Universidad Central “Marta Abreu” de Las Villas.

Facultad Matemática, Física y Computación

Licenciatura en Ciencias de la Computación

Trabajo de Diploma

PREDICCIÓN DE LAS DESERCIONES EN ESTUDIANTES DE PRIMER AÑO DE INGENIERÍA INFORMÁTICA EN LA UNIVERSIDAD DE CAMAGÜEY APLICANDO TÉCNICAS DE MINERÍA DE DATOS

Autor: José Ramón Abadía Lugo

Tutores: Dr. C. Julio Madera Quintana
Dr. C. Yailen Martínez Jiménez

Santa Clara
Junio 2013

Dictamen.

Hago constar que el presente trabajo fue realizado en la Universidad Central “Marta Abreu” de Las Villas como parte de la culminación de los estudios de la especialidad de Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

Firma del tutor

Firma del jefe del Laboratorio

Dedicatoria

Le dedico este Trabajo de Diploma a mi familia y a todos los que de una forma u otra contribuyeron a la realización del mismo y muy especialmente:

A mi abuela Nana, por haberme dado tantos momentos felices.

A mi pequeña Saskia María por poner todos los días una sonrisa en mi rostro.

A mi esposa Ani por haber sido tan paciente y por darme el mejor de los regalos.

A mis padres y hermano, Saskia, Jose Ramon y Joan por confiar siempre en mí y apoyarme en todo.

A mi tía Alla por darme su amor incondicional.

A mi suegra Anita por tanta ayuda recibida.

Agradecimientos

Para el desarrollo de una investigación científica son muchas las personas e instituciones que brindan su apoyo profesional, moral y material, por eso deseo agradecer especialmente:

Al claustro de profesores de la carrera de Ciencias de la Computación de la Universidad Central Martha Abreu de las Villas.

A la Universidad de Camagüey por la asesoría brindada y el apoyo al avance científico de esta investigación.

A mis tutores Dr. Julio Madera Quintana y Dra. Yailen Martínez Jiménez por la conducción científica, por el rigor mostrado, por el apoyo incondicional para llevar a feliz término este proyecto.

A todos aquellos que me brindaron su apoyo profesional, moral y material.

Muchas Gracias

RESUMEN

La deserción escolar es un tema de análisis constante en cualquier tipo de educación, principalmente la superior. Esto está muy relacionado con el futuro desarrollo de un país y la inversión que realiza en la preparación de los futuros profesionales. El análisis de la temática se puede realizar desde varias perspectivas y ramas de la ciencia como son la pedagogía, la estadística, la metodología de la enseñanza y en particular el análisis automático a partir de la aplicación de técnicas de minería de datos. Es por ello que en esta tesis se propone extraer la información de los estudiantes de la carrera de Ingeniería Informática de la Universidad de Camagüey almacenada en el SIGENU y utilizando estas técnicas computacionales caracterizar los principales rasgos que influyen en la deserción y los arrastres del primer año. Para esto se propone la aplicación de una metodología específica a partir de la metodología CRISP-DM para validar los modelos propuestos, con la particularidad del desbalance entre las clases de la base de conocimiento. Por último y no menos importante se validó la propuesta de los mejores modelos con el primer año actual de la carrera, obteniéndose un índice de aciertos elevados con relaciones a los estudiantes que llevan mundiales y su posibilidad de causar baja o pasar a segundo año con arrastres.

ABSTRACT

School dropout is a subject of constant analysis in any type of education, especially higher education. This is closely related to the future development of a country and the investment made in the preparation of future professionals. The thematic analysis can be done from various perspectives and branches of science such as pedagogy, statistics, teaching methodology and in particular the automatic analysis from the application of data mining techniques. That is why in this thesis to extract information from the students of Computer Engineering degree at the University of Camagüey SIGENU stored in the computer and using these techniques to characterize the main features that influence dropping out and drag the first year. For this we propose the application of a specific methodology from the CRISP-DM methodology to validate the proposed models, with the particularity of the imbalance between the classes of the knowledge base. Last but not least validated the proposal of the best models with the first current year career, obtaining a high hit rate with links to world leading students and their ability to cause low or take second year trawl.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

En cualquier proyecto educativo, los temas de la retención escolar y la culminación de estudios de forma exitosa por parte de los estudiantes, son elementos esenciales en su continuidad y calidad. Los problemas más complejos que enfrentan las instituciones de educación son la deserción, la reprobación, el rezago estudiantil y los bajos índices de eficiencia terminal **(ANUIES, 2003)**.

La deserción estudiantil en los programas de pregrado en los centros de educación superior es un problema que tiene un impacto multidimensional en el desarrollo social y económico de cualquier país **(Timarán Pereira, 2010)**.

Se entiende por deserción estudiantil al hecho de que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella, por repetir cursos o por retiros temporales **(UPN, 2009)**.

Las causas de la deserción estudiantil son múltiples y algunas de ellas dependen de las particularidades del contexto, constituyendo una necesidad para las instituciones de la educación superior conocer estos factores causales, para en correspondencia implementar estrategias más flexibles que logren la adaptación del estudiante a las condiciones educativas o a los cambios en su medio. De esta forma, se les ayudaría a enfrentarse a un nuevo entorno de enseñanza-aprendizaje, lleno de retos y competencias.

Por lo cual, lograr identificar las causas reales que provocan tanto la deserción escolar como la culminación exitosa de los estudios universitarios; potenciaría las capacidades de las instituciones universitarias y sus claustros para disminuir las primeras, y potenciar las segundas; en una solución más integral y duradera a esta problemática.

Este tipo de solución, necesariamente requiere de información previa y actual sobre los antecedentes y el desempeño de los estudiantes, que permita detectar, evaluar y corregir dichas causales de forma oportuna.

La necesidad de manejar información oportuna y útil, automatizar los procesos y obtener gran capacidad de almacenamiento de la información, es una premisa indispensable de

la gerencia de la información. El entorno educativo y las universidades, en particular, no escapan a esta realidad. Muchos han sido los sistemas de control de estudiantes y de gestión integral de dicha actividad que existen en el mundo. Sin embargo, aún su explotación eficaz es limitada, así como los impactos en el perfeccionamiento de la gestión universitaria, lo que ha limitado la obtención de resultados, tanto en el proceso formativo de los estudiantes, como en la calidad del graduado.

Existen numerosas herramientas informáticas, que permiten extraer información útil e implícita en grandes volúmenes de datos. No obstante, la gran mayoría no está al alcance de todas las organizaciones, ya que depende mucho de sus recursos económicos; tal es el caso del entorno educativo en general. Debido a esto, muchas veces se emplean como alternativa otras herramientas que permiten, por sus características, brindar información valiosa del estado actual de las organizaciones.

A pesar de ello, dar respuesta a interrogantes de por qué un centro educativo se comporta de una forma determinada, o cómo se comportará en el futuro, requiere de herramientas más potentes y la implementación de técnicas dentro de las ciencias de la computación, que permitan automatizar características de perfiles humanos.

Una de las alternativas más viables, que puede utilizar un Centro de Educación Superior (CES) para recopilar información estratégica es el uso de técnicas de extracción de conocimiento o minería de datos en educación, lo que ha dado lugar a la denominada minería de datos educativa (*Educational Data Mining*, EDM) **(Romero y Ventura, 2007)**. Esta nueva área de investigación se ocupa del desarrollo de métodos para explorar los datos que se dan en el ámbito educativo, así como de la utilización de estos métodos para entender mejor a los estudiantes y los contextos en que ellos aprenden **(Romero y Ventura, 2010)**.

Las técnicas EDM ya se han empleado con éxito para crear modelos de predicción del rendimiento de los estudiantes **(Kotsiantis, *et al.*, 2010)**, obteniendo resultados prometedores que demuestran cómo determinadas características sociológicas, económicas y educativas de los alumnos pueden afectar en el rendimiento académico **(Más-Estellés, *et al.*, 2009)**.

El CES, que será el núcleo para la aplicación del presente trabajo, es la Universidad de Camagüey "Ignacio Agramonte Loynaz" (UC), específicamente la carrera de Ingeniería Informática.

En la UC, a pesar de contar con información de utilidad almacenada, no existe un estudio para predecir la deserción de los estudiantes en ninguna de las carreras que en ella se imparten. El poder determinar qué factores influyen en el fracaso de un estudiante, la baja voluntaria, entre otros, conllevaría a la aplicación de estrategias más efectivas que ayuden a minimizar este fenómeno y contribuyan al mejoramiento de la calidad educativa en esta institución.

Tomando en cuenta todo lo anteriormente expuesto, se plantea como **problema científico:**

Limitaciones en la caracterización de los factores que influyen en al fracaso de los estudiantes de primer año que ingresan a la carrera de Informática en la UC y a partir de los cuales se pueda predecir las posibles bajas académicas y arrastres.

A partir del problema identificado, se reconoce como **objeto de estudio** el proceso de caracterización de factores influyentes en el fracaso estudiantil en carreras universitarias.

Definiéndose como **objetivo general** del presente trabajo:

Aplicar técnicas de Minería de Datos, para caracterizar los factores que influyen en el fracaso de los estudiantes de primer año que ingresan a la carrera de Informática en la UC y predecir las posibles bajas académicas y arrastres.

Para dar cumplimiento al objetivo general de la presente investigación y al problema planteado, se han definido los siguientes **objetivos específicos**:

1. Seleccionar la metodología, adecuando la misma a las particularidades del objeto de estudio, las herramientas y técnicas para el análisis de datos utilizando la minería de datos.
2. Crear la base de conocimientos con la información de los estudiantes de Informática a partir de los datos almacenados en el SIGENU.
3. Validar las técnicas de minería de datos aplicando la metodología seleccionada sobre la base de conocimiento creada.
4. Aplicar las técnicas que mejores resultados arrojaron, al primer año actual de Informática, para verificar la funcionalidad de los modelos obtenidos.

Teniendo en cuenta el problema formulado se plantea como **hipótesis de la investigación** que: La aplicación de técnicas de Minería de Datos permite extraer conocimientos que se encuentran implícitos en la base de datos del SIGENU, contribuyendo a la caracterización de los factores que influyen en el fracaso de los estudiantes de primer año que ingresan a la carrera de Informática en la UC.

Al trabajo se le concede valor desde el punto de vista **metodológico y práctico**; pues la metodología propuesta constituye una herramienta que influye en el estilo de trabajo organizacional, a partir que facilita y agiliza el proceso de información argumentada y oportuna para la toma de decisiones. En lo práctico esta herramienta brinda la posibilidad de utilizar información útil de carácter proyectivo, en correspondencia al cumplimiento de los objetivos a mediano y largo plazo y teniendo en cuenta el perfil de los estudiantes para garantizar su culminación de estudios de forma exitosa. Se establece como novedad científica la definición de una metodología específica para la proyección de la deserción estudiantil en el primer año de la carrera de Ingeniería Informática de la UC a partir de la metodología general CRISP-DM.

El trabajo se ha organizado en Introducción, tres Capítulos, Conclusiones, Recomendaciones, Bibliografía y Anexos. El **primer capítulo**, recoge el marco teórico relacionado con los procesos de Minería de Datos y Descubrimiento del Conocimiento en Bases de Datos. Mencionando las principales características, definiciones, objetivos y los diferentes tipos de problemas a los que da solución la Minería de Datos.

En el **segundo capítulo** se abordan las herramientas utilizadas para dar solución al problema, explicando, para cada una, en que parte del proceso se utilizará. Se diserta sobre la metodología CRISP-DM y se propone una adecuación de la misma para nuestro problema en particular.

En el **tercer capítulo** se describe el procedimiento propuesto para incorporarlo al proceso de descubrimiento de conocimiento en bases de datos. Se detallan las acciones, las técnicas y experimentos realizados con vistas a lograr una propuesta para la caracterización de los factores influyentes en el fracaso de los estudiantes, guiadas por dichas técnicas informáticas.

CAPÍTULO I

ACERCA DE LA MINERÍA DE DATOS Y EL PROCESO DE DESCUBRIMIENTO EN BASES DE DATOS

1. ACERCA DE LA MINERÍA DE DATOS Y EL PROCESO DE DESCUBRIMIENTO EN BASES DE DATOS

El presente capítulo está dedicado a dar los fundamentos teóricos sobre la temática abordada, para lo que se abordará sobre las técnicas informáticas que se utilizan para la caracterización. Particularmente se diserta acerca de la Minería de Datos y los procesos y técnicas que la conforman, por constituir la base de la propuesta realizada para la solución del problema relacionado con la caracterización del perfil de estudiantes en las carreras universitarias, determinando los elementos principales que se presentan en dicha caracterización.

1.1 Minería de Datos (*Data Mining*)

El concepto de *Data Mining* o Minería de datos (MD) data desde los años 60, cuando los estadísticos manejaban términos como: *Data Fishing*, *Data Mining* o *Data Archaeology*. La idea principal era encontrar correlaciones sin una hipótesis previa en Base de Datos (BD) con ruido.

Similar situación se presenta con los modelos estadísticos presentes en la MD, los cuales no son nuevos. Los árboles de decisión y de regresión (*classification and regression trees - CART*) son utilizados desde los años 60 del siglo XX. Las bases de reglas fueron popularizadas durante el auge de los Sistemas de Expertos en los 80 del pasado siglo y las redes neuronales se conocen desde los años 40 del siglo XX, pero han sido necesarios varios años de desarrollo para que fueran utilizables de manera sencilla.

Fue a principios de la década del 80 del pasado siglo que Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de DM y Descubrimiento del Conocimiento en BD.

En las últimas décadas, pocas empresas han hecho uso de la MD. Actualmente existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Los foros de

discusión están integrados por investigadores de más de 80 países, y han sido un punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios. En el ámbito educativo existen experiencias en universidades como: el Instituto Superior Politécnico José Antonio Echeverría (CUJAE), La Habana **(Brito, *et al.*, 2008)**; el Colegio de Bachilleres del Estado de Baja California, México **(Alvares, 2009)**; Universidad Nacional de Misiones, Argentina **(Pautsch, 2009)**; Universidad de Medellín, Colombia **(Colonia, 2010)**; Universidad de San Buenaventura, Sede Cali, Colombia **(Timarán Pereira, 2010)**; entre otras.

1.1.1 Tipos de conocimiento

Antes de describir la MD, dando alguna definición al respecto, es necesario comprender e identificar los tipos de conocimientos que se pueden extraer de una BD.
Se puede clasificar los tipos de conocimiento según las siguientes categorías:

- Evidente: esta información se puede obtener de las BD a través de consultas SQL.
- Multidimensional: modela una tabla con n atributos como un espacio de n dimensiones, lo que permite detectar varias situaciones difíciles de observar. Este tipo de análisis se logra utilizando herramientas OLAP (*On-Line Analytical Processing*).
- Oculto: es la información no evidente, desconocida hasta el momento, pero potencialmente útil, que puede obtenerse a través de técnicas de MD. Esta información tiene un gran valor, ya que hasta el momento se desconoce, y descubrirla permite tener una nueva visión del problema y de su solución (Figura 1.1).

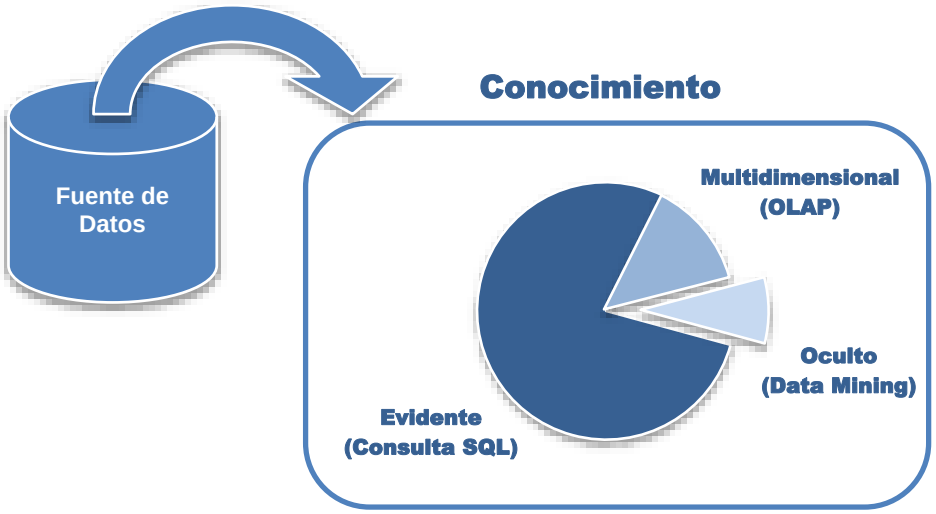

Figura 1.1 - Tipos de Conocimientos

Se estima que un 80% de la información contenida en una BD corresponde al conocimiento evidente (fácilmente recuperable). El otro 20% requiere de técnicas más complejas para su obtención.

Puede que esta cifra parezca despreciable, pero la información oculta en ese pequeño porcentaje puede ser de vital importancia para el éxito de la empresa u organización.

1.1.2 Definición de Minería de Datos

La Minería de Datos se define formalmente como "un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir, de forma automatizada, tendencias o comportamientos y descubrir modelos previamente desconocidos" **(Frawley, *et al.*, 1992)**.

El término Minería de Datos es utilizado indistintamente para referirse al proceso completo de Descubrir Conocimiento en Bases de Datos o KDD *(Knowledge Discovery in Databases)* **(Fayyad, 1996)**, se le puede definir de esta forma: "La integración de un conjunto de áreas que tienen como propósito la identificación de conocimiento obtenido

a partir de las bases de datos que aporten un sesgo hacia la toma de decisión" **(Molina, 2001)**.

De manera general, puede afirmarse que la MD constituye un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos y evaluar e interpretar los resultados **(Chen, *et al.*, 1996; Han y Kamber, 2001; Imielinski y Mannila, 1996)**.

1.1.3 Características y objetivos

En la actualidad, para realizar una investigación con el método científico tradicional, generalmente, primero se formula la hipótesis y luego el experimento, para posteriormente coleccionar los datos necesarios que confirmen o refuten la hipótesis. De esta manera se obtiene el nuevo conocimiento.

Una de las características principales de la MD es que invierte la dinámica del método científico. Es decir, primero se coleccionan los datos y luego se los "escucha" para que de ellos emerjan las hipótesis. Luego se validan esas hipótesis en los datos mismos.

Por lo antes expuesto es que la MD debe presentar un enfoque exploratorio, y no confirmador. Usar la MD para confirmar las hipótesis no sería correcto, ya que se está haciendo una inferencia poco válida y acotando el análisis sólo a la hipótesis elaborada.

No se debe confundir a la MD con un gran software ya que durante el desarrollo de un proyecto de este tipo, deben utilizarse diferentes aplicaciones para cada etapa. Las mismas pueden ser aplicaciones estadísticas, de visualización de datos o de inteligencia artificial.

El objetivo de la MD es extraer la información oculta en las profundidades de las BD, para luego intentar predecir futuras tendencias y comportamientos. De esta forma permiten a las organizaciones tomar decisiones proactivas y así adaptarse a un entorno permanentemente cambiante y sumamente competitivo o complejo.

Las técnicas utilizadas en la MD son el resultado de un largo proceso de investigación y desarrollo de productos, que comenzó cuando los datos de negocio fueron almacenados por primera vez en computadoras y luego, con tecnologías generadas para permitir que los usuarios naveguen entre los datos en tiempo real. La MD engloba todas estas técnicas para brindar información prospectiva y proactiva. La MD está lista para su aplicación ya que está sostenida por cuatro tecnologías que ya se encuentran suficientemente desarrolladas:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Data warehouse.
- Algoritmos de Data Mining.

En términos estrictamente académicos, los términos MD y KDD no deben utilizarse de manera indistinta.

La MD es un paso esencial en el KDD que utiliza algoritmos para generar patrones a partir de los datos pre procesados **(Frawley, *et al.*, 1992)**.

Como se describirá más adelante en este trabajo, la MD produce cinco tipos de información:

- Asociaciones.
- Secuencias.
- Clasificaciones.
- Agrupamientos.
- Pronósticos.

Uno de los factores claves que define la verdadera MD, es que la aplicación misma realiza el análisis sobre los datos. En otros casos, el análisis es guiado por una interacción con el usuario. Las aplicaciones que no son, en algún grado, auto guiadas están realizando análisis de datos y no MD.

1.1.4 Tipos de modelos

La Minería de Datos utiliza modelos predictivos y descriptivos sobre el conjunto de datos existentes, lo que permite el manejo y estructuración eficiente de la información para presentar datos visuales de gran utilidad en la toma de decisiones, generación de datos estadísticos y otras aplicaciones útiles en instituciones y empresas **(Agrawal y Shafer, 1996)**.

Descripción por cada modelo:

- Descriptivos o no supervisados: este modelo aspira a descubrir patrones y tendencias sobre el conjunto de datos sin tener ningún tipo de conocimiento previo de la situación a la cual se quiere llegar. Descubre patrones en los datos analizados. Proporciona información sobre las relaciones entre los mismos.
- Predictivos o supervisados: crean un modelo de una situación, donde las respuestas son conocidas y luego, lo aplica en otra situación de la cual se desconoce la respuesta. Conociendo y analizando un conjunto de datos, intentan predecir el valor de un atributo (etiqueta), estableciendo relaciones entre ellos.

1.1.5 Etapas de la Minería de Datos

En un proyecto de MD se deben tener en cuenta las siguientes etapas (ver Figura 1.2):

Figura 1.2 - Etapas en un proyecto de MD

Selección de datos

Los datos pueden tener un gran volumen y contener una cantidad inmensa de información. En esta etapa se reduce considerablemente el volumen de los datos, seleccionando solo los atributos y tuplas que aporten la información que sea más influyente sobre el tema a tratar.

Existen varios métodos para la selección de este subconjunto de atributos **(R. García, *et al.*, 2005)**. Entre algunos de ellos, se pueden citar:

- Selección por Pasos Hacia Adelante: se comienza con un conjunto vacío de atributos, en cada paso se agrega al conjunto el mejor atributo del conjunto original.
- Eliminación por Pasos Hacia Atrás: se comienza con un conjunto que posee todos los atributos originales, en cada paso se elimina del conjunto el peor atributo.
- Combinación de Selección por Pasos Hacia Adelante y Eliminación por Pasos Hacia Atrás: es una combinación de los dos anteriores. Se puede utilizar un umbral de medición para establecer cuándo detener la eliminación y agregación de los atributos.
- Inducción con árboles de decisión: se utilizan algoritmos como ID3 y C4.5. Los atributos que no son representados en el árbol se consideran irrelevantes y se les descarta. Por el contrario, los atributos que aparecen en el árbol son los elegidos para conformar el subconjunto de atributos.

Pre-Procesamiento de Datos

El formato de los datos de las distintas fuentes por lo general no suele ser apropiado. Esto dificulta que los algoritmos de minería obtengan buenos modelos trabajando sobre estos datos en bruto.

El objetivo del pre-procesamiento es adecuar los datos para que la aplicación a los algoritmos de minería sea óptima. Para esto hay que filtrar, eliminar datos incorrectos, no válidos, crear nuevos valores y categorías para los atributos e intentar completar o descartar los valores desconocidos e incompletos.

Extracción de Conocimiento

Es la aplicación de diferentes algoritmos sobre los datos ya pre-procesados, para extraer patrones.

Evaluación e Interpretación de Patrones

Una vez obtenidos los patrones se debe comprobar su validez. Si los modelos son varios, se debe elegir el que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, se debe volver a las etapas anteriores y modificar alguna entrada para, de esta manera, generar nuevos modelos.

1.1.6 Tipos de problemas de la Minería de Datos

Por lo general, los proyectos de minería de datos implican una combinación de diferentes tipos de problemas, que juntos solucionan el problema del negocio o institucional.

Descripción de datos y resumen

La descripción y el resumen de datos apuntan a la especificación concisa de las características de los datos, típicamente en forma elemental y agregada. Esto da al usuario una exposición de la estructura de los datos. A veces, una descripción y resumen de los datos puede ser el objetivo de un proyecto de minería de datos.

En casi todos los proyectos de minería de datos, sin embargo, la descripción y resumen de los datos son un objetivo subordinado en el proceso, típicamente en sus etapas tempranas. Al principio de un proceso de minería de datos, el usuario a menudo no conoce, ni el objetivo preciso del análisis, ni la naturaleza exacta de los datos. La exploración inicial del análisis de datos puede ayudar a los usuarios a entender la naturaleza de los datos y formar hipótesis potenciales de la información oculta. La estadística descriptiva simple y las técnicas de visualización proporcionan las primeras ideas sobre los datos.

La descripción y el resumen de datos típicamente ocurren en combinación con otros tipos de problemas de minería de datos. Es aconsejable llevar a cabo una descripción y resumen de datos antes de que cualquier otro tipo de problema de minería de datos sea especificado. El resumen también juega un papel importante en la presentación de los resultados finales.

Segmentación

La segmentación apunta a la separación de los datos en subgrupos o clases significativas e interesantes. Todos los miembros de un subgrupo comparten características comunes.

La segmentación puede ser realizada a mano o semiautomáticamente. El analista puede suponer ciertos subgrupos como relevantes para la pregunta de negocio o institucional, basada sobre un conocimiento previo o sobre el resultado de la descripción y el resumen de datos. En adición, hay también técnicas automáticas de agrupamiento (del inglés *clustering*) que pueden descubrir estructuras antes insospechadas y ocultas en datos que permiten la segmentación.

La segmentación a veces puede ser un objetivo de la minería de datos. Entonces la detección de segmentos sería el objetivo principal de un proyecto de minería de datos.

Muy a menudo, sin embargo, la segmentación es un paso hacia la solución de otros tipos de problema. Entonces, el objetivo es el de guardar o mantener el tamaño de los datos manejables o encontrar los subconjuntos de datos homogéneos que son más fáciles para analizar.

Entre las técnicas apropiadas para la segmentación se puede señalar:

- Técnicas de agrupamiento.
- Redes Neuronales.
- Visualización.

Descripciones de concepto

La descripción de concepto apunta a una descripción comprensible de conceptos o clases. El objetivo de las descripciones de concepto no es completar el desarrollo de modelos con predicción de exactitud alta, sino obtener ideas.

Una descripción de concepto tiene una conexión cercana tanto a la segmentación, como a la clasificación. La segmentación puede conducir a una enumeración de objetos que pertenecen a un concepto o clase sin proporcionar cualquier descripción comprensible. Típicamente la segmentación es llevada a cabo antes de que la descripción de concepto sea realizada.

Las descripciones de concepto también pueden ser usadas con objetivos de clasificación. Por otra parte, algunas técnicas de clasificación producen modelos de clasificación

comprensibles, que pueden entonces ser consideradas descripciones de concepto. La distinción importante es que la clasificación apunta a ser completa en algún sentido. El modelo de clasificación tiene que aplicarse a todos los casos en la población seleccionada.

Por otro lado, las descripciones de concepto no tienen que ser completas. Es suficiente si describen las partes importantes de los conceptos o clases. Las técnicas apropiadas de descripción son:

- Métodos de inducción de reglas.
- Agrupamiento conceptual.

Clasificación

La clasificación asume que hay un conjunto de objetos caracterizados por algún atributo o rasgo que pertenece a diferentes clases. La etiqueta de clase es un valor (simbólico) discreto y es conocido para cada objeto. El objetivo es construir los modelos de clasificación (a veces llamados clasificadores), que asignan la etiqueta de clase correcta a objetos no vistos antes y sin etiquetas. Los modelos de clasificación sobre todo son usados para el modelado predictivo.

Las etiquetas de clase pueden ser presentadas antes de la segmentación, o como consecuencia de ella. La clasificación es uno de los tipos de problemas más importantes de minería de datos que están presentes en una amplia gama de aplicaciones. Muchos problemas de minería de datos pueden ser transformados en problemas de clasificación

La clasificación tiene conexiones con casi todos los otros tipos de problemas. Los problemas de regresión pueden ser transformados en problemas de clasificación por discretización de etiquetas de clase continuas, porque estas técnicas permiten transformar rangos continuos en intervalos discretos. Estos intervalos discretos, más que los valores numéricos exactos, son usados como etiquetas de clase, lo que lleva a un problema de clasificación. Algunas técnicas de clasificación producen una clase comprensible o descripciones de concepto. Hay también una conexión al análisis de dependencias porque los modelos de clasificación típicamente explotan y aclaran las dependencias entre atributos.

La segmentación puede también proporcionar las etiquetas de clase o restringir el conjunto de datos para que se puedan construir buenos modelos de clasificación. Un modelo de clasificación también puede ser usado para identificar desviaciones y otros problemas con los datos.

Entre las técnicas de clasificación se encuentran:

- Análisis discriminante.
- Métodos de inducción de reglas.
- Árboles de Decisión.
- Redes neuronales.
- Vecino más cercano.
- Razonamiento basado en casos.
- Algoritmos genéticos.

Regresión

La regresión es muy similar a la clasificación. La única diferencia es que en la regresión el atributo objetivo (la clase) no es un atributo cualitativo discreto, sino que es continuo.

El objetivo de la regresión está en encontrar el valor numérico del atributo objetivo para objetos no vistos. Si la regresión trata con datos de series temporales, entonces a menudo se llama estimación.

Algunas técnicas de regresión destacadas son:

- Análisis de regresión.
- Árboles de regresión.
- Redes neuronales.
- El vecino más cercano.
- Métodos de la Caja-Jenkins.
- Algoritmos genéticos.

Análisis de dependencias

El análisis de dependencias consiste en encontrar un modelo que describa dependencias significativas o asociaciones entre datos o elementos. Las dependencias pueden ser

usadas para predecir el valor de un elemento, dada la información de otros. Las dependencias pueden ser estrictas o probabilísticas.

Las asociaciones son un caso especial de dependencias, que recientemente se han hecho muy populares. Las asociaciones describen las afinidades entre elementos. Los algoritmos para detectar asociaciones son muy rápidos y producen muchas asociaciones. Seleccionar el más interesante es un desafío.

El análisis de dependencias tiene conexiones cercanas a la regresión y a la clasificación, ya que las dependencias implícitamente son usadas para la formulación de modelos predictivos. Hay también una conexión con las descripciones de concepto, que a menudo destacan dependencias. El modelo secuencial es una clase especial de dependencias, en las que el orden de acontecimientos es considerado.

Algunas técnicas de análisis de dependencias destacadas son:

- Análisis de correlación.
- Análisis de regresión.
- Reglas de asociación.
- Redes bayesianas.
- Programación de lógica inductiva.
- Técnicas de visualización

1.2 Descubrimiento del conocimiento en bases de datos (KDD)

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos. Con la Minería de Datos (extracción de patrones) se estima ocupar solo del 15 al 20 por ciento del esfuerzo total del proceso de KDD **(Reyes y García, 2005)**. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas.

En este sentido, KDD implica un proceso interactivo e iterativo, involucrando la aplicación de varios algoritmos de Minería de Datos, siendo precisamente ésta la fase central del proceso, por ser aquí donde se buscan o descubren los patrones ocultos en los datos, los cuales pasan a una etapa de evaluación, donde se determina la validez y confiabilidad de dichos patrones. El proceso KDD concluye con la extracción de conocimiento a partir de la interpretación y evaluación de los patrones extraídos por la Minería de Datos.

En la Figura 1.3 se puede observar la jerarquía entre datos, información y conocimiento **(Molina, 2001)**. Por otra parte, el volumen de datos existente en cada nivel y el valor que los responsables de la toma de decisiones le atribuyen a cada jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. Se refleja también la estrecha unión entre datos e información. El corte de la pirámide representa la brecha existente entre la información y el conocimiento.

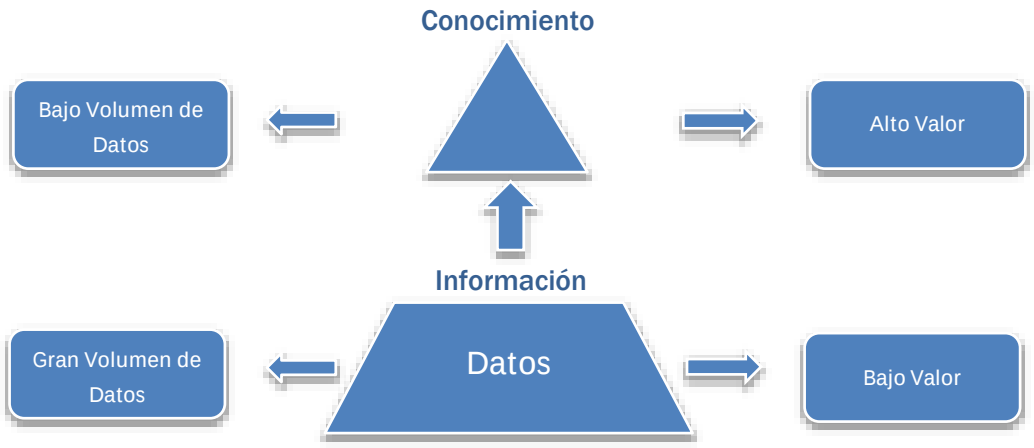

Figura 1.3 - Pirámide del conocimiento

El objetivo principal del KDD es encontrar conocimiento relevante y nuevo mediante la utilización de algoritmos eficientes, en grandes BD.

Una característica importante es la representación de los resultados y su visualización, de forma que su interpretación sea lo más clara posible para el usuario. La calidad de los modelos resultantes no debe verse afectada por el incremento en los volúmenes de datos, por el ruido o datos incompletos, debido a esto los algoritmos de descubrimiento de información deben ser altamente robustos.

1.2.1 Fases del proceso KDD.

El proceso de descubrimiento de conocimiento en bases de datos, KDD, involucra varias fases (ver Figura 1.4). En las primeras fases se deben preparar los datos para luego poder aplicar sobre ellos los métodos de Minería de Datos que permitan extraer el conocimiento.

- En la fase de **integración y recopilación** se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas; se transforman todos los datos a un formato común, y se detectan y resuelven las inconsistencias.
- En la fase de **selección, limpieza y transformación**, se eliminan o corrigen los datos incorrectos, y se decide la estrategia a seguir con los datos incompletos; además, se consideran únicamente aquellos atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería.

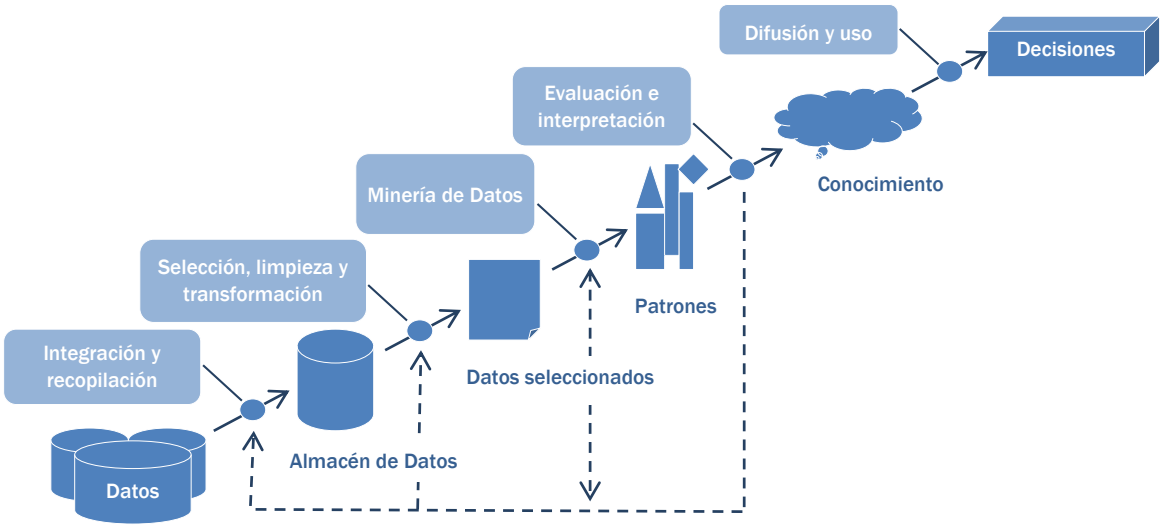

Figura 1.4 - Fases del proceso de KDD

- En la fase de **minería de datos**, proceso esencial donde se aplican diversos métodos (clasificación, clustering, redes neuronales) para extraer patrones de los datos. El modelo resultante dependerá del algoritmo utilizado.

- En la fase de **evaluación e interpretación** se realiza la identificación de patrones interesantes, basándose en algún parámetro de comparación impuesto por el usuario.

  Como resultado, esta etapa puede requerir volver sobre las operaciones y repetirlas, variando parámetros o utilizando otros algoritmos de minería. De esta manera se obtendrán nuevos patrones. Para esto es muy importante contar con conocimiento sobre el dominio abordado.
- Finalmente, en la fase de **difusión** se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles interesados.

Las fases que componen el KDD hacen que su desarrollo sea un proceso interactivo e iterativo. Es iterativo ya que la salida de alguna de las fases puede hacer volver a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario (experto en el dominio del problema) interviene en la toma de muchas decisiones **(Hernández, *et al.*, 2004)**.
Es oportuno señalar que las dos primeras fases se suelen englobar bajo el nombre de preparación de datos; y, por otro lado, en varias ocasiones se incluye previo a las fases descritas, una etapa de entendimiento del dominio para el análisis de las necesidades de la organización, o sea, para definir y priorizar los objetivos del negocio **(Hernández, *et al.*, 2004)**.

1.2.2 Relación con otras disciplinas

En las primeras etapas se deben preparar los datos para luego poder aplicar sobre ellos los métodos de MD que permitan extraer el conocimiento.
Observando con detenimiento los cuatro procesos iniciales descriptos en el apartado anterior (Limpieza, Integración, Selección y Transformación de datos), los mismos son los procesos que deben realizarse para la construcción de una *Data Wearhouse* (DW)
Por otro lado, los procesos de Minería de Datos, Evaluación de Patrones y Presentación de Conocimientos, generalmente se agrupan en otra etapa que se conoce como Minería

de Datos. De ahí la confusión que puede llegar a existir entre los conceptos de Minería de Datos y Descubrimiento del Conocimiento (ver Figura 1.5).

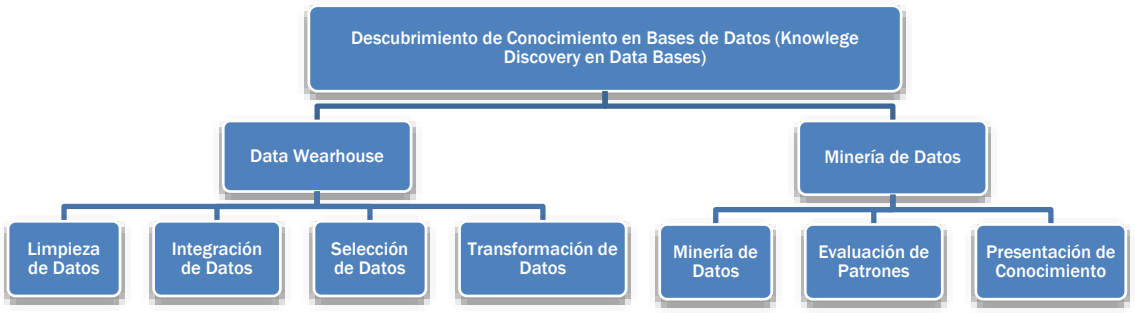

Figura 1.5 - Procesos que involucra KDD

1.2.3 Presentación y Utilización del Nuevo Conocimiento

Constituye una serio problema decisional, en la actualidad, tener mucha información disponible y no saber qué hacer con ella. En esta etapa se debe tener en cuenta los resultados de la evaluación y determinar una estrategia para desarrollar el nuevo conocimiento. Este debe ser utilizado de manera proactiva, reportándolo a los administradores de la organización o incorporándolo a los sistemas OLTP, ya que el valor real de los datos reside en la información que a futuro se pueda extraer de ellos y que aporte en el proceso de toma de decisión o contribuya a la comprensión de los fenómenos que rodean las organizaciones. Esta fase puede incluir la solución de conflictos con el conocimiento ya existente en la organización y sus integrantes.

Se deben elaborar estrategias para el desarrollo del conocimiento obtenido. La misma debe describir cuáles son las etapas que se llevarán a cabo y qué recursos y tareas son necesarias para su realización, debiéndose resumir las experiencias importantes obtenidas durante el desarrollo del proyecto. Un aspecto importante es la supervisión y el mantenimiento de cada etapa. La elaboración cuidadosa de la estrategia contribuirá a la utilización correcta de los resultados arrojados por la MD.

Es recomendable la organización de reuniones, en las cuales los resultados sean presentados a los administradores de la organización.

Los integrantes del proyecto, deben tener claridad que trabajar con esta tecnología, implica cuidar un sinnúmero de detalles, debido a que el producto final involucrado es la "toma de decisiones".

Conclusiones parciales

En el presente capítulo se abordaron de forma teórica lo referente a Minería de Datos y el proceso de Descubrimiento de Conocimiento en Bases de Datos o KDD. Se mencionan sus principales conceptos, características, objetivos, etapas, tipos de modelos, tipos de problemas, además de su relación con otras disciplinas. La utilización de la Minería de datos para dar solución al problema planteado sería la solución más eficiente teniendo en cuenta que lo que se persigue es encontrar una pequeña cantidad de información que sea de gran importancia para la toma de decisiones.

CAPÍTULO II

HERRAMIENTAS Y METODOLOGÍA PARA LLEVAR A CABO LA MINERÍA DE DATOS

2. HERRAMIENTAS Y METODOLOGÍA PARA LLEVAR A CABO LA MINERÍA DE DATOS

El presente capítulo está dedicado al análisis de las herramientas y la metodología que se utilizaron en la realización del proyecto. En el caso de las herramientas se mencionan las principales características y potencialidades para resolver problemas de MD. Se elabora una adecuación de la metodología CRISP-DM con el fin de dar solución al problema.

2.1 Herramientas para la Minería de Datos

Existen diversas herramientas que permiten implementar las técnicas de minería de datos, y que a su vez incluyen el pre-procesamiento de datos que se encuentran disponibles en grandes Bases de Datos. A continuación se relacionan las dos utilizadas, con sus características más relevantes:

2.1.1 KEEL

KEEL *(Knowledge Extraction based on Evolutionary Learning)* **(Alcalá-Fdez*, et al.*, 2011; Alcalá-Fdez*, et al.*, 2009)** es una herramienta de software libre desarrollada completamente en Java. KEEL permite al usuario emplear una gran cantidad de técnicas de aprendizaje automático en diferentes tipos de problemas: Regresión, clasificación, agrupamiento, asociación, etc., incluyendo una gran recopilación de los Sistemas Difusos existentes. Además de ser una herramienta para investigación, KEEL ha sido diseñado también con características educativas. Las características principales de KEEL son las siguientes:

- Posee una librería estadística para análisis de algoritmos. Los test de esta librería permiten analizar la bondad de los resultados obtenidos, realizando comparaciones paramétricas y no paramétricas.

- Incluye algoritmos de aprendizaje de modelos predictivos, de pre-procesamiento (discretización, selección de instancias, selección de características, etc.) y post-procesamiento. También incluye muchas propuestas del estado del arte de diferentes áreas de la minería de datos, como por ejemplo, árboles de decisión, sistemas difusos basados en reglas, etc.
- Ofrece al usuario una interfaz amigable, orientada al análisis de algoritmos.
- Permite crear experimentaciones conteniendo múltiples conjuntos de datos y algoritmos conectados entre sí. Los experimentos son generados mediante scripts independientes de la interfaz de usuario, para permitir una ejecución separada en la misma u otras máquinas.

2.1.1.1 La interface de KEEL

La versión actual de KEEL está compuesta por los siguientes módulos (ver Figura 2.1):

- **Tratamiento de datos** (*Data Management*): Este módulo contiene una serie de herramientas de tratamiento de datos: Importación, exportación, edición y visualización de datos, aplicación de transformaciones, etc.
- **Experimentos** (*Experiments*): Este módulo está dedicado al diseño de experimentos, proporcionando numerosas opciones: Tipo de validación, tipo de aprendizaje (clasificación, regresión, aprendizaje no supervisado), etc.
- **Educativo** (*Educational*): Este módulo permite realizar experimentos interactivos. Con una estructura similar al módulo anterior, permite diseñar experimentos con propósitos educativos.
- **Módulos** (*Modules*): Adicionalmente, KEEL incluye un módulo para datos no-balanceados **(Batista, *et al.*, 2004)**, un módulo de análisis estadístico no paramétrico **(Derrac, *et al.*, 2011; S. García y Herrera, 2008)**, y un módulo de aprendizaje multi-instancia **(Dietterich, *et al.*, 1997)**.

Figura 2.1 - Pantalla principal de KEEL 2.0

2.1.1.2 Módulo de tratamiento de datos

En el módulo de tratamiento de datos de KEEL se pueden realizar las siguientes operaciones (Ver figura 2.2):

- **Importar datos:** esta opción permite transformar archivos de diferentes formatos (*txt*, *excel*, *xml*, etc.) en el formato de datos de KEEL.
- **Exportar datos:** esta opción permite transformar del formato de datos de KEEL al formato de datos deseado (*txt*, *excel*, *xml*, *html table*, etc.).
- **Visualización de datos:** esta opción permite ver información detallada de los datos que se seleccionen. La información cambia de acuerdo a:
    - **DataSet View**: es similar a un editor de texto, sin embargo no se pueden modificar los datos.
    - **Attribute Info**: en esta etiqueta el usuario puede obtener información detallada de cada uno de los atributos definidos en el conjunto de datos.
    - **Charts 2D**: permite comparar atributos diferentes. La comparación se muestra en un gráfico, el cual se puede exportar en formato *pdf* o *png*.

Figura 2.2 - Módulo de tratamiento de datos de KEEL

- **Editar los datos**: esta opción permite modificar un conjunto de datos existente para agregar nuevos atributos, corregir errores, etc.
- **Particionar los datos**: esta opción permite crear particiones en los datos para su posterior procesamiento, cuenta con tres técnicas para particionar: validación cruzada con k particiones, validación cruzada 5x2 y Hold-Out.

2.1.1.3 Módulo para datos desbalanceados

El primer paso para crear un nuevo experimento con el módulo para datos desbalanceados, consiste en seleccionar qué conjuntos de datos se desean emplear (ver Figura 2.3).

La interfaz principal permite el diseño de experimentos de forma gráfica. La Figura 2.4 muestra los diferentes tipos de pre-procesamiento para datos desbalanceados con que cuenta KEEL, aunque también se pueden seleccionar otros tipos de pre-procesamiento. En la Figura 2.5 se observan las diferentes técnicas de clasificación específicas para datos desbalanceados, aunque también se pueden seleccionar otros tipos de técnicas. Arrastrando y colocando los iconos que representan a cada técnica, se puede diseñar fácilmente un experimento sin necesidad de emplear complicados procedimientos para

establecer los parámetros, algoritmos y conjuntos de datos a utilizar típicos de un diseño experimental mayor (ver Figura 2.6).

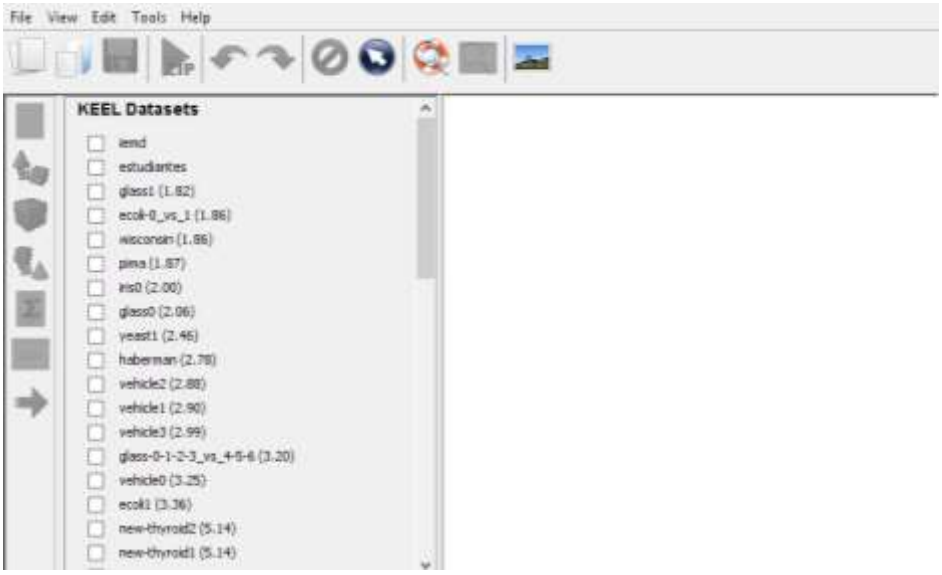

Figura 2.3 - Selección del conjunto de datos

En el módulo para datos desbalanceados una vez creado el experimento, el usuario lo guarda en formato *ZIP* y se ejecuta desde un archivo *.jar* existente en una de las carpetas. Una vez finalizado, se accede a los archivos de resultados obtenidos por cada algoritmo creado en estas carpetas.

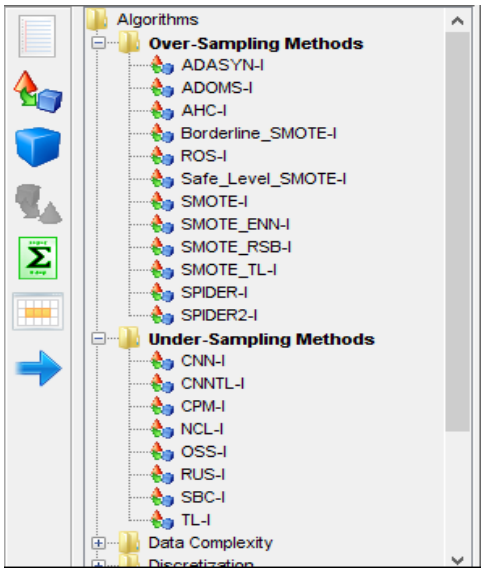

Figura 2.4 - Métodos de pre-procesamiento

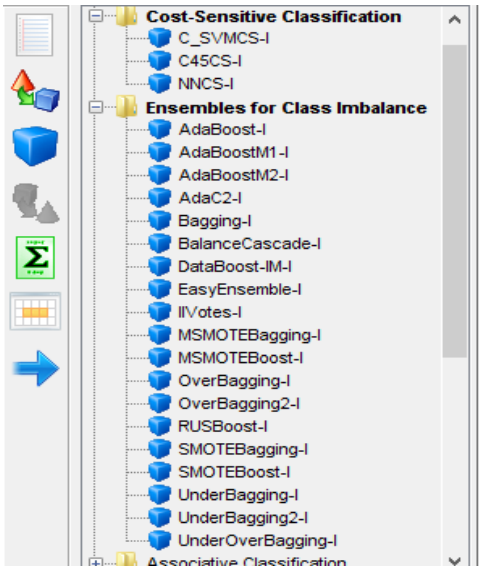

Figura 2.5 - Técnicas de clasificación

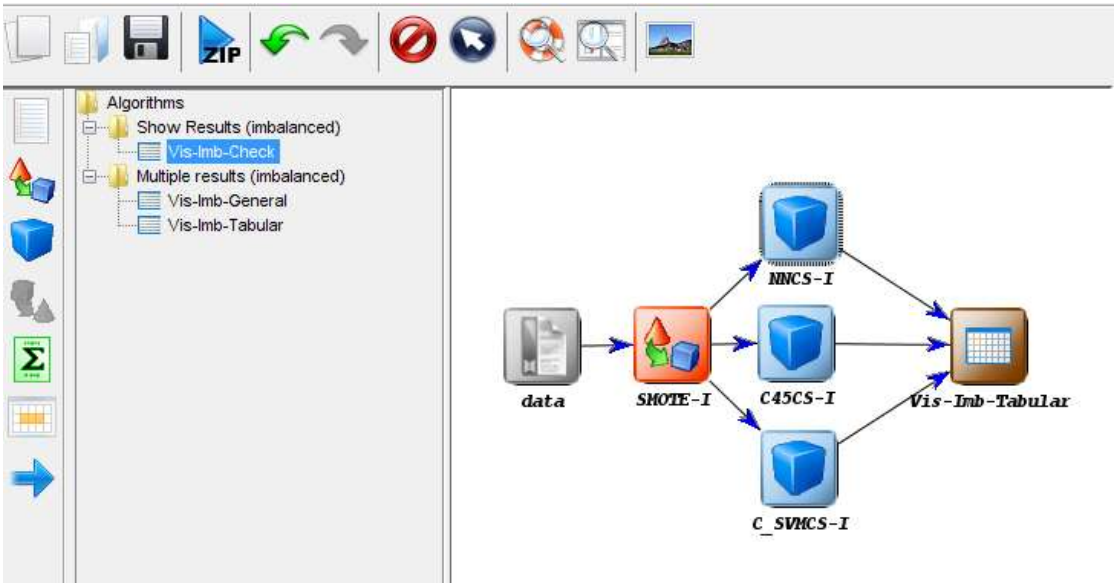

Figura 2.6 - Ejemplo de experimento en KEEL

2.1.2 WEKA

WEKA, acrónimo de *Waikato Environment for Knowledge Analysis* (Entorno para Análisis del Conocimiento de la Universidad de Waikato), es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario **(Kairúz, 2008)**.

Constituye un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software desarrollado bajo licencia GPL, lo cual ha impulsado que sea una de las suites más utilizadas en los últimos años en el área.

Está constituido por una serie de paquetes de código abierto con diferentes técnicas de pre-procesado, clasificación, agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos.

Las ventajas de WEKA son las siguientes:

- Proporciona una alternativa para aquellos que piensan en términos de cómo los datos fluyen a través del sistema.
- Permite guardar los datos en diferentes formatos, para después exportarlos a otra herramienta que disponga de alguna técnica de aprendizaje automático que WEKA no posea.

2.1.2.1 Interfaces de WEKA

WEKA se distribuye como un fichero ejecutable comprimido de java (fichero "*.jar*"), que se invoca directamente sobre la máquina virtual JVM. En las primeras versiones de WEKA se requería la máquina virtual Java 1.2 para invocar a la interfaz gráfica, desarrollada con el paquete gráfico de Java Swing. En el caso de la versión, WEKA 3.7, se utiliza Java 6, donde la herramienta se corre desde el intérprete de Java.
La primera pantalla de WEKA muestra una serie de opciones en su parte lateral derecha, como se muestra en la Figura 2.7. Las más importantes son *Explorer* y *Experimenter*.

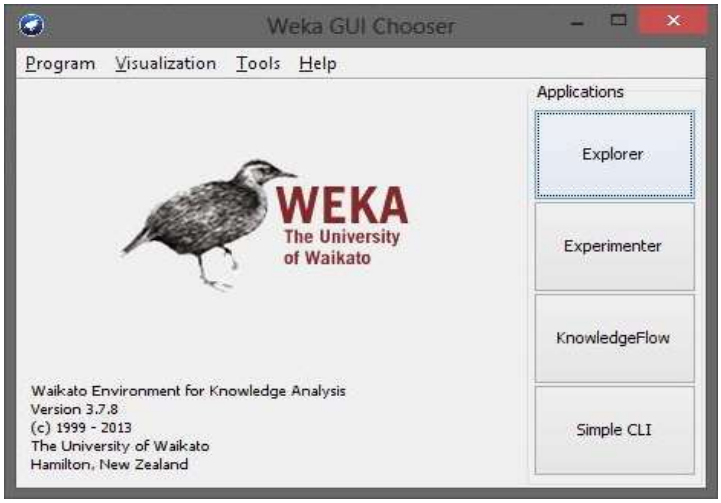

Figura 2.7 - Interfaz gráfica de WEKA 3.7.8

WEKA define cuatro entornos de trabajo:

- ***Explorer*:** Entorno visual que ofrece una interfaz gráfica para el uso de los paquetes.
- ***Experimenter*:** Entorno centrado en la automatización de tareas de manera que se facilite la realización de experimentos a gran escala.

- ***KnowledgeFlow*:** Permite generar proyectos de minería de datos mediante la generación de flujos de información.
- ***Simple CLI*:** Entorno consola para invocar directamente con java a los paquetes de WEKA.

2.1.2.2 Técnicas de Minería de Datos implementadas en la interfaz Explorer

La interfaz Explorer (ver Figura 2.8) permite visualizar y aplicar distintos algoritmos de aprendizaje a un conjunto de datos, admite el acceso a la mayoría de las funcionalidades integradas en WEKA de una manera sencilla. En él cada tarea de la Minería de Datos se representa por pestañas en la parte superior.

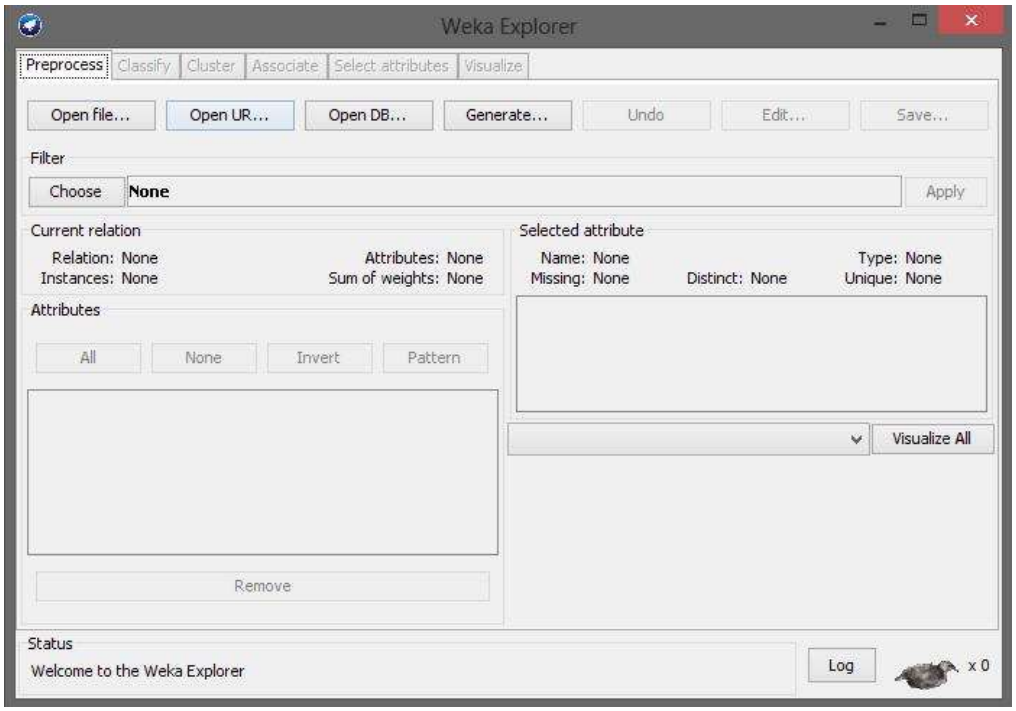

Figura 2.8 - Interfaz WEKA Explorer

- ***Preprocess*:** visualización y pre-procesado de los datos (aplicación de filtros), incluye las herramientas y filtros para cargar y manipular los datos. El *Preprocess*, tiene como objetivo construir un fichero que me permita clasificar la información contenida en la base de datos.

- ***Classify*:** Aplicación de algoritmos de clasificación y regresión permitiendo el acceso a estas técnicas.
- ***Cluster*:** Agrupación, integra varios métodos de agrupamiento.
- ***Associate*:** Asociación, incluye varias técnicas de reglas de asociación
- ***Select Attributes*:** Selección de atributos, permite aplicar diversas técnicas para la reducción del número de atributos
- ***Visualize*:** Visualización de los datos por parejas de atributos, permite estudiar el comportamiento de los datos mediante técnicas de visualización.

2.1.3 Elección de la herramienta a utilizar

Las dos herramientas descritas anteriormente poseen características importantes, dentro del ámbito de la Minería de Datos, para la solución de problemas.

Se decidió utilizar ambas herramientas para llevar a cabo esta minería. En la etapa de pre-procesamiento se determinó utilizar KEEL debido a que cuenta con un módulo dedicado especialmente a este tipo de análisis. Además se utilizaron los métodos de pre-procesamiento de conjuntos de datos desbalanceados contenidos en KEEL para obtener una mejor comparación entre estos. Sin embargo se decidió utilizar WEKA para realizar la clasificación debido a que, aunque KEEL cuenta con un módulo especialmente para trabajo con datos desbalanceados, solo trae muy pocos algoritmos de costo sensitivo, en cambio, la herramienta WEKA contiene un metaclasificador en el cual se le puede asignar costos a las clases independientemente del algoritmo que se utilice.

2.2 Metodología CRISP-DM

La metodología CRISP-DM *(CRoss-Industry Standard Process for Data Mining)* fue creada por un consorcio europeo de grandes empresas SPSS, NCR y Daimer Chrysler, que se unieron con el objetivo de crear una metodología de libre distribución, que se identifica por perseguir el cumplimiento de objetivos desde el punto de vista empresarial. CRISP-DM es actualmente la guía de referencia más utilizada en el

desarrollo de proyectos de Minería de Datos **(Chapman, *et al.*, 2000)** y cuenta con alrededor de 200 miembros del *CRISP-DM Special Interest Group* (SIG), incluidos proveedores de DM, consultores y usuarios finales.

Por ser una metodología de libre distribución, puede trabajar con cualquier herramienta, aplicando así una característica adicional que es el de ser una metodología equitativa.

2.2.1 Distribución jerárquica

La metodología CRISP-DM está descrita en términos de un proceso jerárquico consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de proceso (ver Figura 2.9).

En el nivel superior, el proceso de minería de datos está organizado en un número de fases; cada fase está formada por varias tareas genéricas que forman el segundo nivel. Este segundo nivel se denomina genérico por ser bastante general para cubrir todas las situaciones posibles de la minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible. Se entiende en este caso, que cubre tanto al proceso entero de minería de datos, como a todas las aplicaciones de minería de datos posibles. Estable significa que el modelo debe ser válido para acontecimientos normales y también para el desarrollo de imprevistos como nuevas técnicas de modelado.

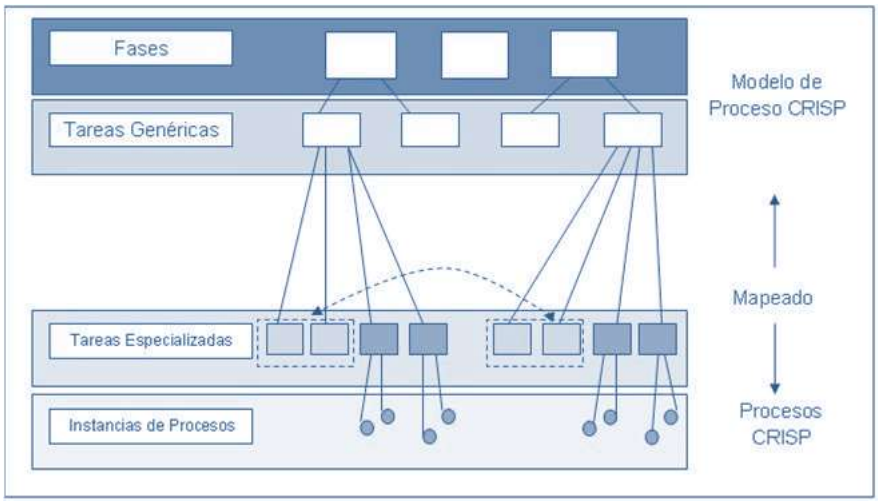

Figura 2.9 - Despliegue de la metodología CRISP-DM

El tercer nivel, el nivel de tareas especializadas, sirve para describir como deberían ser realizadas, en ciertas situaciones específicas, las acciones en las tareas genéricas. Por ejemplo, en el segundo nivel podría existir una tarea genérica llamada limpieza de datos, el tercer nivel describe cómo esta tarea se diferencia en situaciones distintas, distinguiendo por ejemplo la limpieza de valores numéricos, de la limpieza de valores categóricos, o si el tipo de problema es de agrupamiento o predicción.

La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos.

En la práctica, muchas de las tareas pueden ser realizadas en un orden diferente, siendo a menudo necesario volver a realizar tareas anteriores, repitiendo ciertas acciones. El modelo de proceso no intenta capturar todas las posibles rutas del proceso de la minería de datos, porque esto requeriría un modelo demasiado complejo.

El cuarto nivel, las instancias de procesos, es un registro de las acciones, las decisiones y los resultados de la minería de datos real.

Una instancia de proceso se organiza de acuerdo a las tareas definidas en los niveles superiores, pero representa lo que en realidad ocurre en un caso concreto, en lugar de lo que sucede en general.

2.2.2 Modelo de referencia y guía de usuario

Horizontalmente, en la metodología CRISP-DM se distingue entre el modelo de referencia y la guía de usuario. El modelo de referencia presenta una descripción rápida de las fases, tareas y sus salidas, describiendo qué hacer en el proyecto de minería de datos. La guía de usuario da consejos más detallados e indicaciones para cada fase y cada tarea dentro de una fase, también indica cómo realizar un proyecto de minería de datos siguiendo la metodología. Posteriormente, se describirá el modelo de referencia, sin embargo, a lo largo de la realización del proyecto la guía de usuario fue consultada.

2.2.3 Paso de modelos genéricos a especializados

El contexto de minería de datos traza un mapa entre el nivel genérico y el especializado en CRISP-DM. Actualmente, se distingue entre cuatro dimensiones diferentes de contextos de minería de datos:

- **El dominio de aplicación** es el área específica en la que el proyecto de minería de datos toma lugar.
- **Los tipos de problemas de minería de datos** describen la(s) clase(s) específica(s) de objetivo(s) con el que el proyecto de minería de datos trata.
- **El aspecto técnico** cubre cuestiones específicas en la minería de datos que describen las distintas dificultades (técnicas), que por lo general ocurren durante el proyecto de minería de datos.
- **La herramienta y dimensión técnica** especifica qué se aplica como herramienta y/o técnica de minería de datos durante el proceso de minería de datos.

Un contexto específico de minería de datos es un valor concreto para una o más de estas dimensiones.

En CRISP-DM se distinguen dos tipos de correspondencia entre el nivel genérico y el especializado.

- **Correspondencia para el presente**: Si sólo se aplica el modelo de proceso genérico para realizar un proyecto de minería de datos simple, intentando pasar de las tareas genéricas y sus descripciones al proyecto específico según las necesidades, se habla generalmente de una asignación para un sólo uso.
- **Correspondencia para el futuro**: Si sistemáticamente se especializa el modelo de proceso genérico según un contexto predefinido, se habla explícitamente de la sobre escritura de un modelo de proceso especializado en términos de CRISP-DM.

Cualquiera de los tipos de correspondencia es apropiada, según los objetivos, en dependencia del contexto de la minería de datos específico y de las necesidades de la organización.

La estrategia básica para pasar del modelo de proceso genérico al nivel especializado es la misma para ambos tipos de correspondencia:

- Analizar el contexto específico.
- Quitar cualquier detalle no aplicable al contexto.
- Agregar cualquier detalle específico al contexto.
- Especializar (o instanciar) el contenido genérico según las características concretas del contexto.
- Renombrar el contenido genérico si es posible, para proporcionar significados más explícitos en el contexto y de esta forma obtener mayor claridad.

2.2.4 Modelo de referencia de CRISP-DM

El modelo de proceso corriente para la minería de datos proporciona una descripción del ciclo de vida del proyecto de minería de datos. Éste contiene las fases de un proyecto, sus respectivas tareas y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones pero, si cabe mencionar, que podrían existir relaciones entre cualquier tarea de minería de datos según los objetivos, los antecedentes, el interés del usuario y, lo más importante, los datos.

El ciclo de vida del proyecto de minería de datos cuenta con seis fases, sin una secuencia rígida como se puede observar en la Figura 2.10.

El movimiento hacia adelante y hacia atrás entre las distintas fases es siempre necesario. El resultado de cada fase determina qué fase, o tarea particular de una fase, tienen que ser realizadas después. Las flechas indican las más importantes y frecuentes dependencias entre fases.

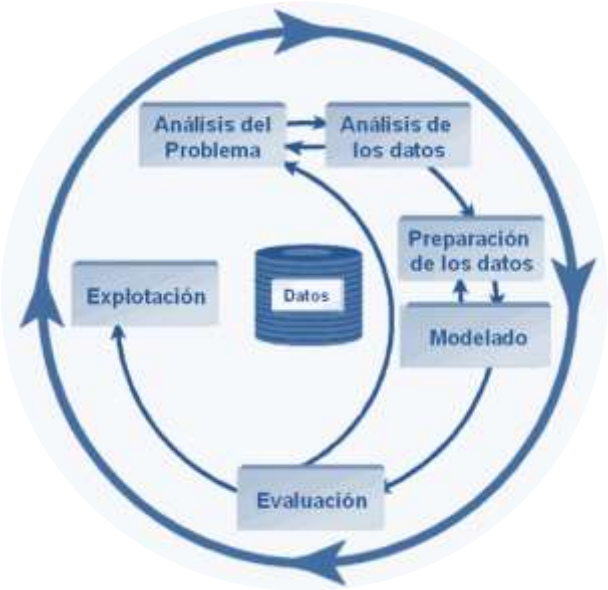

Figura 2.10 - Fases del modelo de referencias CRISP-DM

El círculo externo de la Figura 2.10 simboliza la naturaleza cíclica de la propia minería de datos. La minería de datos no se termina una vez que la solución es desplegada. Las informaciones ocultas durante el proceso y la solución desplegada pueden provocar nuevas preguntas. Los procesos de minería de datos sucesivos se beneficiarán de las experiencias previas.

La metodología propone un orden lógico a sus fases, aunque permite retrocesos entre varias de ellas, pues frecuentemente, a lo largo del desarrollo de un proyecto, es necesario volver atrás en numerosas ocasiones para re-analizar los resultados obtenidos.

Además, el proyecto se torna cíclico, pues, éste no se termina una vez que la solución es desplegada, ya que las informaciones obtenidas pueden provocar nuevas preguntas.

A continuación, se resumen las distintas fases que componen CRISP-DM:

Fase 1 Análisis del problema / Comprensión del Negocio

Esta fase inicial se enfoca en la comprensión de los objetivos y exigencias del proyecto, desde una perspectiva de negocio, para definir un problema de minería de datos y elaborar un plan preliminar diseñado para alcanzar dichos objetivos.

- **Determinación de los objetivos de negocio**: El primer objetivo del analista de datos para un contexto es entender, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. Se describen los criterios de éxito para un resultado acertado o útil para el proyecto desde el punto de vista del negocio.
- **Evaluación de la situación**: Se enuncian los recursos disponibles para el proyecto (personal, datos, recursos computacionales, otros). Se realiza un cronograma del proceso, se enumeran las presunciones, restricciones y disponibilidad de recursos. Se listan los riesgos que podrían retrasar el proyecto y los planes de contingencia correspondientes. Se realiza un análisis de costo-beneficio para el proyecto, tan específico como sea posible.
- **Determinación de los objetivos de la minería de datos**: Esta tarea tiene como fin representar los objetivos del negocio en términos de las metas del proyecto de minería de datos. Se definen los criterios de éxito de un resultado exitoso para el proyecto en términos técnicos, por ejemplo, un cierto nivel de predicción. Además, también puede expresarse en términos subjetivos, y en este caso, deben ser identificadas las personas que hacen el juicio.
- **Elaboración del plan del proyecto**: Se describe el plan para alcanzar los objetivos de minería de datos y con ello los del negocio; dicho plan debe especificar los pasos durante el resto del proyecto, incluyendo la selección inicial de herramientas y técnicas, y una lista de las etapas a ser ejecutadas, junto con su duración, recursos requeridos, entradas, salidas y dependencias.

Fase 2 Comprensión de los datos

Esta fase comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos.

Las principales tareas a desarrollar en esta fase del proceso son:

- **Recolección inicial de los datos:** Se confecciona una lista del conjunto de datos obtenidos, sus localizaciones y los métodos usados para obtenerlos.
- **Descripción de los datos iniciales**: Se describen los datos que han sido adquiridos, incluyendo su formato y cantidad; además, se evalúa si satisfacen las exigencias previstas.
- **Exploración de los datos:** Esta tarea está dirigida a responder interrogantes de minería de datos usando visualización y técnicas de reporte. De ser apropiado pueden ser incluidos gráficos para indicar las características de los datos, de donde se desprenden las conclusiones o hipótesis iniciales del proyecto
- **Verificación de la calidad de los datos:** Se examina la calidad de los datos en relación a si están completos, si son correctos, si contienen errores y qué tan comunes son estos, si existen valores omitidos, etc.

Fase 3 Preparación de los Datos

Esta fase cubre todas las actividades necesarias para conformar el conjunto de datos final (los datos que serán utilizados por las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y sin un orden preestablecido, e incluyen la selección de registros y atributos, así como el proceso de transformación y limpieza.

- **Selección de los datos**: Se decide qué datos serán excluidos y cuáles usados para el análisis, de acuerdo a su importancia respecto a los objetivos de la minería de datos, su calidad y las restricciones técnicas. Cubre la selección de atributos (columnas) así como la selección de registros (filas).
- **Limpieza de los datos**: Se debe elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos de datos limpios, la inserción de datos por defecto adecuados, o técnicas más ambiciosas tales como la estimación de datos ausentes mediante modelado.

- **Construcción de los datos**: Esta tarea incluye la construcción de operaciones de preparación de datos tales como la producción de atributos derivados, el ingreso de nuevos registros o la transformación de valores para atributos existentes.
- **Integración de los datos**: Son los métodos por el cual la información es combinada de múltiples tablas o registros para crear nuevos registros o valores.
- **Formato de los datos:** Se realizan modificaciones principalmente sintácticas a los datos que no cambian su significado, pero que si pueden ser requeridas por la herramienta de modelado.

Fase 4 Modelado

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son ajustados. Es a menudo necesario, de acuerdo a los algoritmos y técnicas seleccionadas, volver a la fase de preparación de los datos.

- **Selección de las técnicas de modelado:**
  - Técnica de modelado: Como primer paso durante el modelado, se debe seleccionar la técnica de modelado que será usada. Si son aplicadas múltiples técnicas, se realiza esta tarea de forma individual para cada una de ellas.
  - Suposiciones del modelado: Se registra cualquier presunción de la técnica de modelado seleccionada, que pueden ser, por ejemplo, que todos los atributos tengan distribuciones uniformes, que el atributo a predecir deba ser simbólico, etc.
- **Generación del diseño del experimento:** se describe el plan intencionado para el entrenamiento, la prueba y la evaluación de los modelos. Un componente primario del plan determina como dividir un conjunto de datos disponible en datos de entrenamiento y datos de validación.
- **Construcción y descripción de los modelos:**
  - Escenario de parámetros: Se listan los parámetros y los valores escogidos para los mismos, así como el razonamiento llevado a cabo para elegirlos.
  - Modelos: Se listan los modelos reales producidos por la herramienta de modelado, no un informe.

  - Descripción de los modelos: Se describen los modelos obtenidos, informándose su interpretación y documentándose cualquier dificultad encontrada con sus significados.
- **Evaluación del modelo**: Se resumen los resultados de esta tarea, listando las calidades de los modelos generados y comparando unos con otros.

Fase 5 Evaluación

En esta etapa, se evalúan los modelos construidos, revisando cada uno de los pasos ejecutados para crearlos, a fin de comprobar si cumplen correctamente con los objetivos del negocio. Un objetivo clave es determinar si hay alguna cuestión importante del negocio que no ha sido considerada suficientemente. Al final de esta fase, se toma una decisión sobre el uso de los resultados obtenidos en el proceso de minería de datos.

- **Evaluación de los resultados:**
  - Valoración de la minería respecto al negocio: Se resumen los resultados de minería de datos en términos de criterios de éxito del negocio.
  - Aprobación de los modelos: Después de la valoración de los modelos, se toma una decisión al respecto.
- **Revisión del proceso:** Se califica el proceso entero de minería de datos con el objetivo de identificar elementos que pudieran ser mejorados.
- **Determinar los próximos pasos:** Si se ha determinado que las fases, hasta este momento, han generado resultados satisfactorios, podría pasarse a la siguiente fase o, en caso contrario, podría decidirse realizar otra iteración desde la fase de preparación de los datos o de modelado. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de minería de datos.

Fase 6 Despliegue

La fase de despliegue puede ser tan simple como la generación de un informe o tan compleja como la repetición del proceso de minería a través de la organización. En muchos casos, es el cliente y no el analista de datos, quien lleva a cabo la fase de despliegue, sin embargo, resulta conveniente la participación de ambos para comprender rápidamente qué acciones ejecutar a fin de emplear los modelos obtenidos.

- **Planificación del despliegue:** De acuerdo al desarrollo de los resultados de la minería en el negocio, se determina una estrategia para su despliegue, donde se incluyen los pasos necesarios y cómo realizarlos.
- **Planificación de la monitorización y el mantenimiento:** Se resume la estrategia de supervisión y mantenimiento, incluyendo los pasos necesarios y cómo realizarlos, a fin de evitar largos periodos innecesarios de uso incorrecto de los resultados de minería de datos.
- **Generación del informe final:** Se redacta un informe escrito final del compromiso de la minería de datos, lo que incluye todo el desarrollo anterior, el resumen y la organización de los resultados. A menudo se realiza una reunión al finalizar en la que los resultados son presentados verbalmente
- **Revisión del proyecto:** De igual modo se resumen las experiencias importantes ganadas durante el proyecto.

2.3 Adecuación de la Metodología CRISP-DM

A partir del estudio teórico realizado sobre la Minería de Datos, la metodología CRISP-DM y las particularidades de aplicación vinculadas a la proyección de la deserción estudiantil en el contexto del primer año de la carrera Ingeniería Informática de la Universidad de Camagüey se define el siguiente esquema lógico metodológico:

Se determinan tres fases para la realización del trabajo de proyección de la deserción estudiantil a partir de la adecuación de las técnicas e instrumentos mencionados anteriormente, las cuales son:

- **Fase 1** Construcción de la Base de Conocimiento
- **Fase 2** Modelado
- **Fase 3** Validación

Dentro de la **Fase 1** se realizarán los siguientes pasos:

1. **Recolección inicial de los datos:** Se confecciona una lista del conjunto de datos obtenidos de la base de datos SIGENU y los datos externos, mencionando los métodos usados para obtenerlos.

2. **Descripción de los datos recolectados**: Se describen los datos que han sido adquiridos, incluyendo su formato y cantidad.
3. **Selección de los datos**: Se decide qué datos serán excluidos y cuáles usados para el análisis. Cubre la selección de atributos (columnas) así como la selección de registros (filas).
4. **Construcción de los datos**: Esta tarea incluye la construcción de operaciones de preparación de datos tales como la producción de atributos derivados, el ingreso de nuevos registros o la transformación de valores para atributos existentes.
5. **Integración de los datos**: Son los métodos por el cual la información es combinada, de la base de datos SIGENU y los datos externos, en la Base de Conocimiento.
6. **Formato de los datos:** Se realizan modificaciones principalmente sintácticas a los datos que no cambian su significado, pero que si pueden ser requeridas por la herramienta de modelado.

En la **Fase 2** se acometerán los siguientes pasos:

1. **Selección de técnicas de modelado**: Se describen las diferentes técnicas utilizadas para realizar el modelado.
2. **Construcción y descripción de los Modelos Experimentales**:
   a. **Conjuntos de datos:** Se describen los métodos utilizados para la creación de los distintos conjuntos de datos a utilizar. Se listan los conjuntos de datos disponibles para realizar la experimentación.
   b. **Minería de datos y experimentación:** Se listan y describen los modelos obtenidos, informándose su interpretación y documentándose cualquier dificultad encontrada con sus significados.
3. **Evaluación de los modelos experimentales:** Se resumen los resultados de esta tarea, listando las calidades de los modelos generados y comparando unos con otros. Lo que posibilita la retroalimentación del sistema ya que si la

evaluación es aceptable se pasa a la siguiente fase, en caso contrario se debe comenzar desde la Fase 1 nuevamente.

En la **Fase 3** se establecen los siguientes pasos:

1. **Construcción de los modelos de validación**:
   a. **Conjuntos de datos:** Se describen los conjuntos de datos utilizados para realizar la validación.
   b. **Modelos de validación:** Se listan y describen los diferentes modelos utilizados para llevar a cabo la validación de la base de conocimiento.
2. **Evaluación de los modelos de validación**: Se comparan los modelos obtenidos en el paso anterior, se determina el mejor modelo, y el mejor algoritmo utilizado. Se especifica a través de los resultados del mejor algoritmo utilizado los factores (atributos) que influyen en la deserción de los estudiantes de primer año que matriculan en la carrera Ingeniería Informática de la Universidad de Camagüey.

Este esquema lógico metodológico constituye la metodología específica a utilizar a partir de la adecuación de la Metodología CRISP-DM.

Conclusiones parciales

En este capítulo se hace una caracterización de las herramientas que se van a utilizar en el proyecto, mostrando para cada una, las potencialidades que posibilitarán la realización del mismo. Además se abordó exhaustivamente la metodología CRISP-DM la cual es un estándar para llevar a cabo proyectos de este tipo y consecuentemente se elaboró una adecuación de la misma.

La adecuación de la metodología CRISP-DM plantea los pasos necesarios para dar solución al problema de la predicción del fracaso en estudiantes universitarios de primer año y su aplicación será llevada a cabo en el siguiente capítulo.

CAPÍTULO III

APLICACIÓN DE LA METODOLOGÍA CREADA EN SOLUCIÓN AL PROBLEMA PLANTEADO

3. APLICACIÓN DE LA METODOLOGÍA CREADA EN SOLUCIÓN AL PROBLEMA PLANTEADO

En el presente capítulo se tratarán cada una de las etapas de la metodología específica creada a partir de CRISP-DM. Se realiza la construcción de la Base de Conocimiento y se prepara para su utilización. Luego se mencionan las técnicas utilizadas para esta minería y se explican cada uno de los experimentos realizados. Por último se evalúan los resultados obtenidos y se hace una validación del modelo propuesto.

3.1 Construcción de la base de conocimiento

En esta fase se presentan las distintas tareas realizadas, con el objetivo de construir una base de conocimientos, la cual se utilizará en la siguiente fase de modelado.

3.1.1 Recolección inicial de los datos

Teniendo en cuenta que en todas las carreras que ofrece la UC, se observa la no culminación de estudios ya sea por cambio de carrera o por deserción; esta información se podrá trabajar con los datos que suministra el SIGENU, listados de matrículas y los registros de calificaciones que se encuentran en secretaría.

La exportación de los datos se llevó a cabo haciendo diferentes consultas al SIGENU y corroborando esta información con los listados de matrículas y los registros de calificaciones de esta facultad. El SIGENU da la posibilidad de exportar la información en archivos con formato xls del software ofimática Excel.

Los archivos suministrados son:

- **Reporte Estudiantes.xls** con la información de los estudiantes inscritos en los años de 2003 a 2012; el archivo contiene 36 atributos, de los cuales, se utilizarán los más relevantes, con un total de 530 registros.

- **Listados de matrículas** de los estudiantes inscritos en los años de 2003 a 2011; los listados tienen diferentes estructuras, por lo general cuentan con 11 atributos de los cuales se utilizaron los más relevantes.
- **Registros de calificaciones**, solo se consultaron para determinar la situación del estudiante al concluir el primer año escolar.

3.1.2 Descripción de los datos recolectados

En este apartado se procede a describir los datos adquiridos en su formato original. Como es normal, estos datos tuvieron que ser tratados para poder formar con ellos una Base de Casos (BC) coherente y consistente con la que se pudiera trabajar a lo largo del proyecto. Como se ha descrito anteriormente los datos fueron obtenidos de dos fuentes diferentes: SIGENU y listados de matrículas. A continuación, se muestran los datos de cada una de ellas.

Datos del SIGENU

El archivo cuenta con 36 atributos (Ver Tabla 2.1).

Atributo	Descripción
Identidad	Número del carnet de identidad del estudiante que ingresa a la facultad.
Nombre	Nombre del estudiante que ingresa a la facultad.
Apellidos	Apellidos del estudiante que ingresa a la facultad.
País	Sexo del estudiante que ingresa a la facultad.
Provincia	Provincia de la cual procede el estudiante que ingresa a la facultad.
Municipio	Municipio del cual procede el estudiante que ingresa a la facultad.
Situación Académica	Situación actual que presenta el estudiante.
Estado	Estado en el que se encuentra el estudiante (pasivo o activo).
Dirección	Dirección particular del estudiante que ingresa a la facultad.
Fecha de nacimiento	Fecha de nacimiento del estudiante que ingresa a la facultad.
Grupo	Número de la brigada en la cual está el estudiante dentro de la facultad.
Carrera	Carrera que cursa el estudiante.
Facultad-Filial	Facultad que pertenece el estudiante.
Tipo de curso	Tipo de curso en el cual está matriculado el estudiante.
Correo	Correo del estudiante dentro de la universidad.
Fuente de ingreso	Fuente por la cual ingresa el estudiante.
Origen académico	Origen de enseñanza de la cual procede el estudiante.
Régimen de estudio	Régimen que de estudio del estudiante.
Natural de	Lugar de nacimiento del estudiante que ingresa a la facultad.
Teléfono	Teléfono particular del estudiante que ingresa a la facultad.
Fecha de ingreso a la ES	Fecha en la cual ingresó el estudiante a la Educación Superior.

Estado civil	Estado civil del estudiante que ingresa al a facultad.
Org. Política	Organización política a la que pertenece el estudiante.
Fecha de ingreso al CES	Fecha de ingreso del estudiante al Centro de Educación Superior.
Fecha de matrícula	Fecha en la que realizó la matrícula en esta facultad.
Sexo	Sexo del estudiante que ingresa a la facultad.
Color de piel	Raza del estudiante que ingresa a la facultad.
Tipo de estudiante	Tipo de estudiante (becado o seminterno).
Año de estudio	Año de la carrera en el cual se encuentra el estudiante.
Centro de trabajo	Centro de trabajo en el caso de ser trabajador.
Nombre del padre	Nombre del padre del estudiante que ingresa a la facultad.
Nivel académico del padre	Nivel académico del padre del estudiante que ingresa a la facultad.
Nombre de la madre	Nombre de la madre del estudiante que ingresa a la facultad.
Nivel académico de la madre	Nivel académico de la madre del estudiante que ingresa a la facultad.
Tipo de servicio militar	Tipo de servicio militar que realizo el estudiante.
Edad	Edad actual del estudiante.

Tabla 3.1 - Descripción de los atributos de los archivos del SIGENU.

Datos de los listados de matrículas (Ver tabla 2.2)

No todos los listados contaban con el mismo formato, en resumen aparecían los siguientes atributos.

Atributo	Descripción
Carnet Identidad	Número del carnet de identidad del estudiante que ingresa a la facultad.
Nombre y Apellidos	Nombre y apellidos del estudiante que ingresa a la facultad.
Sexo	Sexo del estudiante que ingresa a la facultad.
Preuniversitario	Centro preuniversitario del cual proviene el estudiante.
Provincia	Provincia de la cual proviene el estudiante.
Municipio	Municipio del cual proviene el estudiante.
Ind	Índice con el que termina el estudiante la enseñanza media.
His	Nota obtenida por el estudiante en la prueba de ingreso de Historia.
Mat	Nota obtenida por el estudiante en la prueba de ingreso de Matemáticas.
Escal	Escalafón final resultado de las pruebas de ingreso y el índice de cada estudiante.
Valor Opción	La opción en la cual el estudiante pide la carrera.

Tabla 3.2 - Descripción de los atributos de los Listados de matrículas.

3.1.3 Selección de los datos

El objetivo de esta fase es listar los atributos que serán incluidos o excluidos del proceso de Minería de Datos, así como las medidas que se adoptaron para tomar estas decisiones. La selección de los datos se realizó sobre los atributos (columnas) y las tuplas (filas) de la Base de Datos.

En la Tabla 2.3 se describen los motivos referidos al descarte de las tuplas.

Cantidad de Tuplas	Motivos
93	Solo se analizan los estudiantes de nuevo ingreso no siendo así los que reingresan, los arrastres ni los traslados.
78	Pertenecientes a los estudiantes que se encuentran en este momento en el primer año de la carrera.
27	Pertenecientes a estudiantes de la provincia de Las Tunas de la cual en este momento no se reciben ingresos.

Tabla 3.3 - Tuplas descartadas

Debido a que los datos para ser analizados son de la carrera de Informática en el CRD, su selección provoca que algunos atributos permanezcan constantes y no aporten información al proceso de minería de datos.

Por este motivo fueron descartados los siguientes atributos (Ver Tabla 2.4).

Atributos descartados
País
Provincia
Carrera
Facultad/Filial
Tipo de curso
Régimen de estudio
Centro de trabajo

Tabla 3.4 - Atributos descartados

Teniendo en cuenta que algunos de los datos de los estudiantes son de carácter irrelevante para la aplicación de la MD, se procedió a descartar también estos atributos de los datos finales (Ver Tabla 2.5).

Atributos irrelevantes
Identidad
Nombre
Apellidos
Dirección
Grupo
Teléfono
Fecha de ingreso a la ES
Fecha de ingreso al CES
Año de estudio
Nombre del padre
Nombre de la madre
Edad
Correo

Tabla 3.5 - Atributos irrelevantes

Luego del pre-proceso de selección de datos, la lista de atributos a tener en cuenta para la MD se muestra en la Tabla 2.6.

Atributos seleccionados para la MD
Municipio
Situación académica
Fecha de nacimiento
Fuente de ingreso
Origen académico
Estado civil
Org. Política
Fecha de matrícula
Sexo
Color de piel
Tipo de estudiante
Nivel académico del padre
Nivel académico de la madre
Tipo de servicio militar
Ind
Mat
His
Escal
Valor
Preuniversitario

Tabla 3.6 - Atributos seleccionados para la MD

3.1.4 Limpieza de los datos

Para aumentar la calidad de los datos del apartado anterior, se analizó la BD en busca de datos con ruido, faltantes o irrelevantes.

Como resultado del análisis se detectaron 114 tuplas, las cuales tienen valor desconocido en el atributo *Nivel académico del padre*, este número representa aproximadamente un 34.34% del total de datos por lo cual se tomó la decisión de eliminar el atributo de los datos finales.

Además se detectaron 19 tuplas del atributo *Índice*, 40 tuplas de los atributos *Historia*, 32 tuplas del atributo *Matemática*, *Escalafón* y *Valor Opción*, además de 12 tuplas del atributo *Preuniversitario* con valores nulos. Se tomó la decisión de eliminar estas tuplas también del conjunto de datos finales. Hay que tener en cuenta que en varias de las tuplas eliminadas coincidía más de un atributo sin valor.

3.1.5 Construcción de los datos

Esta tarea demanda actividades tales como: La producción de atributos derivados, generación de nuevas tuplas o la transformación de los valores existentes.

Atributo *Situación Académica*

Según lo revelado en los datos obtenidos por el SIGENU, existen cuatro categorías de estudiantes. Estas categorías contienen los siguientes valores (Tabla 2.7).

Categoría existente	Descripción
Promovido	Para el caso de los alumnos que pasan de año sin problema.
Baja	Para el caso de los alumnos que causan baja.
Promovido con arrastre	Para el caso de los alumnos que pasan el año pero llevan un acarreo de las asignaturas ponchadas.
Repitente	Para el caso de los alumnos que desaprueban y repiten el año.

*Tabla 3.7 - Categorías del atributo **Situación académica***

Si bien estas categorías se tomaron en el momento actual que se hizo la consulta, no son relevantes para el presente trabajo, debido a que no reflejan la situación que presentaron los estudiantes al terminar su primer año.

Teniendo en cuenta el análisis anterior y que el objetivo de este trabajo es determinar la situación de estos estudiantes al terminar su primer año universitario y que las categorías descritas no son las que tenían los mismos en ese momento, se decidió construir nuevas categorías como se muestra en la Tabla 2.8.

Nuevas categorías	Descripción
Sin problemas	Para los estudiantes que pasan de año sin problemas.
Con problemas	Para los estudiantes que causan baja, reingresos y llevan arrastres.

*Tabla 3.8 - Nuevas categorías para el atributo **Situación académica***

Atributo *Edad de Ingreso*

Partiendo de la importancia de saber la edad en que ingresa el estudiante, para determinar si este podría ser un factor que influya en su promoción al año siguiente se realizaron, a través de técnicas recogidas en Excel, modificaciones a algunos atributos

(Ver Tabla 2.9) para determinar la edad con que matriculó cada uno de los estudiantes a la carrera.

Atributos existentes	Nuevo Atributo
Fecha de nacimiento	Edad de Ingreso
Fecha de matrícula	

*Tabla 3.9 - Creación de atributo derivado **Edad de ingreso***

3.1.6 Integración de los datos

Teniendo ya los datos extraídos del SIGENU en formato Excel se le incorporaron, de manera manual, los datos extraídos de los listados de matrículas, tomando como referencia el carnet de identidad de los estudiantes para realizar la combinación. Además se poblaron los datos del atributo Situación Académica de manera manual con valores obtenidos de los registros de calificaciones de primer año de cada estudiante. En la Tabla 2.10, se muestra la estructura y los nombres de los atributos de la Tabla IEMD (**I**nformación de **E**studiantes para la **M**inería de **D**atos) creada para el análisis.

Atributos seleccionados	Atributos Tabla IEMD	Tipo
Municipio	mun	Nominal
Situación académica	class	
Fecha de nacimiento	eda_ing	
Fecha de matrícula		
Fuente de ingreso	fue_ing	
Origen académico	ori_aca	
Estado civil	est_civ	
Org. Política	org_pol	
Sexo	sex	
Color de piel	col_pie	
Tipo de estudiante	tip_est	
Nivel académico de la madre	na_mad	
Valor	opc	
Preuniversitario	pre	
Tipo de servicio militar	tip_sm	
Ind	ind	Numérico
Mat	mat	
His	his	
Escal	esc	

Tabla 3.10 - Integración de datos

La BC final cuenta con 18 atributos y 292 instancias, de ellas 217 clasificados como **Sin problemas** y los restantes 75 **Con problemas**. La tabla se guardó en formato CSV (delimitado por coma) para poder ser utilizado con las herramientas seleccionadas.

3.1.7 Formato de datos

El formateo de los datos involucra sólo transformaciones sintácticas o que tengan que ver con el orden en que son presentados los atributos. Algunas herramientas de MD tienen requerimientos respecto al orden de los atributos. Otras herramientas no trabajan con atributos categóricos y sólo lo hacen con atributos continuos.

Para facilitar la lectura de los resultados arrojados por los algoritmos de MD se han formateado los valores de los siguientes atributos:

Atributo *mun*

Con el fin de no dejar espacio entre los nombres de los municipios se realizaron las siguientes modificaciones (Ver Tabla 2.11).

Formato existente	Nuevo formato
SANTA CRUZ DEL SUR	SCRUZ
CARLOS MANUEL DE CESPEDES	CESPEDES

*Tabla 3.11 - Formato del atributo **mun***

Atributo *sex* (Ver tabla 2.12)

Formato existente	Nuevo formato
Masculino	M
Femenino	F

*Tabla 3.12 - Formato para el atributo **sex***

Atributo *col_pie* (Ver Tabla 2.13)

Formato existente	Nuevo formato
Blanco	B
Negro	N
Mestizo(Mulato)	MEZ

*Tabla 3.13 - Formato para el atributo **col_pie***

Atributo *est_civ* (Ver Tabla 2.14)

Formato existente	Nuevo formato
Casado (a)	CAS
Soltero (a)	SOL
Separado (a)	SEP

*Tabla 3.14 - Formato para el atributo **est_civ***

Atributo *org_pol* (Ver Tabla 2.15)

Formato existente	Nuevo formato
Ninguna	NO

*Tabla 3.15 - Formato del atributo **org_pol***

Las demás categoría del atributo *org_pol* no se modificaron.

Atributos *na_mad* (Ver Tabla 2.16)

Formato existente	Nuevo formato
Primaria	PRI
Media	MED
Media Superior	MSUP
Superior	SUP

*Tabla 3.16 - Formato de los atributo **na_mad***

Atributo *tip_sm* (Ver Tabla 2.17)

Formato existente	Nuevo formato
Diferido	DIF
Orden 18	O18
----	DIR

*Tabla 3.17 - Formato del atributo **tip_sm***

Atributo *fue_ing* (Ver Tabla 2.18)

Formato existente	Nuevo formato
Atletas de Alto Rendimiento	AAR
Cadetes MININT	CMININT
Concurso	CON
Curso de preparación de la enseñanza técnica y profesional	CPETP
Instituto Politécnico	IP
Orden 18	O18
Orden 18 Unidades Militares	O18UM
Preuniversitario	PRE
Trabajadores	TRA

*Tabla 3.18 - Formato del atributo **fue_ing***

Atributo *ori_aca* (Ver Tabla 2.19)

Formato existente	Nuevo formato
Enseñanza Técnico Profesional	ETP
Preuniversitario	PRE

*Tabla 3.19 - Formato del atributo **ori_aca***

Atributo *tip_est* (Ver Tabla 2.21)

Formato existente	Nuevo formato
Seminterno	SIN
Becado Nacional	BN

*Tabla 3.20 - Formato del atributo **tip_est***

Atributo *pre* (Ver Tabla 2.20)

Formato existente	Nuevo formato
Abel Santamaría	ABSantamaria
Álvaro Barba	ALBarba
Álvaro Morell	ALMorell
Amado Fernández	AMFernandez
Ángel del Castillo	ANCastillo
Antonio Varona	ANVarona
Aralio Hernández	ARHernandez
Armando Mestres	ARMestres
Asamblea de Guáimaro	ASGuaimaro
Cadete MININT	CADMININT
Campaña de la Reforma	CAReforma
Carlos Rodríguez Carea	CRCarea
Combate Cocal	COCocal
Concurso-CMG	CON-CMG
Dagoberto Rojas	DARojas
EIDE	EIDE
ESPA Inés Loases	ESPA
Félix Sotolongo	FESotolongo
Francisco Agüero	FRAguero
IPUM Jesús Suarez Gayol	JSGayol
IPVCE Gral. Máximo Gómez	IPVCE
Levantamiento de Jucaral	LEJucaral
Manuel de Quesada	MAQuesada
Mario Muñoz	MAMunoz
Rafael Martínez	RAMartinez
República Bolivariana	REBolibariana
Rescate de Sanguily	RESanguily
Roberto Coco Peredo	RCPeredo
UM	UM
UM 1024	UM1024
UM 1390	UM1390
UM 1448	UM1448
UM 1510	UM1510
UM 2010	UM2010
UM 2051	UM2051
UM Tropas Especiales	UMTE

*Tabla 3.21 - Formato del atributo **pre***

Atributo *class* (Ver Tabla 2.22)

Formato existente	Nuevo formato
Sin Problema	SP
Con Problemas	CP

*Tabla 3.22 - Formato del atributo **class***

Los demás atributos que no aparecen en este apartado permanecieron con los valores originales.

3.2 Modelado

En esta fase, se describirán las distintas técnicas de modelado seleccionadas y los parámetros aplicados en cada una de ellas. Se aplicaron tareas de preparación de los datos para generar nuevos ficheros de datos, a los cuales aplicar las técnicas elegidas. Una vez aplicadas las técnicas de modelado a los distintos conjuntos de datos, los resultados obtenidos fueron recopilados y comparados.

3.2.1 Selección de las técnicas de modelado

Las principales tareas en problemas de minería de datos pueden ser divididas en dos grupos: predictivas y descriptivas. En este caso se tomó la decisión de contemplar únicamente las tareas predictivas, a partir de los objetivos principales del proyecto relacionados con la predicción de estudiantes que presentarían problemas al final de su primer año universitario, específicamente la tarea de clasificación.

En los problemas de clasificación normalmente se dispone de una base de datos compuesta por un número N de ejemplos o instancias que están descritos por un número P de atributos y que pertenecen a una clase. En ellos se trata de aprender la forma de distinguir los ejemplos de las distintas clases. El inconveniente planteado en este proyecto puede ser tratado como un problema de clasificación, en el que la clase es el resultado de cada estudiante al final de su primer año. La forma de enunciarlo podría ser la siguiente: Se dispone de una base de datos amplia de información de cada uno de los estudiantes. La clase sería si el estudiante tiene dificultades o no para terminar su año, por tanto, sólo admite dos posibles valores: con problemas o sin problemas. Entonces, el dilema que se afronta en este proyecto puede ser tratado como un problema de clasificación.

Debido a la gran diferencia que existe entre la cantidad de casos pertenecientes a una clase con respecto a la otra se tratará el problema como clasificación sobre conjuntos de datos desbalanceados.

Esta dificultad ocurre cuando el número de instancias de una clase es mucho menor que el número de instancias de la otra clase (o de otras clases). En nuestro caso, de los 292 alumnos, 217 aprobaron y 75 tuvieron problemas. Por lo que se considera que los datos están desbalanceados, es decir, hay una mayoría de alumnos que aprobaron frente a una minoría que no.

El inconveniente de utilizar datos desbalanceados es que los típicos algoritmos de clasificación han sido desarrollados para maximizar la tasa de exactitud total, lo cual es independiente de la distribución de clases, esto provoca que los clasificadores tengan en la etapa de entrenamiento una clase mayoritaria lo que lleva a clasificar en la etapa de prueba con baja sensibilidad a los elementos de la clase minoritaria.

3.2.1.1 Técnicas de balanceo

Una manera de resolver el problema, es actuar en la etapa de pre-procesado de los datos, haciendo un sobre muestreo o balanceo de la distribución de clases, para ello existen varios algoritmos de re-balanceo, en este trabajo se utilizan tres de los más mencionados:

- **SMOTE** (*Synthetic Minority Oversampling Technique*). En términos generales, SMOTE introduce de manera sintética elementos de la clase minoritaria para equilibrar la muestra de datos, basado en la regla del vecino más cercano. Los elementos sintéticos creados son introducidos en el espacio que hay entre los elementos de la clase minoritaria. Dependiendo del tamaño del sobre muestreo requerido, los vecinos más cercanos son elegidos aleatoriamente **(Chawla, *et al.*, 2002)**.

Una forma distinta de abordar el problema de la clasificación de datos desbalanceados es realizar una clasificación sensible al costo. En la clasificación tradicional no se distingue si una de las clases a clasificar es más importante que otra, es decir, si una tiene un costo de clasificación diferente. Optimizar la tasa de clasificación sin tomar en cuenta el costo de los errores a menudo puede conducir a resultados no óptimos debido al alto costo que puede ocasionar la mala clasificación de una instancia minoritaria.

Para llevar a cabo esta tarea se utilizó:

- **CostSensitiveClassifier**: es un metaclasificador que realiza la clasificación utilizando costo-sensible. Pueden usarse dos métodos para introducir la sensibilidad al costo: re-pesado que entrenan los casos según el costo total asignado a cada clase; o prediciendo la clase con el costo mínimo de fallos esperados.

3.2.1.2 Técnicas de selección de atributos

Debido a la cantidad de atributos recopilados (18), se realizó también un análisis o estudio de selección de atributos para determinar cuáles son los que mayormente influyen en la variable de salida o clase a predecir. Para seleccionar las variables de mayor relevancia se utilizaron varios métodos de selección de atributos disponibles en el software WEKA. En general, estos algoritmos de selección pueden ser agrupados por varios criterios. Una categorización popular es aquella en la que los algoritmos se distinguen por su forma de evaluar atributos y se clasifican en: filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje y *wrappers* (envoltorios), los cuales usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar lo deseable de un subconjunto **(Hall y Holmes, 2002)**.
Los algoritmos de selección de atributos empleados son:

- **ChiSquaredAttributeEval**: Calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre la clase y cada atributo.
- **GainRatioAttributeEval**: Evalúa el valor de un atributo midiendo la proporción de ganancia con respecto a la clase.
- **OneRAttributeEval**: Evalúa el valor de un atributo usando el clasificador OneR.
- **ReliefFAttributeEval**: Evalúa el valor de un atributo probando un caso repetidamente y considerado el valor del atributo dado para el caso más cercano de la misma y diferente clase. Puede operar con datos discretos y continuos **(Kononenko, 1994)**.

- **SymmetricalUncertAttributeEval**: Evalúa el valor de un atributo midiendo la incertidumbre simétrica con respecto a la clase.

Como método de búsqueda se utilizó:

- **Ranker**: Elabora un orden de los atributos de acuerdo a la evaluación individual de cada uno.

3.2.1.3 Técnicas de clasificación

Se seleccionaron 10 algoritmos de clasificación entre los disponibles por la herramienta de minería de datos WEKA. Esta selección se ha realizado debido a que estos algoritmos, son todos del tipo "caja blanca", es decir, se obtiene un modelo de salida comprensible para el usuario, porque o se obtienen reglas de clasificación del tipo "Si – Entonces" o árboles de decisión. De esta forma un usuario no experto en minería de datos como un profesor o instructor puede utilizar directamente la salida obtenida por estos algoritmos para detectar a los alumnos con problemas a tiempo y poder tomar decisiones sobre cómo ayudarlos y evitar que suspendan o abandonen.

Las reglas de clasificación del tipo "Si – Entonces" son una manera simple y fácilmente comprensible de representar el conocimiento. Una regla tiene dos partes, el antecedente y el consecuente. El antecedente de la regla (la parte del "Si") contiene una combinación de condiciones respecto a los atributos de predicción. El consecuente de la regla (la parte del "Entonces") contiene el valor predicho para la clase. De esta manera, una regla asigna una instancia de datos a la clase señalada por el consecuente si los valores de los atributos de predicción satisfacen las condiciones expresadas en el antecedente, y por tanto, un clasificador es representado como un conjunto de reglas. Los algoritmos incluidos en este paradigma pueden ser considerados como una búsqueda heurística en un espacio de estados. En este caso, un estado corresponde a una regla candidata, y los operadores corresponden a la generalización y especialización de operaciones que transformen una regla candidata en otra. Los 5 algoritmos de inducción de reglas de clasificación que se usaron son:

- **JRip**, es la implementación que hace WEKA del *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER), propuesto por William W. Cohen como una versión optimizada de IREP **(Cohen, 1995)**.
- **NNge**, parecido al algoritmo del vecino más cercano, usando ejemplos no anidados generalizados(los cuales son hiper-rectángulos que pueden verse como reglas "Si-Entonces") **(Martin, 1995)**.
- **OneR**, implementación de WEKA para construir y usar el clasificador 1R; es decir, usa el atributo de error mínimo para la predicción, discretizando atributos numéricos **(Holte, 1993)**.
- **PART**, genera una lista de decisión PART. Usa divide y vencerás. Construye un árbol de decisión C4.5 parcial en cada iteración y pone el mejor resultado en la regla **(Frank y Witten, 1998)**.
- **Ridor**, una implementación de la regla de aprendizaje RIpple-Down.

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, el cual contiene cero o más nodos internos y uno o más nodos de hoja. Los nodos internos tienen dos o más nodos secundarios y contienen divisiones, los cuales prueban el valor de una expresión de los atributos. Los arcos de un nodo interno a otro secundario (o de menor jerarquía) son etiquetados con distintas salidas de la prueba del nodo interno. Cada nodo hoja tiene una etiqueta de clase asociada. El árbol de decisión es un modelo predictivo en el cual una instancia es clasificada siguiendo el camino de condiciones cumplidas desde la raíz hasta llegar a una hoja, la cual corresponderá a una clase etiquetada. Un árbol de decisión se puede convertir fácilmente en un conjunto de reglas de clasificación **(Quinlan, 1993)**. Los 5 algoritmos de árboles de decisión que se utilizarán son:

- **ADTree**, genera un árbol de decisión que alterna. El algoritmo básico se basa en **(Freund y Mason, 1999)**. Esta versión solo soporta problemas de dos clases.
- **J48**, implementación para generar un árbol de decisión C4.5 podado o no podado **(Quinlan, 1993)**.

- **RamdomTree**, implementación que construye un árbol que considera los k atributos escogidos aleatoriamente como cada nodo. No realiza podado. Además, con una opción para permitir estimación de las probabilidades de la clase basada en un conjunto *hold-out*.
- **REPTree**, árbol rápido de decisión de aprendizaje. Construye un árbol de decisión/regresión usando la información ganancia/varianza y podándolo usando la poda de error reducido. Solo sortea una vez los valores para los atributos numéricos. Los valores perdidos son tratados con un corte de las correspondientes instancias (i.e. as in C4.5).
- **SimpleCart**, implementa poda de complejidad mínima. Nótese que cuando trata con valores perdidos, se usa el método "*fractional instances*" en lugar del método *split method*.

3.2.1.4 Evaluación de la clasificación en problemas con datos desbalanceados

La evaluación de los procesos de aprendizaje con datos desbalanceados tiene sus propias características. Las medidas de la calidad de la clasificación se construyen a partir de una matriz de confusión (como se muestra en la tabla 2.23) que registra correctamente e incorrectamente los ejemplos reconocidos de cada clase.

	Predicciones positivas	**Predicciones negativas**
Clase positiva	Verdaderos positivos (**VP**)	Falsos negativos (**FN**)
Clase negativa	Falsos positivos (**FP**)	Verdaderos negativos (**VN**)

Tabla 3.23 - Matriz de confusión para problemas de dos clases

La medida empírica más usada: ACC descrita en Ec 2.1, no distingue entre el número de etiquetas correctas de diferentes clases, lo cual en el ámbito de los problemas de desbalance puede conducir a conclusiones erróneas. Por ejemplo, un clasificador que obtiene una precisión de 90% en un conjunto de datos con un valor de IR=9 (*Imbalance Rate*, IR) **(Orriols-Puig y Bernad´o-Mansilla, 2009)**, podría no ser válida si no se clasifica correctamente cualquier instancia de la clase minoritaria.

$$ACC = \frac{VP+VN}{VP+FN+FP+VN}$$ Ec 3.1

Debido a ello, en lugar de utilizar ACC, se consideran otras alternativas de medición. En concreto, en la Tabla 2.23 se pueden obtener cuatro indicadores del desempeño que miden la calidad de la clasificación para las clases positivas y negativas de forma independiente:

- Tasa de verdaderos positivos TVP=VP/ (VP+FN): es el porcentaje de casos positivos correctamente clasificados como pertenecientes a la clase positiva.
- Tasa de verdaderos negativos TVN=VN/ (FP+VP): es el porcentaje de casos negativos correctamente clasificados como pertenecientes a la clase negativa.
- Tasa de falsos positivos FP=FP/ (FP+VN): es el porcentaje de casos negativos mal clasificados como pertenecientes a la clase positiva.
- Tasa de falsos negativos FN=FN/ (VP+FN): es el porcentaje de casos positivos mal clasificados como pertenecientes a la clase negativa.

Una medida apropiada que podría ser utilizada para medir el rendimiento de la clasificación de conjuntos de datos con mayor desbalance es el área bajo la curva ROC (*Receiver Operating Characteristic*) **(Bradley, 1997)**. En estos gráficos se reconoce el hecho de que la capacidad de cualquier clasificador no puede aumentar el número de verdaderos positivos sin aumentar los falsos positivos. El área bajo la curva ROC (AUC) **(Huang y Ling, 2005)** proporciona un número único de resumen para el desempeño de algoritmos de aprendizaje.

Para calcular AUC sólo hay que obtener el área bajo la curva como se muestra en la expresión Ec 2.2:

$$AUC = \frac{1+TVP-TFP}{2}$$ Ec 3.2

3.2.2 Construcción y descripción de los modelos experimentales

En este apartado se llevará a cabo la descripción de los experimentos, para la realización de los mismos se utilizaron las herramientas KEEL y WEKA. Estas herramientas dan la posibilidad de configurar gran cantidad de parámetros, en general en este proyecto se trabajó con los parámetros por defecto debido a que ajustar de forma óptima todos los parámetros podría haber ocasionado un retraso elevado en el desarrollo del proyecto.

3.2.2.1 Conjuntos de datos

Después de haber realizado las anteriores tareas de pre-procesamiento se dispone de un primer fichero de datos con 18 atributos/variables sobre 292 alumnos. Este fichero se importó en la herramienta KEEL y se procedió a particionar el fichero (5 particiones) para realizar la validación cruzada en las pruebas de clasificación. Una partición es la división aleatoria del fichero original de datos en otros dos, uno para la etapa de entrenamiento (*training*) y el otro para la etapa de prueba (*test*). Resultaron de este proceso 5 ficheros de entrenamiento y 5 ficheros de prueba con los datos originales de la BC.

Con el fin de disminuir la cantidad de atributos para obtener una posible mejoría en los resultados se realizó también un análisis o estudio de selección de atributos para determinar cuáles son los que mayormente influyen en la variable de salida o clase a predecir. Para seleccionar las variables de mayor relevancia se utilizaron los métodos antes mencionados disponibles en WEKA. Después de haber aplicado cada algoritmo se procedió a construir una tabla (ver Anexo A) de algoritmos contra atributos en la cual los valores serían la posición en el ranking devuelto por cada algoritmo, hallando los totales por cada atributo, el de menor valor sería el atributo más relevante de todos los determinados por cada algoritmo de selección.

Luego de obtener estos valores se procedió a realizar una reducción de atributos. Para ello se seleccionaron sucesivamente cada uno de los atributos y para cada selección se le calculó el área bajo la curva con cada método de clasificación, luego se construyó una

tabla (ver Anexo B) de método de clasificación por selección, se obtuvo el promedio por cada selección y se escogió el de mayor ratio, resultando ser con los 13 mejores atributos.

Con estos resultados se construyó un nuevo fichero solo con los 13 atributos seleccionados al cual se le realizó el mismo procedimiento de particionado que a la base completa, resultando de este, 5 ficheros de entrenamiento y 5 de pruebas.

Como se mencionó anteriormente el conjunto de datos está desbalanceado, para corregir este problema se procedió, en la herramienta KEEL, a re-balancear los datos. Para ello se aplicó el método de re-balanceo mencionado anteriormente, sobre el conjunto de datos originales y sobre el conjunto de datos seleccionados. Cada fichero de entrenamiento se re-balanceó, de forma que tuvieran el 50% de instancias por cada clase, dejando los ficheros de prueba sin re-balancear.

Después de realizar todas las tareas de pre-procesado de datos, se cuenta con:

- 5 ficheros de entrenamiento y testeo con toda la base (17 atributos).
- 5 ficheros de entrenamiento y testeo con solo los 13 atributos seleccionados.
- 5 ficheros de entrenamiento y testeo con toda la base, donde los ficheros de entrenamiento están re-balanceados con SMOTE.
- 5 ficheros de entrenamiento y testeo con solo los 13 atributos seleccionados donde los datos están re-balanceados con SMOTE.

Finalmente la herramienta KEEL transforma los datos en ficheros .dat muy similar a los ficheros que utiliza WEKA (.arff) la única diferencia es que en los ficheros de KEEL se relacionan las variables de entrada (input) y las variables de salida (output), por lo tanto, se hicieron los cambios pertinentes para que los ficheros pudieran utilizarse en WEKA.

3.2.2.2 Minería de datos y experimentación

Hemos realizado varios experimentos con el objetivo de obtener la mejor clasificación de los datos. En cada experimento hemos ejecutado los 10 algoritmos de clasificación mencionados anteriormente, sobre los distintos conjuntos de datos obtenidos en la etapa de pre-procesamiento.

Hemos realizado una validación cruzada con 5 particiones por cada experimento. En este tipo de validación cruzada, se realiza el entrenamiento y el testeo cinco veces con las diferentes particiones y el resultado es la media de los valores obtenidos.

3.2.2.2.1 Experimento 1: IEMD

En el primer experimento, se han ejecutado los 10 algoritmos utilizando toda la información disponible, es decir, los ficheros de datos con los 18 atributos de los 292 alumnos. Los resultados obtenidos (la media de las 5 ejecuciones) con los ficheros de prueba/test de la aplicación de los algoritmos de clasificación se muestran en la Figura 3.1.

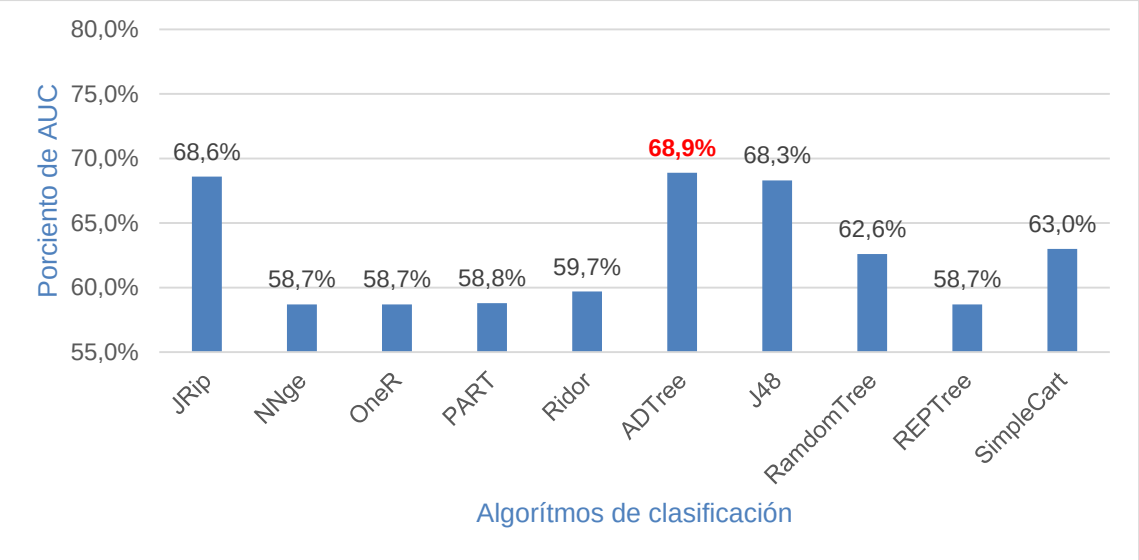

Figura 3.1 - Resultados de la validación cruzada de IEMD

Para este primer experimento que es el más sencillo de todos se observan resultados discretos. Sin embargo es el punto de referencia para realizar los demás experimentos. En este caso el algoritmo con mejores resultados fue el árbol de decisión ADTree con un 68,9% de área bajo la curva.

3.2.2.2.2 Experimento 2: IEMD + SMOTE

Para el segundo experimento se han utilizado los ficheros a los cuales se le aplicó el sobre muestreo SMOTE con toda la información disponible, se volvieron a ejecutar los 10 algoritmos de clasificación obteniendo los siguientes resultados (ver Figura 3.2).

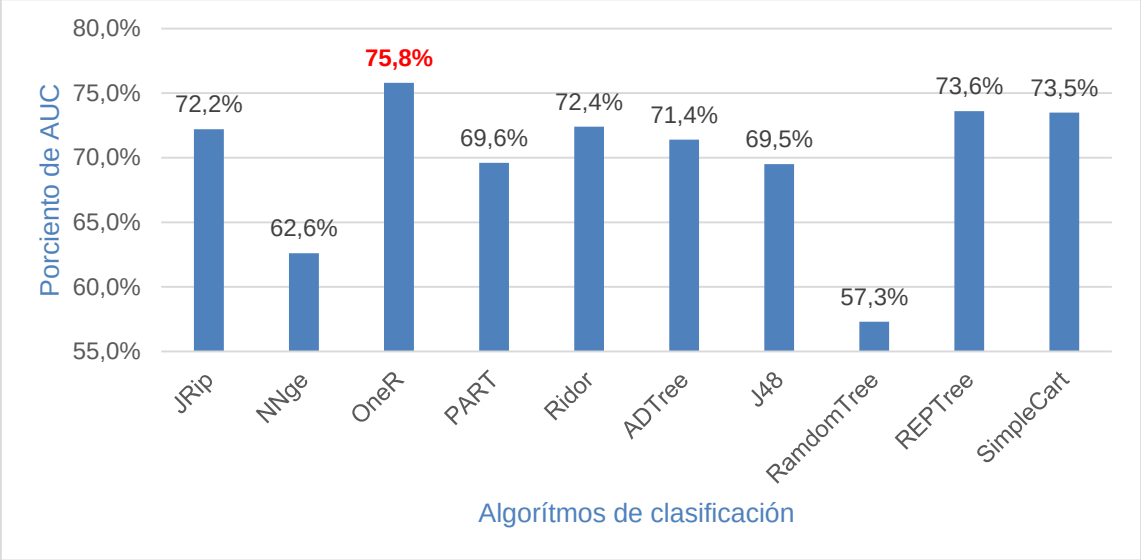

Figura 3.2 - Resultados de la validación cruzada de IEMD+SMOTE

Si comparamos los resultados de las Figuras 3.1 y 3.2 se observa un mejoramiento considerable de los porcientos. Solo en el caso del árbol de decisión RamdomTree se manifiesta una disminución en el porcentaje. Para este experimento, resultó con mejor porciento de área bajo la curva la regla de clasificación OneR con 75,8%.

3.2.2.2.3 Experimento 3: IEMD + Costo Sensitivo

Para este experimento se utilizó el metaclasificador de WEKA **CostSensitiveClassifier** sobre el fichero original con los 18 atributos. En el problema existente estamos mucho más interesados en la clasificación de los alumnos Con Problemas (clase minoritaria). Después de hacer varias pruebas con diferentes costos, se encontró que utilizando la matriz [0, 1; 4, 0], se obtuvieron los mejores resultados de clasificación, lo cual indica que al realizar la clasificación se tiene en cuenta que es 4 veces más importante clasificar de manera correcta los casos de Con Problemas que los casos de Sin Problemas. Los resultados están en la siguiente tabla (ver Figura 3.3).

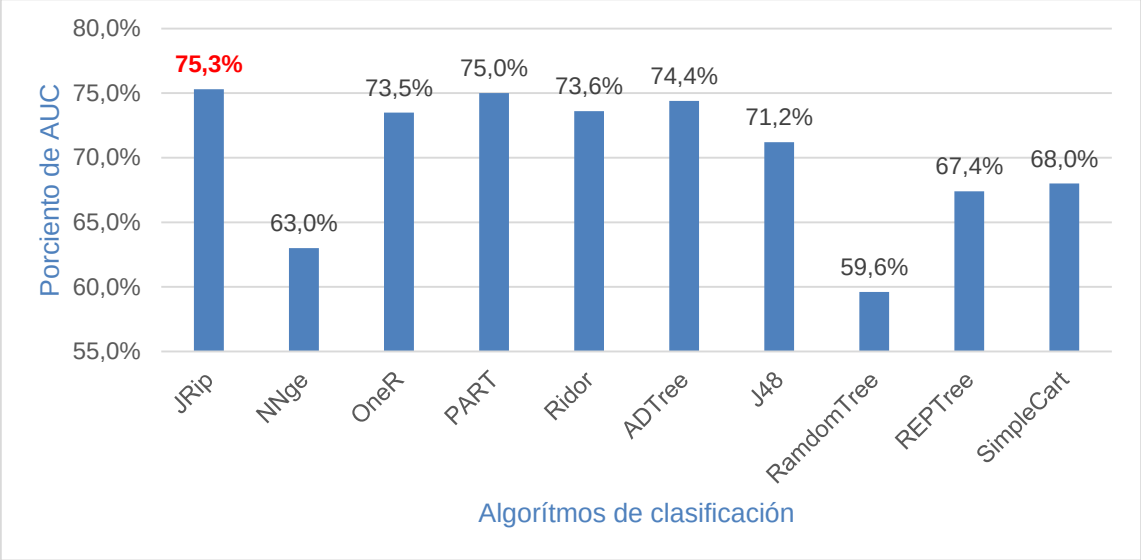

Figura 3.3 - Resultados de la validación cruzada de IEMD + Costo Sensitivo

Con respecto a los anteriores no alcanza un máximo en el porciento de área bajo la curva. Sin embargo, la media por experimento es superior a los otros. Todos los algoritmos de clasificación logran superar los resultados del primer experimento. En este caso, el mejor algoritmo de clasificación resultó ser la regla de clasificación JRip con 75,3%.

3.2.2.2.4 Experimento 4: Experimento 1 + Selección de atributos

Las pruebas sobre los atributos seleccionados se realizaron con el fin de buscar una mejoría en relación con las pruebas que se hicieron sobre el conjunto de datos original. Para este experimento se utilizaron los ficheros con los 13 mejores atributos y consiste en ejecutar nuevamente los 10 algoritmos de clasificación para comprobar cómo ha afectado la selección de atributos en la predicción. La Figura 3.4 muestra los resultados de la validación cruzada (la media de las 10 ejecuciones) de los algoritmos de clasificación utilizando solamente los 13 mejores atributos.

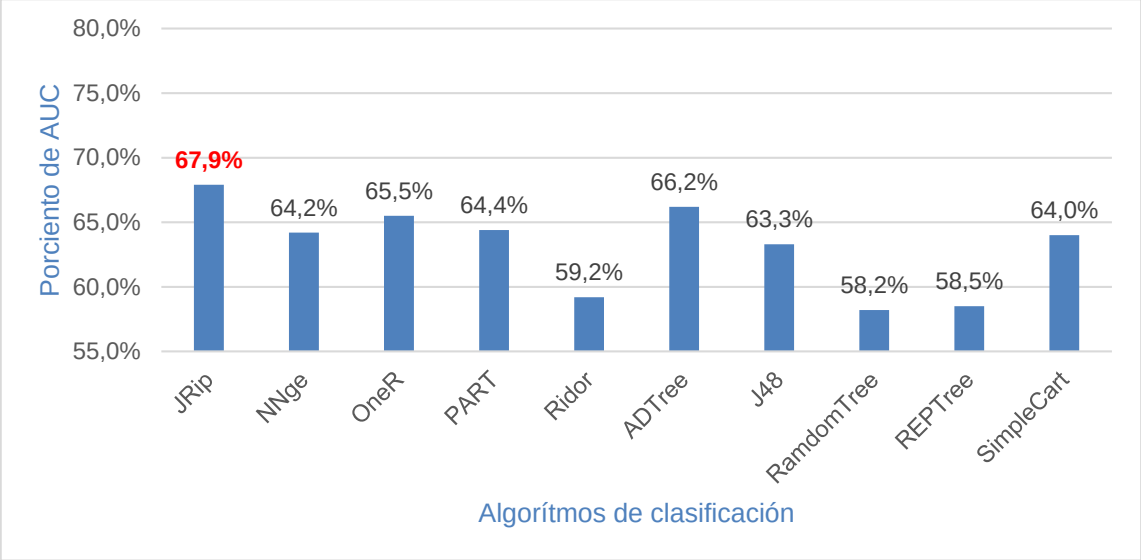

Figura 3.4 - Resultados de la validación cruzada de Experimento1 + Selección de atributos

Para este experimento se decidió realizar una comparación con los resultados del primer experimento. Aunque no mejora el porciento máximo, en general, se manifiesta una mejoría en cuanto a todos los algoritmos. Para el mismo, el mejor porciento de área bajo la curva lo aporta, como en el experimento anterior, la regla de clasificación JRip con un 67,9%.

3.2.2.2.5 Experimento 5: Experimento 2 + Selección de atributos

Para este experimento se utilizaron los ficheros de los 13 mejores atributos re-balanceados con la técnica de sobre muestreo SMOTE. La Figura 3.5 muestra los resultados de la validación cruzada (la media de las 10 ejecuciones) de los algoritmos de clasificación, utilizando solamente los 13 mejores atributos.

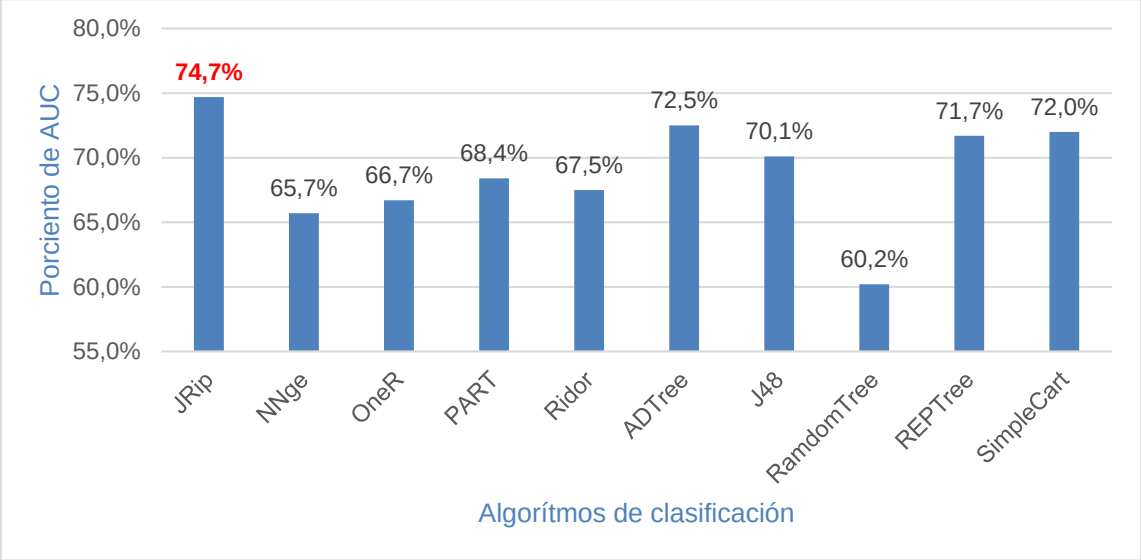

Figura 3.5 - Resultados de la validación cruzada de Experimento2+Selección de atributos

Al comparar los resultados con los del segundo experimento no se observa ninguna mejoría. El mejor resultado lo obtiene nuevamente la regla de clasificación JRip con un 74,7%.

3.2.2.2.6 Experimento 6: Experimento 5 + Selección de atributos

Se utilizó nuevamente el metaclasificador de WEKA **CostSensitiveClassifier** sobre los ficheros con los 13 mejores atributos. Después de hacer varias pruebas con diferentes costos, se encontró que utilizando nuevamente la matriz [0, 1; 4, 0], se obtienen los mejores resultados de clasificación, lo cual indica que al realizar la clasificación se tiene en cuenta que es cuatro veces más importante clasificar de manera correcta los casos de Con Problemas que los casos de Sin Problemas. Los resultados están en la siguiente figura (ver Figura 3.6).

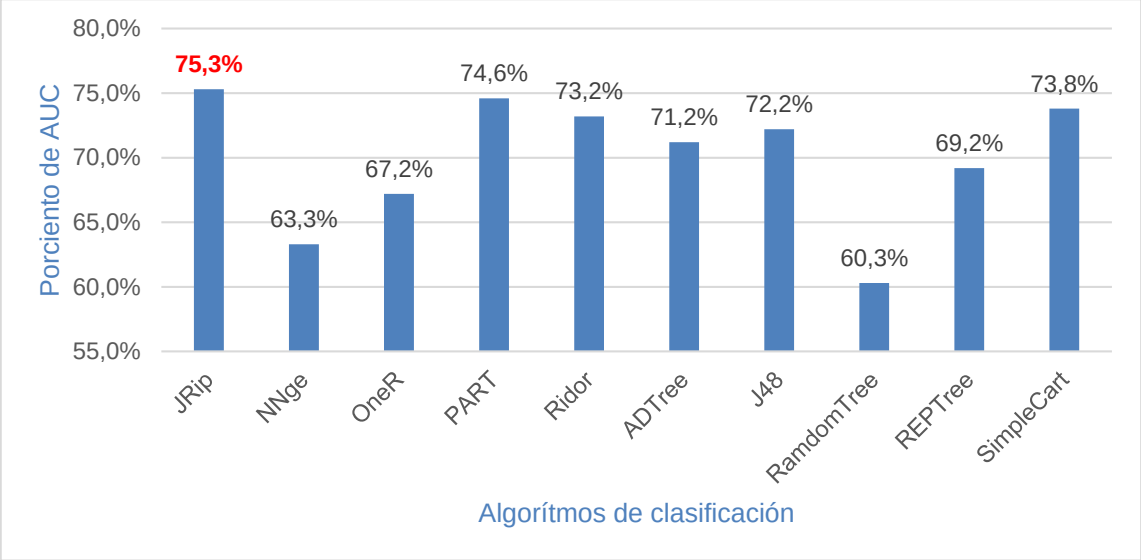

Figura 3.6 - Resultados de la validación cruzada de Experimento5+Selección de atributos

Haciendo una comparación con los resultados del tercer experimento los máximos alcanzados, en cada uno de ellos, tienen el mismo valor. Aunque la media global disminuya, lo hace de una forma muy discreta en comparación con el tercer experimento. Nuevamente la regla de clasificación JRip es el que se lleva el crédito de mejor porciento de área bajo la curva.

3.2.3 Evaluación de los modelos experimentales

En este apartado se realiza un resumen de los resultados obtenidos en todos los experimentos realizados comparando unos con otros. El siguiente gráfico, indica los mejores resultados obtenidos en cada uno de los mismos (ver Figura 3.7).

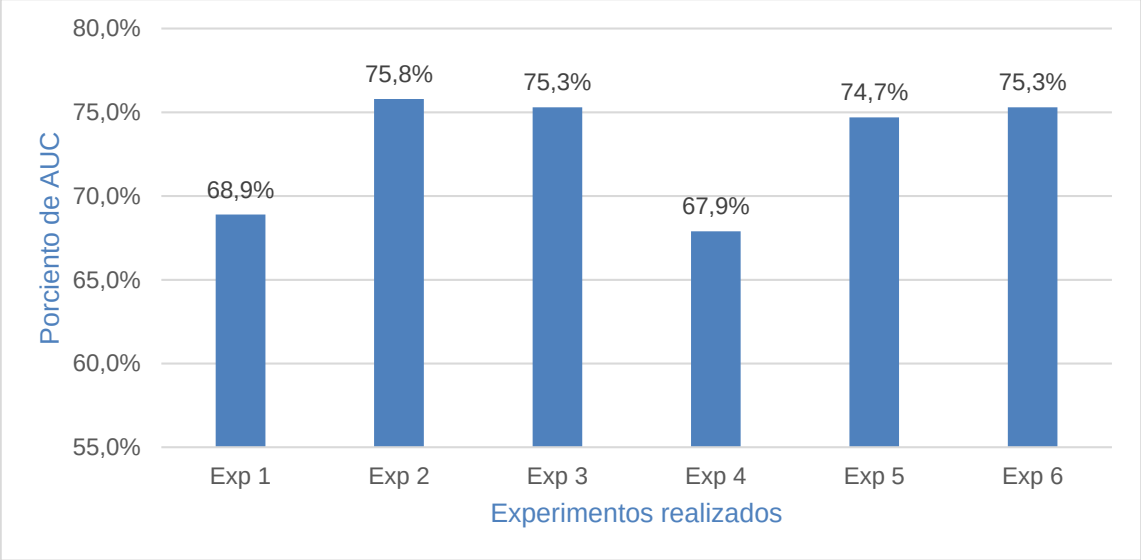

Figura 3.7 - Mejores resultados por cada experimento

Al realizar una comparación en cuanto a los mejores resultados obtenidos se observa, a simple vista, que en los experimentos realizados, sin aplicar ninguna técnica que diera solución al problema del desbalance entre las clases, se obtienen resultados más discretos en contraste con los obtenidos de aplicar el sobre muestreo SMOTE o clasificar teniendo en cuenta los diferentes costos para las clases. El mejor resultado se obtuvo con el segundo experimento que es el resultado de aplicar la técnica de sobre-muestreo SMOTE sobre los 18 atributos de la base original, sin embargo la diferencia es casi imperceptible cuando analizamos estos resultados con los obtenidos por el análisis de costos entre clases del tercer y sexto experimento. En general, se obtienen mejores valores en los experimentos realizados con la base original que con los experimentos resultantes de aplicar la selección de los 13 mejores atributos.

Se realizaron, además, una comparación entre las medias de los porcientos obtenidos por cada experimento, debido a que de esta forma tendríamos una imagen de en cual experimento se obtuvieron mejores resultados globales (ver Figura 3.8).

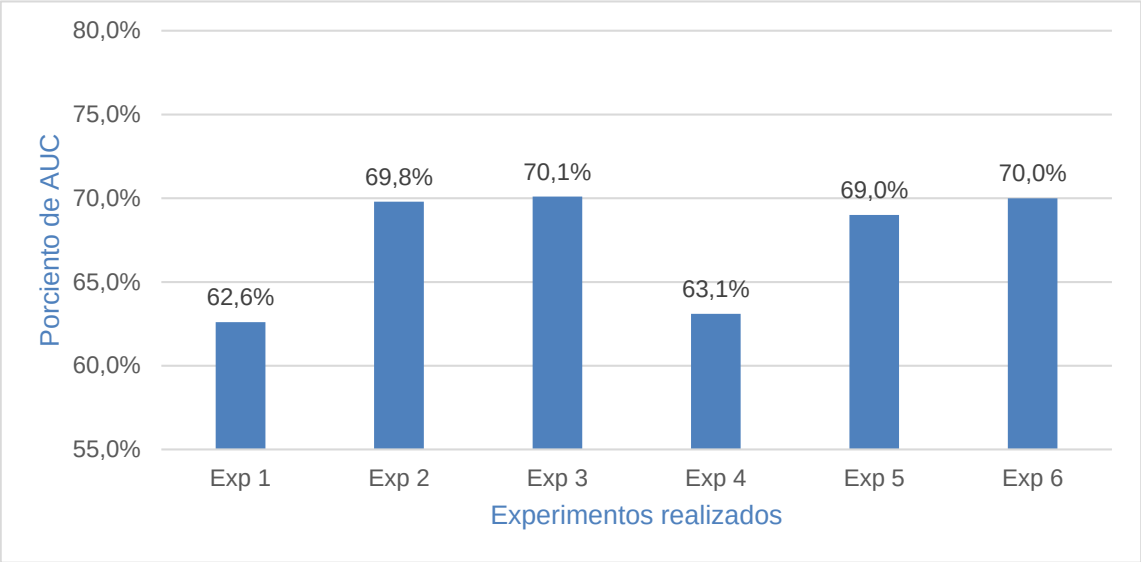

Figura 3.8 - Media obtenida por cada experimento

Al analizar los resultados, se puede observar cierta similitud en cuanto al comportamiento, en comparación con la figura anterior, las medias para el primer y cuarto experimento son inferiores con respecto a las demás, sin embargo a la hora de analizar la media de los resultados obtenidos por las dos variantes para tratar datos desbalanceados nos damos cuenta que donde mejor se comportaron los algoritmos de clasificación empleados resultó ser con el análisis de costo realizado en el tercer y sexto experimento, aunque seguidos muy de cerca por los experimentos donde se aplicaron técnicas de sobre muestreo. Podemos ratificar además que, como en el análisis de los mejores valores obtenidos, hay un mejor desempeño de los algoritmos en los experimentos realizados sobre la base original.

Finalmente se determinó hacer un análisis del comportamiento de los algoritmos de clasificación en cada uno de los experimentos, teniendo en cuenta que de éste se podrían extraer conclusiones más veraces de cuáles algoritmos son los que hacen una mejor clasificación de la base de casos en sentido general. En la Figura 3.9 se muestra los valores medios de cada algoritmo a lo largo de todos los experimentos.

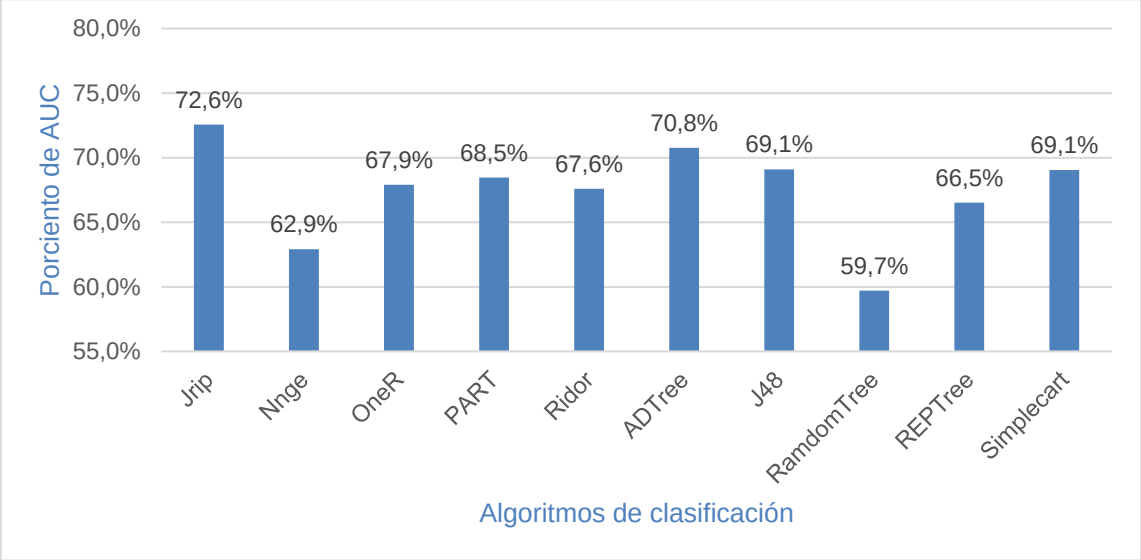

Figura 3.9 - Media obtenida por cada algoritmo de clasificación

Después de haber analizados los resultados medios de cada algoritmo se obtuvo una panorámica de cuales de éstos son los que mejor se comportaron en general para cada experimento. En primer lugar tenemos la regla de clasificación Jrip, en segundo lugar tenemos el árbol de decisión ADTree y finalmente un tercer lugar compartido entre los árboles de decisión J48 y SimpleCart. Esta información será de vital importancia para la realización de la siguiente fase.

3.3 Validación

En esta fase se hace una validación de la base de conocimiento construida utilizando para ello como conjunto de entrenamiento la base como tal y como conjunto de prueba la información de los estudiantes que se encuentran concluyendo su primer año en la carrera.

Como es lógico no se utilizaron ni todos los modelos, ni todos los algoritmos de clasificación para realizar esta validación, solo se emplearon los modelos y algoritmos en los que se obtuvieron mejores resultados.

3.3.1 Construcción y descripción de los modelos de validación

En este apartado, como en el de la fase anterior, se llevará a cabo la descripción de los experimentos de validación, se mantuvo la utilización de las herramientas KEEL y WEKA. Nuevamente, por un problema de tiempo, se determinó mantener los parámetros por defecto contenidos en ambas herramientas.

3.3.1.1 Conjunto de datos

Para la realización de la validación de los datos se decidió utilizar la información descartada en el apartado 3.1.3 referente a los estudiantes que se encuentran en este momento cursando su primer año en la carrera.

Para construir la base de conocimiento de estos estudiantes se procedió a aplicar el mismo procedimiento mencionado anteriormente en el epígrafe 3.1, para logar obtener la mayor concordancia, en cuanto al formato de la base ya existente.

Como resultado de este procedimiento se cuenta con una base con 18 atributos sobre 35 estudiantes, de ellos 15 clasificados Con Problemas y el resto clasificados Sin Problemas. Con respecto a los estudiantes clasificados Con Problemas nos basamos en que estos estudiantes están pendientes a realizar una evaluación extraordinaria que define su pase a segundo año, el resto se encuentra con todas las evaluaciones favorables para iniciar el siguiente curso su segundo año.

3.3.1.2 Modelos de validación

3.3.1.2.1 Validación 1

Para realizar este experimento se tuvo en cuenta el segundo experimento de la pasada fase, donde se obtuvo el mejor porciento de área bajo la curva. Para llevarlo a cabo, se procedió a re-balancear la base de conocimientos con la técnica de sobre muestreo SMOTE, de forma que cada clase contara con un 50% de existencia en la base de conocimiento.

Como resultado del proceso anterior se cuenta con un fichero con los 18 atributos originales, re-balanceado con la técnica SMOTE, que se utilizará como entrenamiento y un fichero con los 18 atributos de la información de los estudiantes que se encuentran en el primer año de la carrera.

Para realizar la clasificación solo se usarán los cuatro algoritmos de clasificación que mejor se comportaron en los experimentos de la pasada fase, los cuales se mencionan en el apartado anterior: JRip, ADTree, J48 y SimpleCart.

Para analizar los resultados de este experimento además de analizar el porciento del área bajo la curva, medida que se ha utilizado hasta este momento, se analizará además los porcientos de acierto para cada una de las clases.

Los resultados del modelo en cuanto al porciento de área bajo la curva se pueden observar en la Figura 3.10.

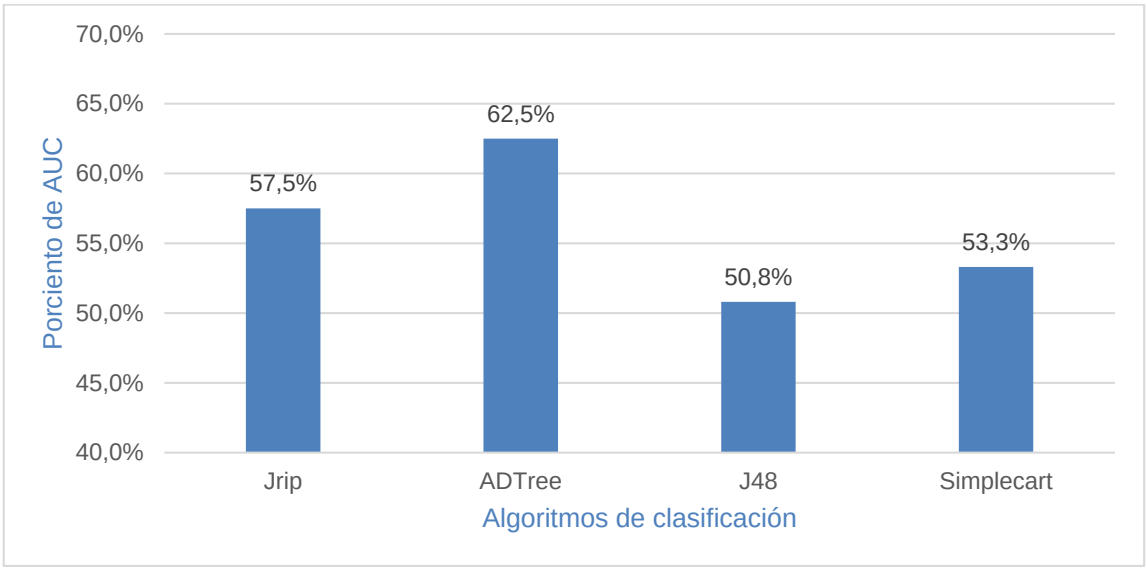

Figura 3.10 - Resultados del porciento de AUC para la Validación 1

Evidentemente los porcientos de AUC obtenidos son inferiores a los obtenidos en el segundo experimento de la pasada fase, sin embargo analizando los resultados referentes al porciento de acierto de cada clase (ver Figura 3.11) y teniendo en cuenta que la clase que deseamos que tenga un mejor porciento de aciertos es Con Problema se obtuvieron resultados bastante favorables en este caso.

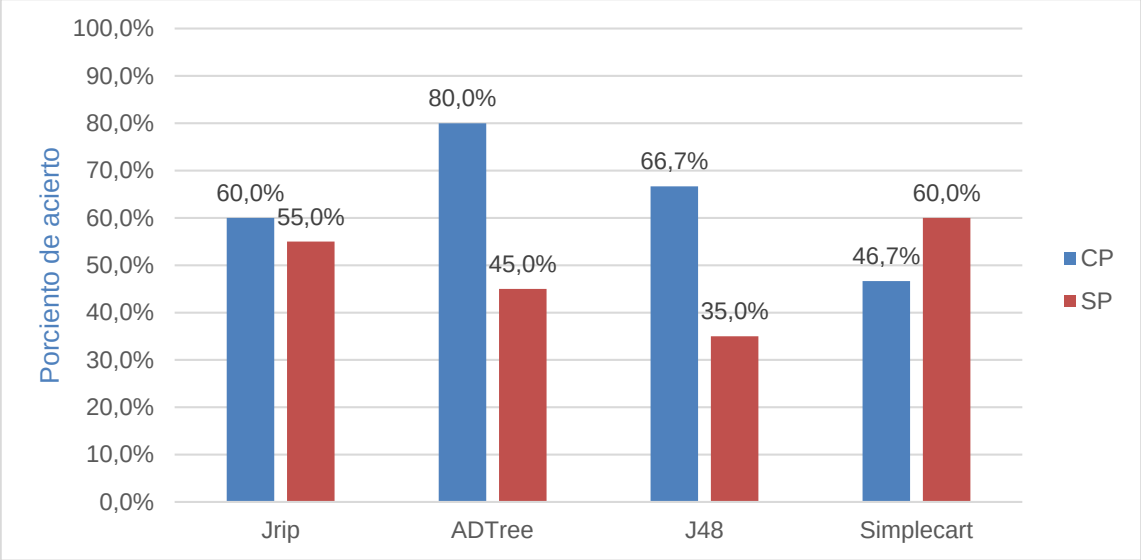

Figura 3.11 - Resultados de los porcientos de aciertos por cada clase para la Validación 1

3.3.1.2.2 Validación 2

En este caso se decidio utilizar el modelo del tercer experimento de la pasada fase, el cual resultó ser donde había una mejor eficiencia en general de los algoritmos de clasificación.

Para llevar a cabo este experimento se utilizó la base original con 18 atributos como fichero de entrenamiento y se utilizó nuevamente el fichero construido con la información de los estudiantes que se encuentran el primer año de la carrera.

Para este experimento se utilizó nuevamente el metaclasificador de WEKA **CostSensitiveClassifier** sobre el fichero original con los 18 atributos. Después de hacer varias pruebas con diferentes costos, se encontró que utilizando la misma matriz utilizada para los experimentos de la pasada fase: [0, 1; 4, 0], se obtuvieron los mejores resultados de clasificación, lo cual indica que al realizar la clasificación se tiene en cuenta que es 4 veces más importante clasificar de manera correcta los casos de Con Problemas que los casos de Sin Problemas.

Como en el experimento anterior se determinó utilizar los clasificadores con mejor desempeño en los experimentos de la pasada fase.

Los resultados del modelo en cuanto al porciento de área bajo la curva se pueden observar en la Figura 3.12.

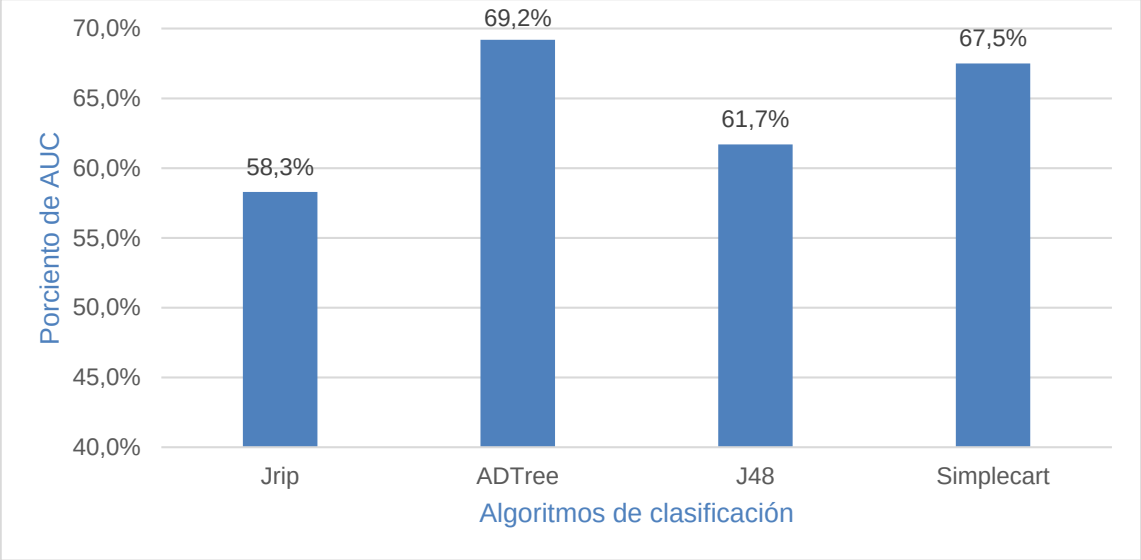

Figura 3.12 - Resultados del porciento de AUC para la Validación 2

En este experimento se observa una mejoría en cuanto a los porcientos de AUC con relación al anterior, comparables casi con los experimentos realizados en la pasada fase. También para éste analizamos el por ciento de acierto de cada clase (ver Figura 3.13). En este caso los resultados son muy buenos obteniendo porcientos de aciertos para la clase que deseamos predecir superiores al 90%, y solo con una disminución muy leve de los porcientos de la otra clase, en comparación con los obtenidos en el experimento anterior.

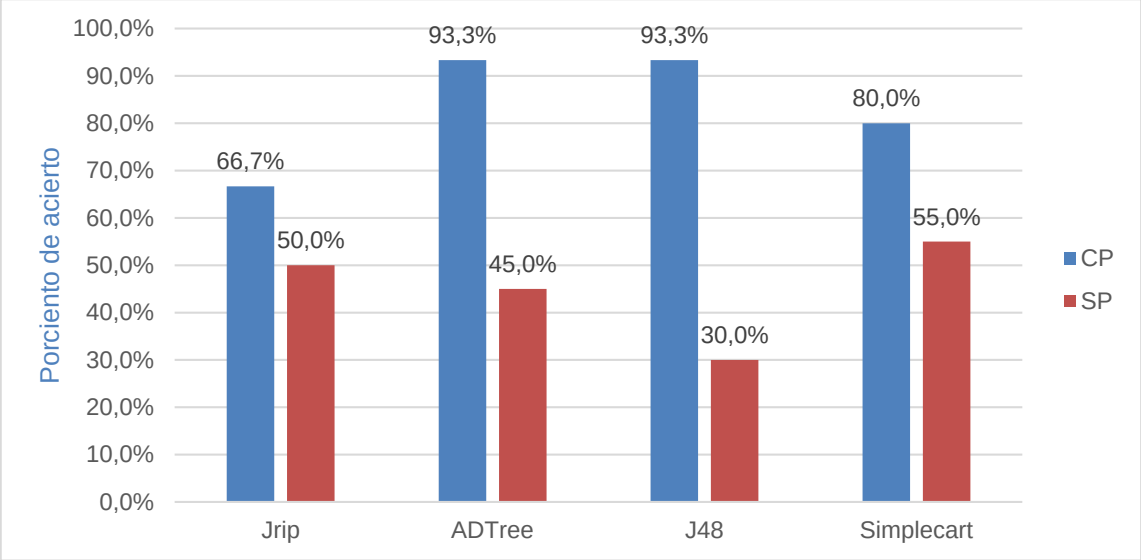

Figura 3.13 - Resultados de los porcientos de aciertos por cada clase para la Validación 2

3.3.2 Evaluación de la validación

En este apartado se realiza el resumen de los experimentos de validación realizados anteriormente, comparando cada uno de los resultados obtenidos, lo mismo para el porciento de área bajo la curva como para el porciento de aciertos de cada clase. Además se muestran los resultados del modelo del mejor clasificador, para determinar los principales factores que inciden en el fracaso estudiantil.

El siguiente gráfico muestra la comparación entre los resultados obtenidos de experimento con un sobre-muestreo de los datos utilizando la técnica SMOTE y los resultados obtenidos de aplicar diferentes costos a las clases (ver Figura 3.14).

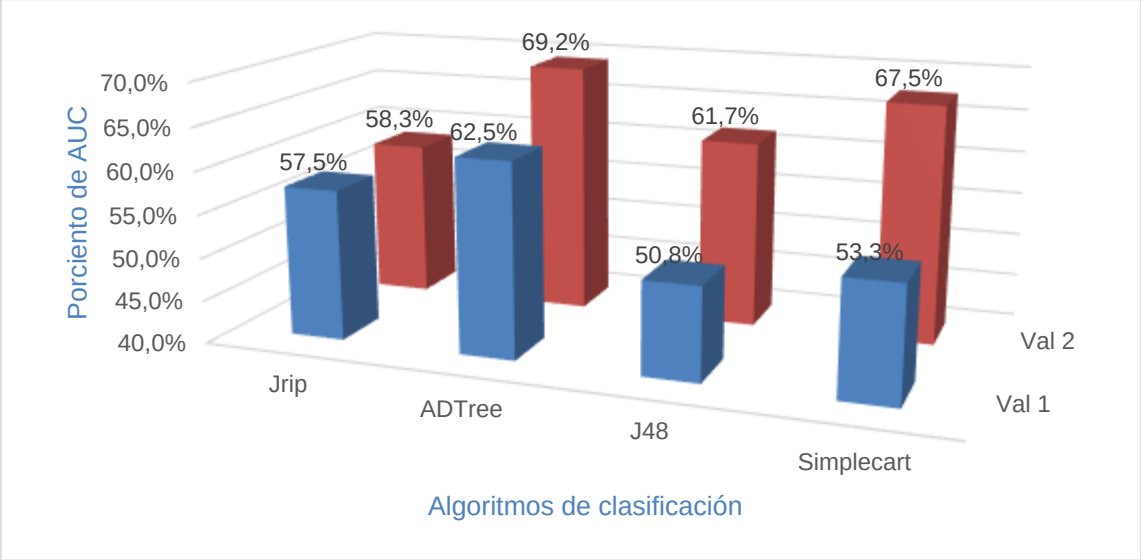

Figura 3.14 - Comparación de los resultados del porciento de AUC para la validación

Como resultado de la comparación entre los dos experimentos se observa evidentemente que la segunda validación realizada obtuvo mejores resultados que la primera en cuanto a todos los algoritmos de clasificación utilizados. Se destaca como mejor algoritmo de clasificación el árbol de decisión ADTree, para ambos experimentos, obteniendo en el primero un 62,5% de área bajo la curva y en el segundo eleva la cifra a 69,2%.

De la comparación de los resultados obtenidos para ambos experimentos en cuanto a porciento de aciertos para la clase Con problemas, que es la que nos interesa clasificar (ver Figura 3.15), se obtiene que los resultados de haber sobre-balanceado la base de conocimiento son buenos y mejoran considerablemente al hacer un análisis con diferentes costos para las clases. Se ratifica el árbol de decisión ADTree como el mejor clasificador para esta base de conocimiento porque a pesar de tener el mismo porciento de clasificación para la clase Con Problemas que el J48 tiene mejor porciento de clasificación para la clase Sin Problemas que el J48, lo que se evidencia en los resultados de los porcientos de área bajo la curva.

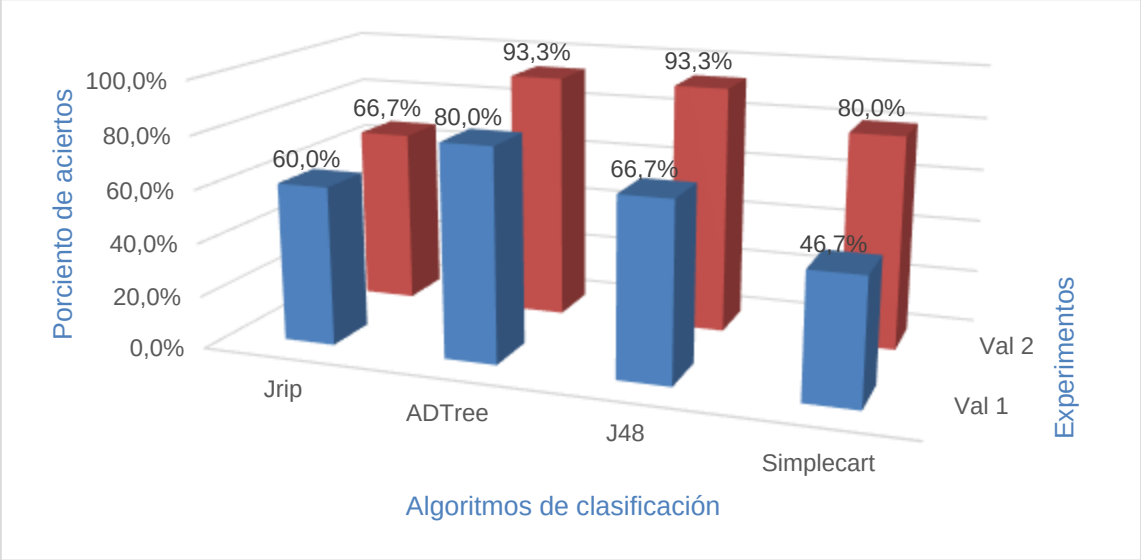

Figura 3.15 - Comparación de los resultados de los porciento de aciertos de la clase CP para la validación

A continuación se visualiza el árbol de decisión construido por el clasificador ADTree, resultado de hacer la clasificación analizando los costos referentes a cada clase y teniendo en cuenta la necesidad de obtener una buena clasificación para la clase Con Problemas.

Classifier Model
Alternating decision tree:

: -0.161
| (1)ind < 98.195: -0.577
| (1)ind >= 98.195: 0.594
| (2)mun = CAMAGUEY: 0.283
| (2)mun != CAMAGUEY: -0.231
| (3)esc < 85.735: -0.429
| (3)esc >= 85.735: 0.116
| | (4)his < 92.75: 0.306
| | (4)his >= 92.75: -0.167
| (5)col_pie = B: -0.055
| | (6)his < 71.25: -0.69
| | (6)his >= 71.25: 0.071
| (5)col_pie != B: 0.622
| (7)fue_ing = AAR: 0.798
| (7)fue_ing != AAR: -0.046

```
|  (8)fue_ing = TRA: -0.923
|  (8)fue_ing != TRA: 0.034
|  |  (9)mat < 61.5: 0.402
|  |  (9)mat >= 61.5: -0.059
|  |  |  (10)pre = ALBarba: -0.467
|  |  |  (10)pre != ALBarba: 0.059
Legend: -ve = CP, +ve = SP
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
```

Luego de observar el árbol construido por ADTree se puede afirmar que los principales atributos utilizados para la clasificación de los estudiantes que ingresan a la carrera Ingeniería Informática en el primer año del Curso Regular Diurno de la Universidad de Camagüey son los siguientes: Índice Académico, Municipio, Escalafón, Historia, Color de Piel, Fuente de Ingreso, Matemáticas y Preuniversitario.

Conclusiones parciales

En este capítulo se realizó la aplicación de las diferentes fases de la metodología propuesta en el capítulo anterior, con datos reales obtenidos por el autor en la UC, comenzándose desde el análisis exploratorio inicial de los datos, con su pre-procesamiento y limpieza hasta la aplicación de la minería de datos.

Se validaron los algoritmos de caja blanca seleccionados realizando un estudio comparativo entre 10 clasificadores, donde el JRip perteneciente al conjunto de reglas de clasificación, superó en calidad a los patrones obtenidos por los demás algoritmos.

Se aplicaron los algoritmos que mejores resultados arrojaron para diagnosticar la situación del actual del primer año de Informática, obteniéndose un índice de aciertos del 93,3% con el árbol de decisión ADTRee de la clase minoritaria, destacándose la aplicación del costo sensitivo. A su vez, la validación permitió detectar cuáles son los atributos más importantes que influyen en el fracaso de los alumnos de primer año.

La información brindada por la minería de datos del problema objeto de estudio es muy diversa, real y cumple con su fin de orientar a los decisores a aplicar medidas que mitiguen las fuentes que causan la deserción estudiantil en el primer año de la carrera.

CONCLUSIONES

Cada uno de los capítulos anteriores contiene una discusión detallada de las conclusiones correspondientes por lo que en este apartado se plantean los resultados y conclusiones de la tesis desde una perspectiva más general.

La investigación desarrollada en esta tesis se ha centrado en la caracterización y clasificación de los estudiantes del primer año de Informática de la UC. Desde el punto de vista de los datos, se ilustra una vez más, que en un conjunto desbalanceado hay que aplicar técnicas de pre-procesamiento y algoritmos que no sean sensibles a esta problemática. Además, se muestra la utilidad de recurrir a los clasificadores de caja blanca, de tal manera que permitan a los profesores entender el modelo obtenido de la minería de datos.

Los resultados y conclusiones más importantes de la tesis son los siguientes:

1. Se seleccionaron la metodología, con su correspondiente adecuación, y las herramientas para el análisis de datos utilizando algoritmos de minería de datos para ser aplicado al problema presente en la carrera de Informática de la UC con su primer año. La selección favoreció a las técnicas de caja blanca por su fácil entendimiento por los humanos.
2. Se creó una base de conocimientos con la información de los estudiantes de Informática a partir de los datos almacenados en el SIGENU, para lo cual se realizó un proceso de limpieza y eliminación de atributos no relevantes.
3. Como resultado de los experimentos realizados se concluye que se obtienen mayor valor de AUC realizando la clasificación con todos los atributos. Se demuestra que la aplicación de técnicas para el tratamiento de datos desbalanceados superan los resultados a cuando no se consideran dichas técnicas.
4. En el diagnóstico de la situación docente del actual primer año de Informática, se logró obtener un índice de aciertos elevados con respecto a los estudiantes que

llevan mundiales del primer semestre y su posibilidad de causar baja o arrastrar al segundo año alguna de estas asignaturas.

5. Con esta investigación se logró un paso de avance en la caracterización de los factores que influyen en el fracaso de los estudiantes de primer año de Informática que servirá de base para investigaciones venideras en dicha carrera y en otras de la Universidad de Camagüey.

RECOMENDACIONES

Las recomendaciones fundamentales para la comunidad que trabaja en el área de la minería de datos educacional, es que debe tener en cuenta las potencialidades de los algoritmos de caja blanca y las técnicas de pre-procesamiento cuando los datos son desbalanceados. Además, que al trabajar con la deserción escolar con minería de datos es preferible utilizar algoritmos que clasifiquen con mayor acierto la clase minoritaria, que de forma general representa a los estudiantes con problemas y con los cuales hay que trabajar de manera diferenciada por parte de los profesores.

Se pueden identificar como líneas de trabajo futuro las siguientes:

1. Extender la investigación a otras carreras de la universidad y de la propia carrera de Informática, aplicando una encuesta que permita obtener datos sociales, económicos y demográficos de los estudiantes.
2. Desarrollar una aplicación Web que permita a los profesores hacer estudios de las características de los estudiantes a partir de los modelos estudiados en esta tesis. La misma debe permitir la incorporación de nuevos casos a la base de conocimiento.
3. Validar el comportamiento de otros clasificadores que no sean de caja blanca como las redes neuronales, las redes Bayesianas, las máquinas de soporte vectorial, entre otros.

REFERENCIAS BIBLIOGRÁFICAS

1. Agrawal, R. y Shafer, J. C., (1996) "Parallel Mining of Association Rules" en *IEEE Transactions on Knowledge and Data Engineering*.
2. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L. y Herrera, F., (2011) "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework" en *Journal of Multiple-Valued Logic and Soft Computing*. Vol. 17, número 2–3, pp. 255–287.
3. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J. y Herrera, F., (2009) "KEEL: A software tool to assess evolutionary algorithms to data mining problems" en *Soft Computing*. Vol. 13, número 3, pp. 307–318.
4. Alvares, L. A., (2009) "Comportamiento de la Deserción y Reprobación en el Colegio de Bachilleres del Estado de Baja California: Caso Plantel Ensenada", *X Congreso Nacional de Investigación Educativa*, México.
5. ANUIES, (2003) "El significado de la tutoría académica en estudiantes de primer ingreso a la licenciatura" en *Revista de la Educación Superior*. Vol. 3, número 127.
6. Batista, G. E. A. P. A., Prati, R. C. y Monard, M. C., (2004) "A study of the behaviour of several methods for balancing machine learning training data" en *SIGKDD Explorations*. Vol. 6, número 1, pp. 20–29.
7. Bradley, A. P., (1997) "The use of the area under the ROC curve in the evaluation of machine learning algorithms" en *Pattern Recognition*. Vol. 30, número 7, pp. 1145-1159.
8. Brito, R., Rosete, A. y Acosta, R., (2008) "Desarrollo de un proceso de KDD en el ámbito docente: Preparación de los datos", *14 Convención Científica de Ingeniería y Arquitectura*, La Habana. Cuba.
9. Cohen, W. W., (1995) "Fast Effective Rule Induction", *Twelfth International Conference on Machine Learning*, pp. 115-123.
10. Colonia, S., (2010) *Caracterización del perfil de los estudiantes de posgrado a partir de la información de admisión y el desempeño académico.* Unpublished Tesis de Maestría. Universidad de Medellín, Medellín, Colombia.
11. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R., (2000) "CRISP-DM 1.0. Guía paso a paso de Minería de Datos. Manual de uso". *DataPrix*, disponible en: http://www.dataprix.com/modelo_crisp-dm [Accesado: Febrero 2013].
12. Chawla, N. V., Bowyer, K. W., Hall, L. O. y Kegelmeyer, W. P., (2002) " SMOTE: synthetic minority over-sampling technique" en *Journal of Artificial Intelligence Research*. Vol. 16, pp. 321-357.

13. Chen, M.-S., Han, J. y Yu, P. S., (1996) "Data mining: an overview from a database perspective" en *Knowledge and data Engineering, IEEE Transactions on*. Vol. 8, número 6, pp. 866-883.
14. Derrac, J., García, S., Molina, D. y Herrera, F., (2011) "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms" en *Swarm and Evolutionary Computation*. Vol. 1, número 1, pp. 3–18.
15. Dietterich, T., Lathrop, R. y Lozano-Pérez, T., (1997) "Solving the multiple instance problem with axis-parallel rectangles" en *Artifical Intelligence*. Vol. 89, número 1–2, pp. 31–71.
16. Fayyad, U. M., (1996) "Data Mining and Knowledge Discovery: Making Sense out of Data" en *IEEE Intelligent Systems*. Vol. 11, número 5.
17. Frank, E. y Witten, I. H., (1998) "Generating Accurate Rule Sets Without Global Optimization", *Fifteenth International Conference on Machine Learning*, pp. 144-151.
18. Frawley, Piatesky-Shapiro y Matheus, (1992) "Knowledge Discovery in Databases: an Overview" en *AI Magazine*.
19. Freund, Y. y Mason, L., (1999) "The alternating decision tree learning algorithm", *Proceeding of the Sixteenth International Conference on Machine Learning*, Bled, Slovenia, pp. 124-133.
20. García, R., Britos, P. V., Hossian, A. y Sierra, E., (2005) Minería de datos Basada en Sistemas Inteligentes. In (Primera edición ed.).
21. García, S. y Herrera, F., (2008) "An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons" en *Journal of Machine Learning Research*. Vol. 9, pp. 2579–2596.
22. Hall, M. A. y Holmes, G., (2002) "Benchmarking Attribute Selection Techniques for Data Mining". *University of Waikato, Department of Computer Science*, Hamilton, New Zealand, disponible en: http://www.cs.waikato.ac.nz/~ml/publications/2000/00MH-GH-Benchmarking.pdf [Accesado].
23. Han, J. y Kamber, M., (2001) *Data mining concepts and techniques.* San Francisco (CA): Morgan Kaufmann Publishers.
24. Hernández, O. J., Ramírez, Q. M. y Ferri, R. C., (2004) *Introducción a la minería de datos.* Madrid (España): Editorial Pearson Prentice Hall.
25. Holte, R. C., (1993) Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. In R. C. Holte (Ed.), *Machine Learning* (pp. 63-91). Ottawa.
26. Huang, J. a. y Ling, C. X., (2005) "Using AUC and accuracy in evaluating learning algorithms" en *IEEE Transactions on Knowledge and Data Engineering*. Vol. 17, número 3, pp. 299-310.
27. Imielinski, T. y Mannila, H., (1996) "A database perspective on knowledge discovery" en *Communications of the ACM*. Vol. 39, número 11, pp. 58-64.
28. Kairúz, G., (2008) *Tutorial para el aprendizaje del Weka.* Nueva Zelanda: Universidad de Zelanda.

29. Kononenko, I., (1994) "Estimating Attributes: Analysis and Extensions of RELIEF", *European Conference on Machine Learning*, pp. 171-182.
30. Kotsiantis, S., Patriarcheas, K. y Xenos, M., (2010) "A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students´ Performance in Distance Education" en *Knowledge Based System*. Vol. 23, número 6, pp. 529-535.
31. Martin, B., (1995) *Instance-Based learning: Nearest Neighbor With Generalization.* New Zealand: Hamilton.
32. Más-Estellés, J., Alcover-Arándiga, R., Dapena-Janeiro, A., Valderruten-Vidal, A., Satorre-Cuerda, E., Llopis-Pascual, F., Rojo-Guillén, T., Mayo-Gual, R., Bermejo-Llopis, M., Gutiérrez-Serrano, J., García-Almiñana, J., Tovar-Caro, E. y Menasalvas-Ruiz, E., (2009) "Rendimiento Académico de los Estudios de Informática en Algunos Centros Españoles", *XV Jornadas de Enseñanza Universitaria de la Informática*, Barcelona.
33. Molina, J., (2001) *Torturando a los Datos Hasta que Confiesen.* Unpublished Coordinador del programa de Data Mining. Universidad Oberta de Catalunya (UOC).
34. Orriols-Puig, A. y Bernad´o-Mansilla, E., (2009) "Evolutionary rule –based systems for imbalanced datasets" en *Soft Computing*. Vol. 13, número 1, pp. 21.
35. Pautsch, G., (2009) *Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación.* Unpublished Tesis de Grado. Universidad Nacional de Misiones, Argentina.
36. Quinlan, J. R., (1993) *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann Publishers.
37. Reyes, J. y García, R., (2005) *El proceso de descubrimiento de conocimiento en Bases de Datos.* Unpublished Apuntes de investigación. Universidad Autónoma de Nuevo León, México.
38. Romero, C. y Ventura, S., (2007) "Educational data mining: A Survey From 1995 to 2005" en *Expert System with Applications*. Vol. 33, pp. 135-146.
39. Romero, C. y Ventura, S., (2010) "Educational Data mining: A Review of the State of the Art" en *IEEE Transactions on Systems, Man, and Cybernetics*.
40. Timarán Pereira, R., (2010) "Una Lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos" en *Revista Científica Guillermo de Ockham*. Vol. 8, número 1, pp. 121-130.
41. UPN, (2009) "La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN", disponible en: http//www.mineducacion.gov.co/1621/articles-85600_Archivo_pdf3.pdf. [Accesado: Febrero 2013].

ANEXOS

Anexo A: Resultado del proceso de selección de atributos, donde se determina un nuevo ranking partiendo de la información resultante de aplicar los cinco algoritmos de selección.

	mun	fue_ing	ori_aca	est_civ	org_pol	sex	col_pie	tip_est	na_mad	tip_sm	eda_ing	pre	ind	his	mat	esc	opc
ChiSquared	4	7	15	17	11	14	16	6	12	13	9	1	2	5	8	3	10
GainRatio	7	6	14	16	9	15	17	4	13	12	10	5	1	2	8	3	11
OneR	17	10	8	7	4	3	5	6	13	14	9	2	1	15	16	12	11
ReliefF	7	3	11	16	4	6	17	9	10	5	2	1	15	12	14	13	8
Symmetrical Uncert	6	7	15	17	10	14	16	5	12	13	9	4	1	3	8	2	11
Total	**41**	**33**	**63**	**73**	**38**	**52**	**71**	**30**	**60**	**57**	**39**	**13**	**20**	**37**	**54**	**33**	**51**
Nuevo Ranking	**9**	**5**	**15**	**17**	**7**	**11**	**16**	**3**	**14**	**13**	**8**	**1**	**2**	**6**	**12**	**4**	**10**

Anexo B: Resultado de la experimentación realizada de aplicar una reducción de atributos utilizando los diez algoritmos de clasificación mencionados anteriormente.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Jrip	0,610	0,675	0,694	0,681	0,701	0,704	0,679	0,699	0,664	0,653	0,671	0,723	0,671	0,660	0,642	0,655
Nnge	0,568	0,646	0,646	0,592	0,590	0,601	0,610	0,615	0,610	0,579	0,599	0,577	0,588	0,592	0,587	0,581
OneR	0,613	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646	0,646
PART	0,500	0,657	0,670	0,634	0,652	0,653	0,642	0,626	0,680	0,650	0,587	0,591	0,642	0,644	0,657	0,657
Ridor	0,515	0,583	0,625	0,648	0,663	0,595	0,617	0,617	0,604	0,604	0,590	0,635	0,635	0,635	0,635	0,635
ADTree	0,603	0,713	0,708	0,642	0,604	0,620	0,600	0,602	0,644	0,644	0,644	0,715	0,715	0,703	0,703	0,703
J48	0,500	0,500	0,548	0,692	0,695	0,708	0,693	0,708	0,697	0,701	0,681	0,681	0,679	0,672	0,672	0,618
RamdomTree	0,613	0,640	0,635	0,601	0,630	0,626	0,681	0,659	0,622	0,643	0,634	0,650	0,666	0,605	0,594	0,657
REPTree	0,580	0,572	0,572	0,615	0,611	0,614	0,607	0,602	0,602	0,637	0,637	0,634	0,634	0,639	0,639	0,639
SimpleCart	0,591	0,664	0,660	0,675	0,675	0,667	0,664	0,664	0,664	0,669	0,669	0,667	0,667	0,669	0,669	0,669
Total	**0,569**	**0,630**	**0,640**	**0,643**	**0,647**	**0,643**	**0,644**	**0,644**	**0,643**	**0,643**	**0,636**	**0,652**	**0,654**	**0,647**	**0,644**	**0,646**

Anexo C: Gráfico que evalúa el desempeño de los resultados del Anexo B.