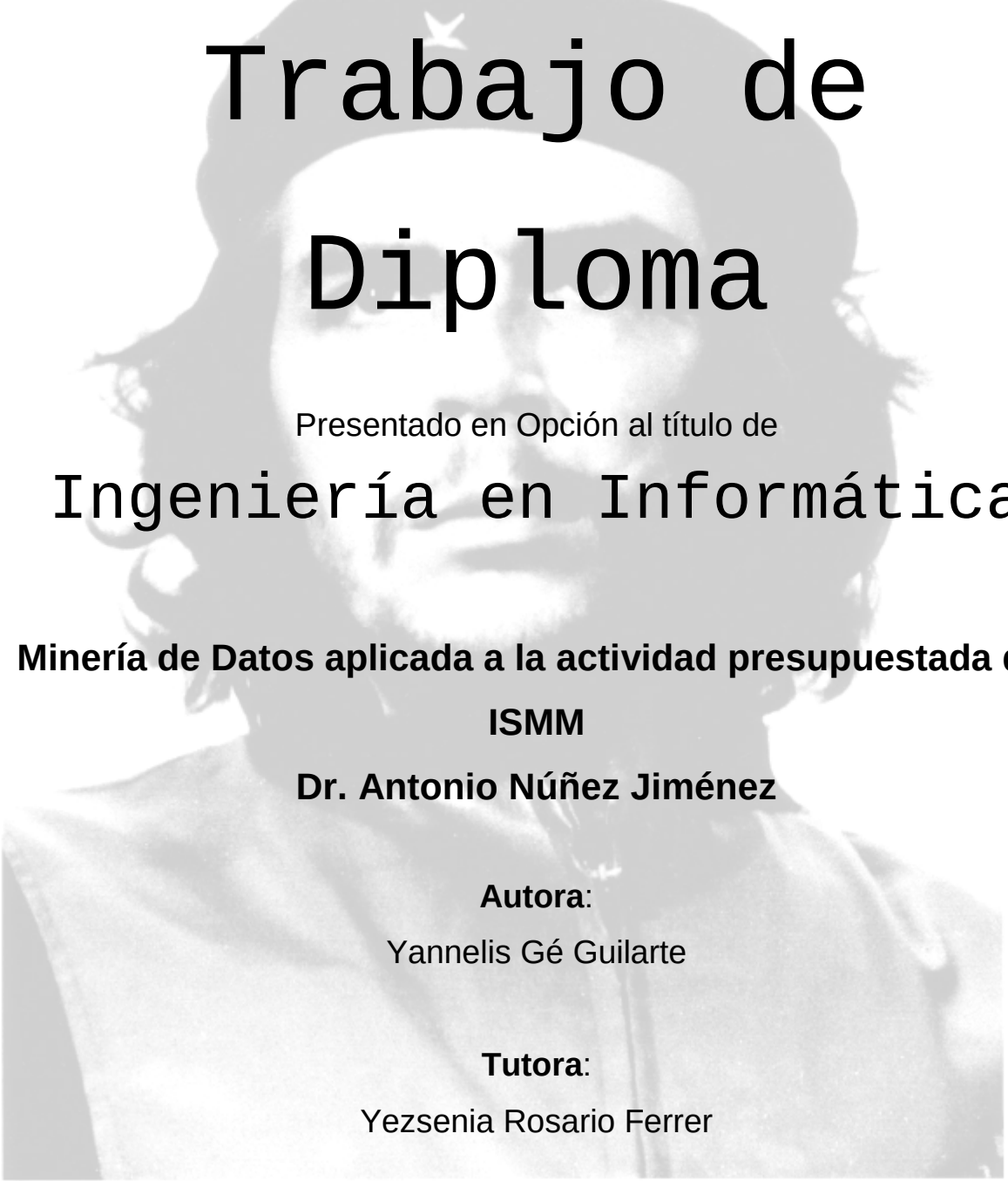
INSTITUTO SUPERIOR MINERO METALÚRGICO

**“Dr. Antonio Núñez Jiménez”.**

**Facultad de Metalurgia - Electromecánica**

**Moa, Holguín**



# Trabajo de Diploma

Presentado en Opción al título de

## Ingeniería en Informática

**Minería de Datos aplicada a la actividad presupuestada del ISMM Dr. Antonio Núñez Jiménez**

**Autora**:

Yannelis Gé Guilarte

**Tutora**:

Yezsenia Rosario Ferrer

Moa, Holguín, Cuba

Junio, 2011

"Año 53 de la Revolución"

**DECLARACIÓN DE AUTORÍA**

Yo, Yannelis Gé Guilarte, declaro que soy la única autora de este trabajo y autorizo al INSTITUTO SUPERIOR MINERO METALURGICO "Dr. Antonio Núñez Jiménez" para que hagan el uso que estimen pertinente con este trabajo.

Para que así conste firmo la presente a los _____ días del mes de_____ del 2011

Yannelis Gé Guilarte

Nombre completo del primer autor

Yezsenia Rosario Ferrer

Nombre completo del primer tutor

# Agradecimientos

Gracias:

A las personas que más han deseado que logre mis sueños:

mis padres: Mariela Guilarte y Wilson Gé, les regalo este momento,

A su ayuda idónea y amigos:

mi madrastra y padrastro,

A quien con este logro pienso servir de ejemplo y guía:

mi hermana Sianna,

A mis viejitos más queridos y especiales:

mis abuelos: Lula, Toña y Virgilio, se que este también ha sido su sueño,

A mis amigos, en donde estén, por su apoyo incondicional y palabras de aliento que nunca faltan,

A mi tutora por su paciencia,

A mi familia toda por siempre haber creído en mí,

A mis compañeros de aula,

A todos los que de una forma u otra me han ayudado y brindado sus servicios para que este, mi deseo, se hiciera realidad,

A todos, muchas gracias !!!.

Y, por sobre todas las cosas y nombres, doy gracias a Dios por ustedes y por este que ha sido mi sueño, por su gracia, misericordias y amor infinito a mí y a todos.

- Gracias porque me viste y me guardaste como a la niña de tus ojos.

“Le halló en tierra de desierto,
Y en yermo de horrible soledad;
Lo trajo alrededor, lo instruyó,
Lo guardó como a la niña de su ojo.”

Deuteronomio 32:10

## RESUMEN

En los últimos años ha existido un gran crecimiento en nuestras capacidades de generar y almacenar datos, de ahí que el volumen y la variedad de información han ido en aumento. Esto ha traído como consecuencia la incapacidad humana para analizar y transformar la información en conocimiento útil. La necesidad de análisis de los datos almacenados ha motivado el empleo de técnicas y herramientas de minería de datos, que posibiliten la extracción de conocimiento, en forma de reglas o patrones a partir de dichos datos.

La investigación tuvo como objetivo fundamental la obtención de mejoras en el proceso de elaboración del presupuesto en el Instituto Superior Minero Metalúrgico (ISMM), a partir del descubrimiento de conocimiento oculto en los datos del presupuesto almacenados desde el 2004 registrados en documentos en formato Excel.

En este sentido, se detallan algunos aspectos relacionados con la Minería de Datos y su aplicación. Se selecciona y describe a la metodología CRISP-DM (Cross – Industry Standard Process for Data Mining) y a la herramienta de análisis de datos WEKA (Waikato Environment Knowledge Analysis), y se obtienen modelos que permitan apoyar determinadas actividades orientadas a la gestión económica.

## ABSTRACT

A great growing in our capabilities of generating and storage data has existed in the last years, therefore the density and variety of information are in raise. This has brought as a consequence the human inability to analyze and transform the information in useful knowledge. The necessity of data analysis storage has motivated the employment of techniques and tools of data mining that allow the extraction of knowledge, in form of rules or patterns from the named data.

The research had as main objective the obtain of improvement in the elaboration process of the budgets in the High Institute of Metallurgy and Mining from the discovery of hidden knowledge in the budgets data storage since 2004 registered in Excel format documents.

In this way, are detach some aspects related to the Data Mining and its application. It is selected and described to the methodology CRISP-DM (Cross – Industry Standard Process for Data Mining) and to the tool of data analysis WEKA (Waikato Environment Knowledge Analysis) and models are obtained that allow to support determinate activities oriented to the economic management.

# ÍNDICE DE CONTENIDO

# INTRODUCCIÓN

La información que se genera diariamente dentro de la organización es uno de sus activos principales, por lo que se debe orientar los recursos tecnológicos de manera que ayuden a tomar decisiones estratégicas y oportunas.

La capacidad de solucionar problemas de decisión, y la calidad de las decisiones tomadas, tienen grandes repercusiones en la organización y en su correcto funcionamiento, de modo que actualmente se enfrentan a la paradoja de que, cuantos más datos se tienen disponibles, menos información se tiene. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

La toma de decisiones es el proceso mediante el cual se realiza una elección entre las alternativas o formas para resolver diferentes situaciones de la vida, estas se pueden presentar en diferentes contextos: a nivel laboral, familiar, sentimental, empresarial (utilizando metodologías cuantitativas que brinda la administración) [Wikipedia 1, 2011].

En una época de cambios en la gerencia moderna, la toma de decisiones exige presión y rapidez, y por lo tanto el factor de predicción y control de los presupuestos es de vital importancia como una eficiente herramienta administrativa.

Se le denomina presupuesto al cálculo anticipado de los ingresos y gastos de una actividad económica durante un período, por lo general en forma anual. Es un plan de acción dirigido a cumplir una meta prevista, expresada en valores y términos financieros que debe cumplirse en determinado tiempo y bajo ciertas condiciones previstas, este concepto se aplica a cada centro de responsabilidad de la organización. El presupuesto es el instrumento de desarrollo anual de las empresas o instituciones cuyos planes y programas se formulan por término de un año.

De ahí que elaborar el presupuesto permite a las empresas, los gobiernos, las

organizaciones privadas o las familias establecer prioridades y evaluar la consecución de sus objetivos. Para alcanzar estos fines, puede ser necesario incurrir en déficit (que los gastos superen a los ingresos) o, por el contrario, puede ser posible ahorrar, en cuyo caso el presupuesto presentará un superávit (los ingresos superan a los gastos) [Wikipedia 2].

El Instituto Superior Minero Metalúrgico (ISMM) cuenta con una Dirección de Economía en la que una de sus tareas es la elaboración del presupuesto, el mismo se hace de forma manual por una trabajadora del centro, la cual ha registrado y almacenado los presupuestos desde 2004 en libros de Microsoft Excel. El conocimiento o predicción del comportamiento de las variables presupuestadas podría ser de suma importancia en la elaboración y desagregación del presupuesto en años futuros pero la cantidad de datos almacenados es tan vasta que excede la habilidad para reducirla y analizarla sin el uso de técnicas de análisis de datos automatizadas, lo que constituye una **situación problémica**.

Por lo que se plantea como **problema de investigación** la ausencia de un modelo para el apoyo a la toma de decisiones respecto al presupuesto asignado a los centros de costo del ISMM.

En la actual sociedad de la información, donde día a día se multiplica la cantidad de datos almacenados casi de forma exponencial, la Minería de Datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz para los objetivos de cualquier organización [López Arévalo, 2005].

La Minería de Datos (DM, por las siglas en inglés de Data Mining) es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Las herramientas de Minería de Datos predicen futuras tendencias y comportamientos, permitiendo en los negocios la toma de decisiones. Un término relacionado con la minería de datos es la extracción

o "descubrimiento de conocimiento en bases de datos" (Knowledge Discovery in Databases o KDD).

El **objeto de estudio** de esta investigación es el descubrimiento de conocimiento en bases de datos económicas.

El **campo de acción** está dado por la aplicación de métodos inteligentes, con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u “ocultos” en los datos de presupuesto obtenidos del departamento de finanzas del ISMM , teniendo como nombre dicho proceso de Minería de Datos.

Para facilitar la solución práctica de la problemática planteada se determina la siguiente **hipótesis:** con la utilización de los métodos, técnicas, algoritmos y tareas de minería de datos, así como la aplicación de una herramienta adecuada para procesar los datos y una metodología que guíe el proceso, se podrá construir un modelo que apoye en la toma de decisiones en cuanto a qué cambios realizar para una mejor asignación del presupuesto en el ISMM.

Para darle solución al problema antes expuesto se trazó como **objetivo general**: elaborar un modelo que permita describir la relación entre las partidas y elementos del presupuesto para la extracción del conocimiento oculto en la información con el fin de apoyar la toma de decisiones en la elaboración del presupuesto en el ISMM.

De acuerdo a esta propuesta se derivan los siguientes **objetivos específicos:**

- Seleccionar una metodología para el proceso de KDD y herramientas para el análisis de los datos que serán procesados.
- Determinar los métodos, técnicas, tareas y algoritmos de Minería de Datos apropiados para las características del problema.
- Desarrollar un proceso de KDD para obtener un modelo que caracterice el

comportamiento de las variables presupuestarias aplicando la metodología seleccionada.

Para darle cumplimiento a los objetivos antes definidos se plantean como **tareas a desarrollar:**

1. Elaboración de los fundamentos teóricos de la investigación.
2. Estudio y documentación de la metodología y herramienta de Minería de Datos seleccionada para el desarrollo del proceso de KDD.
3. Estudio y documentación de los métodos, técnicas, tareas y algoritmos de Minería de Datos.
4. Selección, limpieza y transformación de los datos que se van a analizar.
5. Desarrollo de la etapa de Minería de Datos dentro del proceso de KDD con los métodos, técnicas, tareas y algoritmos de Minería de Datos seleccionados.

Para la realización de las tareas antes propuestas se utilizaron los **métodos empíricos y teóricos** de investigación científica.

Entre los **métodos empíricos** de investigación utilizados para la obtención de información sobre los datos recopilados que serán analizados está **la encuesta**, entre las técnicas que esta emplea para la obtención de información se utilizan:

- **La entrevista** porque permite obtener la mayor información posible acerca del tema de investigación, además de las experiencias, las ideas y los puntos de vistas de los entrevistados que aportan conocimientos específicos del tema.

Entre los **métodos teóricos** de investigación, utilizados para descubrir y relevar la esencia del objeto y sus relaciones, se emplean:

- **Método Analítico-Sintético** porque permite analizar, estudiar en partes e interpretar la teoría con el fin de extraer los elementos más importantes que se relacionan con el objeto.

- **Método Inductivo-Deductivo** porque permite a través de un razonamiento llegar

a un grupo de conocimientos particulares y generales.

- **Método de Modelación** porque permite la creación de modelos, es decir representar lo que se quiere estudiar de forma más simple, explicando lo que pasa de una manera lógica.

Otro de los métodos es el **análisis documental** ya que la recopilación y análisis de documentos ahorra esfuerzo y rentabiliza el trabajo, indicando situaciones y hechos para estudiar.

El presente trabajo consta de Introducción, 2 Capítulos, Conclusiones, Recomendaciones, Bibliografía y Anexos:

En el **Capítulo 1: Fundamentación teórica:** se ofrece una breve descripción de diferentes conceptos imprescindibles que le dan base a la investigación. Se introducen temas relacionados con el proceso de descubrimiento de conocimiento, la minería de datos, herramientas utilizadas en proyectos donde es aplicada, así como, metodologías. Se justifica la selección de la metodología y de la herramienta, CRISP-DM y WEKA respectivamente.

En el **Capítulo 2: Minería de Datos aplicada a la actividad presupuestada del ISMM Dr. Antonio Núñez Jiménez:** Se da a conocer el desarrollo de las fases de la metodología CRISP-DM en el problema a desarrollar. Se aplican los modelos, técnicas y algoritmos de Minería de Datos, utilizando WEKA, relacionados con la extracción de conocimiento.

Finalmente se muestran los resultados obtenidos y las Conclusiones a las que se arribaron. Se establece si se cumplieron los objetivos planteados, así como las Recomendaciones que se proponen de acuerdo al resultado obtenido y los Anexos con información necesaria sobre el trabajo.

# 1

# Capítulo 1. Fundamentación Teórica

## Introducción

En los últimos años ha existido un gran crecimiento en nuestras capacidades de generar y almacenar datos, esto se debe básicamente al bajo costo de almacenamiento y al poder de procesamiento que las máquinas han ido adquiriendo, de ahí que el volumen y la variedad de información han ido en aumento.

Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido, la cual es útil en cuanto explicar el pasado, entender el presente y predecir la información futura.

Dentro de estas enormes masas de datos existe una gran cantidad de información oculta a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información es posible gracias a la Minería de Datos, que nos permite la creación de modelos (es decir, representaciones abstractas de la realidad) a partir de patrones y relaciones encontrados dentro de los datos.

Este estudio está dirigido fundamentalmente a la obtención de mejoras en el proceso de elaboración del presupuesto en el ISMM así como en el perfeccionamiento a la hora de la toma de decisiones.

El presente capítulo abordará temas relacionados con el proceso de descubrimiento de conocimiento, la minería de datos así como herramientas, técnicas y metodologías de la misma.

## 1.1 Descubrimiento de Conocimiento en Bases de Datos

La idea general de descubrir "conocimiento" en grandes bases de datos es intuitiva y llamativa, pero técnicamente hablando es todo un desafío.
¿Qué es *conocimiento*? Desde el punto de vista de las organizaciones, se define el conocimiento como aquella información que permite generar acciones asociadas a satisfacer las demandas del mercado, y apoyar las nuevas oportunidades a través de la explotación de las competencias centrales de la organización. El conocimiento es una combinación de valores, información contextualizada y experiencias que proporcionan un marco para evaluar e incorporar nuevas experiencias e información. El conocimiento se origina y aplica en la mente de las personas. En las organizaciones, el conocimiento reside en documentos y bases de datos y también en los procesos, prácticas y normas corporativas.

El concepto de Descubrimiento de Conocimiento en los Datos suele indicarse en inglés con las siglas KDD, abarcando con ello tanto a las bases de datos (Knowledge Discovery in Databases), como a la Minería de Datos (Knowledge Discovery and Data Mining). Una definición clásica de este concepto es el siguiente: “El descubrimiento de conocimiento en las bases de datos es el proceso no trivial de identificación de patrones válidos, nuevos, potencialmente útiles y, finalmente, comprensibles en los datos” [Fallad et al, 1996].

Se entiende actualmente a KDD como el proceso, que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales permiten que dichos datos adquieran un sentido y aporten un nuevo conocimiento.

Surge como necesidad de analizar grandes volúmenes de información almacenados en bases de datos ya sea que se encuentren en programas como SQL SERVER, EXCEL, ACCESS, ORACLE, MySQL [Gómez]. Implica un proceso interactivo e iterativo, involucrando la aplicación de métodos de minería de datos, para extraer o identificar lo que se considera como conocimiento de acuerdo a la especificación de ciertos

parámetros usando una base de datos, junto con pre-procesamientos, muestreo y transformaciones de la base de datos. La meta de este proceso es justamente procesar automáticamente grandes cantidades de datos crudos, identificar los patrones más significativos y relevantes, y presentarlos como conocimiento apropiado para satisfacer las metas del usuario [Acosta].

### *1.1.1 El proceso de KDD y La Minería de Datos*

La minería de datos revela *patrones* o *asociaciones* que son desconocidos para el usuario, por esta razón, entra o se asocia con el contexto de Descubrimiento de Conocimientos en las Bases de Datos (KDD, por las siglas en inglés de Knowledge Discovery in Database). Este término es originado de la *Inteligencia Artificial* (AI).

Aunque algunos autores usan los términos Minería de Datos y KDD indistintamente, como sinónimos, existen claras diferencias entre los dos. Así la mayoría de los autores coinciden en referirse al KDD como un proceso que consta de un conjunto de fases, una de las cuales es la minería de datos [Berthold, M.; Hand, D.J. (eds.) 2003]. De acuerdo con esto, el proceso de minería de datos consiste únicamente en la aplicación de un algoritmo para extraer patrones de datos y se llamará KDD al proceso completo que incluye pre-procesamiento, minería y post-procesamiento de los datos.

El enfoque de adquirir conocimientos de un conjunto de datos fue separado por Fayyad en pasos individuales. Según Fayyad, hay cinco pasos: Selección, pre-procesamiento, transformación, minería de datos e interpretación.

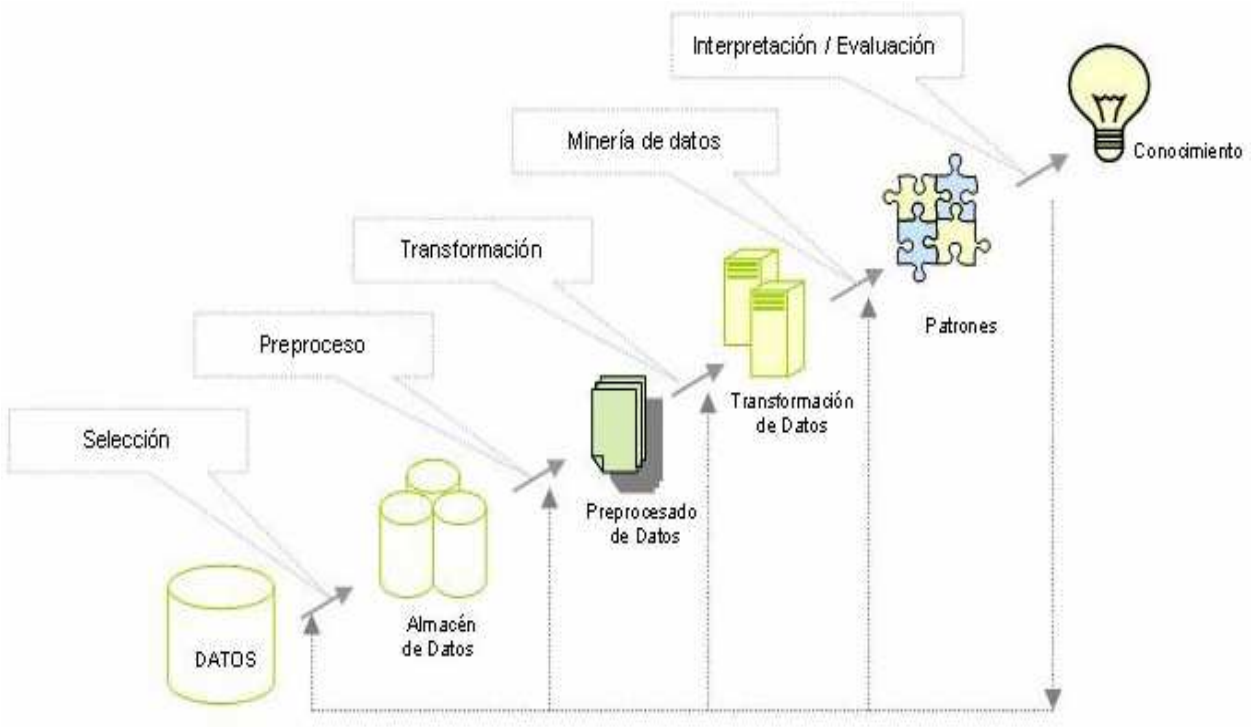



**Fig. 1 Pasos en el proceso de KDD (proceso por Fayyad 1996)**

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

**Pre-procesamiento:** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

**Transformación:** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

**Minería de Datos:** Es la fase de modelamiento propiamente tal, en donde métodos

inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.

**Interpretación y Evaluación:** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

## 1.2 Minería de Datos

La Minería de Datos es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Tiene como objetivo analizar los datos para extraer conocimiento. Aunque se basa en varias disciplinas, algunas de ellas más tradicionales (como la estadística), se distingue de ellas en la orientación más hacia el fin que hacia el medio. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido a la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones [Zamarrón, 2006]. Es un proceso que invierte la dinámica del método científico, el cual consiste en formular una hipótesis y luego se diseña el experimento para confirmarla o refutarla; y en minería de datos primero se diseña y realiza el experimento y finalmente se obtiene el nuevo conocimiento [Vallejos, 2006], en otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Las definiciones dadas al concepto Minería de Datos por numerosos autores [Vallejos, 2006; Zamarrón, 2006; Toledano; Fallad et al, 1996] varían de acuerdo a las opiniones de cada uno de ellos. Teniendo en cuenta las definiciones se puede concluir que la Minería de Datos es un proceso que integra diferentes áreas sirviendo como mecanismo de explotación para identificación de información valiosa, novedosa y útil;

así como para predicción de comportamientos. Por tanto el objetivo fundamental de esta es aprovechar el valor de la información localizada y usar patrones preestablecidos para que se tomen decisiones más confiables. El resultado de la minería será un modelo que se tendrá que evaluar para ver qué tan certero será con respecto a sus predicciones y posteriormente se utilizará para predecir el patrón de comportamiento de cualquier dato nuevo (esto se hace calificando los nuevos datos basándose en el modelo generado) que llegue a la base de datos.

### *1.2.1 Fases en el proceso de Minería de Datos*

Al igual que la extracción de conocimiento la Minería de Datos es un proceso que está compuesta por etapas. Vallejos (2006) indica que los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

El proceso de minería de datos pasa por las siguientes fases:

- **Filtrado de datos:** El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse...) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar algún algoritmo de minería sobre los datos iniciales sin que requieran alguna transformación. En este paso se filtran los datos con el objetivo de eliminar valores incorrectos, no válidos o desconocidos; según las necesidades y el algoritmo a utilizar. Además se obtienen muestras de los datos en busca de mayor velocidad y eficiencia de los algoritmos, o se reducen el número de valores posibles (mediante redondeo, clustering,...) para los atributos de análisis.
- **Selección de Variables:** Después de realizar la limpieza de los datos, en la mayoría de los casos se tiene una gran cantidad de variables o atributos. La selección de características reduce el tamaño de los datos, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería; seleccionando las variables más influyentes en el problema.

Los métodos para la selección de los atributos que más influencia tienen en el problema son básicamente dos: aquellos basados en la elección de los mejores atributos del problema y aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

En la minería de datos casi nunca se menciona el tiempo que se invierte en la limpieza y la verificación de los datos, así como la definición de las variables, pero este proceso es muy importante ya que por lo regular las bases de datos de los sistemas operacionales contienen datos duplicados, a veces erróneos, superfluos o incompletos. A esto se le suman los errores por la operación de los sistemas.

- **Extracción de Conocimiento:** La extracción del conocimiento es la esencia de la minería de datos donde mediante una técnica, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Los modelos que se generan son expresados de diversas formas:

  Reglas, árboles y redes neuronales.

  También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre-procesado diferente de los datos.

- **Interpretación y Evaluación:** Una vez obtenido el modelo, se procede a su validación; donde se comprueba que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos para buscar el que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Una vez validado el modelo, si resulta ser aceptable (proporciona salidas adecuadas y/o con márgenes de error admisibles) este ya está listo para su explotación. Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones, e incluso, en los sistemas

transaccionales.

Las limitaciones de la minería de datos son los primeros datos o datos puros, y no tanto la tecnología o herramientas para el análisis, es decir depende mucho de la limpieza de los datos y de la definición de las variables, si los datos no están correctos el modelo creado no servirá. Del mismo modo la validez de los patrones descubiertos depende de cómo se apliquen al mundo real o a las circunstancias.

### *1.2.2 Ventajas y Dificultades en el uso de Minería de Datos*

¿Por qué usar Minería de Datos?, la minería de datos proporciona múltiples ventajas en su uso, entre las que encontramos:

- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- Contribuye a la toma de decisiones tácticas y estratégicas.
- Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de la mejor forma.
- Genera modelos descriptivos: permite a empresas, explorar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.
- Genera modelos predictivos: permite que relaciones no descubiertas a través del proceso del DM sean expresadas como reglas de negocio.

Aunque las ventajas son varias existen factores que pueden crear un descrédito a esta tecnología. Entre las dificultades que se presentan en la aplicación de la Minería de Datos encontramos [Martínez de Pisón, 2003]:

- Uno de los mayores problemas es que el número de posibles relaciones es demasiado grande, y resulta prácticamente imposible validar cada una de ellas.

Para resolver este problema, se utilizan estrategias de búsqueda, extraídas del área de aprendizaje automático.

- Además todas estas herramientas siguen funcionando mejor fijándoles objetivos de búsqueda concretos. Si bien la Minería de Datos da la impresión de que se puede simplemente aplicar como herramienta a los datos, se debe tener un objetivo, o al menos una idea general de lo que se busca.

- El coste de esta prospección de datos debe ser coherente con el beneficio esperado. Si bien las herramientas (hardware y software) han bajado su precio, el coste en tiempo, personal y consultoría se ha incrementado, llegando en algunos casos a hacer inviable el proyecto.

- Es necesario trabajar en estrecha colaboración con expertos en el negocio para definir los modelos.

### *1.2.3 Modelos de Minería de Datos*

Como se ha visto la minería de datos tiene como función principal encontrar en los datos patrones para extraer información oculta, con el fin de descubrir conocimiento. Para lograr este objetivo se deben concretar relaciones entre los datos que constituyen el modelo, para ello existen formas diferentes de representar dichos modelos y con ello también diferentes técnicas que se utilizan para deducirlos.
De acuerdo a sus características los modelos pueden clasificarse en dos tipos: descriptivos y predictivos.

Los **modelos descriptivos** caracterizan las propiedades generales de los datos, realizan un análisis preliminar de los datos (resumen, características de los datos, casos extremos, etc.). Con esto, el usuario se sensibiliza con los datos y su estructura. Busca derivar descripciones concisas de características de los datos (medias, desviaciones estándares, etc.). Algunas de las tareas que producen estos modelos son el agrupamiento, la asociación y la correlación [Brito, 2008].

Los **modelos predictivos** estiman o predicen valores futuros de la variable objetivo del análisis, también conocida como variable dependiente, partiendo de otros datos que se consideran influyentes en su comportamiento. Dentro de las tareas que producen este tipo de modelo están la clasificación y la regresión. Cada tarea puede ser realizada utilizando distintas técnicas y algoritmos, aunque estos pueden ser empleados para distintos propósitos [Brito, 2008].

### *1.2.4 Tareas de Minería de Datos*

Brito (2008) plantea que el proceso de minería de datos requiere en un principio, establecer objetivos para el análisis de los datos disponibles. De ahí que sean necesarias varios tipos de tareas.
En dependencia del tipo de búsqueda empleado para obtener conocimiento, las tareas se pueden clasificar en directas o indirectas.
Entre las tareas que podemos encontrar en un proyecto de Minería de Datos están:

- La clasificación.
- La regresión (predicción o estimación).
- El agrupamiento (clustering o segmentación).
- La asociación.

La clasificación y la regresión son tareas directas, pues se conoce claramente lo que se busca. El agrupamiento, la asociación y la correlación son indirectas, y se emplean para descubrir patrones que describan los datos sin un objetivo concreto definido.

**Clasificación**

La clasificación es considerada una de las tareas predictivas más importantes dentro de la minería de datos [Brito, 2008]. Se agrupan todos los elementos de una entrada de datos y revisa sus resultados y a partir de estos establece elementos de predicción. Analiza un conjunto de datos de entrenamiento cuya clasificación de clase se conoce y

construye un modelo de objetos para cada clase. Dicho modelo puede representarse con árboles de decisión o con reglas de clasificación, que muestran las características de los datos. El modelo puede ser utilizado para la mayor comprensión de los datos existentes y para la clasificación de los nuevos datos que se agreguen a la base de datos.

**Regresión**

La regresión o estimación como también es conocida predice una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos [Brito, 2008]. Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística) [Varcárcel Asencios, 2004]. Se usan algoritmos como por ejemplo, árboles de regresión, regresión lineal, redes neuronales, kNN, etc. En general, es muy similar a la clasificación, la diferencia consiste en que la clase no es un atributo discreto sino numérico [Brito, 2008]. El objetivo en este caso es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real.

**Agrupamiento**

El agrupamiento (clustering) es la tarea descriptiva por excelencia y consiste en obtener grupos o clusters (donde un cluster es una colección de datos “similares”) a partir de los datos [Brito, 2008]. Los datos son agrupados basados en sus características en el principio de maximizar la similitud entre los elementos de un grupo y minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo [Reyes Saldaña, 2005].

**Asociación**

La asociación es una tarea descriptiva. Se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar

análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las reglas de asociación, como también se les conoce, no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados y la formulación más común es "si el atributo 'X' toma el valor 'a' entonces el atributo 'Y' toma el valor 'b' [Hernández, 2004]. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos.

### *1.2.5 Técnicas y Algoritmos de Minería de Datos*

En la fase de Minería de Datos es donde se aplica las tareas a utilizar en la extracción de conocimiento. Para la realización de estas tareas se han definido varios métodos o algoritmos que permiten el uso de técnicas capaces de conocer las anomalías presentes en los datos que sean procesados.

Las técnicas de Minería de Datos intentan obtener patrones o modelos a partir de los datos recopilados. Estas se clasifican en dos grandes categorías: supervisadas o predictivas (predicen un dato o un conjunto de ellos desconocidos a priori, a partir de otros conocidos) y no supervisadas o descriptivas (se descubren patrones y tendencias en los datos) [Molina, 2006]:


**Fig. 2. Técnicas de Minería de Datos**

Una técnica constituye el enfoque conceptual para extraer la información de los datos, y, en general es implementada por varios algoritmos. Cada algoritmo representa, en la práctica, la manera de desarrollar una determinada técnica paso a paso, de forma que es preciso un entendimiento de alto nivel de los algoritmos para saber cual es la técnica más apropiada para cada problema. Asimismo es preciso entender los parámetros y las características de los algoritmos para preparar los datos a analizar. De la correcta elección del algoritmo de minería dependerá la forma de representación del conocimiento obtenido, bien en forma de relaciones, patrones y reglas o en forma de una descripción más concisa [Molina, 2006].

Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de

entidad mientras que una descripción puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones. De esta forma, hay algoritmos o técnicas que pueden servir para distintos propósitos, por lo que la figura anterior únicamente representa para qué propósito son más utilizadas las técnicas. Por ejemplo, las redes de neuronas pueden servir para predicción, clasificación e incluso para aprendizaje no supervisado [Molina, 2006].

Entre las técnicas más representativas y los algoritmos más utilizados para cada una de ellas encontramos:

**Regresión:** La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión. Se pueden resolver muchos problemas por medio de la regresión lineal, y puede conseguirse todavía más aplicando las transformaciones a las variables para que un problema no lineal pueda convertirse a uno lineal [Molina, 2006]. Es la más utilizada para formar relaciones entre datos, es rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

Algunos algoritmos que lo implementan en WEKA son: Algoritmo de Regresión lineal.

✓ **Algoritmo Regresión Lineal** [Molina, 2006]:

La clase en la que se implementa el algoritmo regresión lineal múltiple en la herramienta WEKA es weka.classifers.LinearRegression.java.

Algunas propiedades de la implementación son:

- Admite atributos numéricos y nominales. Los nominales con k valores se convierten en k-1 atributos binarios.
- La clase debe ser numérica.
- Se permite pesar cada ejemplo.

**Árboles de Decisión:** Un árbol de decisión [Molina, 2006] puede interpretarse esencialmente como una serie de reglas compactadas para su representación en forma de árbol. Es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. De manera general, los árboles de decisión pueden clasificarse en dependencia del tipo de variables que predicen. Si el árbol de decisión es usado para predecir variables nominales, recibe el nombre de árbol de clasificación y si su uso está determinado en la predicción de variables numéricas, se denomina árbol de regresión o predicción. Entre las ventajas más importantes de los árboles de decisión es que se construyen rápidos y son fáciles de interpretar. Cada camino de la raíz a las hojas forma una regla. La predicción basada en árboles de decisión es eficiente. Cuando un caso cae en un nodo hoja del árbol, el valor a predecir en el caso se basa en las estadísticas del nodo. Como los datos que se emplearán para la búsqueda de conocimiento en la investigación son numéricos el árbol a utilizar es el de regresión o predicción. El procedimiento para generar un árbol de decisión consiste en seleccionar un atributo como raíz del árbol y crear una rama con cada uno de los posibles valores de dicho atributo. Con cada rama resultante (nuevo nodo del árbol), se realiza el mismo proceso, esto es, se selecciona otro atributo y se genera una nueva rama para cada posible valor del atributo. Este procedimiento continúa hasta que los ejemplos se clasifiquen a través de uno de los caminos del árbol. El nodo final de cada camino será un nodo hoja, al que se le asignará la clase correspondiente. Así, el objetivo de los árboles de decisión es obtener reglas o relaciones que permitan clasificar a partir de los atributos.

Algunos algoritmos que lo implementan en WEKA son: Algoritmo ID3 y Algoritmo C4.5.

- ✓ **Algoritmo ID3** [Molina, 2006]:

El sistema ID3 es un algoritmo simple y, sin embargo, potente, cuya misión es la elaboración de un árbol de decisión. La clase en la que está codificado el algoritmo ID3

es *weka.classifiers.ID3.java*.

En cuanto a la implementación, no permite ningún tipo de configuración. Las ideas básicas son:

• En el árbol de decisión cada nodo corresponde con un atributo no-categórico y un arco con el valor posible de ese atributo. Una hoja del árbol especifica el valor esperado del atributo para los valores descritos en el camino desde el nodo inicial hasta la hoja.

• En el árbol de decisión cada nodo debe estar asociado al atributo no-categórico que da más información acerca de los atributos que aún no hayan sido considerados.

• Se utiliza la entropía para medir la información que tiene un nodo.

**Inducción de Reglas:** Las técnicas de Inducción de Reglas permiten la generación y contraste de árboles de decisión, o reglas y patrones a partir de los datos de entrada. La información de entrada será un conjunto de casos donde se ha asociado una clasificación o evaluación a un conjunto de variables o atributos. Con esa información estas técnicas obtienen el árbol de decisión o conjunto de reglas que soportan la evaluación o clasificación. En los casos en que la información de entrada posee algún tipo de “ruido" o defecto (insuficientes atributos o datos, atributos irrelevantes o errores u omisiones en los datos) estas técnicas pueden habilitar métodos estadísticos de tipo probabilístico para generar árboles de decisión recortados o podados. También en estos casos pueden identificar los atributos irrelevantes, la falta de atributos discriminantes o detectar "gaps" o huecos de conocimiento. Esta técnica suele llevar asociada una alta interacción con el analista de forma que éste pueda intervenir en cada paso de la construcción de las reglas, bien para aceptarlas o bien para modificarlas [Molina, 2006]. La inducción de reglas se puede lograr fundamentalmente mediante dos caminos: Generando un árbol de decisión y extrayendo de él las reglas.

Algunos algoritmos que lo implementan en WEKA son: Algoritmo 1R, Algoritmo PRISM, Algoritmo PART

✓ **Algoritmo 1R** [Molina, 2006]:

Este algoritmo genera un árbol de decisión de un nivel expresado mediante reglas. Consiste en seleccionar un atributo (nodo raíz) del cual nace una rama por cada valor, que va a parar a un nodo hoja con la clase más probable de los ejemplos de entrenamiento que se clasifican a través suyo. La clase debe ser simbólica, mientras los atributos pueden ser simbólicos o numéricos. También admite valores desconocidos, que se toman como otro valor más del atributo. La clase *weka.classifers.OneR.java* implementa el algoritmo 1R.

### *1.2.6 Antecedentes y Áreas de aplicación de la minería de datos*

La minería de datos surge a principios de los 80's cuando la Administración de Hacienda Estadounidense desarrolló un programa de investigación para detectar fraudes en la declaración y evasión de impuestos, mediante lógica difusa, redes neuronales y técnicas de reconocimiento de patrones. Sin embargo, su expansión se produce hasta los 90's originada principalmente por tres factores:

-Incremento en la potencia de procesamiento de las computadoras, así como en la capacidad de almacenamiento.

-El crecimiento de la cantidad de datos almacenados se ve favorecido no sólo por el abaratamiento de los discos y sistemas de almacenamiento masivo, sino también por la automatización de muchos trabajos y técnicas de recogida de datos (observación con nuevas tecnologías, entrevistas más prácticas, encuestas por Internet, etcétera)

-Aparición de nuevos métodos de técnicas de aprendizaje y almacenamiento de datos, como las redes neuronales, la Inteligencia artificial, el surgimiento del almacén de datos (Data Ware House) [Artículos estadísticos].

La minería de datos surge por la necesidad de obtener estrategias de negocio, conocer

a los clientes, obtener información de productos, interpretar información valiosa para la toma de decisiones, etcétera.

Desde los años sesenta los estadísticos manejaban términos como data fishing, minería de datos o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos.

A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de minería de datos y descubrimiento de conocimiento en base de datos.

La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales:

- Colección de datos (1960).
- Acceso a Datos (1980).
- Almacén de Datos y Apoyo a las Decisiones (principio de la década de 1990).
- Minería de Datos inteligente (finales de la década de 1990).

A finales de los años ochenta sólo existían un par de empresas dedicadas a ésta tecnología; en 2002 existían más de 100 empresas en el mundo que ofrecían alrededor de 300 soluciones, ahora se ven áreas dedicadas a la minería de datos dentro de cada empresa, ya que, es una herramienta ideal para obtener información valiosa e importante de manera rápida y eficaz, a través de procesos especializados y sistemáticos.

Existen numerosos dominios donde la minería de datos puede ser aplicada, en principio en todas las áreas o actividades que generen datos. Desde la década de los años 90, del pasado siglo, se vienen aplicado intensamente técnicas de minería de datos con diversos fines: apoyo a la toma de decisiones, gestión de procesos industriales,

investigación científica, soporte al diseño de bases de datos y mejora de la calidad de los datos, entre otros. Como ven la minería de datos ofrece un gran valor en un amplio espectro de industrias:

**En Internet:**

E-bussines: Perfiles de clientes, publicidad dirigida, fraude.

Buscadores Inteligentes: Generación de jerarquías, bases de conocimiento web.

Gestión del Tráfico de la Red: Control de eficiencia y errores.

**El Mundo de los Negocios:**

Banca: Grupos de clientes, préstamos, oferta de productos.

Compañías de Seguros: Detección de fraude, administración de recursos.

Marketing: Publicidad dirigida, estudios de competencia.

**En Mundo de la Ciencias:**

Meteorología: Teleconexiones (asociaciones espaciales), predicción.

Física: Altas energías, datos de colisiones de partículas (búsqueda de patrones).

Bio-Informática: Búsqueda de patrones en ADN, proyectos científicos como genoma humano, datos geofísicos, altas energías, etc.

Medicina: para predecir la eficacia de los procedimientos quirúrgicos, exámenes médicos o medicamentos.

La universidad cubana ha sufrido notables transformaciones en los últimos años, haciéndose necesario una mayor utilización de la información con que se cuenta. La aplicación de la minería de datos no ha estado exenta de los centros universitarios del país. El ISMM cuenta con una serie de trabajos de este tipo, aunque no precisamente en el ámbito de gestión económica. Estos trabajos han girado, en torno al ámbito educacional: en la Gestión Docente del ISMM [López, 2010], en el ámbito de la generación de energía distribuida: en la obtención de mejoras en la explotación de los Grupos Electrógenos Diesel [Bisset] y en los estudios de riesgo por contaminación atmosférica: en la estimación de parámetros meteorológicos [Batista].

## 1.3 Metodologías

El desarrollo de un proyecto de KDD provoca usualmente desviaciones y retrasos en su planificación, lo que se debe de forma general a que no es posible extraer conclusiones por adelantado, unido al hecho de que una gran parte del esfuerzo se produce en la preparación de los datos [Gondar].

Ante la necesidad existente en el mercado de una aproximación sistemática para la realización de los proyectos de minería de datos, diversas empresas y consultorías han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión de pasos que le dirijan a obtener buenos resultados. La selección de una metodología puede ser muy diversa, encontrándose algunos que no se rigen por ninguna, otros que adoptan una propia, y finalmente otro grupo que se rige por alguno de los estándares más difundidos. Dentro de las principales metodologías utilizadas por los analistas en los proyectos de minería de datos se tiene la metodología SEMMA (Sample, Explore, Modify, Model, Assess ) propuesta por SAS y la metodología de libre distribución CRISP-DM (Cross-Industry Standard Process for Data Mining), estas se encuentran actualmente entre los estándares metodológicos más difundidos para desarrollar un proceso de KDD.

Existen otras metodologías menos usuales como las metodologías CRITIKAL y Metodología de las 5 A´s.

### *1.3.1 CRISP-DM*

El modelo de proceso CRISP-DM, (CRoss-Industry Standard Process for Data Mining) Proceso Construcción Cruzada Estándar Industrial para Minería de Datos, que incluye seis pasos se propuso por primera vez a principios de 1996. Lo que trata ésta metodología es desarrollar los proyectos de minería de datos bajo un proceso estandarizado de definición y validación de tal forma que se desarrollen proyectos minimizando los costos que impliquen y con un alto impacto en el negocio.

La metodología CRISP-DM proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de minería de datos: el modelo de referencia y la guía del usuario.

El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de minería en general.

La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de minería de datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase.

CRISP-DM está definida como un proceso jerárquico, que consiste en un conjunto de tareas descritas en cuatro niveles de abstracción, desde el general hasta el específico: fase, tareas generales, tareas específicas e instancias de proceso.

En la actualidad la metodología CRISP-MD (CRoss-Industry Standard Process for Data Mining) es una de las más usadas por su generalización y su practicidad, además de su libre utilización.

### *1.3.2 SEMMA*

SAS Institute desarrollador de ésta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.

El nombre de esta terminología es el acrónimo correspondiente a los cinco pasos básicos del proceso: Sample (Muestreo), Explore (Exploración), Modif. (Manipulación), Model (Modelado) and Assess (Valoración).

La metodología consiste en los siguientes pasos: tomar los datos o una muestra en caso de que la cantidad de datos sea muy grande, se exploran, modifican, modelan y se evalúan en el modelo o los modelos resultantes para elegir el más adecuado.

La ejecución de sus fases no está descrita de forma rígida, por lo que no es necesario terminar una antes de comenzar otra, conservando así, la iterabilidad y ciclicidad del proceso. Es necesario señalar que no toma en cuenta los objetivos del negocio al que se esté aplicando la minería de datos, ni el despliegue o explotación de los modelos resultantes, por tanto, en este sentido limita la visión del proceso de KDD [Gondar].

### *1.3.3 Otras Metodologías*

Existen también otras metodologías que nos ayudan a guiar el proceso de KDD tal es el caso de:

**CRITIKAL**

La metodología CRITYKAL (Client-Server Rule Induction Technology for Industrial Knowledge Acquisition from Large Date Base), fue creada y adoptada por un grupo de empresas y universidades europeas y no es de completa distribución libre. Esta metodología considera fases muy similares a las de CRISP-DM y se caracteriza por una fuerte integración con las técnicas de los almacenes de datos.

**Metodología de las 5 A´s**

La metodología de las 5 A's, definida también por SPSS, toma su nombre de los cinco términos siguientes Assess, Access, Analyze, Act y Automate (Asesorar, Acceder, Analizar, Actuar, Automatizar). Se considera en alguna medida padre de CRISP-DM pues de manera global abarca los mismos aspectos radicando la diferencia en que en la sucesora se encuentran mucho más detallados y formalizados [Martínez, 2003].

### *1.3.4 Metodología seleccionada*

Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de minería de datos en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso en iterativo e interactivo.

La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de minería de datos donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema. Entonces la metodología CRISP-DM está más cercana al concepto real de proyecto, integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas [Rodríguez y otros].

Otra diferencia significativa entre la metodología SEMMA y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo de minería de datos siendo su distribución libre y gratuita [Rodríguez y otros].

Partiendo de las razones antes expuestas y las características de cada metodología se decidió el empleo de la metodología CRISP-DM para el desarrollo de esta investigación teniendo a su favor el que además de estructurar un proyecto de KDD en fases que se encuentran interrelacionadas entre sí, y lo describen de forma iterativa e interactiva, trata el proyecto de forma global y estrechamente relacionado al negocio en cuestión tomando en cuenta los aspectos empresariales de este permitiendo el contacto con los expertos en el contenido a trabajar, los cuales brindan un conocimiento necesario para

la comprensión de los datos, la validación de los resultados obtenidos y la explotación de los modelos resultantes. Está diseñada de forma neutral a las herramientas que se utilicen para el desarrollo del proyecto, y su distribución es libre, encontrándose en continuo perfeccionamiento.

Presenta una precisa y sólida distribución de tareas de carácter general con sus resultados, así como una guía para su desarrollo agilizando el trabajo ya que el tiempo con el que se cuenta para el desarrollo de la investigación es escaso. La metodología deja abierta la posibilidad de seguir desarrollando la investigación a partir de descubrir otros conocimientos, otras relaciones entre variables o utilizando el modelo que se obtiene.

Tiene a su favor además, un preciso y sólido repertorio de tareas de propósitos generales, por lo que goza de una importante popularidad, siendo, por tanto, frecuentemente empleada.

### 1.4 Herramientas

En la actualidad existe un número amplio de herramientas de apoyo al análisis de datos durante un proceso de KDD. Generalmente, estas herramientas disponen de sus propios entornos gráficos y suelen permitir al usuario hacer múltiples tareas, pero siempre limitados a las especificaciones de cada aplicación. El grado de eficiencia de cada herramienta depende de múltiples factores, entre los que podemos mencionar: tipos de algoritmos, funciones de tratamiento de la información, eficiencia de los algoritmos, etc.

Son muchas las herramientas que sobresalen por su popularidad entre los desarrolladores de proyectos de MD, a continuación breves características de algunas de estas herramientas:

### *1.4.1 Rapid Miner*

RAPID MINER, anteriormente YALE (Yet Another Learning Envoronment) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usan en investigación y en aplicaciones empresariales.

Está desarrollada sobre el lenguaje java y funciona en los sistemas operativos más conocidos, constituyendo un software de código abierto y de libre distribución, además, se retroalimenta de las librerías de funciones de WEKA en su entorno de aprendizaje.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización. También permite utilizar los algoritmos incluidos en WEKA.

### *1.4.2 Excel*

Entre los productos que Microsoft comercializa, están los mundialmente conocidos: procesador de texto (Word), el programa de base de datos (Access), el programa para hacer presentaciones (PowerPoint) y la hoja de cálculo (Excel). Estos pueden adquirirse por separado o como parte de Office, un paquete integrado de programas informáticos.

Desde que en 1979 apareció la primera aplicación de este tipo, las hojas de cálculo se convirtieron en los programas estrella del ordenador personal; a esta aplicación siguieron otras que también alcanzaron gran popularidad, como SuperCalc, Multiplan, Lotus 1-2-3 y Excel. Una sola hoja de cálculo puede contener miles o millones de celdas, y en ellas, cientos y cientos de datos. Excel permite vincular una hoja de cálculo a otra que contenga información relacionada y puede actualizar de forma automática los datos de las hojas vinculadas. Además, se puede utilizar para crear y ordenar bases de datos. Cuenta con capacidades gráficas donde se obtienen representaciones de los

datos en forma de gráficos de líneas, barras, pastel y otros, los cuales facilitan su lectura e interpretación. Proporciona un buen número de opciones de formato tanto para las páginas y el texto impreso como para los valores numéricos y las leyendas de los gráficos. Es por demás un excelente programa para el análisis estadístico de datos dando la posibilidad de obtener cientos de respuestas del análisis de un mismo dato [Vila, 2005].

### *1.4.3 WEKA*

WEKA (Waikato Environment for Knowledge Analysis) es una colección de algoritmos de aprendizaje automático para tareas de minería de datos, y un software de libre distribución. Fue desarrollado por La Universidad de Waikato (Nueva Zelanda) implementado en Java; útil para ser aplicado sobre datos mediante las interfaces que ofrece o para embeberlos dentro de cualquier aplicación.

Entre sus principales características se encuentra el poseer una interfaz gráfica de usuario compuesta de cuatro entornos que permiten diferentes funcionalidades y formas de análisis. Una de las ventajas fundamentales de esta herramienta es que su desarrollo sobre el lenguaje java la hace multiplataforma. Además, el hecho de ser de código abierto unido a su prestigio, hace que se encuentre en constante evolución por parte de la comunidad internacional.

Debemos comentar también la gran diversidad de algoritmos incluidos en WEKA que se pueden utilizar según queramos obtener unos u otros objetivos. Los algoritmos se aplican directamente en la data set o llamado desde el propio código java. Permite la experimentación de análisis de información, formado por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, asociación y visualización de datos.

### *1.4.4 Herramienta seleccionada*

Para el desarrollo de esta investigación se decidió trabajar con una herramienta de código abierto, libre distribución y que no comprometan su uso con una metodología en particular, las que más se asemejan por sus características son Rapid Miner y WEKA.

La diferencia entre ambas es que WEKA exhibe altas prestaciones en lo referente al pre-procesado de los datos y a la modelación de los mismos, tiene asociado un número considerablemente superior de proyectos en comparación con Rapid Miner y presenta una mayor documentación, siendo así que este último emplea librerías de funciones propias de WEKA en su entorno de aprendizaje.

Se opta por la herramienta de análisis WEKA no solo por todos los beneficios antes expuestos sino por las ventajas y facilidades que por sus características brinda: este programa es de libre distribución y difusión, además, ya que está programado en Java, es muy portable siendo independiente de la arquitectura, funcionando en cualquier plataforma sobre la que haya una máquina virtual Java disponible; contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado y además es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

## Conclusiones del Capítulo

Como se ha visto el análisis de datos es una tarea que consiste en buscar o encontrar tendencias o variaciones de comportamiento en los mismos, de tal manera que esta información resulte de utilidad para los usuarios finales. A estas tendencias o variaciones se le conocen como patrón, los cuales si son de importancia y útiles para el dominio en cuestión se le denomina conocimiento.

Existe en la actualidad un conjunto de herramientas y técnicas que soportan la extracción de conocimiento útil a partir de los datos disponibles, y que se agrupan bajo el calificativo de “minería de datos”, como resultado no obtiene datos sino conocimiento. Esta técnica resulta muy útil en situaciones donde el volumen de datos es muy grande o complejo por la cantidad de variables que se manipulan, o donde los especialistas no están disponibles para el análisis de los datos y la extracción de conocimiento. Tal caso es el del ISMM, de ahí la importancia del empleo de esta técnica donde se pretende descubrir conocimiento de los datos presupuestarios cuyo uso ayude a la toma de decisiones más seguras que reporten algún tipo de beneficio a la organización.

El desarrollo de este capítulo ha sentado las bases para la solución del problema a resolver siguiendo la metodología y herramientas seleccionadas.

# 2

# Capítulo 2: Minería de Datos aplicada a la actividad presupuestada del ISMM Dr. Antonio Núñez Jiménez

## Introducción

El presente capítulo está dedicado al desarrollo de un proceso de KDD para la extracción de conocimiento de los datos recogidos del departamento de economía del ISMM. Utilizando la metodología CRISP-DM y la herramienta WEKA, y empleando técnicas de clasificación, asociación y agrupamiento se pretende encontrar reglas que describan las relaciones entre las variables almacenadas, de tal modo que podamos predecir el comportamiento de las mismas que ayuden en la elaboración de un plan que se ajuste más a las necesidades del ISMM.

### 2.1 Comprensión del negocio: análisis del problema

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos [CHAPMAN et al, 2000].

#### *2.1.1 Objetivos del negocio*

1. Conocer las relaciones que se establecen entre las variables: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital empleando los datos de los tres años registrados.

2. Determinar cómo influye el comportamiento de los epígrafes y partidas en el resultado final de las variables.

### *2.1.2 Criterios de éxito del negocio*

El cumplimiento de los objetivos del negocio será reflejado en:

- Que más del 80 % del coeficiente de correlación de los resultados obtenidos esté por encima o igual al 0.80.
- Obtener modelos que brinden una comprensión profunda y provechosa del comportamiento de las variables: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital.
- El aumento de los elementos que apoyan la toma de decisiones de los directivos.
- El informe sobre los problemas encontrados en los datos y las recomendaciones para un mejor proceso.

### *2.1.3 Recursos Disponibles*

**Software:** Se cuenta con el Paquete Office completo para el procesado de Textos y la herramienta de minería de datos WEKA versión 3.7.1.
**Hardware:** Se dispone de una PC de escritorio con sistema operativo Windows XP Profesional y 256 MB de memoria RAM.
**Datos:** La información fue recopilada de los registros almacenados en formato Excel en el Departamento de Economía.

### *2.1.4 Requerimientos, supuestos y restricciones*

Se lista los requerimientos, supuestos y restricciones relativos tanto a la planificación del proyecto como a los datos y recursos disponibles.

- Se necesita un gran volumen de información con un alto porcentaje de fiabilidad y libre acceso a ella.
- Lograr una adecuada calidad de los datos.
- Todo el proceso será debidamente documentado.
- Se consultarán con el experto los resultados obtenidos.

### *2.1.5 Riesgos y contingencias*

Se listan las circunstancias que pueden retrasar o impedir la realización del proyecto:

- Las herramientas de software y hardware con las que se cuenta deben ser capaces de tratar gran cantidad de información con tiempos reducidos de procesamiento.
- Se necesita una comunicación fluida y flexible con el personal involucrado en el proyecto.
- Los datos deben ser fiables y de buena calidad para obtener resultados satisfactorios.

### *2.1.6 Beneficios*

Mayor conocimiento y dominio del comportamiento del presupuesto asignado, el cual permitirá a la administración, si lo considera necesario, replantearse objetivos y metas. En otro sentido, aporta experiencias novedosas en el campo de descubrir conocimientos en bases de datos, siguiendo la metodología CRISP-DM y empleando la herramienta de análisis WEKA.

### *2.1.7 Objetivos de la Minería de Datos*

Los objetivos de la minería de datos van a estar dados por:

- Con la utilización de técnicas estadísticas y el empleo de la herramienta WEKA

obtener una descripción inicial de los datos para una mejor comprensión de los mismos y determinar las variables más significativas.

- Realizar el preprocesado de los datos eliminando o rellenando los datos en vacíos, eliminando los ruidos o la información errónea, construir y/o transformar atributos.
- Obtener modelos que describan el comportamiento de las variables a analizar: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital entrenando y probando el modelo con los datos.

### *2.1.8 Criterios de éxito de la Minería de Datos*

Los criterios de éxito de la minería de datos serán:

- Para determinar las variables más significativas se tiene en cuenta la opinión de los expertos según su experiencia, y el por ciento de influencia que presenten para determinar o predecir un atributo.
- Para la extracción de los patrones de comportamiento se toma en consideración la valoración de los expertos, y el uso que aporta el conocimiento descubierto.
- En la construcción de los modelos se tiene en cuenta su precisión; en este sentido se toman para su entrenamiento y prueba, conjuntos distintos de datos, a fin de no sobrestimarlos.

### *2.1.9 Elaborar el plan del proyecto*

Dentro de la planificación del proyecto se plantearon las siguientes tareas:

1. Obtención de los datos de la fuente de información.

- Entrevista con los expertos.
- Recogida de los datos.
- Estudio de la documentación disponible.

2. Estudio descriptivo para la comprensión de los datos.

- Empleo de técnicas estadísticas: descriptores estadísticos, gráficos.

3. Estudio exploratorio de los datos: filtrado, limpieza y transformación.

- Selección de las variables más importantes.
- Construcción de los datos para llevarlos al formato .arff.
- Filtrado horizontal y vertical de los datos.
- Manejo de la herramienta WEKA.

4. Extracción de patrones de comportamiento.

- Empleo de técnicas de clasificación.
- Manejo de la herramienta WEKA.

5. Interpretación de los patrones obtenidos.

- Consulta con los expertos.

## 2.2 Comprensión de los Datos

La comprensión de datos está relacionada con la recolección y descripción de la información inicial con la que se comienza el proceso de KDD, una vez establecidos los objetivos a seguir. Además, se desarrollan actividades que permiten su exploración, a fin de identificar problemas con su calidad.

### *2.2.1 Recopilación y Descripción de los datos*

La información recopilada fue obtenida de una única fuente, que de manera centralizada es la empleada en el Instituto para registrar todos los datos relacionados con el presupuesto. Los datos recopilados reflejan una serie de valores de las variables presupuestadas del Instituto.

Los datos contenidos en las hojas de cálculo de Microsoft Excel 2003 contienen

históricos desde 2004 hasta el 2010. La información más completa y útil para la investigación contiene los años desde el 2008 hasta el 2010, de estos se tienen registros de las variables (Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital).

### *2.2.2 Exploración de los datos utilizando WEKA*

Para realizar un análisis exploratorio inicial de los datos se utiliza la herramienta WEKA con el fin de utilizar los histogramas y las descripciones de los atributos que nos brinda. Para ello fue necesario emplear una de las tareas de la preparación de los datos "Construcción de los datos".

**Construcción de datos**

Para la tarea siguiente de limpieza de los datos fue necesario llevar los datos al formato .arff que es el que acepta Weka pues están almacenados en Microsoft Excel. Para ello fue necesario exportar un archivo de texto desde el Excel de tipo CSV (delimitado por coma) obteniéndose un archivo de forma:

Fecha (representando el plan, real y porciento de ese año), Variables (Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital). Los datos fueron declarados numéricos y los valores desconocidos se sustituyeron por un signo de interrogación (?) que es como lo reconoce WEKA ya que al cargar los datos en WEKA dio un error donde existían espacios en blanco.

**Exploración y descripción de los datos**

Al cargar los datos en WEKA se obtuvo la siguiente información de cada variable:

**Total de Ingresos, Total de Gastos y Total de Gastos Corrientes:** de cada una de estas variables existen 6 instancias con valores entre 162.2 y 22393114.89, 98.16 y

16156596.95, 94.56 y 16627600 respectivamente.

**Gastos de Personal:** De esta variable existen 9 instancias con valores entre 86.67 y 9466200.

**Gastos de Bienes y Servicios:** Existen 9 instancias con valores entre 97.69 y 2790000.

**Otras Transferencias Corrientes:** De las 9 instancias de este atributo 7 se encuentran entre los valores mínimos (96.33 y 2586050.57) y 2 entre los valores máximos (2586050.57 y 5172004.81).

**Gastos de Capital:** La variable presenta 9 instancias registradas con dos grupos de valores. Estos grupos varían uno con 6 instancias entre 67.03 y 471079.2 y el otro con 3 instancias entre 471079.2 y 942091.37.

### *2.2.3 Verificación de la calidad de los datos*

Debido a los diversos factores que afectan a la información recogida del mundo real, como el ruido, redundancia, valores ausentes y otros factores que pueden influir en la calidad de los datos es que se realiza la tarea “Verificación de la calidad de los datos” la que tiene como objetivo fundamental identificar los problemas en la calidad de los datos.

Los datos almacenados datan del 2004 en adelante. Durante los años 2004 a 2007 la información recopilada contiene datos faltantes que influyen en el resultado final de las variables a analizar. También se tiene que algunos de estos años están planificados de forma mensual y otros de una forma más general (anual) teniéndose incoherencias en los datos. Lo anteriormente planteado puede ser debido a la forma manual en la que se trabaja y a lo histórico de los datos, no teniéndose en ocasiones los cambios en las planificaciones de estos años, por tanto la actualización de estos datos no se tiene y tienden a ser incorrectos.

En los restantes años no se encontraron campos con valores atípicos ni datos redundantes, aunque si atributos que coincidían en estar vacíos todos los años y que por tanto no influían en el resultado final de las variables a analizar.

## 2.3 Preparación de los datos

La preparación de datos está relacionada con las actividades necesarias para conformar el conjunto de datos final, que será utilizado por la herramienta de modelado WEKA, a partir de la información inicial.

### *2.3.1 Selección de los datos*

En la tarea “Selección de los datos” se parte de los obtenidos en la fase anterior por lo que se decide excluir de los atributos iniciales a:
Medicinas y Materiales Afines, Otros Gastos, Donaciones Corrientes, De la Asistencia Social, Otras Transferencias Corrientes, Compra de Activos Fijos: debido a que no proporciona información pues no se almacenaron datos de estas variables.
Se decide eliminar los registros del 2004 al 2007 y trabajar con los del 2008 al 2010, dejando los del 2011 para ser utilizados en realizar las pruebas.

### *2.3.2 Limpieza y transformación de los datos*

A los datos seleccionados no fue necesario aplicarles ningún filtro debido a que todos los datos con los que se disponen son necesarios y guardan relación entre ellos.

## 2.4 Modelado

En esta fase, se seleccionan las técnicas y algoritmos de modelado que serán aplicados a los datos, y sus parámetros son calibrados. Los modelos encontrados son descritos brevemente, antes de ser evaluados completamente [CHAPMAN et al., 2000].

### *2.4.1 Selección de las técnicas de modelación*

Las técnicas de modelado seleccionadas están enfocadas a los objetivos del negocio definidos. La tabla 1 resume las tareas y técnicas planificadas, los objetivos que

persiguen y los algoritmos de WEKA que se utilizarán para desarrollarlas.

**Tabla 1. Tareas, técnicas y algoritmos seleccionados.**

| Tarea | Técnica | Algoritmo | Objetivos |
|---|---|---|---|
| Predicción | Regresión | Regresión Lineal | Obtener un modelo de regresión lineal para predecir el valor de: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital. |
| | Árboles de Predicción | M5P | Obtener un árbol y a partir de este reglas con el fin de predecir el valor de: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital. |
| Clasificación | Inducción de Reglas | M5Rules | Obtener reglas para predecir el valor de: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital. |

### *2.4.2 Generar el diseño del experimento*

El diseño de los experimentos consiste en esbozar como se construirán los modelos y como se validarán, empleando para el entrenamiento y la prueba conjuntos distintos de datos.

En WEKA las tareas de Predicción y Clasificación se encuentran dentro de la ventana Classify. Una vez elegido el clasificador y sus características el próximo paso es la configuración del modo de entrenamiento (Test Options).

Para entrenamiento y prueba de los modelos se emplean conjuntos distintos con el fin

de no sobreestimar su precisión. Para lo cual se utilizará entre los modos de evaluación del clasificador la opción con la que Weka entrenará el método con todos los datos disponibles y luego lo aplicará otra vez sobre los mismos, teniendo como nombre esta opción de Use training set.

### *2.4.3 Escenario de parámetros y modelos*

Dentro de la tarea "Construcción y descripción de los modelos" se describen los parámetros de los algoritmos empleados.

Para cada algoritmo de las tareas seleccionadas anteriormente mencionadas (ver Tabla 1) se emplean los parámetros descritos en las tablas 2, 3, 4.

**Tabla 2. Regresión Lineal**

| Parámetros | Valor | Descripción |
|---|---|---|
| AttributeSeleccionMethod | M5 Method | Método de selección del atributo a eliminar de la regresión. |
| debug | False | Muestra el proceso de construcción del clasificador. |
| eliminateColinearAttributes | True | |
| ridge | 1.0E-8 | |

**Tabla 3. M5P**

| Parámetros | Valor | Descripción |
|---|---|---|
| buildRegressionTree | True | Permite construir un árbol de regresión |
| Debug | False | Muestra el proceso de construcción del clasificador |
| minNumInstances | 4.0 | Número mínimo de instancias por hojas |
| unpruned | False | Si se activa no se realiza la poda del árbol |
| saveInstances | False | Una vez finalizada la creación del árbol de decisión se eliminan todas las instancias que se clasifican en cada nodo, que hasta el momento se mantenían almacenadas. |
| useUnsmoothed | False | Indica si se realizará el proceso suavizado |

**Tabla 4. M5Rules**

| Parámetros | Valor | Descripción |
|---|---|---|
| buildRegressionTree | True | Permite construir un árbol de regresión. |
| Debug | False | Muestra el proceso de construcción del clasificador. |
| minNumInstances | 4.0 | Número mínimo de instancias por hojas |
| unpruned | False | Si se activa no se realiza la poda del árbol |
| useUnsmoothed | False | Indica si se realizará el proceso de suavizado. |

**Modelos generados por cada tarea:**

**Tarea 1:** Regresión Lineal. (Total de experimentos: 11)

Árboles de Predicción. (Total de experimentos: 11)

**Tarea 2:** Inducción de Reglas. (Total de experimentos: 11)

Con el fin de cumplir con las tareas planteadas se realizaron un total de 33 experimentos.

### *2.4.4 Descripción de los modelos*

Como parte de la tarea “Construcción y descripción de los modelos” se hace necesario describir los modelos obtenidos para una mejor comprensión de los mismos.

Luego de realizados los experimentos se decide escoger los modelos generados por los árboles de predicción, debido a que con la regresión lineal no se cumplía con que más del 80 % del coeficiente de correlación de los resultados obtenidos esté por encima o igual al 0.80. Con las reglas, con las cuales se logran resultados muy parecidos a los árboles, este criterio de éxito si se cumplía pero requieren de un mayor costo de tiempo para obtenerlas. Por las razones antes expuestas se decide utilizar los modelos obtenidos por los árboles de predicción que además se construyen rápidos, son fáciles de interpretar y proporcionan reglas que podrían ser empleadas en el desarrollo de un Prototipo de Sistema Experto.

**Tarea: Predicción con Árboles.**

**Objetivo:** Obtener un árbol y a partir de este reglas con el fin de predecir el valor de las variables: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital. De forma general el 100 % de los modelos construidos se encuentran entre el 0,7 y 0,9 del coeficiente de correlación.

A continuación se describen los árboles de predicción obtenidos. Además se mencionan las reglas de los modelos lineales obtenidos que clasificaron una mayor cantidad de instancias correctamente. Cada uno de estos modelos tiene como objetivo estimar el valor de las variables cuyos resultados se almacenan en los nodos hojas de los árboles (Anexos 1A – 1F).

**Total de Ingresos:**

En el árbol de predicción obtenido para esta variable (Anexo 1A) se muestra un modelo lineal que presenta un coeficiente de correlación igual 0. Pero al analizarlo con el conjunto de datos completo obtenemos dos modelos lineales. Este modelo presenta un coeficiente de correlación igual a 0.7709 e incluye un total de 6 instancias logrando clasificarlas en los siguientes modelos lineales (LM):

LM1: este modelo clasificó 4 instancias y en él se cumple que si la partida Otros no Especificados Previamente es menor o igual que 96055.18 entonces el valor estimado del Total de Ingresos sería:

Total de Ingresos = 110.7922 * Otros no Esp. Previamente - 981179.7847

LM2: modelo que clasificó 2 instancias y en el que se cumple que si la partida Otros no Especificados Previamente es mayor que 96055.18 entonces el valor estimado del Total de Ingresos sería:

Total de Ingresos = 123.8266 * Otros no Esp. Previamente + 1398515.4459

**Total de Gastos:**

Para el Total de Gastos se obtuvo un árbol de predicción (Anexo 1B) en el cual el coeficiente de correlación del modelo obtenido es de 0.9952. Se obtuvieron 2 modelos lineales y se puede observar que la variable influyente es el Total de Ingresos. El modelo contiene 6 instancias las que se clasifican en los siguientes modelos lineales (LM):

LM1: clasificó 2 instancias, en este se cumple que: si el Total de Ingresos es menor o igual que 62589.57 entonces el valor estimado del Total de Gastos sería:

Total de Gastos = 0.8779 * Total de Gastos Corrientes + 106368.1161

LM2: clasificó 4 instancias, en el mismo se cumple que: si el Total de Ingresos es mayor que 62589.57 entonces el valor estimado del Total de Gastos sería:

Total de Gastos = 0.7855 * Total de Gastos Corrientes + 3329497.2018

**Total de Gastos Corrientes:**

Para el Total de Gastos Corrientes se obtuvo un árbol de predicción (Anexo 1C) construido con un coeficiente de correlación igual a 0.979 en el cual se obtuvieron 2 modelos lineales. Como se muestra en el árbol el atributo más influyente es el Total de Ingresos. Las 6 instancias de entrada del modelo se clasificaron en los modelos lineales (LM) que a continuación se muestran:

LM1: clasificó 2 instancias cumpliéndose que: si el Total de Ingresos es menor o igual que 62589.57 entonces el valor estimado del Total de Gastos Corrientes sería:

Total de Gastos Corrientes = 5.0807 * Gastos de Bienes y Servicios + 346955.2107

LM2: clasificó 4 instancias y se cumple que: si el Total de Ingresos es mayor que

62589.57 entonces el valor estimado del Total de Gastos Corrientes sería:

Total de Gastos Corrientes = 3.6822 * Gastos de Bienes y Servicios + 5723205.9802

**Gastos de Personal:**

En el árbol de predicción (Anexo 1D) el atributo más importante son los Gastos de Bienes y Servicios. Se obtuvo un coeficiente de correlación igual a 0.9817. Los modelos lineales que se lograron fueron 2 con 9 instancias en total:

LM1: clasificó 3 instancias, en este se cumple que si Gastos de Bienes y Servicios es menor o igual que 960020.53 entonces el valor estimado de Gastos de Personal sería:

Gastos de Personal = 3.0209 * Gastos de Bienes y Servicios + 308635.879

LM2: clasificó 6 instancias y se cumple que si Gastos de Bienes y Servicios es mayor que 960020.53 entonces el valor estimado de Gastos de Personal sería:

Gastos de Personal = -0.0108 * Total de Ingresos + 2.5894 * Gastos de Bienes y Servicios + 2955763.7978

**Gastos de Personal (usando el conjunto de datos con sus epígrafes y partidas)**

Se obtuvo un modelo lineal con un coeficiente de correlación igual a 0.9985 y se cumple que:

Gastos de Personal = 1.3015 * Retribuciones Salariales + 28722.7476

**Gastos de Bienes y Servicios:**

El árbol de predicción (Anexo 1E) obtenido muestra que los atributos más influyentes son Gastos de Personal y Gastos de Capital. El modelo se obtuvo con un coeficiente de correlación de 0.9936 y con 9 instancias de entrada. Los modelos lineales, 3 en total, se describen a continuación:

LM1: clasificó 3 instancias y se cumple que si Gastos de Personal es menor o igual que 4102182.965 entonces el valor estimado de Gastos de Bienes y Servicios sería:

Gastos de Bienes y Servicios = 0.1498 * Gastos de Personal + 1.0671 * Gastos de Capital + 10087.96

LM2: clasificó 4 instancias y se cumple que si Gastos de Personal es mayor que 4102182.965 y Gastos de Capital es menor o igual que 689350 entonces el valor estimado de Gastos de Bienes y Servicios sería:

Gastos de Bienes y Servicios = 0.0043 * Total de Gastos + 0.1284 * Gastos de Personal + 1.1956 * Gastos de Capital + 457678.0535

LM3: clasificó 2 instancias y se cumple que si Gastos de Personal es mayor que 4102182.965 y Gastos de Capital es mayor que 689350 entonces el valor estimado de Gastos de Bienes y Servicios sería:

Gastos de Bienes y Servicios = 0.1284 * Gastos de Personal + 1.2286 * Gastos de Capital + 521712.1285

**Gastos de Bienes y Servicios (usando el conjunto de datos con sus epígrafes y partidas)**

Para la relación entre partidas y elementos de Gastos de Bienes y Servicios se obtuvo un árbol de predicción (Anexo 1e) construido con 9 instancias y con un coeficiente de correlación de 0.9831. Los modelos lineales construidos se describen a continuación:

LM1: clasificó 3 instancias y se cumple que si Viáticos es menor o igual que 42075.89 entonces el valor estimado de Gastos de Bienes y Servicios sería:
Gastos de Bienes y Servicios = 4.0568 * Alimentación - 5075.4148

LM2: este modelo clasificó 2 instancias y se cumple que si Viáticos es mayor que 42075.89, Alimentación es menor o igual que 490929.315 y Viáticos es menor o igual

que 87963.25 entonces el valor estimado de Gastos de Bienes y Servicios sería:

Gastos de Bienes y Servicios = 42.3859 * Viáticos + 12.5122 * Alimentación
- 7403071.9814

LM3: clasificó 2 instancias y se cumple que si Viáticos es mayor que 42075.89, Alimentación es menor o igual que 490929.315 y Viáticos es mayor que 87963.25 entonces el valor estimado de Gastos de Bienes y Servicios sería:

Gastos de Bienes y Servicios = 42.3859 * Viáticos + 12.5122 * Alimentación
- 7400380.1225

LM4: este modelo clasificó 2 instancias y se cumple que si Viáticos es mayor que 42075.89 y Alimentación es mayor que 490929.315 entonces el valor estimado de Gastos de Bienes y Servicios sería:

Gastos de Bienes y Servicios = 42.8995 * Viáticos + 13.5751 * Alimentación
- 7930841.3983

**Otras Transferencias Corrientes:**

Para Otras Transferencias Corrientes se obtuvo un árbol de predicción (Anexo 1F) en el cual el coeficiente de correlación del modelo obtenido es de 0.8387. Se obtuvieron 3 modelos lineales. El modelo contiene un total de 9 instancias logrando clasificarse con los siguientes modelos lineales (LM).

LM1: clasificó 3 instancias y se cumple que si Gastos de Personal es menor o igual que 4102182.965 entonces el valor estimado de Otras Transferencias Corrientes sería:

Otras Transferencias Corrientes = 0.205 * Total de Gastos - 383790.2054

LM2: clasificó 4 instancias y se cumple que si Gastos de Personal es mayor que 4102182.965 y Total de Gastos Corrientes es menor o igual que 15156148.215

entonces el valor estimado de Otras Transferencias Corrientes sería:

Otras Transferencias Corrientes = 0.2742 * Total de Gastos - 840777.7442

LM3: clasificó 2 instancias y se cumple que si Gastos de Personal es mayor que 4102182.965 y Total de Gastos Corrientes es mayor que 15156148.215 entonces el valor estimado de Otras Transferencias Corrientes sería:

Otras Transferencias Corrientes = 0.2768 * Total de Gastos - 761145.3413

**Otras Transferencias Corrientes (usando el conjunto de datos con sus epígrafes y partidas)**

Se obtuvo un modelo lineal con un coeficiente de correlación igual a 1 y se cumple que:

Otras Transferencias Corrientes = 1.0022 * Al Presup. de la Seguridad Social + 1.0039 * Estipendio a Estudiantes - 125.0315

**Gastos de Capital:**

Se obtuvo un modelo lineal con un coeficiente de correlación de 0.7893 y clasifica 9 instancias. Se cumple que:

Gastos de Capital = 0.0438 * Total de Gastos - 84846.0635

**Gastos de Capital (usando el conjunto de datos con sus epígrafes y partidas)**

Se obtuvo un modelo lineal con un coeficiente de correlación igual a 0.9999 y se cumple que:

Gastos de Capital = 0.9986 * Inversiones Materiales U.P + 2179.3015

## 2.5 Evaluación

En esta etapa, se evalúan los modelos construidos revisando cada uno de los pasos ejecutados para crearlo, a fin de comprobar si cumple correctamente con los objetivos

del negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido considerada suficientemente. En el final de esta fase, se toma una decisión para el uso de los resultados de minería de datos.

### *2.5.1 Evaluar los resultados*

Para evaluar los resultados del proceso de minería es necesario valorar los resultados obtenidos en términos de criterios de éxitos del negocio.

De forma general el 100 % de los modelos construidos está por encima del 0,7 del coeficiente de correlación y entre ellos el 87% por encima del 0,8 cumpliéndose con el primero de los criterios de éxitos del negocio.

Para el cumplimiento de los 3 objetivos del negocio planteados se realizaron 11 experimentos en los que se obtuvieron modelos representados por árboles de predicción que estiman el valor de las variables a analizar.

De los modelos obtenidos que fueron descritos anteriormente, siete experimentos dan cumplimiento al objetivo #1 pues se obtuvo la relación que se establece entre las variables: Total de Ingresos, Total de Gastos, Total de Gastos Corrientes, Gastos de Personal, Gastos de Bienes y Servicios, Otras Transferencias Corrientes y Gastos de Capital empleando los datos de los tres años registrados tomando como atributos dichas variables (Anexos 1A – 1F).

Cuatro experimentos dan cumplimiento al objetivo #2 ya que se determina la influencia del comportamiento de los epígrafes y partidas en el resultado final de las variables, tomando como datos a relacionar cada variable con los epígrafes y partidas que la identifican (Anexo 1e).

En las pruebas realizadas a los datos recogidos del año en curso, entre los modelos de

las variables Total de Gastos y Otras Transferencias Corrientes se encontró que los resultados obtenidos están cercanos al plan del presupuesto propuesto por el Instituto.

Los modelos lineales con resultados positivos para las variables fueron: para el Total de Gastos el modelo lineal (LM2), teniendo una diferencia de miles de pesos entre los resultados (LM=13 741 299 y Total de Gastos en 2011 = 13 255 000); para Otras Transferencias Corrientes con un valor de 2 898 000 el modelo LM2 se obtuvo de usar el conjunto de datos de las variables finales y tuvo un valor de 2 793 743, y el otro modelo obtenido del conjunto de datos independiente que relaciona la variable con sus subconceptos (epígrafes y partidas) que arrojó un valor de 2 881 256.

En resumen, se han realizado un total de 33 experimentos y se han cumplido todos los objetivos de negocio trazados, por lo que pueden considerarse los modelos obtenidos como aceptados, desde el punto de vista analítico, para apoyar la toma de decisiones en la planificación del presupuesto del Instituto, teniéndose en cuenta que para lograr mejores resultados se necesitaría de una data histórica de diez años o más.

## 2.6 Despliegue

La etapa de despliegue final del proyecto, en este caso, recae completamente sobre los directivos del Departamento de Economía del Instituto, que son los encargados de emprender acciones y determinar, si así lo estiman conveniente, una estrategia a seguir, que basadas en la información descubierta por los modelos construidos ayuden a una mejor planificación del presupuesto. De similar modo, deberán planificar la supervisión y el mantenimiento del proceso, a fin de evitar largos periodos innecesarios, de uso incorrecto de los resultados de la minería de datos.

## Conclusiones del capítulo

En el desarrollo de las etapas de Modelación y Evaluación de la metodología CRISPDM se plantea la realización de dos tareas de la minería de datos, la predicción y la clasificación, dentro de la predicción la técnica de los árboles de predicción fue la escogida para darle cumplimiento a los objetivos del negocio propuestos los cuales se consideran fueron cumplidos ya que de forma general más del 80 % de los modelos presentan un coeficiente de correlación mayor o igual que el 0.8.

# CONCLUSIONES

Con el desarrollo de la investigación se cumplieron los objetivos planteados con el fin de darle respuesta al problema de investigación propuesto. A partir de los resultados es posible llegar a las siguientes conclusiones:

- Se lograron resultados satisfactorios en el proceso de KDD como resultado de emplear una metodología para guiar el proceso y una herramienta para apoyar el análisis de datos, CRISP-DM y WEKA respectivamente, pues poseen las características necesarias para aprovecharlas en el proyecto.

- El proceso de KDD se desarrolló sobre un conjunto inicial de datos formado por 7 variables.

- En correspondencia a las características del problema y teniendo en cuenta los objetivos del negocio se plantearon dos tareas de Minería de Datos, la predicción y la clasificación, relacionadas con las técnicas de regresión, árboles de predicción e inducción de reglas, realizándose un total de 33 experimentos, de ellos 11 que corresponden a los árboles de predicción que fueron los escogidos para modelar el problema.

- Las reglas obtenidas representan los modelos lineales construidos aceptados para ser útiles en la toma de decisiones en la estimación del presupuesto de años posteriores.

- La investigación realizada aporta nuevos conocimientos que permiten redefinir algunos objetivos del negocio y replantearse un nuevo proceso de KDD.

# RECOMENDACIONES

Como trabajos futuros y a fin de consolidar los resultados obtenidos, se recomienda:

- Investigar sobre nuevas tendencias y aplicaciones de la minería de datos en el ámbito de la actividad presupuestada.

- Experimentar con otros entornos de trabajo de la herramienta de análisis WEKA que pudieran resultar interesantes para las tareas de minería de datos.

- Estudiar las vías para optimizar de forma dinámica los parámetros de los algoritmos empleados, a fin de mejorar los resultados encontrados.

- Experimentar con una data histórica con registros de diez años o más para obtener mejores resultados en los modelos que se construyan.

- Explorar otras técnicas de minería de datos para la extracción de conocimiento.

- Realizar con el conocimiento adquirido un sistema para la elaboración del plan del presupuesto para años posteriores.

# BIBLIOGRAFÍA

ACOSTA AGUILERA, M.E. Minería de datos y descubrimiento de conocimiento. La Habana: Universidad de La Habana. Disponible en: http://www.google.com.

ARTÍCULOS ESTADÍSTICOS. Disponible en: http://www.estadistico.com/arts.html

BATISTA ALDANA, Tamara. Análisis de las variables meteorológicas en el Municipio de Moa aplicando Minería de Datos. Instituto Superior Minero Metalúrgico "Dr. Antonio Núñez Jiménez", 2010.

BERTHOLD, M.; Hand, D.J. (eds.) Intelligent Data Analysis. An Introduction, Springer, 2ndEdition, 2003.

BISSET HECHAVARRÍA, Kirenia. Obtención de mejoras en la explotación de los Grupos Electrógenos Diesel de Moa aplicando Minería de Datos. Instituto Superior Minero Metalúrgico "Dr. Antonio Núñez Jiménez", 2010.

BRITO SARASA, Raycos. Minería de datos aplicada a la gestión docente del Instituto Superior Politécnico "José Antonio Echeverría". Alejandro Rosete Suárez (Tutor). Tesis Maestría. Instituto Superior Politécnico "José Antonio Echeverría", 2008.

CHAPMAN, P. et al., CRISP-DM 1.0: Step-by-step data mining guide. USA: SPSS Inc., CRISP-DM Consortium, 2000.

CORRÍA RAMÍREZ, Isidro M.; SHELTON NADAL, Ronald. Estrategia de trabajo para el desarrollo del módulo de Minería de Datos de un CALL CENTER, aplicando la metodología CRISP-DM. Lic. Gabriel Zerquera Guerra (tutor). Tesis de Grado. Universidad de las Ciencias Informáticas, 2004.

FAYYAD, U. et al., "Advanced in Knowledge Discovery and Data Mining," MIT Press, MA, 1996.

FERNÁNDEZ ALDANA, Luís Antonio. Principios de data mining. 2005. Disponible en

http: www.monografias.com visitado en enero/2011

GARCÍA MORATE, Diego. Manual de WEKA. Disponible en: http://www.google.com.

GÓMEZ RAMOS, J.L. Descubrimiento del conocimiento en base de datos utilizando herramienta de software libre, caso: DAIS. Universidad Juárez Autónoma de Tabasco. Disponible en: http://www.google.com.

GONDAR NORES, J. E. Metodologías para la Realización de Proyectos de Data Mining. España, Madrid: Data Mining Institute, 2004. Disponible en: http://www.estadistico.com

HERNÁNDEZ, et al., Introducción a la minería de datos. Madrid, Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación: Ed. Pearson Educación, S.A., 2004.

IBM Software Group. Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark. IBM Press, USA, 2003.

REYES SALDAÑA, et al. El proceso de descubrimiento de conocimiento en bases de datos, 2005.

LÓPEZ ARÉVALO, Iván Dr., Línea de Investigación, Laboratorios de Tecnologías de Información. Cinvestav – Tamaulipas, 2005.

LÓPEZ PUPO, Eliannys. Minería de datos aplicada a la gestión docente del Instituto Superior Minero Metalúrgico "Dr. Antonio Núñez Jiménez". Instituto Superior Minero Metalúrgico "Dr. Antonio Núñez Jiménez", 2010.

MARTÍNEZ DE PISÓN ASCACIBAR, Javier. Optimización mediante técnicas de minería de datos del ciclo de recocido e una línea de galvanizado. Dr. D. Joaquín Bienvenido Ordieres Meré. Tesis Doctoral. Universidad de La Rioja, 2003.

MINERÍA DE DATOS. Disponible en:

http://catarina.udlap.mx/u_dl_a/tales/documentos/msp/gonzalez_r_l/apendiceC.pdf

MOLINA LÓPEZ, José M.; GARCÍA HERRERO, Jesús. Técnicas de análisis de datos. Universidad Calor III Madrid, 2006. Disponible en: http://www.google.com.

MONOGRAFÍAS 1, La Minería de Datos y el Descubrimiento de Conocimiento en Bases de Datos, 2004. Disponible en: http://www.monografias.com

RODRÍGUEZ MONTEQUÍN, Mª Teresa; ÁLVAREZ CABAL, J. Valeriano. Metodologías para la realización de proyectos de data mining, Universidad de Oviedo.

TOLEDANO MUÑOZ, J. Disponible en: http://datamining.iespana.es/datamining_enfoque.htm

VALLEJOS, Sofía J. Minería de Datos. Corrientes, Argentina: Universidad Nacional del Nordeste, 2006. Disponible en: http://www.google.com

VARCÁRCEL ASENCIOS, Violeta. Data Mining y el Descubrimiento del Conocimiento. Universidad Nacional Mayor de San Marcos. Perú, 2004

VILA, Amparo. Introducción a la Extracción del Conocimiento y a la Minería de Datos. Universidad de Granada. 2005.

VILCHES GONZÁLEZ, Erika; Escobar Broitman, Iván A. Minería de datos, 2007

WIKIPEDIA 1, Toma de decisiones, 2011. Disponible en: http://es.wikipedia.org/wiki/Toma_de_decisiones

WIKIPEDIA 2, Presupuesto. Disponible en: http://es.wikipedia.org/wiki/Presupuesto

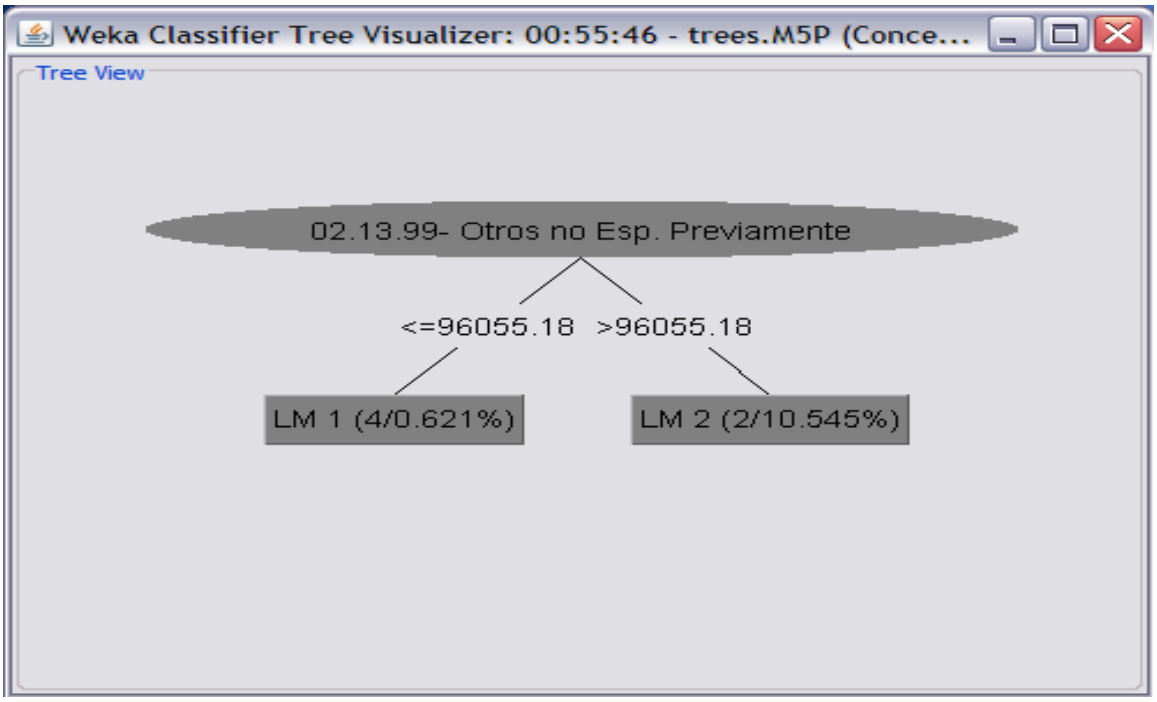
WIKIPEDIA 3, Minería de datos, Disponible en: http://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos

ZAMARRÓN SANZ, Carlos; et al. Aplicación de la minería de datos al estudio de las alteraciones respiratorias durante el sueño. Santiago de Compostela: Servicio de

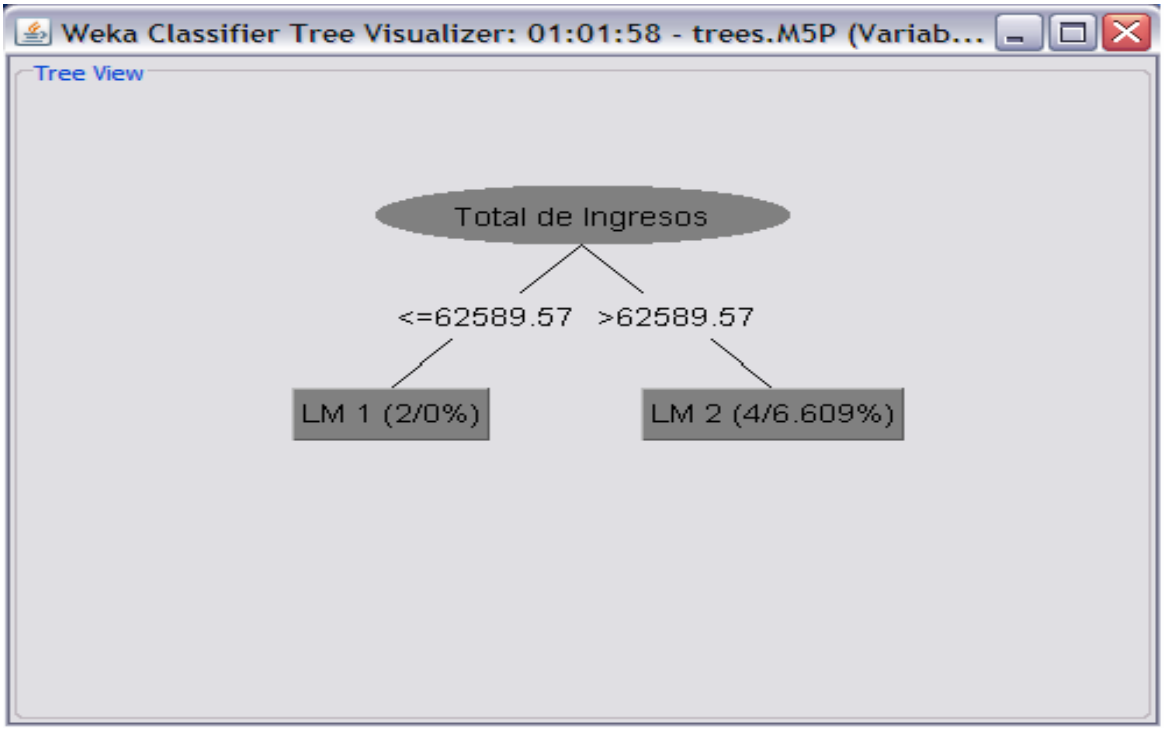
Neumología, Hospital Clínico Universitario, 2006. Disponible en: http://www.google.com.

# ANEXOS

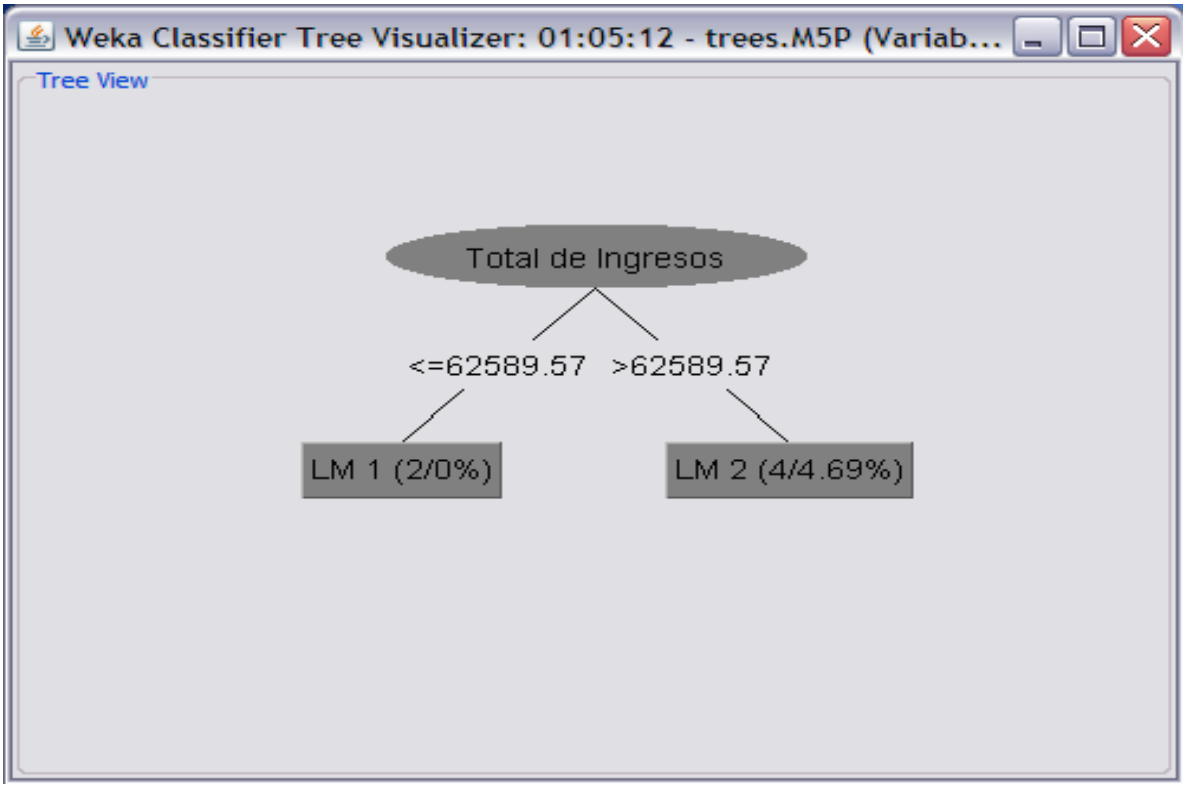
## Anexo 1A. Total de Ingresos.



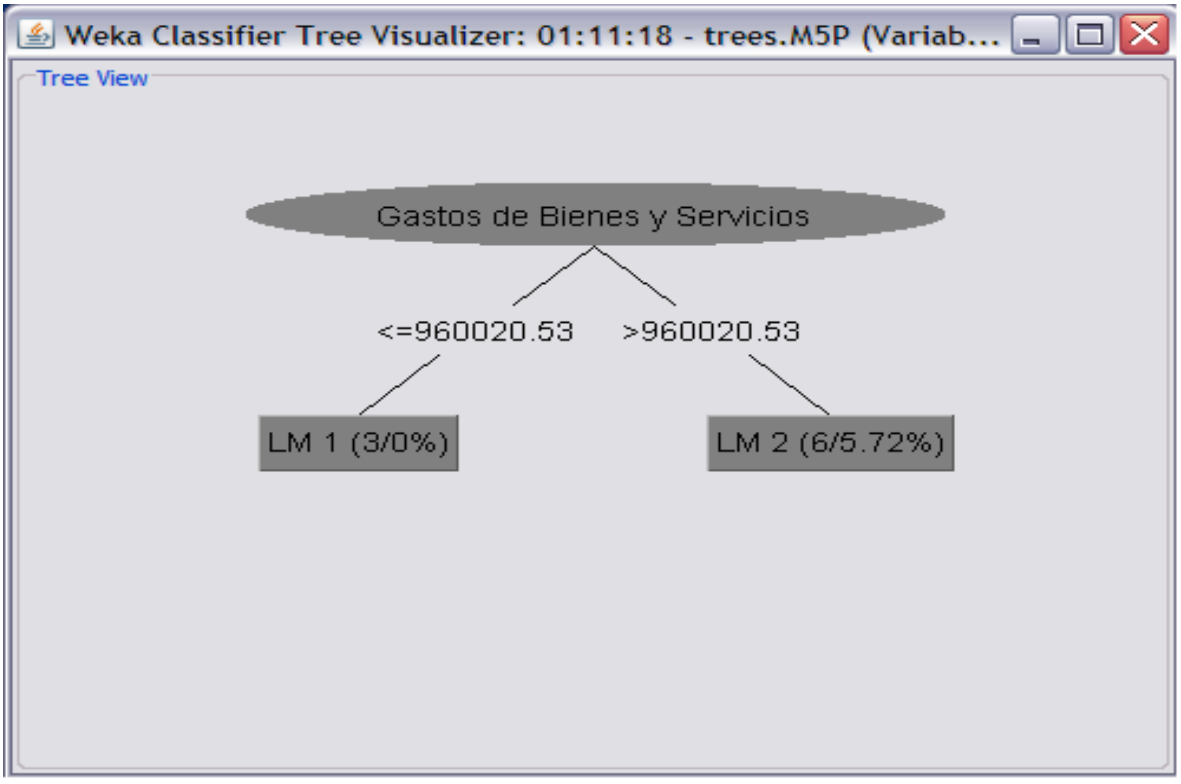

## Anexo 1B. Total de Gastos.



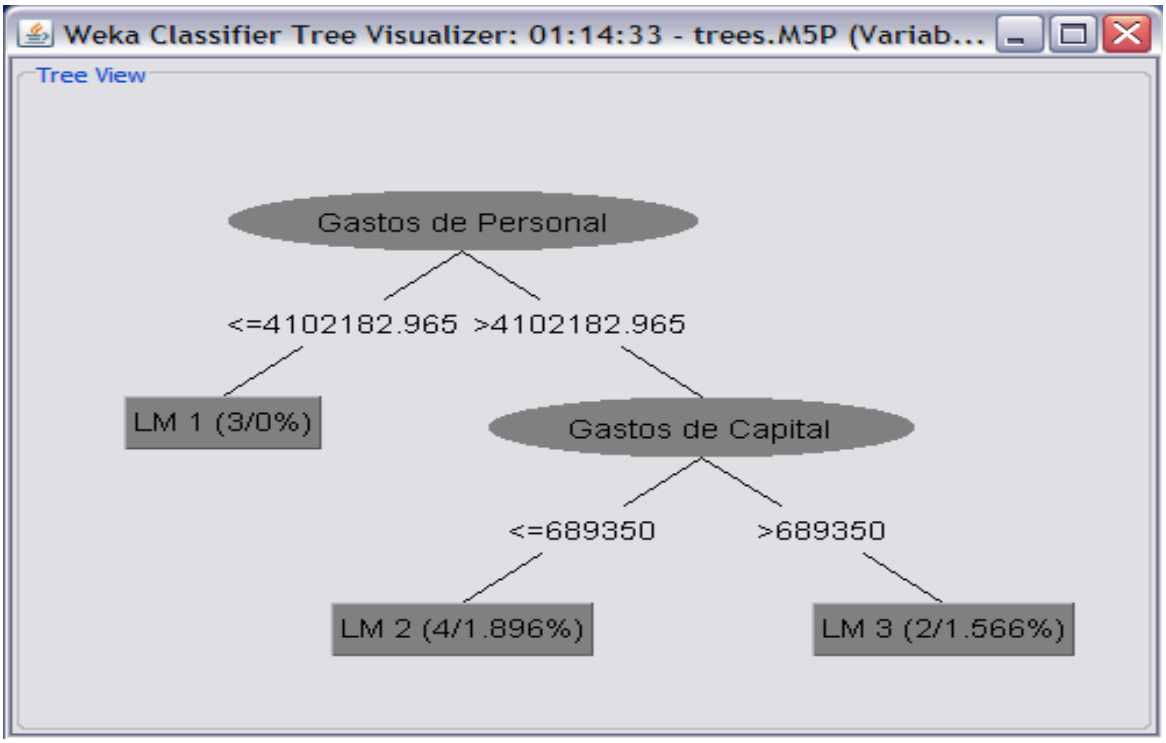
**Anexo 1C. Total de Gastos Corrientes.**



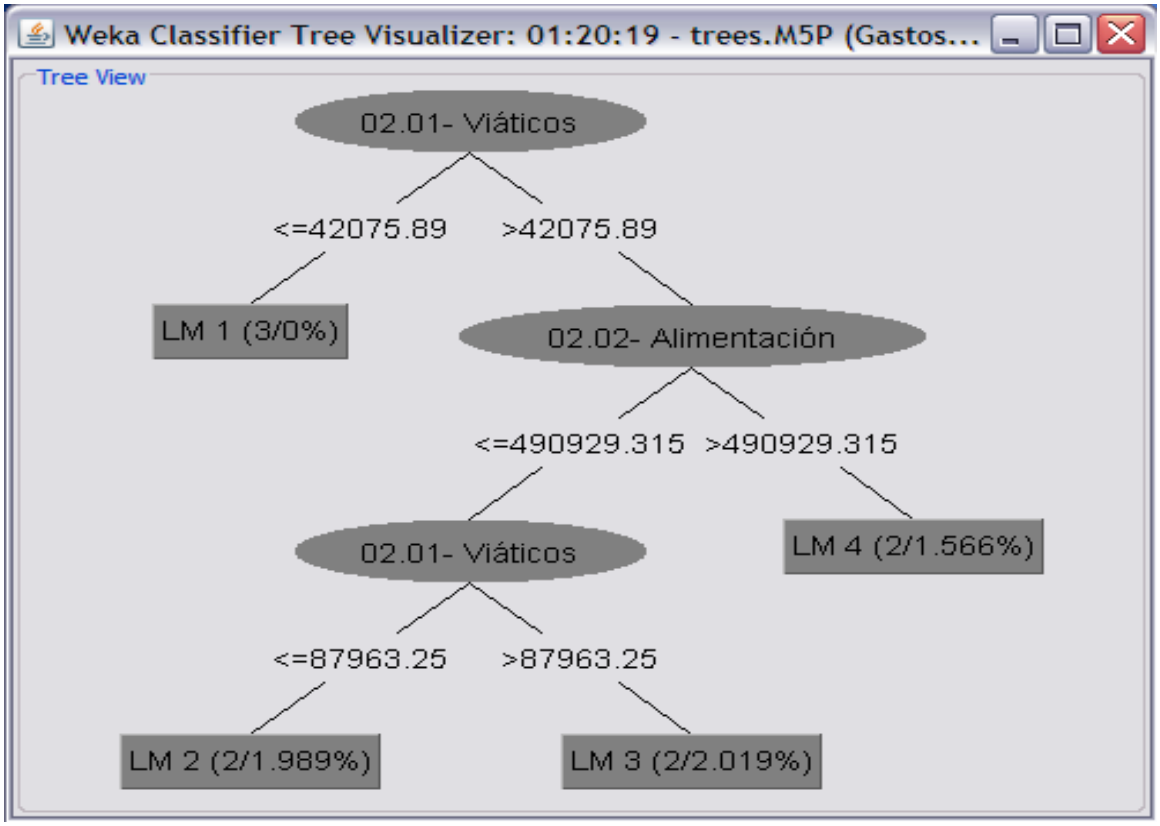

**Anexo 1D. Gastos de Personal.**



**Anexo 1E. Gastos de Bienes y Servicios.**




**Anexo 1e. Gastos de Bienes y Servicios.**



**Anexo 1F. Otras Transferencias Corrientes.**