The 30th Latin-American Conference on Informatics

Carlos Araya Pacheco
Universidad Católica del Norte
Departamento de Ciencias Empresariales
Antofagasta,Chile.
caaraya@ucn.cl

Monique Olmos Carrasco
Universidad de Antofagasta
Departamento de Ingeniería de Sistemas
Antofagasta, Chile
molmos@uantof.cl

Predicción del Rendimiento de los Alumnos de Plan Común de las Carreras de Ingeniería Civil Industrial de la Universidad de Antofagasta a través de Minería de Datos.

Abstract:

The knowledge discovery in large databases is a process which can turn out competitive advantages to the companies, useful to their business models. Under this viewpoint, the outdoing of the information management is looked at as a must to the companies, in the never ending searching of newer and powerful useful behavioral patterns.

The goal of the research is to forecast the general achievement of the students who belong to junior industrial engineering program, in order to help the strategic objectives of the Faculty, giving it light about the successful criteria and factors related with, to set up the wished entry level behavior and help reducing the student drop-off and failing.

This research looks insight and describes the process, and was developed in the Engineering Faculty of the University of Antofagasta, Chile, using to evaluate the algorithms of the decision tree and the neuronal networks the methodology CRISP-DM.

The outcomes and conclusions of the research show a 95 % forecasting successful to the Calculus-I and Algebra-I subjects and a 70% forecasting successful to the school-grade scores and schooling type.

Árboles Decisión, Redes Neuronales, Minería Datos, Asociación de Reglas, Gestión Universitaria.

Introducción

En la actualidad, el descubrimiento de Conocimiento en Bases de Datos Masivas, es un proceso que puede entregar a las organizaciones educacionales ventajas competitivas para su modelo de negocio. Siendo fundamental el apoyo de los directivos en la definición de objetivos claros, los cuales guiarán esta búsqueda de conocimiento.

Existe en la comunidad científica muy pocos estudios del rendimiento de los estudiantes, con apoyo de técnicas de minería de datos, dentro de ellos se muestra una tendencia a definir el proceso como proyecto tecnológico o un proceso de gestión educacional.

En la primera visión, Dhammika [1] define subconjuntos de características que responden a rangos de variables asociados al tipo de organización. En la misma línea Salpeter [2]., valoriza la conducción de los datos, para cumplir los objetivos de Data Mining.

En cambio Thomas [3], plantea que la utilización de Data Mining es para identificar aspectos específicos de los estudiantes, que permita crea subconjuntos de sus características, utilizando árboles de decisión. Gornitza [4] plantea una reestructuración administrativa de las fuerzas de trabajo en las universidades, para impactar el rendimiento de los estudiantes a través de evaluaciones académicas.

En este contexto, esta investigación se basa en un proyecto desarrollado en la Facultad de Ingeniería, de la Universidad de Antofagasta (Chile), con la Metodología CRISP-DM, que presenta una metodología de procedimientos jerárquicos considerando una serie de tareas [5].

Las actividades a realizar en la fase de "comprensión del negocio", es entender los objetivos del proyecto y los requerimientos desde la perspectiva del negocio. En este caso se realizaron entrevista y reuniones con diferentes stakeholders de la Universidad de Antofagasta.

En segundo lugar se procedió a la "comprensión de los Datos", en la cual pretende obtener una familiaridad con los datos, descubriendo los primeros hechos o subconjuntos de datos interesantes, a fin de formular algunas hipótesis. Para ello se analizo el Modelo de Datos de la BD Simbad de Alumnos y Bienestar, perteneciente a la Universidad de Antofagasta.

Posteriormente en la tercera etapa de "Preparación de los Datos", se agregaron, eliminaron o modificaron algunos atributos de los datos para conformar el conjunto y subconjuntos de datos ha explorar. Se construyo un conjunto de datos de las notas de los alumnos y sus antecedentes más relevantes.

Una vez definido el conjunto de datos ha utilizar, se realizó el "Modelamiento", que consistió en la definición de las técnicas de modelamiento iniciales y finales, a través de múltiples secuencias de pruebas, hasta encontrar un modelo, con un grado de confianza y soporte adecuado. En este caso se utilizo las técnicas de Árboles de Decisión y Redes Neuronales

Finalmente conseguido un modelo adecuado de esta perspectiva para el análisis de los datos, se realizó la fase de la "Evaluación", donde se cerciora si también cumple con los objetivos de negocio, para su posterior implementación, en el caso puntual del proyecto, la evaluación esta siendo efectuada por los directivos de la Facultad de Ingeniería de la Universidad de Antofagasta.

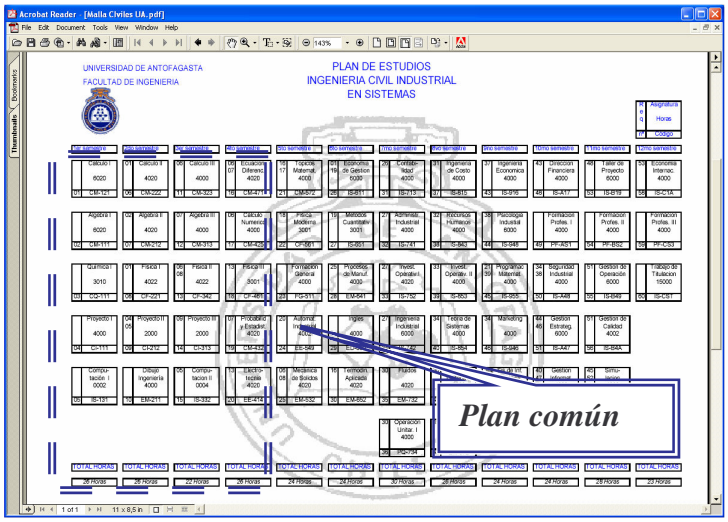
1. Comprensión del Negocio

En el inicio del Proyecto de Minería de Datos se debe individualizar la institución y los propósitos del negocio. En este caso particular se refiere a la Universidad de Antofagasta, específicamente a la Facultad de Ingeniería, en donde se imparten las Carreras de Ingeniería Civil Industrial, siendo el tronco común del cual se desprenden las Carreras de Ingeniería Civil Industrial Electricidad, Ingeniería Civil Industrial Electrónica, Ingeniería Civil Industrial Mecánica, Ingeniería Civil Industrial Minas, Ingeniería Civil Industrial Química y Ingeniería Civil Industrial Sistemas. Siendo el número de ingreso de alumnos a Ingeniería Civil Plan Común de aproximadamente 450 alumnos, los cuales después de 4 semestres académicos[1] en su avance curricular, deberán elegir las especialidades detalladas anteriormente.

El plan de estudios para los alumnos que optan a la Carrera de Ingeniería se puede visualizar a través de la siguiente malla curricular, donde se destaca el plan común con una línea punteada. (ver Figura 1).

[1] Semestre académico para Ingeniería: Se le denomina a un conjunto de 5 meses, los que pueden estar contenidos de marzo a julio para el primer semestre o de agosto a diciembre para el segundo semestre en la cual se dictan asignaturas, difiere para otras carreras donde las asignaturas son de marzo a diciembre.

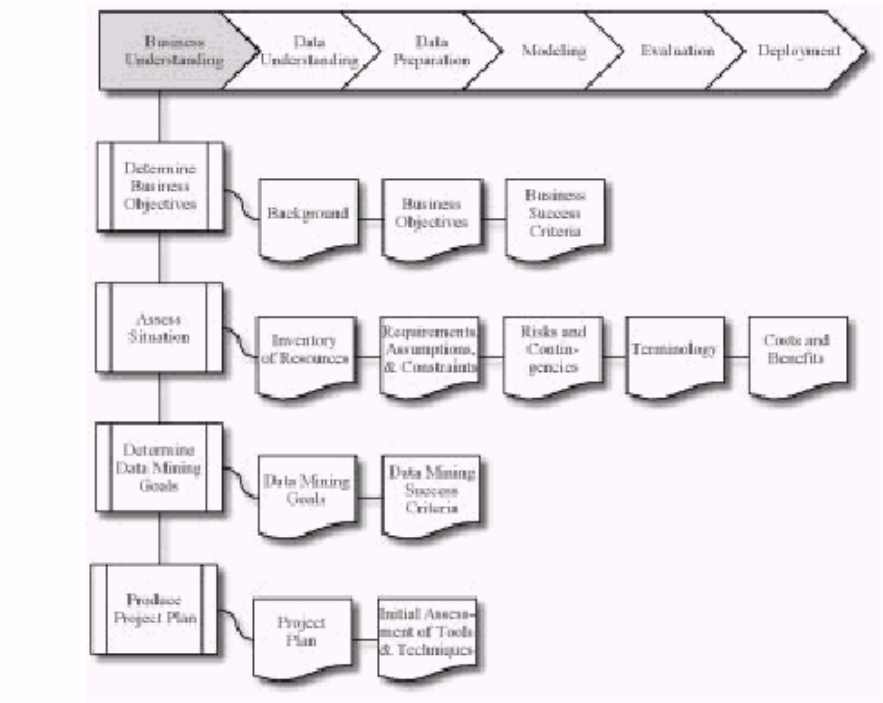
Figura 1. Plan de estudios de Ing. Civil Industrial, destacando área común.



Fuente: Web Site de la Universidad de Antofagasta

Para este proyecto se utilizó la Metodología Crisp-DM, que entrega una visión integral del proceso de Minería de Datos. En la Figura 2, se presenta un diagrama con los pasos a desarrollados durante el proyecto.

Figura 2: "Comprensión del Negocio". Modelo Crisp-DM 1.0



Fuente: Crisp-DM Step by Step Data Mining Guide. Página 16

1.1 Determinación de los Objetivos del Negocio

La Determinación de los objetivos de la investigación de Minería de Datos, esta relacionada con la comprensión del comportamiento de los alumnos en su ciclo básico, específicamente en cuales son los factores de éxito para superar el plan común de Ingeniería en la Universidad de Antofagasta, siendo una de las aristas de la evaluación. Además como complemento se necesitó analizar el rol del académico y los perfiles de egreso que se desean implementar a través del avance curricular, con un énfasis en un cluster de competencias definidas.

1.1.1 Background

La investigación se definirá como una "Predicción del Rendimiento de los alumnos en el Ciclo Básico de Ingeniería Civil Plan Común de la Universidad de Antofagasta", para definir los factores claves del éxito, que permitan definir las competencias de entrada al ciclo. Esta información podrá utilizarse para medir con mayor eficiencia los aspectos cuantitativos en el proceso de selección, tales como puntaje de ingreso y promedios de notas de enseñanza media, tipo de institución, etc.

1.1.2 Objetivos del Negocio

El Objetivo del negocio está definido en el Plan de Desarrollo de la Universidad de Antofagasta, que es una Institución de Educación Superior del Estado, independiente, autónoma y con personalidad jurídica propia, creada por Decreto Nº 11 del 10 de marzo de 1981. La Universidad de Antofagasta es la sucesora y continuadora legal de la Universidad de Chile y Universidad Técnica del Estado.

Dentro de la Declaración de la Misión Institucional se menciona en el punto 1,1 de los objetivos estratégicos, el "Aumento de la excelencia académica", como una línea de acción a desarrollar. Ello implica, por ejemplo, hacer el mayor esfuerzo por acreditar y ampliar la docencia de postgrado, y un esfuerzo importante para imbuir a los alumnos de pregrado con una clara conciencia de la calidad de la formación que reciben y mejorar su cultura general, y un esfuerzo razonable para mantener o incluso elevar la calidad de la docencia de pregrado.

Estos se resume en el punto 1.5 de la misión institucional , donde la definición de los factores claves de éxito para los alumnos de primer año de una carrera relevante dentro de la institución, tendrá un gran impacto en los indicadores definidos en los planes de acción. Considerando la inversión en aspectos que signifiquen una mejora cuantitativa y cualitativa del proceso de enseñanza, logrando la optimización de los recursos que siempre son escasos.

1.1.3 Criterios de Éxito en el Negocio

Los Criterios de Éxito, son en definitiva lo que determino el curso de la investigación a través de un análisis de las variables que afectan el proceso de enseñanza, considerando un periodo de 2 años en el Plan Común.

Para ello se realizó un estudio horizontal a través de líneas de conocimiento[2], que a continuación se detallan:

- Línea Cálculo: corresponde a Calculo 1, Calculo 2, Calculo 3, Ecuaciones Diferenciales, Calculo Numérico.
- Línea Álgebra: Álgebra 1, Álgebra 2, Álgebra 3, Probabilidad y Estadística, Ecuaciones Diferenciales.
- Línea Cispi: Proyecto1, Proyecto 2, Proyecto 3, Computación 1.
- Línea de Computación: Computación 1, Dibujo Ingeniería, Computación 2
- Línea Física: Química, Física 1, Física 2, Física 3.

[2] Líneas de Conocimiento: Se le denomina a un conjunto de ramos, que entre ellos tienen que cumplir un pre requisito para poder continuar con el siguiente.

Las definiciones de éxito, son dos, una de ellas es Éxito 1(Ex1), que son todos aquellos alumnos que tengas notas (nr1…nr10) iguales o mayores a 4, en las asignaturas de plan común. Ya que con estas condiciones cumplen con el requisito básico. Pero para obtener el Éxito 2 (Ex2), que es el conjunto óptimo, se utilizará la variable tiempo, en función al número de veces que los alumnos, deben cursar sus asignaturas obligatorias de segundo año (4 semestres), en un periodo menor o igual del tiempo máximo permitido (8 semestres). Este enfoque esta asociado a los objetivos de negocios de la economicidad de recursos, con la consiguiente disminución de costos en recursos humanos (académicos), debido a la alta repitencia, y los costos escondidos tales como ayudantes, instalaciones, becas, etc.

$$Ex1 = \{ nr_1, nr2, nr3, nr4, nr5, n6, nr7, nr8, nr9, nr10. / \frac{\sum nr_m}{10} \geq 4 \}$$

Donde:
Ex1= Éxito 1
nr1= Nota Álgebra1, nr2= NotaCalculo1, nr3= NotaQuimica1, nr4= NotaProyecto1, nr5= NotaComp1, nr6= NotaÁlgebra2, nr7= NotaCalculo 2, nr8=NotaFisica1, nr9=NotaProyecto2, nr10= Computación2.

El éxito 2 se descompone en 5 éxitos individuales por las líneas de conocimiento, asociadas a la sumatoria de la cantidad de veces que el alumno realiza la asignatura, para llegar con éxito al último ramo de la línea.

$Ex2 = \{ Ex2ecuadif , Ex2calnum , Ex2fis3, Ex2prob, Ex2proy, Ex2comp2 / \Sigma Ex2 \leq 48 \}$

$Ex2ecuadif = \{Nalg1 , Nalg2, Nalg3, Nedif, Ncal1, Ncal2, Ncal3 / \Sigma Nm \leq 8\}$

$Ex2calnum = \{Nalg1 , Nalg2, Nalg3, Ncnum, Ncal1, Ncal2, Ncal3 / \Sigma Nm \leq 8\}$

$Ex2fis3 = \{Ncal1 , Nqui, Nfis1, Nfis2, Nfis3, Ncnum / \Sigma Nm \leq 8 \}$

$Ex2prob = \{Nalg1 , Nalg2, Nalg3, Nprob / \Sigma Nm \leq 8\}$

$Ex2proy = \{Nproy1 , Nproy2, Nproy3, Ncomp1 / \Sigma Nm \leq 8\}$

$Ex2comp2 = \{ Ncomp1, Ndibing, Ncomp1 / \Sigma Nm \leq 8\}$

Donde
Nalg1 =Numero Veces Algebra1, Nalg2 =Numero Veces Álgebra 2, Nalg3 =Numero Veces Álgebra 3, Nedif = Numero Veces Diferenciales, Nprob = Numero Veces Probabilidad, Ncal1 = Numero Veces Calculo1, Ncal2 = Numero Veces Calculo2, Ncnum = Numero Veces CalculoNumerico , Nproy1=Numero Veces Proyecto1 , Nproy2= Numero Veces Proyecto2, Nproy3= Numero Veces Proyecto3, Ncomp1= Numero Veces Computación1 , Ncomp2=Numero Veces Computación2, Ndibing= Numero Veces DibujoIngeniería, Nfis1= Numero Veces Fisica1, Nqui= Numero Veces Química, Nfis2= Numero Veces Fisica2, Nfis3= Numero Veces Fisica3.

1.1.4 Riesgos y Contingencias

Los riesgos intrínsecos que estaban asociados, fueron un elemento a evaluar en los procesos de gestión del proyecto. En el caso particular el trabajo con Bases de Datos nos presento los siguientes riesgos inherentes

1. Inconsistencia de Datos (formato): Los Datos originales del modelo de datos, no coincidía con las utilizadas actualmente para la gestión de la información. Se realizaron procesos de evaluación y comprensión de datos
2. Incompatibilidad de Bases de Datos: Las bases de Datos que se manejaban en el Departamento de Informática y Admisión, no eran las mismas, por esta razón se utilizaron los datos de Informática, que contenían mayor cantidad de registros históricos.

1.2 Determinación de Metas de Minería de Datos

La investigación considero como metas de negocio ha satisfacer, la disminución de la cantidad de deserciones y repitencia de las asignaturas del Plan Común, que esta acordes a una institución sin fines de lucro, como es la Universidad de Antofagasta, buscando definir patrones de conducta e identificando cuales son las variables que más influencia en las distintas asignaturas.

1.2.1 Metas de Minería de Datos

En cuanto a metas específicas del proyecto se definió como medida cuantitativa los siguientes aspectos

1. Definir el rango de puntajes mínimos como conducta de entrada de los alumnos a la carrera de Ingeniería Plan Común
2. Definir las variables de mayor influencia en el comportamiento de los alumnos en las asignaturas de primer año de la carrera de Ingeniería Plan Común (Factores Críticos de Éxito)

2 Comprensión de los Datos

2.1 Colección Inicial de Datos

2.1.1 Reporte de la Colección de Datos

Los datos utilizados para el presente proyecto, como colección inicial fueron entregados por la Dirección y Administración de Registro Curricular (DARC)[3] de la Universidad Antofagasta, el modelo de datos a utilizar fue entregado por el Departamento de Informática[4].

Dentro de los Criterios de selección de los datos, podemos enumerar los siguientes :

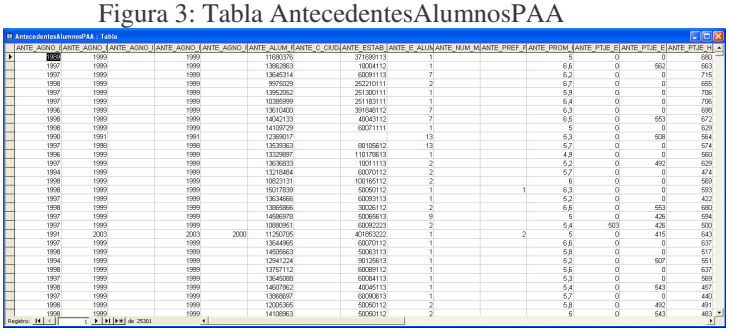
- Los datos de los alumnos que ingresa vía PAA[5] a la carrera de Ingeniería
- Datos históricos de 5 años de Ingresos de alumnos.
- Seguimiento curricular correspondiente a dos semestres, con un máximo de tiempo de aprobación (doble de la cantidad de años).

2.2 Descripción de los Datos

2.2.1 Reporte de Descripción de los Datos

En la mayoría de las tablas, la información estaba incompleta o simplemente era inexistente. El volumen de datos ascendían a 25.301 (ver figura 3).

Figura 3: Tabla AntecedentesAlumnosPAA



Fuente: Adaptación de Tabla en Access

---

[3] DARC: Departamento de Administración y Registro Curricular

[4] Departamento de Informática: Dentro de la Universidad Antofagasta esta inserto en el área de planificación, uno de los objetivos es mantener funcionando en forma optima los diferentes Sistemas de Información con que cuenta la Universidad, como así también mantener los Sistemas en líneas.

[5] Prueba Actitud Académica (PAA): Es un examen que evalúa las aptitudes de los alumnos que han terminado la Enseñanza Media, para la Facultad de Ingeniería esta evaluación debe ser superior a 450 puntos para poder postular (cuerdo tomado por Consejo de Rectores de Chile para las Universidades que pertenecen a el)

En una primera etapa se consideraron para el análisis preliminar, las tablas de Anteced_Postul (considera los datos de ingreso de los alumnos, p.e. puntaje de PAA), la tabla NotasRamos(considera las notas de los alumnos de todas las carreras de los años 1999 al 2001). Para este efectos se realizaron diversas Querys, para obtener la información de los alumnos de Ingeniería Civil, referente a los cuatros semestres con sus respectivas asignaturas y los datos de la PAA (análisis horizontal) y las notas de las líneas de conocimiento con los datos de la PAA (análisis vertical). Ambos análisis se consolidaron en la ConsultaFinal.

Los datos recogidos en la consulta final de Access, fueron consolidados, para finalmente migrar la tabla a un archivo en el Software SPSS versión 11.0, llamado ConsultaFinal.SAV .

2.3 Exploración de los Datos

Para iniciar la exploración de los datos, se importo el archivo ConsultaFinal.SAV al software de Data Mining SPSS Clementine, versión 6.5, herramienta que permite realizar los estudios estadísticos de los datos de acuerdo a su tipo y para cada campo, obteniendo ocurrencias, media, desviación estándar, error típico de la media y correlaciones.

En esta etapa de exploración primero se verifico la calidad de los datos, del total de registro del universo en estudio (2887 registros), obteniendo un 100 %.

En una segunda etapa se realizo los estudios estadísticos, para todos los campos, entregando información relevante sobre las desviaciones estándar y correlaciones, que en muchos casos permitieron encontrar relaciones interesantes entre los datos, que posteriormente se transformaron en hipótesis a confirmar.

Una de las hipótesis planteadas, se relaciona con las líneas de Cálculo y Álgebra, debido a la presencia de una correlación positiva alta, la cual es presente en todas las asignaturas de ambas líneas, y que se relacionan con las deserciones y la eliminación.

En paralelo en el Software SPSS 11.0 se amplio el estudio, para obtener las frecuencias que se presentan en los datos. El análisis se efectuó para cada unos de los campos, y se complemento con histogramas

3 Preparación de los Datos

3.1 Selección de los datos

Finalmente en la exploración de los datos se definió iniciar la selección para el proceso de minería de datos, considerando las metas de éxito 1 y éxito 2, para la hipótesis de la implicancia del éxito de las líneas de cálculo y álgebra. Para este efecto se definieron 2 conjuntos de 3 muestras para probar el modelo de 1000 registros cada una. Para el primer conjunto se definió el criterio de 1 de cada 2 registros, 1 de cada 3 registros, 1 de cada 4 registros y para el segundo conjunto se crearon 3 grupos de registros con una selección aleatoria.

3.2 Limpieza de Datos

La limpieza de datos es un proceso, en el cual se utilizan una gran cantidad de recursos, según Piramuthu [6], corresponden a un 80%, en el caso del proyecto tuvo una incidencia del 60% del tiempo con un costo asociado del 50 %. Los datos seleccionados, tenían un proceso de selección (querys), para tener solo los alumnos del plan común, como metodología de limpieza se eliminaron los alumnos que no tenían información de las asignaturas, y aquellos que no tenían cursadas las asignaturas se rellenaron los campos con valores ceros, para diferenciarlos de los alumnos que cursando la asignatura reprobaron con notas 1. La misma regla se utilizo para los campos relacionados con la Prueba de Aptitud Académica.

3.3 Construcción de Datos

En esta etapa se derivaron nuevos atributos, para preparar los datos para los algoritmos seleccionados, que para este caso, son los árboles de decisión y redes neuronales en función al problema de negocio presentado, que se puede definir como predictivo.

Este proceso de definición se realizo para que el algoritmo de árbol de decisión pueda generar un modelo más óptimo, a través de una derivación en un campo de tipo booleano. Esto entrega un grupo de alumnos que cumplen con parte de lo definido en éxito 1, que sirve para determinar la constante disminución en la cantidad de alumnos que pueden llegar al final del plan común.

En la exploración y manipulación de Datos surgieron algunas hipótesis, que se esperaban probar a través de las técnicas de modelado a utilizar, dichas técnicas necesitaron que se derivaran campos, es por esto que los campos iniciales se filtraron para no utilizarlos dos veces en el algoritmo.

4 Modelado

4.1 Selección de la técnica elegida

Las técnicas elegidas en función al problema de negocio presentado, inicialmente son:

a. Las redes neuronales
b. Los árboles de decisión

Estas técnicas se pueden definir como predictivos, con datos que son discretos (notas) y datos continuos (PAA), por ende un algoritmo de clasificación puede examinar las características de un nuevo alumno y asignarlo a una clase dentro de comportamiento de sus notas para el Plan Común.

4.2 Generación de diseño de prueba

El entrenamiento de los datos de prueba se presentará a través de las dos técnicas previstas, con las muestras seleccionadas en el modelo:

a. red neuronal
b. árboles de decisión

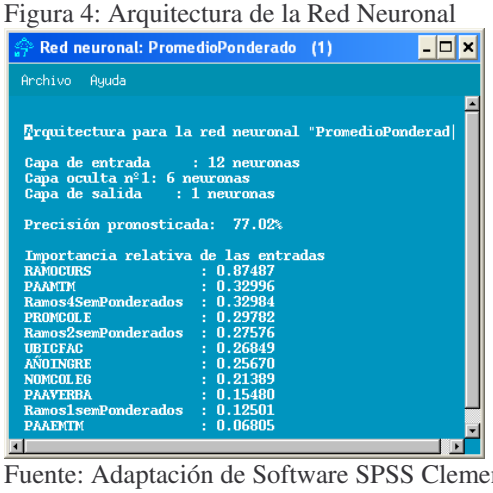
4.2.1 Red neuronal.

El entrenamiento de prueba de las muestras se ha realizado tras ejecutar el algoritmo de la red neuronal C5. Con ello obtendremos un modelo, que se espera sea potencialmente útil para el usuario final.

La información sobre la topología de la red, las estimaciones y la exactitud se puede visualizar en la figura 19, ella nos indica el número de neuronas en las capas de entrada, teniendo solo una salida. Así mismo la precisión es de un 77,02 %, para la salida PromedioPonderado.

En el análisis de sensibilidad, se tiene que el campo RamoCurs tiene una mayor importancia relativa con un 87%, lo sigue el puntaje de la PAA Matemáticas con una importancia relativa de un 32%.

Figura 4: Arquitectura de la Red Neuronal



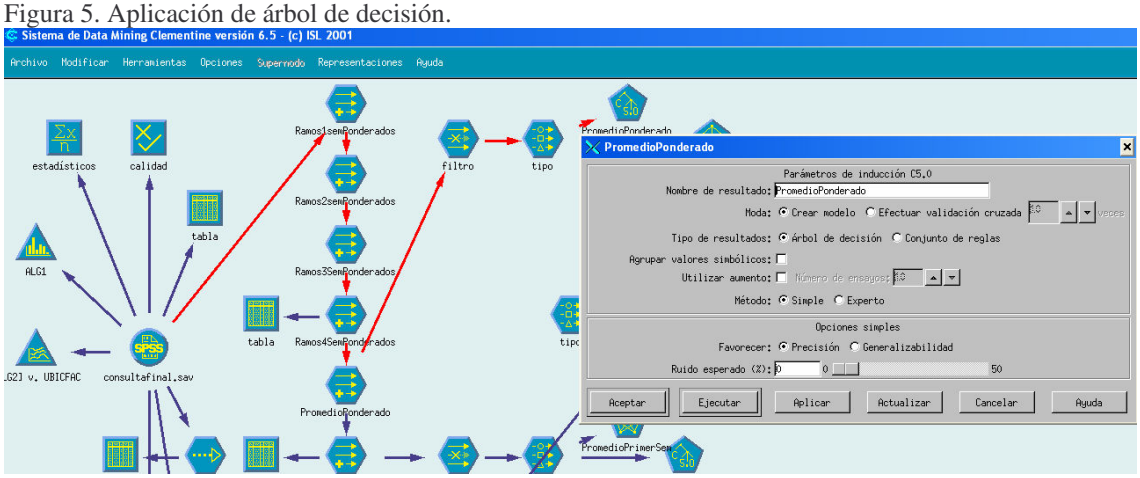
Fuente: Adaptación de Software SPSS Clementine

4.2.2 Árboles de Decisión.

El entrenamiento de los datos de prueba a través de la técnica árbol de decisión C5, en un proceso de descubrimiento supervisado, esta basado en los campos que nos proporcionarán el máximo de información posible (para el PromedioPonderado), esta nos entregará un conjunto de reglas, que para el usuario final sean potencialmente útiles. En esta etapa se evaluaron diversos algoritmos, optando por el C5, siendo el que ofrecía un mejor rendimiento, según Osei [7], esta fase de analizar diversos algoritmos, se podría automatizar a través de un proceso de multi-criterios.

En la figura 5, se muestra la ejecución (a través de la secuencia de flechas rojas) del algoritmo de Árbol de Decisión C5.

Figura 5. Aplicación de árbol de decisión.

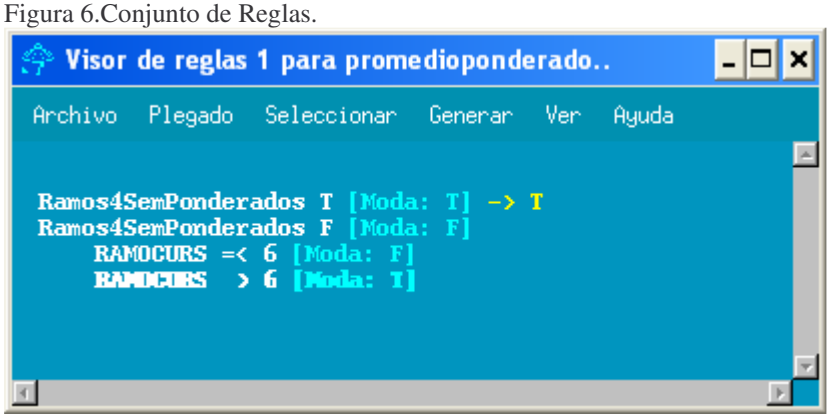


Fuente: Adaptación de Software SPSS Clementine

Tras aplicar el algoritmo, para construir un árbol de decisión C5, este nos da un conjunto de reglas (ver Figura 6), se puede interpretar que cuando el alumno cursa a los menos 6 ramos es muy factible que logre aprobar la totalidad de los ramos de plan común.

Los parámetros de entrada son los que se han definido en el punto 3.4 y se visualizan en la figura 6.

Figura 6.Conjunto de Reglas.



Fuente: Adaptación de Software SPSS Clementine

Se puede mencionar que respecto a los rendimientos a las variable de tiempo, los árboles de decisión son mucho más eficientes y entregan un resultado más claro para el usuario final.

5 Evaluación

Finalmente los algoritmos presentaron el siguiente conjunto de regla (se presentan las más relevantes) para Éxito 1,Éxito 2.

Regla 1 : Si en PAA Matemática >= 657 el Promedio Final Ponderado será verdadero.
Regla 2: Si en Promedio Colegio >= 6,1 el Promedio Final Ponderado será verdadero
Regla 3: Si Nota Colegio >= 5,6 Y PAA Matemática >=614 el Promedio Final Ponderado será verdadero
Regla 4: Si Álgebra 1 >=4 entonces Calculo 1 será Aprobado.

Donde: Éxito 1 = Promedio Final Ponderado.

Para las reglas anteriormente mencionadas, el análisis de los resultados son los siguientes:
Regla 1 → 61% Datos Correctamente Clasificados
Regla 2 → 70% Datos Correctamente Clasificados
Regla 3 → 65% Datos Correctamente Clasificados
Regla 4 → 95% Datos Correctamente Clasificados.

6 Consideraciones Finales

La investigación de Predicción del Rendimiento de los Alumnos de la Facultad de Ingeniería de la Universidad de Antofagasta, a través de Minería de Datos, nos permite presentar las siguientes consideraciones finales:

i. Existe una relación importante entre las asignaturas de Cálculo y Álgebra, que inciden directamente en el rendimiento de los estudiantes de ingeniería. Debido a que la asignatura de Cálculo es el prerrequisito más importante para acceder a las asignaturas finales de cuarto semestre de Ecuaciones Diferenciales, Calculo Numérico, Física III, que corresponde a un 60% de los requisitos para éxito 1, con capacidad de predicción del 95% de los casos.

ii. Las notas y el tipo de colegio, tiene un alto grado de incidencia en el rendimiento de los estudiantes, con una capacidad de predicción del 70 % de los casos.

iii. Los resultados relacionados a la Prueba de Aptitud Académica, no necesariamente pueden presentar resultados similares con la actual PSU.

7 Referencias

[1] Dhammika, A. Mining data to find subset of high activity. Journal of Statistical of Planing & Inference.Vol. 122 Issue ½ , (May 2004), pp 19-34.

[2] Salpeter, J. Data: Mining with a Mission. Technology & Learning. Vol. 24, Issue 8, (March 2004), pp30-36.

[3] Thomas, E. What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis. Vol. 45, Issue 3, (May 2004), pp 30-36.

[4] Gornitzka, A. Towards professionalisation ? Restructuring of administrative work force in universites. Vol. 47, Issue 4, pp 445-472.

[5] Chapman, P. Crisp DM 1.0 Step by Step Data mining guide. Crisp DM Consortium.

[6] Piramuthu, S. Evaluating feature selection methods for learning in data mining aplications. European Journal of Operational Research. Vol. 156 Issue 2, pp 483-495.

[7] Osei B. Evaluation of decision trees: a multi-criteria approach. Vol 31. Issue 11, (September 2003), pp 1933-1946.