos los atributos, el algoritmo identifica el atributo único que separa más claramente los resultados. Este punto de la mejor separación se mide usando una ecuación que calcula la obtención de información. El atributo que tiene la mejor puntuación para la obtención de información se usa para dividir los casos en subconjuntos, que posteriormente son analizados de forma recursiva por el mismo proceso hasta que no sea posible dividir más el árbol.

La ecuación exacta empleada para evaluar la obtención de información depende de los parámetros establecidos al crear el algoritmo, del tipo de datos de la columna de predicción y del tipo de datos de la entrada.

12.1.2. Entradas discretas y continuas

Cuando tanto el atributo de predicción como las entradas son discretas, el recuento de los resultados por entrada se realizará creando una matriz y generando puntuaciones para cada celda de la matriz.

Sin embargo, cuando el atributo de predicción es discreto y las entradas son continuas, la entrada de las columnas continuas se discretiza automáticamente. Puede aceptar el valor predeterminado y dejar que Analysis Services busque el número óptimo de bandejas, o puede controlar la forma en la que se discretizan las entradas continuas estableciendo las propiedades DiscretizationMethod y DiscretizationBucketCount. Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

Cuando el atributo de predicción es un tipo de datos numéricos continuo, la selección de características también se aplica a las salidas para reducir el número de resultados posibles y generar más rápidamente el modelo. Puede cambiar el umbral para la selección de

características, incrementando o disminuyendo de esta manera el número de valores posibles, estableciendo el parámetro MAXIMUM_OUTPUT_ATTRIBUTES.

12.1.3. Métodos de puntuación y selección de características

El algoritmo de árboles de decisión de Microsoft proporciona tres fórmulas para puntuar la obtención de información: la entropía de Shannon, la red bayesiana con prioridad K2 y la red bayesiana con una distribución Dirichlet uniforme de prioridades. Los tres métodos están bien consolidados en el campo de la minería de datos. Se recomienda que experimente con parámetros y métodos de puntuación diferentes para determinar cuáles son los que proporcionan mejores resultados.

Todos los algoritmos de minería de datos de Analysis Services utilizan automáticamente la selección de características para mejorar el análisis y reducir la carga de procesamiento. El método usado para la selección de características depende del algoritmo empleado para generar el modelo. Los parámetros del algoritmo que controlan la selección de características para el modelo de árboles de decisión son MAXIMUM_INPUT_ATTRIBUTES y MAXIMUM_OUTPUT.

Algoritmo	Método de análisis	Comentarios
Árboles de decisión	Puntuación de grado de interés Entropía de Shannon Bayesiano con prioridad K2 Dirichlet bayesiano con prioridad uniforme (predeterminado)	Si alguna columna contiene valores continuos no binarios, se utiliza la puntuación interestingness (grado de interés) en todas las columnas para asegurar la coherencia. En caso contrario, se utiliza el método predeterminado o el especificado.
Regresión lineal	Puntuación de grado de interés	La regresión lineal solo utiliza la puntuación interestingness porque solo admite columnas continuas.

12.1.4. Escalabilidad y rendimiento

La clasificación es una estrategia de minería de datos importante. Generalmente, la cantidad de información necesaria para clasificar los casos crece en proporción directa al número de registros de entrada. Esto limita el tamaño de los datos que se pueden clasificar. El algoritmo de árboles de decisión de Microsoft usa los métodos siguientes para resolver estos problemas, mejorar el rendimiento y eliminar las restricciones de memoria:

- Selección de características para optimizar la selección de atributos.

- Puntuación bayesiana para controlar el crecimiento del árbol.
- Optimización de bandejas para los atributos continuos.
- Agrupación dinámica de valores de entrada para determinar los valores más importantes.

El algoritmo de árboles de decisión de Microsoft es rápido y escalable, y se ha diseñado para ser usado en paralelo, es decir, con todos los procesadores funcionando juntos para generar un modelo único y coherente. La combinación de estas características convierte al clasificador de árboles de decisión en una herramienta ideal para la minería de datos.

Si las restricciones de rendimiento son graves, podría mejorar el tiempo de procesamiento durante el entrenamiento de un modelo de árbol de decisión usando los métodos siguientes. Sin embargo, si lo hace, tenga en cuenta que la eliminación de atributos para mejorar el rendimiento del procesamiento cambiará los resultados del modelo, y es posible que éste sea menos representativo de la población total.

- Aumente el valor del parámetro COMPLEXITY_PENALTY para limitar el crecimiento del árbol.
- Limite el número de elementos de los modelos de asociación para limitar el número de árboles que se generan.
- Aumente el valor del parámetro MINIMUM_SUPPORT para evitar el sobreajuste.
- Restrinja a 10 o menos el número de valores discretos para todos los atributos. Puede intentar agrupar valores de distintas maneras en modelos diferentes.

12.2. Personalizar el algoritmo de árboles de decisión

El algoritmo de árboles de decisión de Microsoft admite parámetros que afectan al rendimiento y la precisión del modelo de minería de datos resultante. También puede establecer marcadores de modelado en las columnas del modelo de minería de datos o de la estructura de minería de datos para controlar la manera en que se procesan los datos.

12.2.1. Establecer los parámetros del algoritmo

En la tabla siguiente se describen los parámetros que puede usar con el algoritmo de árboles de decisión de Microsoft.

COMPLEXITY_PENALTY

Controla el crecimiento del árbol de decisión. Un valor bajo aumenta el número de divisiones y un valor alto lo reduce. El valor predeterminado se basa en el número de atributos de un modelo concreto, como se describe en la lista siguiente:

- De 1 a 9 atributos, el valor predeterminado es 0,5.
- De 10 a 99 atributos, el valor predeterminado es 0,9.
- Para 100 o más atributos, el valor predeterminado es 0,99.

FORCE_REGRESSOR

Fuerza al algoritmo a utilizar las columnas indicadas como regresores, independientemente de su importancia según los cálculos del algoritmo. Este parámetro sólo se utiliza para árboles de decisión que predicen un atributo continuo.

Nota: Establezca este parámetro si desea que el algoritmo intente usar el atributo como un regresor. Sin embargo, el atributo se usará realmente como regresor en el modelo final en función de los resultados del análisis. Para averiguar las columnas que se usaron como regresores, consulte el contenido del modelo.

[SQL Server Enterprise]

MAXIMUM_INPUT_ATTRIBUTES

Define el número de atributos de entrada que puede controlar el algoritmo antes de invocar la selección de características.

El valor predeterminado es 255.

Establezca este valor en 0 para desactivar la selección de características.

[SQL Server Enterprise]

MAXIMUM_OUTPUT_ATTRIBUTES

Define el número de atributos de salida que puede controlar el algoritmo antes de invocar la selección de características.

El valor predeterminado es 255.

Establezca este valor en 0 para desactivar la selección de características.

[SQL Server Enterprise]

MINIMUM_SUPPORT

Determina el número mínimo de casos de hoja necesarios para generar una división en el árbol de decisión.

El valor predeterminado es 10.

Es posible que necesite aumentar este valor si el conjunto de datos es muy grande, para evitar el sobreentrenamiento.

SCORE_METHOD

Determina el método usado para calcular el resultado de la división. Las siguientes opciones están disponibles:

Id.	Nombre
1	Entropía
2	Bayesiano con prioridad K2
3	Equivalente Dirichlet bayesiano (BDE) con prioridad (predeterminado).

El valor predeterminado es 3.

SPLIT_METHOD

Determina el método usado para dividir el nodo. Las siguientes opciones están disponibles:

Id.	Nombre

1	**Binary:** indica que, independientemente del número real de valores para el atributo, el árbol se debería dividir en dos bifurcaciones.
2	**Complete:** indica que el árbol puede crear tantas divisiones como valores de atributo existan.
3	**Both:** especifica que Analysis Services puede determinar si se debe usar una división binaria o completa para generar los mejores resultados.

El valor predeterminado es 3.

12.2.2. Marcadores de modelado

El algoritmo de árboles de decisión de Microsoft admite los marcadores de modelado siguientes. Al crear la estructura o el modelo de minería de datos, se definen los marcadores de modelado que especifican cómo se tratan los valores de cada columna durante el análisis.

Marcador de modelado	Descripción
MODEL_EXISTENCE_ONLY	Significa que la columna se tratará como si tuviera dos estados posibles: **Missing** y **Existing**. Un valor NULL es un valor ausente. Se aplica a las columnas del modelo de minería de datos.
NOT NULL	Indica que la columna no puede contener un valor NULL. Se producirá un error si Analysis Services encuentra un valor NULL durante el entrenamiento del modelo. Se aplica a las columnas de la estructura de minería de datos.

12.2.3. Regresores en modelos de árbol de decisión

Aun cuando no use el algoritmo de regresión lineal de Microsoft, cualquier modelo de árbol de decisión que tenga entradas y salidas numéricas continuas puede incluir nodos que representan una regresión en un atributo continuo.

No es necesario especificar que una columna de datos numéricos continuos representa un regresor. El algoritmo de árboles de decisión de Microsoft usará automáticamente la columna

como un regresor potencial y dividirá el conjunto de datos en regiones con patrones significativos aunque no se establezca el marcador REGRESSOR en la columna.

Sin embargo, puede usar el parámetro FORCED_REGRESSOR para garantizar que el algoritmo empleará un regresor determinado. Este parámetro solo se puede usar con los algoritmos de árboles de decisión de Microsoft y de regresión lineal de Microsoft. Al establecer el marcador de modelado, el algoritmo intentará buscar ecuaciones de regresión con el formato *a*C1 + b*C2 + ...* que se ajusten a los patrones de los nodos del árbol. Se calcula la suma de los valores residuales y, si la desviación es demasiado grande, se fuerza una división en el árbol.

Por ejemplo, si está prediciendo los hábitos de compra de los clientes usando **Income** como atributo y ha establecido el marcador de modelado REGRESSOR en la columna, el algoritmo intentará en primer lugar ajustar los valores de **Income** mediante una fórmula de regresión estándar. Si la desviación es demasiado grande, se abandona la fórmula de regresión y el árbol se dividirá de acuerdo con otro atributo. A continuación, el algoritmo de árboles de decisión intentará ajustar un regresor para los ingresos en cada una de las ramas después de la división.

Requisitos

Un modelo de árbol de decisión debe contener una columna de clave, columnas de entrada y al menos una columna de predicción.

12.2.4. Columnas de entrada y de predicción

El algoritmo de árboles de decisión de Microsoft admite las columnas de entrada y de predicción específicas que se incluyen en la tabla siguiente.

Columna	Tipos de contenido
Atributo de entrada	Continuous, Cyclical, Discrete, Discretized, Key, Ordered, Table
Atributo de predicción	Continuous, Cyclical, Discrete, Discretized, Ordered, Table
Nota: Se admiten los tipos de contenido cíclico y ordenado, pero el algoritmo los trata como valores discretos y no realiza un procesamiento especial.	

13. Anexo - Referencia técnica del algoritmo de clústeres.

En esta sección se explica la implementación del algoritmo de clústeres de Microsoft, incluidos los parámetros que se pueden usar para controlar el comportamiento de los modelos de agrupación en clústeres. Además, incluye instrucciones sobre cómo mejorar el rendimiento durante la creación y el procesamiento de modelos de agrupación en clústeres.

13.1. Implementación del algoritmo de clústeres.

El algoritmo de clústeres de Microsoft proporciona dos métodos para crear clústeres y asignar puntos de datos a dichos clústeres. El primero, el algoritmo *K-medianas*, es un método de agrupación en clústeres duro. Esto significa que un punto de datos puede pertenecer a un solo clúster, y que únicamente se calcula una probabilidad de pertenencia de cada punto de datos de ese clúster. El segundo, el método *Expectation Maximization* (EM), es un método de *agrupación en clústeres blando*. Esto significa que un punto de datos siempre pertenece a varios clústeres, y que se calcula una probabilidad para cada combinación de punto de datos y clúster.

Puede elegir el algoritmo que desee utilizar estableciendo el parámetro *CLUSTERING_METHOD*. El método predeterminado para la agrupación en clústeres es el método EM escalable.

13.1.1. Agrupación en clústeres EM

En el método de agrupación en clústeres EM, el algoritmo refina de forma iterativa un modelo de clústeres inicial para ajustar los datos y determina la probabilidad de que un punto de datos exista en un clúster. El algoritmo finaliza el proceso cuando el modelo probabilístico ajusta los datos. La función utilizada para determinar el ajuste es el logaritmo de la probabilidad de los datos dado el modelo.

Si durante el proceso se generan clústeres vacíos, o si la pertenencia de uno o varios de los clústeres cae por debajo del umbral especificado, los clústeres con poblaciones bajas se reinicializan en los nuevos puntos y vuelve a ejecutarse el algoritmo EM.

Los resultados del método de agrupación en clústeres EM son probabilísticos. Esto significa que cada punto de datos pertenece a todos los clústeres, pero cada asignación de un punto de

datos a un clúster tiene una probabilidad diferente. Dado que el método permite que los clústeres se superpongan, la suma de los elementos de todos los clústeres puede superar la totalidad de los elementos existentes en el conjunto de entrenamiento. En los resultados del modelo de minería de datos, las puntuaciones que indican soporte se ajustan para tener en cuenta este hecho.

El algoritmo EM es el algoritmo predeterminado utilizado en los modelos de agrupación en clústeres de Microsoft. Este algoritmo se utiliza como algoritmo predeterminado porque proporciona numerosas ventajas comparado con la agrupación en clústeres K-medianas:

- Requiere examinar la base de datos como máximo una vez.
- Funciona incluso si la cantidad de memoria (RAM) es limitada.
- Tiene la capacidad de utilizar un cursor de sólo avance.
- Sus resultados superan los obtenidos por los métodos de muestreo.

La implementación de Microsoft proporciona dos opciones: EM escalable y no escalable. De forma predeterminada, en EM escalable, los primeros 50.000 registros se utilizan para inicializar el examen inicial. Si esta operación se realiza correctamente, el modelo sólo utiliza estos datos. Si el modelo no se puede ajustar con 50.000 registros, se leen otros 50.000. En EM no escalable, se lee el conjunto de datos completo independientemente de su tamaño. Este método puede crear clústeres más precisos, pero los requisitos de memoria pueden ser significativos. Dado que EM escalable funciona en un búfer local, recorrer los datos en iteración es mucho más rápido, y el algoritmo hace un mejor uso de la caché de memoria de la CPU que EM no escalable. Es más, EM escalable es tres veces más rápido que EM no escalable, incluso si todos los datos caben en la memoria principal. En la mayoría de casos, la mejora en el rendimiento no significa una reducción de la calidad del modelo completo.

13.1.2. Agrupación en clústeres K-medianas

La agrupación en clústeres K-medianas es un método muy conocido para asignar la pertenencia al clúster que consiste en minimizar las diferencias entre los elementos de un clúster al tiempo que se maximiza la distancia entre los clústeres. El término "mediana" hace referencia al *centroide* del clúster, que es un punto de datos que se elige arbitrariamente y que se refina de forma iterativa hasta que representa la verdadera media de todos los puntos de datos del clúster. La "K" hace referencia a un número arbitrario de puntos que se utilizan para inicializar el proceso de agrupación en clústeres. El algoritmo K-medianas calcula las distancias

euclidianas cuadradas entre los registros de datos de un clúster y el vector que representa la media de clústeres, y converge en un conjunto final de K clústeres cuando la suma alcanza su valor mínimo.
El algoritmo K-medianas asigna cada punto de datos a un solo clúster y no permite la incertidumbre en la pertenencia. En un clúster, la pertenencia se expresa como una distancia desde el centroide.

Normalmente, el algoritmo K-medianas se utiliza para crear clústeres de atributos continuos, donde el cálculo de la distancia a una media se realiza de manera sencilla. Sin embargo, la implementación de Microsoft adapta el método K-medianas a atributos discretos de clúster mediante el uso de probabilidades. Para los atributos discretos, la distancia de un punto de datos desde un clúster determinado se calcula de la manera siguiente:

1 - P(punto de datos, clúster)

Nota:

El algoritmo de clústeres de Microsoft no expone la función de distancia utilizada para calcular la K-medianas, y las medidas de la distancia no están disponibles en el modelo completado. Sin embargo, se puede utilizar una función de predicción para devolver un valor que corresponda a la distancia, donde la distancia se calcula como la probabilidad de que un punto de datos pertenezca al clúster.

El algoritmo K-medianas proporciona dos métodos para realizar un muestreo en el conjunto de datos: K-medianas no escalable, que carga el conjunto de datos completo y realiza una pasada de agrupación en clústeres, y K-medianas escalable, donde el algoritmo utiliza los primeros 50.000 casos y lee más casos únicamente si necesita más datos para lograr un buen ajuste del modelo a los datos.

13.2. Personalizar el algoritmo de clústeres

El algoritmo de clústeres Microsoft admite varios parámetros que afectan al comportamiento, el rendimiento y la precisión del modelo de minería de datos resultante.

13.2.1. Establecer los parámetros del algoritmo

En la tabla siguiente se describen los parámetros que se pueden utilizar con el algoritmo de clústeres de Microsoft. Estos parámetros afectan tanto al rendimiento como a la precisión del modelo de minería de datos resultante.

CLUSTERING_METHOD

Especifica el método de agrupación en clústeres que va a utilizar el algoritmo. Los métodos de agrupación en clústeres disponibles son:

Id.	Método
1	EM escalable
2	EM no escalable
3	K-medianas escalable
4	K-medianas no escalable

El valor predeterminado es 1 (EM escalable).

CLUSTER_COUNT

Especifica el número aproximado de clústeres que será generado por el algoritmo. Si no se puede generar el número aproximado de clústeres a partir de los datos, el algoritmo genera tantos clústeres como sea posible. Si CLUSTER_COUNT se establece en 0, el algoritmo utiliza la heurística para determinar el mejor número de clústeres que debe generar.

El valor predeterminado es 10.

CLUSTER_SEED

Especifica el número de inicialización utilizado para generar clústeres aleatoriamente para la fase inicial de generación del modelo.

Cambiando este número, se puede cambiar la manera en que se generan los clústeres iniciales y, a continuación, comparar modelos que se han generado utilizando inicializaciones diferentes. Si se cambia la inicialización pero los clústeres hallados no cambian en gran medida, el modelo se puede considerar relativamente estable.

El valor predeterminado es 0.

MINIMUM_SUPPORT

Especifica el número mínimo de casos requeridos para generar un clúster. Si el número de casos del clúster es inferior a este número, el clúster se trata como vacío y se descarta.

Si se establece un número demasiado alto, se pueden perder clústeres válidos.

> **Nota** Si se utiliza EM, que es el método de agrupación en clústeres predeterminado, algunos clústeres pueden tener un valor de soporte más bajo que el valor especificado. Esto se debe a que cada caso se evalúa según su pertenencia a todos los clústeres posibles, y para algunos clústeres puede haber sólo un soporte mínimo.

El valor predeterminado es 1.

MODELLING_CARDINALITY

Especifica el número de modelos de ejemplo que se construyen durante el proceso de agrupación en clústeres.

Reducir el número de modelos candidatos puede mejorar el rendimiento, pero se corre el riesgo de perder algunos modelos buenos.

El valor predeterminado es 10.

STOPPING_TOLERANCE

Especifica el valor que se utiliza para determinar cuándo se alcanza la convergencia y el algoritmo termina de generar el modelo. La convergencia se alcanza cuando el cambio general de las probabilidades del clúster es inferior al resultado de dividir el parámetro STOPPING_TOLERANCE entre el tamaño del modelo.

El valor predeterminado es 10.

SAMPLE_SIZE

Especifica el número de casos que el algoritmo utiliza en cada paso si el parámetro CLUSTERING_METHOD está establecido en uno de los métodos de agrupación en clústeres escalables. Si se establece el parámetro SAMPLE_SIZE en 0, todo el conjunto de datos se agrupará en un único paso. Cargar el conjunto de datos completo en un paso único puede causar problemas de memoria y de rendimiento.

El valor predeterminado es 50000.

MAXIMUM_INPUT_ATTRIBUTES

Especifica el número máximo de atributos de entrada que el algoritmo puede procesar antes de invocar la selección de características. Si se establece este valor en 0, se especifica que no existe un número máximo de atributos.

Aumentar el número de atributos puede degradar significativamente el rendimiento.

El valor predeterminado es 255.

MAXIMUM_STATES

Especifica el número máximo de estados de atributo admitido por el algoritmo. Si un atributo tiene más estados que el máximo permitido, el algoritmo utiliza los estados más conocidos y pasa por alto los estados restantes.

Aumentar el número de estados puede degradar significativamente el rendimiento.

El valor predeterminado es 100.

13.2.2. Marcadores de modelado

El algoritmo admite los marcadores de modelado que se indican a continuación. Los marcadores de modelado se definen al crear la estructura de minería de datos o el modelo de minería de datos. Los marcadores de modelado especifican cómo se procesan los valores de cada columna durante el análisis.

Marcador de modelado	Descripción
MODEL_EXISTENCE_ONLY	La columna se tratará como si tuviera dos estados posibles: ausente y existente. Un valor NULL es un valor ausente. Se aplica a la columna del modelo de minería de datos.
NOT NULL	La columna no puede contener valores NULL. Se producirá un error si Analysis Services encuentra un valor NULL durante el entrenamiento del modelo. Se aplica a la columna de la estructura de minería de datos.

Requisitos

Un modelo de agrupación en clústeres debe contener una columna de clave y columnas de entrada. También se pueden definir columnas de entrada como columnas de predicción. Las

columnas establecidas en **Predict Only** no se utilizan para generar clústeres. La distribución de estos valores en los clústeres se calcula después de que se hayan generado los clústeres.

13.2.3. Columnas de entrada y de predicción

El algoritmo de clústeres de Microsoft admite las columnas de entrada y de predicción específicas que se enumeran en la tabla siguiente:

Columna	Tipos de contenido
Atributo de entrada	Continuous, Cyclical, Discrete, Discretized, Key, Table, Ordered
Atributo de predicción	Continuous, Cyclical, Discrete, Discretized, Table, Ordered

Nota Se admiten los tipos de contenido cíclicos y ordenados, pero el algoritmo los trata como valores discretos y no realiza un procesamiento especial.