UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 15

TÍTULO: Propuesta de un Sistema de Información para la Gestión de Datos Climáticos.

Trabajo de Diploma para optar por el Título de Ingeniero en Ciencias Informáticas

AUTORES:

José Isvet Barlia Bernal

Jorge Fonseca Córdova

TUTORA: **Águeda Librada García Martín**

CONSULTANTE: **Alejandro Rodríguez Pupo**

Ciudad de La Habana, junio 2010

DECLARACIÓN DE AUTORÍA

Declaramos ser los únicos autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso exclusivo del mismo en su beneficio. Para que así conste firmamos la presente a los _ días del mes de _____ del año _____.

_____
José Isvet Barlia Bernal (Autor)

_____
Jorge Fonseca Córdova (Autor)

_____
M.Sc. Águeda Librada García Martín (Tutor)

_____
Alejandro Rodríguez Pupo (Consultante)

AGRADECIMIENTOS

La realización de este trabajo ha sido posible gracias al amor y el apoyo de muchas personas, pero en especial quisiera agradecer a mis padres y hermanas, los que me han brindado su apoyo incondicional, y han estado siempre en el momento que los he necesitado.

A mis verdaderos amigos, los que se encuentran todavía junto a mí en este tren, "el tren que es la vida" por ser y estar.

Y de forma muy especial a nuestros tutores Águeda y Alejandro, que han sabido encaminarnos y servirnos de guía de forma incondicional en este trabajo, soportando nuestros tropiezos, y a todas las personas que han contribuido con su granito de arena para poder llegar hasta este momento.

Muchas Gracias.

José

Quiero agradecer en especial con todo mi amor a la persona que más quiero en la vida, mi madre, por ser la persona que siempre ha estado conmigo en todo momento, siempre confiando en mí, ofreciéndome su apoyo incondicional de una forma u otra, guiándome por el camino correcto y haciendo de mí una mejor persona, te quiero.

A mi padre por la confianza depositada en mí y por sus consejos que me han servido durante este tiempo de estudios.

A mis tías que siempre se preocuparon por mí y también son partícipes de este sueño.

A nuestros tutores Águeda y Alejandro los que nos dedicaron su tiempo y sus conocimientos para realizar con éxito este trabajo.

A mis amigos y hermanos de estudios que vienen conmigo desde primer año, y los demás que se fueron incorporando en el transcurso del tiempo de estos cinco años.

Muchas Gracias.

Jorge

DEDICATORIA

Dedico mi trabajo de diploma a mis padres y hermanas, los que me han sabido encaminar durante la vida, me han amado con mis defectos y virtudes, me han enseñado a tener paciencia, a ser fuerte y no amedrentarme por duro que parezca el camino, a saber que todo es posible y que todo lo que sucede por una buena razón es, aún cuando sea difícil de ver para nuestros ojos.

José

Dedico especialmente este trabajo a mis padres quienes me dieron un impecable ejemplo a seguir durante esta etapa de estudiante, a mi hermanita pequeña que es una de las cosas más grandes que me dio mi madre. A mis tías por ser las mejores tías del mundo y por el apoyo que me brindaron.

Jorge

RESUMEN

En la actualidad, en el Sistema de Administración de Datos Climáticos se encuentran almacenados grandes volúmenes de datos de todas las estaciones meteorológicas del país, con lo cual el Centro Nacional del Clima brinda servicios a los diferentes departamentos de la institución, así como a empresas y entidades de la nación, usando técnicas automatizadas y manuales.

En estos momentos, el centro cuenta con la necesidad de obtener patrones de comportamiento y relaciones entre la enorme mina de datos almacenados, aplicando las técnicas computacionales y algoritmos provistos por la Minería de Datos, los cuales permiten proporcionar nueva información que los investigadores y especialistas del centro y entidades relacionadas, pudieran usar para obtener aportes prácticos de diversa índole.

Estos conocimientos comprendidos por los mismos pudieran ser expuestos en un espacio colaborativo e intercambio, red de conocimientos, posibilitando el avance simétrico del conocimiento.

Por lo tanto el objetivo desarrollado en la presente investigación es la propuesta del diseño del Sistema de Información para la Gestión de Datos Climáticos.

TABLA DE CONTENIDOS

INTRODUCCIÓN

La observación del tiempo atmosférico por métodos instrumentales y visuales genera un gran volumen de información, el cual se incrementa en el tiempo exponencialmente, se hace imprescindible el empleo de recursos computacionales para su tratamiento, procesamiento, conservación y consulta. Esta realidad se hace más evidente dado que muchos de los instrumentos o tecnologías de medición fabricados en la actualidad tienen salida digital.

El empleo de Bases de Datos constituye la solución más adecuada para la conservación y consulta de la información, incluyendo la meteorológica pues estas brindan gran capacidad de almacenamiento, seguridad y protección de la información y ofrecen posibilidades de gestión (búsqueda, actualización y eliminación) utilizando determinadas herramientas estándar (lenguajes de datos) y modelos matemáticos para manejar los datos relacionados entre sí.

Esto permite que los Sistemas Informáticos que utilizan Bases de Datos cuenten con propiedades importantes, tales como: facilidad en el manejo y consulta de grandes volúmenes de datos relacionados, con rapidez y seguridad; estos sistemas, al estar basados en estándares, tienen escasa dependencia del creador del sistema, de los propios datos e incluso de la tecnología computacional utilizada. (Date, 2003)

Con respecto a la información meteorológica, en la Nota Técnica (OMM – 902, 1998), la **Organización Meteorológica Mundial (OMM)**, en el capítulo "Consideraciones y Recomendaciones", establece las siguientes "Consideraciones Generales", donde se expone de forma clara y precisa acerca de las "Bases de Datos Climáticas":

> *"Para las Bases de Datos Climatológicas puede que se desee comparar datos con diferentes restricciones en las condiciones reportadas, por ejemplo, determinados valores de velocidad del viento bajo determinados rangos de temperatura y otros de humedad. Para esto, es necesario acceder a los valores de forma no codificada y disponer de la capacidad de efectuar operaciones lógicas sobre los valores de diferentes tablas o columnas. Para realizar estos tipos de operaciones, está claro que es necesario organizar los datos en una estructura que dependa de tres componentes básicos: donde (lugar) está el dato, cuando se obtuvo, y de qué dato se trata."*

Más adelante, en ese mismo capítulo, la OMM insiste en el concepto de *"Donde*", indicando que se trata de las coordenadas geográficas de la estación o sitio de medición, su altitud y otras características que

determinan el entorno de medición. Enfatizando posteriormente en el "*Qué"* se mide, indicando la conveniencia de reservar variables auxiliares para el control de calidad de cada elemento.

Es evidente la importancia que le concede la OMM a todo el conjunto de metadatos que rodea el historial de las mediciones en cada estación, desde el régimen de trabajo de la estación, los instrumentos utilizados y su confiabilidad, hasta el entorno medio – ambiental y físico – geográfico de la estación, según lo expresado en variadas notas técnicas, recomendaciones y reuniones de expertos, como la celebrada en Málaga del 24-26 de Febrero del 2003 (Gobierno de Australia. Departamento de Meteorología, 2005).

Considerando todos estos factores, la OMM (TD No. 1025, 2000) y (TD- No. 1130, 2002) ha prestado especial atención a que cada país miembro disponga de un Sistema Informático que le permita almacenar y gestionar la información meteorológica proveniente de su red de estaciones y a la vez, le permita efectuar diferentes consultas a esta información, con múltiples intereses, así como que dicho Sistema permita consultar la climatología básica de estas estaciones. A este tipo especial de Sistemas Informáticos, se les denomina **Sistemas de Administración de Datos Climáticos (SADC)**, e internacionalmente se les conoce por las siglas **CDMS**, tomadas de su nombre en inglés**: Climate Database Management Systems** (TD No. 1025, 2000)**.**

Técnicamente, estos sistemas consisten en una o varias bases de datos estándar y un conjunto de programas de aplicación y otras herramientas de software anexas que garantizan la adquisición y validación de la información meteorológica, su almacenamiento, gestión, realización de cálculos climáticos, administración del sistema, seguridad, protección y otras tareas.

Cuba no ha estado ajena a la necesidad de contar con una herramienta de este tipo. A finales de la década del 80 del pasado siglo se desarrolló en el Centro del Clima una versión preliminar de un CDMS nombrado SADCLIM (Sistema de Administración de Datos Climáticos), con avanzadas características y buen desempeño.

También fue adquirido el CLICOM (Climate Computing Project), recomendado por la OMM. Sin embargo, el CLICOM no fue instalado completamente en Cuba y solo fue posible utilizar parte de sus funcionalidades, las cuales muy pronto se hicieron obsoletas con el Sistema Operativo Windows. Algunas soluciones paliativas fueron utilizadas, empleando Microsoft Access (sistema de gestión de bases de datos relacional creado y modificado por Microsoft para uso personal en pequeñas organizaciones) u hojas de cálculo (Microsoft Excel) para almacenar y manipular una parte de los datos climáticos.

En el 2003, tras la participación de Cuba en el Taller Regional sobre Rescate y Gerencia, Monitoreo, Aplicación y Predicción de Datos Climáticos, se orientó por la Dirección de la institución disponer en un plazo corto de un CDMS, preferiblemente de producción nacional.

Por tal razón, se decidió por el Centro del Clima (Institución que radica dentro del Instituto Nacional de Meteorología) diseñar y programar un CDMS con las siguientes características:

- Capacidad de guardar y gestionar de manera automatizada, toda la información de metadatos geográficos, instrumentales y ambientales asociados a las estaciones meteorológicas del país y al proceso de las mediciones.
- Almacenar los datos provenientes de estaciones automáticas a intervalos de tiempo de 10 o 15 minutos.
- Ofrecer información climatológica actualizada de manera rápida y sencilla.
- Brindar salidas y servicios especializados a los diferentes Centros y Departamentos de la institución.
- Seguir las recomendaciones de la Organización Meteorológica Mundial respecto a las características anheladas en un Sistema de Administración de Datos Climáticos.
- Garantizar un óptimo control de la calidad de los datos y metadatos, así como la estabilidad de los mismos ante actualizaciones, tratamientos estadísticos, reubicación de estaciones y otras transformaciones.

En la actualidad en el **Sistema de Administración de Datos Climáticos** se encuentran almacenados grandes volúmenes de datos de todas las estaciones meteorológicas del país aproximadamente 4 millones de registros procesados y digitalizados a lo largo de decenas de años y se continúa con el rescate de las observaciones meteorológicas de años anteriores, pues existen observaciones de la primera década del siglo XX.

Con lo cual el **Centro Nacional del Clima** brinda servicios a los diferentes departamentos de la institución, así como a empresas y entidades de la nación, usando técnicas automatizadas y manuales. Además es posible contar con las posibilidades que ofrece la red empresarial al diseñar un cliente Web que permite acceder a la climatología tradicional (o estándar) a los investigadores de la institución.

Una vez obtenido este logro, el SADC del Instituto de Meteorología está entrando en una nueva fase, la de aplicar las técnicas de inteligencia artificial y **Minería de Datos (MD, data mining -término en inglés-)** en nuevos servicios donde ya no sólo el sistema brindaría información sino conocimiento para ayudar a la toma de decisiones por parte de los investigadores de la institución y demás clientes.

Por todo lo anterior, se hace necesaria la búsqueda de nuevos modelos, algoritmos y técnicas computacionales que permitan, a través de procesos inteligentes, obtener resultados que sean de interés para los especialistas del centro y que les permita nuevos conocimientos sobre la información generada.

La Minería de Datos engloba una serie de procesos y técnicas muy novedosas basadas en la Inteligencia Artificial y el Análisis Estadístico, encaminadas a la extracción de conocimiento procesable, nuevo y útil, implícito en grandes almacenes de datos y/o bases de datos.

Para la realización de la presente investigación se tiene como **problema científico**: *"La inexistencia de una herramienta o aplicación en la institución para procesar los grandes volúmenes de datos existentes en el Sistema de Administración de Datos Climáticos y convertirlo en un **Sistema de Información para la Gestión de Datos Climáticos (SIGDC)** que brinde conocimiento para los usuarios potenciales en sus investigaciones"*

Para dar respuesta a la interrogante anterior se tiene como **objeto de estudio** *los modelos matemáticos, algoritmos y técnicas computacionales asociadas a la Minería de Datos como base de los Sistemas de Información* y como **campo de acción** *los modelos matemáticos, algoritmos y técnicas computacionales asociadas a la Minería de Datos que permitan convertir el Sistema de Administración de Datos Climáticos del **Instituto de Meteorología (INSMET)** en un Sistema de Información para la Gestión de Datos Climáticos.*

Se formula la siguiente **hipótesis** *si se contara con un Sistema de Información para la Gestión de Datos Climáticos se podrían ofrecer diversos servicios especializados a investigadores que tendrían a disposición un conjunto de modelos matemáticos, algoritmos y técnicas computacionales pertinentes a los diferentes problemas de investigación con la data climática*.

Se plantea como **objetivo general**: *Proponer un SIGDC aplicando la Minería de Datos que permita ampliar los servicios actuales del SADC*.

Un sistema de información procesa los datos a través de procesos automatizados, lo cual unido a la gestión de una base de datos hace posible acceder a los datos integrados que atraviesan los límites operacionales, funcionales u organizacionales. Al definirlo como un sistema garantizamos la relación entre todas las partes y su vinculación como un todo. Para adoptar un Sistema de Información es necesario comprender hasta sus últimas consecuencias, la importancia de la información y la tecnología que lo soportan, además de los Recursos Humanos, sus más valiosos activos (Blanco, 2004).

En la actualidad a la gestión de las bases de datos se le atribuye una importancia propia porque este es el proceso que permitirá además de administrar, gestionar todos los datos en información reveladora con la utilización de modelos matemáticos, que procesen los datos existentes y generen a su vez nuevos datos e indicadores.

Y como objetivos específicos:

- Realizar un estudio de los modelos matemáticos, algoritmos y técnicas computacionales de la Minería de Datos.
- Seleccionar los algoritmos más adecuados para la solución de los problemas identificados de acuerdo a la tecnología existente en el centro.
- Proponer una herramienta de desarrollo para la aplicación de la Minería de Datos en el Sistema de Administración de Datos Climáticos.

A continuación se proponen las siguientes tareas de la investigación como vía de solución a los objetivos propuestos:

- Analizar el estado del arte de la Minería de Datos. Establecer un diagnóstico de las tendencias actuales.
- Estudiar los algoritmos existentes con el objetivo de actualizar sus ventajas y limitaciones.
- Investigar el desarrollo actual de herramientas para la aplicación de la Minería de Datos.
- Analizar las posibles formas de presentación de los resultados teniendo en cuenta los algoritmos seleccionados y las condiciones de la tecnología en la institución.
- Validar la solución propuesta.

Con la presente investigación científica se espera fundamentar una propuesta de desarrollo de un sistema integrado de bases de datos, modelos y herramientas para el procesamiento y visualización de la data

meteorológica requerida para los diversos trabajos de investigación que desarrolla el INSMET como una Red de Conocimientos.

La estructuración del contenido del trabajo de diploma está dada por tres capítulos.

Capítulo 1: **Fundamentación Teórica y Antecedentes. Estado del Arte.** El objetivo con este capítulo es realizar un estudio bibliográfico y una caracterización del estado del arte en los temas referentes al Descubrimiento de Conocimientos en Bases de Datos, y dentro de éste el proceso de Minería de Datos, con sus técnicas y herramientas para su aplicación.

Capítulo 2: **Análisis de las técnicas y herramientas de la Minería de Datos.** Este capítulo tiene como objetivo principal realizar un análisis, desde los modelos matemáticos hasta las técnicas y las principales y más importantes herramientas para la aplicación de la Minería de Datos contextualizado al campo de interés: el Clima y la Meteorología, y fundamentar el diseño del Sistema de Información para la Gestión de Datos Climáticos.

Capítulo 3: **Validación de la Propuesta.** En este capítulo se analiza la factibilidad del diseño propuesto, desde el punto de vista de su futura implementación así como de la satisfacción de los usuarios finales en algunos ejemplos o casos de estudio.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA Y ANTECEDENTES. ESTADO DEL ARTE.

1.1 INTRODUCCIÓN

"Algo peor que no tener información disponible es tener mucha información y no saber qué hacer con ella"

Anónimo

Las Tecnologías de la Información y las Comunicaciones (TIC) han estado revolucionándose aceleradamente en los últimos años, lo cual ha influido notablemente en un crecimiento en las capacidades de generar y colectar datos, al aumentar la capacidad de procesamiento de las máquinas y bajar el costo de almacenamiento de los mismos.

En este gran volumen de datos que es almacenado a diario existe oculta una enorme cantidad de información, la cual en su mayor parte tiene importancia estratégica que no puede ser accedida por las técnicas tradicionales de recuperación de la información.

El descubrimiento de esta información oculta se hizo posible gracias a la Minería de Datos, que entre otras sofisticadas técnicas aplica la inteligencia artificial y el análisis estadístico para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad (Vallejos, 2006).De modo que el objetivo principal de la Minería de Datos es generar patrones de comportamiento en los datos y a partir de ellos generar conocimiento útil.

Las técnicas de Minería de Datos son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de los negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los mismos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. La Minería de Datos toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva.

Los datos son considerados la materia prima en bruto, que se convierten en información cuando un usuario los analiza y les atribuye algún significado especial. Además se puede hacer referencia a estos como conocimiento, desde el momento en que los especialistas elaboran o encuentran un modelo para interpretar esta información y con este modelo obtienen un valor agregado.

1.2 BASES DE DATOS

En la actualidad, y debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos están en formato digital, lo cual ofrece un amplio rango de soluciones al problema de almacenar datos. Las bases de datos son ampliamente utilizadas en entornos científicos con el objeto de almacenar la información experimental.

Un sistema de bases de datos es básicamente un sistema computarizado para llevar registros. Es posible considerar a la propia base de datos como una especie de armario electrónico para archivar; es decir, es un depósito o contenedor de una colección de archivos de datos computarizados cuya finalidad general es almacenar información y permitir a los usuarios recuperar y actualizar esa información con base en peticiones. La información en cuestión puede ser cualquier cosa que sea de importancia para el individuo u organización; en otras palabras, todo lo que sea necesario para auxiliarle en el proceso general de su administración (Date, 2003).

Las Bases de Datos se clasifican en dos grupos según la utilidad de las mismas:

- Bases de datos estáticas: Éstas son bases de datos de sólo lectura, utilizadas primordialmente para almacenar datos históricos que posteriormente se pueden utilizar para estudiar el comportamiento de un conjunto de datos a través del tiempo, realizar proyecciones y tomar decisiones.
- Bases de datos dinámicas: Éstas son bases de datos donde la información almacenada se modifica con el tiempo, permitiendo operaciones como actualización, borrado y adición de datos, además de las operaciones fundamentales de consulta.

Un modelo de datos es básicamente una "descripción" de un contenedor de datos -lugar donde se guarda la información-, así como de los métodos para almacenar y recuperar información de esos contenedores. Los modelos de datos no son cosas físicas: son abstracciones que permiten la implementación de un sistema eficiente de base de datos.

El **modelo relacional** es uno de los más utilizados en la actualidad para modelar problemas reales y administrar datos dinámicamente. Tras ser postulados sus fundamentos en 1970 por Edgar Frank Codd, de los laboratorios IBM en San José (California), no tardó en consolidarse como un nuevo paradigma en los modelos de bases de datos.

Su idea fundamental es el uso de "relaciones". Estas relaciones podrían considerarse en forma lógica como conjuntos de datos llamados "tuplas". Pese a que ésta es la teoría de las bases de datos relacionales creadas por Codd, la mayoría de las veces se conceptualiza de una manera más fácil de imaginar. Esto es pensando en cada relación como si fuese una tabla que está compuesta por registros (las filas de una tabla), que representarían las tuplas y campos (las columnas de una tabla).

En este modelo, el lugar y la forma en que se almacenen los datos no tienen relevancia, a diferencia de otros modelos como el jerárquico y el de red. Esto tiene la considerable ventaja de que es más fácil de entender y de utilizar para un usuario esporádico de la base de datos. La información puede ser recuperada o almacenada mediante "consultas" que ofrecen una amplia flexibilidad y poder para administrar la información.

Structured Query Language o Lenguaje Estructurado de Consultas (**SQL**, siglas en inglés) es un estándar implementado por los principales motores o sistemas de gestión de bases de datos relacionales, el cual es utilizado para construir las consultas a bases de datos relacionales.

Durante su diseño, una base de datos relacional pasa por un proceso al que se le conoce como normalización de una base de datos.

Al realizar un primer análisis de una base de datos con las sentencias SQL, se puede obtener aproximadamente un 80% de la información. El 20 % restante, por lo regular es el que contiene la información más valiosa para el usuario, el cual requiere de la utilización de técnicas más avanzadas, para poder extraer la información requerida.

1.3 DESCUBRIMIENTO DE CONOCIMIENTOS EN BASES DE DATOS

Relacionado con la Minería de Datos aparecen comúnmente otros términos que son utilizados frecuentemente por las distintas bibliografías especializadas en el tema como lo es la extracción o **Descubrimiento de Conocimiento en Bases de Datos** (Knowledge Discovery in Databases o **KDD**, según sus siglas en inglés) (Orallo Hernández, Quintana Ramírez, & Ramírez Ferri, 2004), considerándose uno de los términos que presentan una estrecha relación con esta novedosa técnica.
A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, empezaron a consolidar los términos de Minería de Datos y KDD (Molina Félix, 2002).

Muchos autores usan estos dos términos indistintamente, como si fueran sinónimos, pero es un error pues existen claras diferencias entre ambos. Es común referirse a KDD como un proceso que consta de un conjunto de fases, una de las cuales es la Minería de Datos (Berthold & Hands, 2003). Tomando como premisa esto, el proceso de Minería de Datos consiste únicamente en la aplicación de uno o más algoritmos para extraer patrones de datos, mientras que KDD es el proceso completo que incluye Preprocesamiento, minería y post-procesamiento de los datos.

Una aproximación a la definición de KDD sería el proceso de la extracción automatizada de conocimiento o patrones interesantes, no triviales, implícitos, previamente desconocidos, potencialmente útiles y predictivos de la información de grandes Bases de Datos (Fayyad, Piatetsky-Shapiro, Smith, & R., 1996).

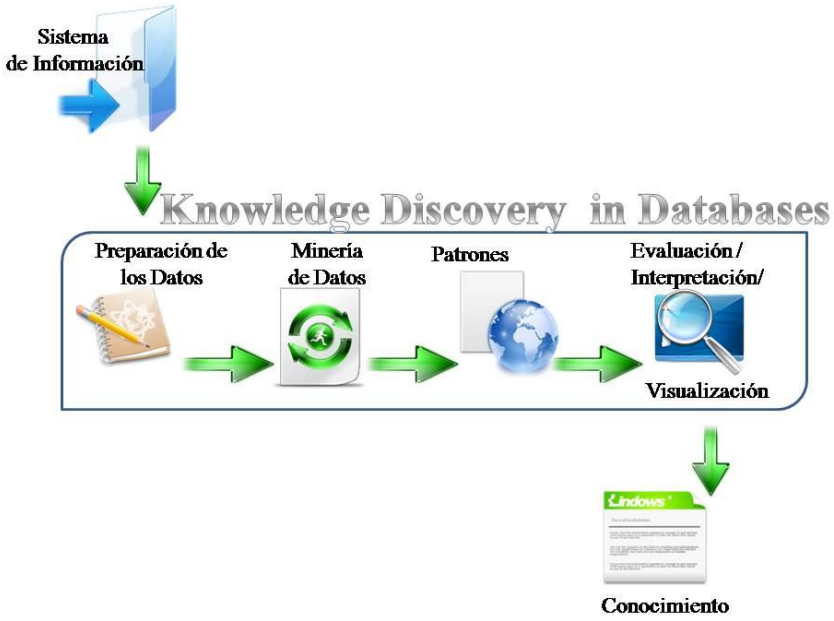


Figura 1 Proceso de KDD

Las investigaciones en temas de KDD incluyen análisis estadístico, técnicas de representación del conocimiento y visualización de datos, entre otras. Algunas de las tareas más frecuentes en procesos de KDD son la clasificación y **agrupamiento** (más conocido por su equivalente en inglés, **clustering**), el reconocimiento de patrones, las predicciones y la detección de dependencias o relaciones entre los datos.

1.4 MINERÍA DE DATOS

> *"Gracias a la Minería de Datos, las computadoras se encargan de seleccionar vastos almacenes de datos. Con una incansable e incesante búsqueda, será posible encontrar la diminuta pepita de oro en una montaña de datos de desperdicio"*
>
> **Edmun De Jesus**

Es imprescindible, para la toma de decisiones, convertir en experiencia, conocimiento y sabiduría los grandes volúmenes de datos, especialmente en las grandes organizaciones y los proyectos científicos. Siempre es útil para las organizaciones la búsqueda de datos relevantes, porque pueden aportar las respuestas más apropiadas a las necesidades de información.
Al contrario del proceso de la investigación científica, en el que se genera una hipótesis para corroborarla o contradecirla con los resultados obtenidos, en la Minería de Datos se captan y procesan los datos, con la esperanza que de ellos surja una hipótesis apropiada. Se espera que los datos describan o indiquen el porqué de la configuración y el comportamiento que presentan (Jesús & Zaiane, 2009).

La Minería de Datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Frank & Witten, 2005). Es la etapa de descubrimiento en el proceso de KDD, consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados (Fayyad, Piatetsky-Shapiro, Smith, & R., 1996). Además es prácticamente, el único proceso analítico que genera nueva información en la capa de acceso.

La misma reúne las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo de datos, principalmente usando como materia prima las Bases de Datos y/o Almacenes de Datos.

Asociado a la MD aparecen comúnmente otros términos que son utilizados frecuentemente por las distintas bibliografías especializadas en el tema como lo es: “análisis inteligente de datos” (Berthold & Hands, 2003), que se especializa y realiza un mayor hincapié en las técnicas de análisis estadístico, realizando un conjunto de actividades o eventos (coordinados u organizados) que se realizan o suceden (alternativa o simultáneamente) con un fin determinado.

La MD toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Esta se encuentra lista para su aplicación en la comunidad de negocios porque está soportada por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de MD.

Sus algoritmos utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más confiables que métodos estadísticos clásicos.

La Minería de Datos comprende una serie de etapas:

- **Seleccionar el conjunto de datos**, tanto en lo que se refiere a las variables dependientes, como a las variables objetivo.
- **Analizar las propiedades de los datos**, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
- **Transformar el conjunto de datos de entrada**, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de Minería de Datos que mejor se adapte a los datos y al problema. En este paso uno de los principales objetivos a tener en cuenta es eliminar todo dato e información que pueda ser irrelevante.
- **Seleccionar y aplicar la técnica** de Minería de Datos, se construye el modelo predictivo, de clasificación o segmentación. Un factor a tener en cuenta antes de generar cualquier tipo de modelo, es entender cuál es la meta del proyecto de Minería de Datos. ¿Vamos a crear un proyecto para clasificar, para generar una asociación, para establecer una segmentación, o para pronosticar?
- **Evaluar los resultados** contrastándolos con un conjunto de datos previamente reservado para validar la generalidad del modelo. Existen unas pocas herramientas para evaluar la calidad de un modelo de Minería de Datos, entre las más conocidas se encuentra el gráfico de rendimiento, la

cual utiliza los datos de un modelo entrenado para predecir los valores de un conjunto de datos de evaluación.

- **Realizar un reporte o generar informes**, una de las tareas claves considerada en muchas organizaciones como la meta final de la Minería de Datos, para los responsables de los diferentes departamentos.

Figura 2 Fases del Proyecto de Minería de Datos

Muchas veces en estos proyectos, la búsqueda de patrones ocultos entre los datos responde a la necesidad de generar predicciones, por tanto es común ver la predicción como una de las tareas claves dentro de la MD para la toma de decisiones.

En algunas ocasiones, los patrones hallados no contienen información útil. Esto por lo regular es consecuencia de algunos factores como:

- Los datos utilizados son totalmente aleatorios.
- Que las variables utilizadas en el modelo no son las más precisas.
- La limpieza de datos no ha sido suficiente, y se ha de volver a la fase anterior.

El proceso de la Minería de Datos puede presentar numerosas ventajas para los usuarios decididos a aplicarla, entre las más significativas se puede mencionar:

- **Los modelos son fáciles de entender:** Personas sin un conocimiento importante de estadísticas - un analista financiero o ejecutivos en general- pueden interpretar el modelo y compararlo con sus propias ideas, es válido señalar que en las organizaciones de hoy es común que los ejecutivos entren en contacto directo con las bases de datos para obtener la información que necesitan.
- **Enormes bases de datos pueden ser analizadas:** Estas Bases de datos pueden ser enormes tanto en largo como en ancho. Por ejemplo, para cada cliente se puede tener cientos de atributos que contienen información detallada; y además tener miles de registros de clientes.
- **La Minería de Datos descubre información que no se esperaba obtener:** Con este proceso muchos modelos diferentes son validados y por consecuencia algunos resultados inesperados tienden a aparecer. En muchos estudios, se ha descubierto que combinaciones particulares de factores producen efectos inesperados que son de valor para la compañía.
- **Los modelos son confiables:** El modelo es probado y validado usando técnicas estadísticas antes de ser usado, por lo que luego las predicciones que se obtienen por el modelo son válidas y confiables.
- **Los modelos se construyen de manera rápida:** La Minería de Datos permite construir y generar modelos en sólo unos minutos u horas.

1.4.1 Procesamiento en paralelo

El procesamiento en paralelo es una técnica que ha sido utilizada durante varios años, desarrollada significativamente, desde sistemas con un único procesador hasta sistemas multiprocesadores. Los sistemas de multiprocesamiento pueden estar formados por sistemas distribuidos o por sistemas centralizados de multiprocesadores con memoria compartida, o con multiprocesadores sin memoria compartida.

Recientemente los sistemas en paralelo se han empezado a utilizar para las aplicaciones comerciales, debido a que los almacenes de datos y las técnicas de Minería de Datos hacen un fuerte uso de los mismos, tanto para acelerar el proceso de las consultas como para optimizar el rendimiento de los algoritmos respectivamente (Aular, Josefina, & Talavera Pereira, 2007).

Cuando las herramientas de Minería de Datos son utilizadas en sistemas de procesamiento en paralelo de alto rendimiento, pueden analizar bases de datos masivas en minutos. Esta capacidad de procesamiento más rápido permite a los usuarios experimentar automáticamente con más modelos para entender datos de gran complejidad, y la alta velocidad permite a los usuarios analizar inmensas cantidades de datos en poco tiempo.

El procesamiento en paralelo consiste en la ejecución simultánea de instrucciones desde el mismo programa pero en diferentes procesadores, ya sea en una computadora con múltiples procesadores o en una red de estaciones de trabajo o PCs. Implica la división del programa en múltiples procesos manejados en paralelo a fin de reducir el tiempo de ejecución.

Para el uso de este tipo de tecnología se necesita mucha sincronización y comunicación entre los diversos procesos, pues el funcionamiento del sistema puede verse seriamente afectado en gran medida si la comunicación entre los distintos servicios internos no funciona correctamente (Kioskea.net, 2009).

Entre las principales ventajas que pueden ser aprovechadas a partir de los sistemas en paralelo se encuentran el calcular gran cantidad de datos -con monoprocesadores no sería posible dado sus limitaciones- y que ese cálculo se realice de una forma acelerada. Además en el procesamiento vectorial se puede trabajar con arreglos en donde cada procesador “n” se encargue de trabajar con cada elemento del vector “+/- filas” sin necesidad de hacer recorridos en todo el arreglo, clásico del mono-procesamiento, al igual que para el cálculo matricial.

Sin embargo, el procesamiento en paralelo puede implicar una serie de dificultades a nivel de programación del software, ya que es difícil lograr una optimización en el aprovechamiento de los recursos de todos los procesadores con los que se trabajan sin que se formen cuellos de botella. Además en muchas ocasiones no es posible el trabajar con equipos multiprocesadores dado el elevado costo que esto representa, así que solo se dedica a ciertas áreas de investigación especializadas o proyectos gubernamentales o empresariales (AstroSeti, 2003).

1.4.2 Aplicaciones

Cada año, en los diferentes congresos, simposios y talleres que se realizan en el mundo se reúnen investigadores con aplicaciones muy diversas de la Minería de Datos. Principalmente en los Estados

Unidos, la Minería de Datos se ha ido incorporando a la vida de empresas, gobiernos, universidades, hospitales y diversas organizaciones que están interesadas en explorar sus bases de datos.

En los últimos tiempos ha existido un incremento en la competitividad en el marco empresarial. Este incremento de la competitividad es en gran medida resultado del marketing actual, de los nuevos canales de distribución que se disponen como Internet y de las telecomunicaciones y la aplicación de las TIC a la actividad empresarial.

Las empresas se enfrentan a una economía globalizada, y el éxito empresarial depende de la capacidad de mantener a los clientes actuales y conseguir nuevos. La Minería de Datos contiene tecnologías que permiten a las empresas analizar los factores que influyen decisivamente en estos temas y tomar decisiones exitosas para el futuro de la misma.

La Minería de Datos puede contribuir significativamente en las aplicaciones de administración empresarial basada en la relación con el cliente. En lugar de contactar con el cliente de forma indiscriminada a través de un centro de llamadas o enviando cartas, sólo se contactará con aquellos que se perciba que tienen una mayor probabilidad de responder positivamente a una determinada oferta o promoción.

Otro típico ejemplo de problema en el cual es de gran utilidad la Minería de Datos es el marketing apuntado a objetivos (en inglés, targeted marketing). La Minería de Datos permite usar datos de correos promocionales anteriores para identificar posibles objetivos, lo cual ayuda a maximizar los resultados de la inversión en futuros correos.

Este proceso genera una serie de modelos que son útiles para la toma de diferentes decisiones como por ejemplo, en la determinación de qué clientes van a ser rentables durante una ventana de tiempo -una quincena, un mes...- y sólo enviar las ofertas a las personas que es probable que sean rentables.

El ejemplo clásico de aplicación de la Minería de Datos tiene que ver con la detección de hábitos de compra en supermercados.

Otro de los servicios brindados es en la detección de patrones de fuga. En muchas industrias —como la banca, las telecomunicaciones, entre otras— existe un comprensible interés en detectar cuanto antes aquellos clientes que puedan estar pensando en rescindir sus contratos para, posiblemente, pasarse a la competencia. A estos clientes —y en función de su valor— se les podrían hacer ofertas personalizadas, ofrecer promociones especiales, etc., con el objetivo último de retenerlos. La Minería de Datos ayuda a

determinar qué clientes son los más proclives a darse de baja estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que, efectivamente, se dieron de baja en el pasado.

Una compañía operadora de telefonía móvil española, se lanzó a la tarea de realizar un estudio de su personal basándose principalmente en dos puntos: el análisis del perfil de los clientes que se dan de baja y la predicción del comportamiento de sus nuevos clientes, el cual arrojó el siguiente resultado dando una muestra de los valiosos resultados que se pueden obtener de la aplicación de una correcta herramienta de Minería de Datos.

Se analizaron los diferentes históricos de clientes que habían abandonado la operadora (12,6%) y de clientes que continuaban con su servicio (87,4%). También se analizaron las variables personales de cada cliente (estado civil, edad, sexo y nacionalidad). De igual forma se estudiaron, para cada cliente, la morosidad, la frecuencia y el horario de uso del servicio, los descuentos y el porcentaje de llamadas locales, interprovinciales, internacionales y gratuitas. Al contrario de lo que se podría pensar, los clientes que abandonaban la operadora generaban ganancias para la empresa, sin embargo, una de las conclusiones más importantes radicó en el hecho de que los clientes que se daban de baja recibían pocas promociones y registraban un mayor número de incidencias respecto a la media. De esta forma se recomendó a la operadora hacer un estudio sobre sus ofertas y analizar profundamente las incidencias recibidas por esos clientes. Al descubrir el perfil que presentaban, la operadora tuvo que diseñar un trato más personalizado para sus clientes actuales con esas características.

Un caso análogo es el de la detección de transacciones de blanqueo de dinero o de fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil e, incluso, en la relación de los contribuyentes con el fisco. Generalmente, estas operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con un alto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para combatirlas de una forma rápida y efectiva.

En el 2001, las instituciones financieras a escala mundial perdieron más de 2.000 millones de dólares estadounidenses en fraudes con tarjetas de crédito y débito. El Falcon Fraud Manager es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para detectar y mitigar fraudes. En un principio estaba pensado, en instituciones financieras sólo de Norteamérica, para detectar fraudes en tarjetas de crédito. Sin embargo, actualmente se le han incorporado funcionalidades de análisis

en las tarjetas comerciales, de combustibles y de débito. El sistema Falcon ha permitido ahorrar más de seiscientos millones de dólares estadounidenses cada año y protege aproximadamente más de cuatrocientos cincuenta millones de pagos con tarjeta en todo el mundo –aproximadamente el 65% de todas las transacciones con tarjeta de crédito (Molina Félix, 2002).

En estos últimos años se ha podido presenciar la revolución de las tecnologías de la información y de las comunicaciones, y aparejado a esta, el desarrollo y auge del gigante de la red "La Internet". Sin duda uno de los factores desconocidos hasta hace no muchos años era el comportamiento seguido por los usuarios en los sitios web, algo sin dudas de mucha utilidad para el creador.

En la actualidad, podemos contar con la Minería de Datos para realizar un análisis y estudio del mismo, permitiendo obtener el comportamiento de estos visitantes, sobre todo, cuando son clientes potenciales. En muchos casos se hace posible además ofrecerles propaganda adaptada específicamente a su perfil, de acuerdo a la información manejada por los mismos dentro del sitio web, y hasta la sugerencia de nuevos productos a partir de la adquisición de uno, de acuerdo a la información histórica disponible acerca de los clientes que han adquirido ese tipo de producto y su preferencias por otros.

La Minería de Datos también se ha estado utilizando ampliamente en estas últimas décadas en diversas áreas relacionadas con la ciencia y la ingeniería.

Uno de los ejemplos de aplicación en estos campos lo es la Biotecnología. La Minería de Datos es usada en el estudio de la genética humana, el objetivo principal es entender la relación cartográfica entre las partes y la variación individual en las secuencias del ADN humano y la variabilidad en la susceptibilidad a las enfermedades. En términos más comunes, se intenta saber cómo los cambios en las secuencias de ADN de un individuo afectan al riesgo de desarrollar enfermedades comunes (como por ejemplo el cáncer). Esto es muy importante para ayudar a mejorar el diagnóstico, prevención y tratamiento de las enfermedades. La técnica de Minería de Datos aplicada para realizar esta tarea se conoce como "Reducción de Dimensionalidad Multifactorial" (Zhu & Davidson, 2007).

En el ámbito de la ingeniería eléctrica, las técnicas de Minería de Datos han sido ampliamente utilizadas para monitorizar las condiciones de las instalaciones de alta tensión. La finalidad de esta monitorización es obtener información valiosa sobre el estado del aislamiento de los equipos. Por ejemplo, se utilizan técnicas de agrupación de datos para la vigilancia de las vibraciones o el análisis de los cambios de carga

en transformadores, lo cual sirve para detectar condiciones anormales y para estimar la naturaleza de dichas anomalías.

En investigaciones espaciales, también se puede encontrar sus aplicaciones. Durante seis años, el Second Palomar Observatory Sky Survey (POSS-II) coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23.040 x 23.040 píxeles por imagen. El objetivo era formar un catálogo de todos esos objetos. El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación y árboles de decisión para poder clasificar los objetos en estrellas, planetas, sistemas y galaxias con una alta confiabilidad (Fayyad, Piatetsky-Shapiro, Smith, & R., 1996). Los resultados han ayudado a los astrónomos a descubrir dieciséis nuevos quásares con corrimiento hacia el rojo que los incluye entre los objetos más lejanos del universo y, por consiguiente, más antiguos. Estos quásares son difíciles de encontrar y permiten saber más acerca de los orígenes del universo.

La aplicación de técnicas de Minería de Datos en el ámbito bibliotecario se conoce con el nombre de bibliomining (Nicholson, 2006). La llegada de las nuevas tecnologías de la Información y las comunicaciones a las bibliotecas ha potenciado la búsqueda de patrones de comportamiento en los datos que se manejan.

1.5 SISTEMAS DE INFORMACIÓN

La información reduce la incertidumbre sobre algún aspecto de la realidad y por tanto, nos permite tomar mejores decisiones.

Un sistema de información está integrado por un conjunto de elementos capacitados para el tratamiento y administración de datos e información, organizados y listos para su posterior uso, generados con el propósito de satisfacer una necesidad u objetivo. Dichos elementos se encuentran dentro de algunas de las siguientes categorías:

- Personas.
- Datos.
- Actividades o técnicas de trabajo.

- Recursos materiales en general, típicamente recursos informáticos y de comunicación

Todos estos elementos interactúan entre sí para procesar los datos, ya sea a través de procesos manuales o automáticos dando lugar a información más elaborada y distribuyéndola de la manera más adecuada posible en una determinada organización en función de sus objetivos.

Inicialmente la finalidad de los sistemas de información era recopilar información para ayudar en la toma de decisiones, pero actualmente, con la informatización de las organizaciones y la aparición de aplicaciones de software, la finalidad principal se ha convertido en dar soporte a los procesos básicos de la organización, dígase ventas, producción, personal; pasando a reconocerse los mismos como *Sistemas de Información para la Gestión*.

En los últimos 50 años ha evolucionado el análisis de la información:

- Década 1960: Informes Batch, la información es difícil de encontrar y analizar, poco flexible y se necesita reprogramar cada petición.
- Década 1970: Primeros **DSS (Decision Support Systems)** y **EIS (Executive Information Systems)**, basados en terminal y no integrados con el resto de herramientas.
- Década 1980: Acceso a datos y herramientas de análisis integradas, conocidas como intelligent business tools. Herramientas de consultas e informes, hojas de cálculo, interfaces gráficos e integrados y fáciles de usar.
- Década 1990: Almacenes de Datos y herramientas **OLAP (On-Line Analyitical Processing).**
- Década 2000: Herramientas de Minería de Datos y Simulación.

Según esta cronología se reconoce a la Minería de Datos como el último escalón en este proceso.

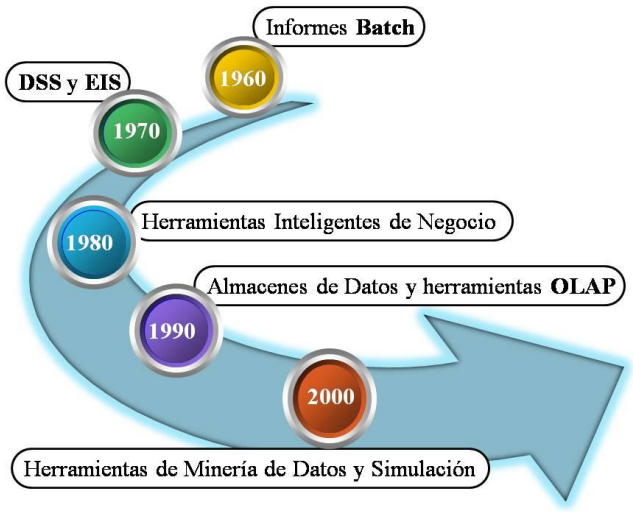


Figura 3 Evolución del análisis de la información en los últimos 50 años

Como se aprecia en la figura han aparecido diferentes herramientas de negocio o sistemas de decisión que coexisten, considerándose las más reconocidos: EIS, OLAP, y la MD.

Un EIS (Executive Information System) es un sistema de información y un conjunto de herramientas asociadas:

- Proporciona a los directivos acceso a la información de estado y sus actividades de gestión.
- Está especializado en analizar el estado diario de la organización, mediante indicadores clave, para informar rápidamente sobre cambios a los directivos.
- La información solicitada suele ser, en gran medida, numérica -ventas semanales, nivel de stocks, balances parciales- y representada de forma gráfica al estilo de las hojas de cálculo.

Las herramientas OLAP (On-Line Analyitical Processing) son más genéricas:

- Funcionan sobre un sistema de información transaccional o almacén de datos.
- Permiten realizar agregaciones y combinaciones de los datos de maneras complejas y ambiciosas, con objetivos de análisis estratégicos.
- Proporcionan facilidades para manejar y transformar los datos.
- Están basadas, generalmente, en sistemas o interfaces multidimensionales.

- Utilizando operadores específicos, además de los clásicos: drill, roll, pivot, slice and dice.
- El resultado se presenta de una manera matricial o híbrida.

La Minería de Datos se diferencia del resto de las otras herramientas presentadas en el sentido de que no transforma la información, sino que la analiza, facilitando el acceso a esta para que el usuario la analice fácilmente.

1.6 CONCLUSIONES

El desarrollo de la tecnología de Minería de Datos está en un punto de inflexión, con respecto a su consolidación, en las aplicaciones. Existen un grupo de elementos que la hacen aplicable, y una realidad que la demanda; sin embargo, existe una serie de retos que atentan contra su credibilidad:

- Los productos comercializados son costosos, por tanto los consumidores pueden hallar una relación coste/beneficio improductiva.
- Se necesita mucha experiencia para utilizar herramientas de la tecnología, o que sea fácil hallar patrones equívocos, triviales o no interesantes.
- Es posible no hallar patrones en tiempo o en espacio.
- No se establece una adecuada comunicación en los equipos multidisciplinarios para elegir la herramienta adecuada y que, por lo tanto, no se alcancen los resultados esperados.
- Pueden existir razones organizativas, éticas o de otro carácter que impidan la utilización de toda la información necesaria para la aplicación de estas herramientas.

Existen grandes expectativas en estos primeros años del siglo XXI con su impacto, pues este es un período de enorme importancia para las aspiraciones de aplicar las herramientas de Minería de Datos a nivel mundial, pues el proceso en sí ofrece una serie de ventajas imposibles de no apreciar:

- 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.
- Los modelos descriptivos que se generan permiten a empresas, explorar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.

- Los modelos predictivos generados permiten que relaciones no descubiertas a través de este proceso sean expresadas como reglas del negocio.

CAPÍTULO 2: ANÁLISIS DE LAS TÉCNICAS Y HERRAMIENTAS DE LA MINERÍA DE DATOS.

2.1 INTRODUCCIÓN

El presente capítulo tiene entre sus objetivos de mayor importancia mostrar las técnicas y algoritmos de Minería de Datos, las cuales suelen ser muy diversas, así como las principales herramientas implementadas para la aplicación de esta novedosa técnica.

Sus algoritmos utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables y entendibles que consistentemente son más confiables que métodos estadísticos clásicos.

La Minería de Datos puede ser dividida para su estudio esencialmente en dos grandes ramas, la Minería de Datos Predictiva, usada principalmente en técnicas estadísticas, y la Minería de Datos para el Descubrimiento de Conocimiento, la cual se centra en técnicas de inteligencia artificial (Haag, 2007).

Los algoritmos de Minería de Datos realizan en general tareas de descripción -de datos y patrones-, de predicción -datos desconocidos- y de segmentación –datos-. Otras tareas como análisis de dependencias e identificación de anomalías se pueden utilizar tanto para descripción como para predicción.

Es muy interesante abordar que para aplicar cualquiera de los algoritmos tratados en este capítulo se debe tener en cuenta qué, la posible elección del mejor algoritmo para una tarea específica puede convertirse en una tarea muy difícil, a veces un desafío, debido a una serie de términos requeridos para cumplimentar un objetivo dado.

Aunque se puede utilizar diferentes algoritmos para realizar la misma tarea, cada uno de ellos genera un resultado diferente, y algunos pueden generar más de un tipo de resultado. Por ejemplo, se puede usar el algoritmo árboles de decisión de Microsoft no sólo para la predicción, sino también como una forma de reducir el número de columnas de un conjunto de datos, ya que el árbol de decisión puede identificar las columnas que no afectan al modelo de Minería de Datos final.

Tampoco es necesario usar los algoritmos de modo independiente. En una única solución de Minería de Datos se pueden usar algunos algoritmos para explorar datos y, posteriormente, usar otros algoritmos para predecir un resultado específico a partir de esos datos. Por ejemplo, puede utilizar un algoritmo de

agrupación en clústeres, que reconoce patrones, para dividir los datos en grupos que sean más o menos homogéneos, y luego usar los resultados para crear un mejor modelo de árbol de decisión (Microsoft Corporation, 2010).

2.2 TÉCNICAS Y ALGORITMOS DE MINERÍA DE DATOS

Las técnicas de la Minería de Datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos sofisticados que se aplican sobre un conjunto de datos para obtener resultados.

Las funciones de la Minería de Datos pueden ser clasificadas generalmente siguiendo varios criterios: supervisadas y no supervisadas, predictivas y descriptivas, transparentes u opacas. En algunos niveles cuando las funciones de la Minería de Datos son implementadas usando uno o más algoritmos, la elección de estos algoritmos se hace, tomando en cuenta las características de las dimensiones a las cuales pertenecen (Hornick, Marcadé, & Venkayala, 2007).

La Minería de Datos Descriptiva muestra los resultados del modelo de una forma transparente, lo que se traduce, en la habilidad de entender porque el modelo trabaja en la forma en que lo hace.

Hasta que nivel un modelo puede ser descriptivo depende en cierta manera del algoritmo usado para su producción. Por ejemplo, un árbol de decisión típicamente proporciona reglas interpretables explicando las razones que dieron origen a la predicción; mientras que si se usa una red neuronal, con los mismos datos, no se puede discernir prontamente ningún entendimiento del porque se realizó de esa forma, por lo que este tipo de algoritmos son clasificados como opacos (“black-box”, término designado en inglés).

La mayoría de las funciones no supervisadas, como las reglas de agrupamiento o asociación, son consideradas por definición descriptivas. Algunas técnicas y algoritmos pueden pertenecer a ambas categorías descriptivas y predictivas, como por ejemplo los árboles de decisión.

Las técnicas predictivas realizan inferencia en los datos disponibles e intentan predecir resultados o asignaciones para los nuevos datos. Estas además son capaces de proveer una probabilidad o confidencia de la predicción basadas en la fortaleza de lo que aporta el modelo. Por ejemplo, a un modelo de agrupamiento se le puede asignar un caso con 5 grupos con un 95 por ciento de probabilidad –una fuerte asignación –, o un modelo de clasificación puede predecir si un cliente compra con un 55 por ciento de probabilidad – una débil predicción –.

Las técnicas con aprendizaje supervisado por definición son de tipo predictivo. Su función es predecir el valor de un atributo de un conjunto de datos – atributo descriptivo – a partir de estos datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida (Bressán, 2003).

Este tipo de técnicas son usadas típicamente para predecir un valor y exigir al usuario la especificación del conocimiento resultante u objetivo propuesto a partir de cada modelo construido. Alguno de los valores posibles a obtener, pueden ser a partir de atributos binarios –valores simples–, los cuales podrían ser por ejemplo si se compró/no compró, si se tuvo éxito/fracaso, o a partir de atributos multievaluados, los cuales como su nombre dice, pueden tomar varios valores, por ejemplo, el nivel de aceptación del incremento del salario en $30.00, la reacción esperada ante una droga como altamente favorable/sin reacción/desfavorable/bajamente favorable, o el color favorito de los usuarios del Yahoo. Estos valores posibilitan al algoritmo supervisado aprender a partir de los resultados esperados.

Muchos de estos algoritmos valoran sus predicciones a partir de los valores propiciados por sus modelos construidos y reajustan los modelos resultantes acorde con los mismos.

El aprendizaje supervisado consta principalmente de dos fases:

- Entrenamiento: Se llama entrenamiento a la construcción de un modelo usando un subconjunto de datos con etiqueta conocida.
- Prueba: Son las pruebas del modelo sobre el resto de los datos.

Cuando una aplicación no es lo suficientemente madura y por lo tanto no tiene el potencial necesario para una solución predictiva, es necesario recurrir a los métodos de aprendizaje no supervisados o de Descubrimiento del Conocimiento, los cuales descubren patrones y tendencias en los datos actuales, pues no utilizan datos históricos. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio, ya sea científico o de negocio de la misma (Moreno García, Miguel Quintales, García Peñalvo, & Polo Martín).

Las funciones no supervisadas de cierta forma no usan blancos específicos, pues su objetivo es encontrar la estructura intrínseca, las relaciones y afinidades entre el conjunto de datos. Estas cubren una gama amplia de capacidades analíticas.

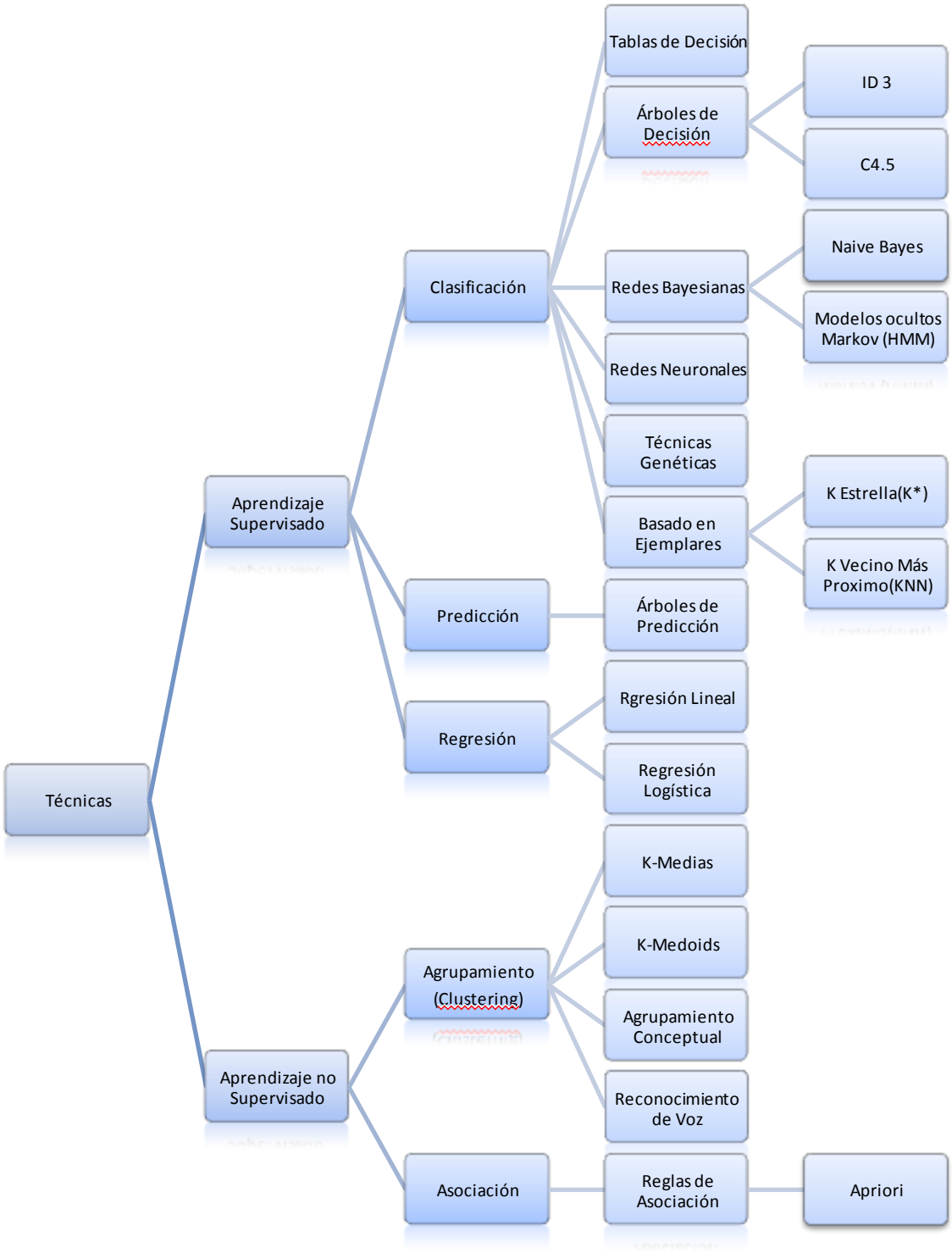


Figura 4 Técnicas y Algoritmos de Minería de Datos

Como se puede apreciar en la figura 4 las técnicas de aprendizaje supervisado se dividen en:

- Clasificación, predice los valores categóricos.
- Regresión, predice los valores continuos.
- Predicción, predice los valores en los datos y en las clases.

La clasificación y la regresión presentan diferencias en cuanto a la evaluación que se tiene de la calidad en los modelos.

Mientras que las de aprendizaje no supervisado lo hacen en:

- Agrupamiento: identifica ocurrencias naturales dentro de un grupo de datos -por ejemplo proteínas similares o células cancerosas-.
- Asociación: las reglas obtenidas a partir de sus modelos pueden servir para establecer relaciones del tipo producto – cliente en sitios de comercio electrónico.

Normalmente un algoritmo tiene tres componentes esenciales:

- El modelo.
- Criterio de preferencia o elección.
- El algoritmo de búsqueda.

Los algoritmos según las topologías del modelo pueden ser clasificados de acuerdo a dos categorías:

- Por su función: pueden ser de clasificación, regresión, agrupamiento, de generación de reglas, reglas de asociación y modelos de dependencia o análisis de secuencias.
- Por su representación: pueden ser redes neuronales, árboles de decisión y discriminación lineal.

2.2.1 Técnicas de Aprendizaje Supervisado.

2.2.1.1 Clasificación

La clasificación se inicia con un conjunto de datos preclasificado, es decir se cuenta con un conjunto de datos que no sólo se conocen las variables a utilizar en la clasificación, sino que también se conocen las clases a las que pertenecen estas variables. El objetivo es sobre la base de esta información crear un modelo capaz de predecir la clase de un nuevo registro (Marante Jacas & Marante Jacas, 2009).

Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas, es decir obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta.

Consiste en definir una serie de clases, donde se pueda agrupar a los diferentes clientes. Por ejemplo: definida unas variables de entrada se produce una determinada salida que clasifica al cliente en un grupo o en otro.

Algunos de los principales algoritmos que cumplen con estas características son:

- Tabla de Decisión
- Árboles de Decisión
- Reglas de Inducción
- Redes Bayesianas
- Redes Neuronales
- Lógica Borrosa
- Técnicas Genéticas
- Aprendizaje Basado en Ejemplares

Tablas de Decisión

La tabla de decisión constituye la forma más simple y rudimentaria de representar la salida de un algoritmo de aprendizaje, que es justamente representarlo como la entrada.

Esta técnica consiste en seleccionar subconjuntos de atributos y calcular su precisión para predecir o clasificar los ejemplos. Una vez seleccionado el mejor de los subconjuntos, la tabla de decisión estará formada por los atributos seleccionados junto a la clase, en la que se insertarán todos los ejemplos de entrenamiento únicamente con el subconjunto de atributos elegido.

Si hay dos ejemplos con exactamente los mismos pares atributo-valor, para todos los atributos del subconjunto, la clase que se elija será la media de los ejemplos, en el caso de una clase numérica; o la que mayor probabilidad de aparición tenga, en el caso de una clase simbólica (Molina López & García Herrero, 2004).

Redes Bayesianas

La clasificación Bayesiana se basa en el teorema de Bayes, y los clasificadores Bayesianos han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos. Diferentes estudios comparando los algoritmos de clasificación han determinado que un clasificador Bayesiano sencillo conocido como el clasificador "Naive Bayes" es comparable en rendimiento a un árbol de decisión y a clasificadores de redes neuronales.

El algoritmo de redes bayesianas es el primero que suele utilizarse para la exploración inicial de los datos. Básicamente este algoritmo busca correlaciones entre atributos.

Cuando no se sabe qué atributo se puede predecir en función de otros, una técnica muy habitual es tratar de utilizar el algoritmo de Redes Bayesianas tratando de predecir el valor de todos los atributos en función de todos los atributos.

Con el teorema de Bayes se puede estimar la probabilidad posterior de la hipótesis dados los datos – P (H/D)-, a continuación la ecuación:

$$P\left(\frac{H}{D}\right) = \frac{P\left(\frac{D}{H}\right)P(H)}{P(D)}$$

Donde:

P (D/H): es la probabilidad de los datos dada una hipótesis

P (D): la probabilidad a priori de los datos -cuales datos son más probables que otros-.

El clasificador Bayesiano asume que los valores de los atributos son condicionalmente independientes dado el valor de la clase, por lo que el efecto de un valor del atributo en una clase dada es independiente de los valores de los otros atributos. Esta suposición se llama "independencia condicional de clase".

Esto hace más simplificados los cálculos involucrados y, en este sentido, es considerado "ingenuo". Esta asunción es una simplificación de la realidad. A pesar de la simplificación realizada, las redes bayesianas funcionan muy bien, sobre todo cuando se filtra el conjunto de atributos seleccionado para eliminar redundancia, con lo que se elimina también dependencia entre datos.

Es una de las técnicas más populares para clasificación de textos.

Modelos ocultos de Markov

Un modelo oculto de Markov (HMM, en inglés) es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. Es considerado la red bayesiana más simple.

El objetivo es determinar los parámetros ocultos a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo sucesivos análisis, por ejemplo en aplicaciones de reconocimiento de formas.

Aprendizaje Basado en Ejemplares

El aprendizaje basado en ejemplares o instancias tiene como principio de funcionamiento, en sus múltiples variantes, el almacenamiento de ejemplos: en unos casos todos los ejemplos de entrenamiento, en otros sólo los más representativos, en otros los incorrectamente clasificados cuando se clasifican por primera vez.

La clasificación posterior se realiza por medio de una función que mide la proximidad o parecido. Dado un ejemplo para clasificar se le clasifica de acuerdo al ejemplo o ejemplos más próximos. El sesgo que rige este método es la proximidad; es decir, la generalización se guía por la proximidad de un ejemplo a otros.

Se han enumerado ventajas e inconvenientes del aprendizaje basado en ejemplares, pero se suele considerar no adecuado para el tratamiento de atributos no numéricos y valores desconocidos. Las mismas medidas de proximidad sobre atributos simbólicos suelen proporcionar resultados muy dispares en problemas diferentes (Silveira Martineaux & Fernández Pérez, 2008).

Algunos de los algoritmos que pertenecen a esta clasificación son:

- **K vecinos más próximos (KNN, K-Nearest Neighbor)**
- K Estrella (K*)

K vecinos más próximos

Es considerado como un buen representante de este tipo de aprendizaje, y es de gran sencillez conceptual. Se suele denominar método porque es el esqueleto de un algoritmo que admite el intercambio de la función de proximidad dando lugar a múltiples variantes.

La función de proximidad puede decidir la clasificación de un nuevo ejemplo atendiendo a la clasificación del ejemplo o de la mayoría de los k ejemplos más cercanos. Admite también funciones de proximidad que consideren el peso o coste de los atributos que intervienen, lo que permite, entre otras cosas, eliminar los atributos irrelevantes.

Una función de proximidad clásica entre dos instancias X_i y X_j es la distancia euclidiana[1].

Generalmente se utiliza en bases de datos históricas

Pseudocódigo:

Inicio
Entrada: D= {(X_1,C_1),…, (X_n,C_n)}
X= (X_1,…,X_n) nuevo caso a clasificar
Para todo objeto clasificado (X_i,C_i) calcular D_i = D(X_i,X)
Ordenar D_i (i=1,…,n) de menor a mayor
Quedarse con los K casos D_X^K ya clasificados más cercanos a X
Asignar a X la clase más frecuente en D_X^K

Algoritmo K Estrella

Es un algoritmo basado en ejemplares en el que la medida de la distancia entre ejemplares se basa en la teoría de la información, donde la distancia entre dos ejemplares se define como la complejidad de transformar un ejemplar en el otro.

El cálculo de la complejidad se basa en definir un conjunto de transformaciones T= {t1, t2,..., tn, σ} para pasar de un ejemplo -valor del atributo- α a uno β. La transformación σ es la de parada y es la transformación identidad -σ (α)=α-.

Técnicas Genéticas

Los algoritmos genéticos son algoritmos de búsqueda basados en los mecanismos de selección natural y genética natural, son una robusta herramienta de optimización que emula la evolución biológica, pueden ser usados para resolver un amplio banco de problemas de forma eficiente y precisa.

[1] ***Ver Distancia Euclidiana dentro de la Sección Agrupamiento en Técnicas de Aprendizaje no Supervisado***

Es importante recalcar, que el uso de algoritmos genéticos no asegura hallar el óptimo total de un problema, pero pueden proporcionan soluciones bastante próximas al mismo (Rojas Figueroa, 2009).

Son métodos de búsqueda dirigida basados en probabilidad. El algoritmo trabaja bajo una condición muy débil, debe mantener elitismo, es decir, guardar siempre al mejor elemento de la población sin hacerle ningún cambio.

A partir de ellos se realizan programas y optimizaciones que pueden ser usadas en la construcción y entrenamiento de otras estructuras como las redes neuronales.

Árboles de Decisión

Los árboles de decisión pueden ser clasificados como herramientas de clasificación y predicción, pero sobre todo son una potente herramienta de clasificación, probablemente el mejor algoritmo que se pueda utilizar para la clasificación.

Un árbol de decisión es un algoritmo de toma de decisiones, que representa la información en forma de conocimiento. También es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, el cual a partir de una base de datos construye diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que representan y categorizan una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema (Molina López & García Herrero, 2004).

Un árbol de decisión puede tener como entradas un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta, la cual en ciertas circunstancias llega a ser una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos[2] o continuos[3], aunque se utilizan más los valores discretos por su simplicidad. Es importante señalar que cuando se utilizan valores discretos se denomina clasificación y cuando se utilizan los continuos se denomina regresión.

Su estructura como bien puede ser deducida de su nombre, es en forma de árbol donde cada nodo es una decisión, los cuales a su vez generan reglas para la clasificación de un conjunto de datos. Este contiene distintos tipos de nodos y las ramas:

- Nodos internos: contienen una prueba sobre algún valor de una de las propiedades

[2] **Un atributo discreto *tiene un número finito o contable de valores***
[3] **Un atributo continuo *tiene un número infinito de valores posibles***

- Nodos de probabilidad: indican que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, estos tipos de nodos son redondos, los demás son cuadrados.
- Nodos hojas: representa el valor que devolverá el árbol de decisión
- Ramas: brindan los posibles caminos que se tienen de acuerdo a la decisión tomada.

En estos árboles se crean varias particiones y siempre se intentará tomar la partición que se encuentre más homogénea en torno a la variable de salida. Para evaluar la calidad de las distintas particiones se utilizan diferentes medidas, las más comunes son:

- Coeficiente de Gini P1 (1-P1): Normalmente asociado al algoritmo **Classification And Regression Trees (CART)**.
- Criterio de reducción de entropía -P1logP1-P2logP2: Normalmente asociado al algoritmo C4.5.

Si se desea medir su posible taza de error, es decir, la calidad que presenta el árbol, sólo es necesario aplicar un conjunto de datos nuevos para el árbol y después medir el porcentaje de registros correctamente clasificados. Dicho proceso es conocido como validación cruzada.

A pesar del interés de los árboles de decisión como herramientas de representación de conocimiento, el principal interés desde el punto de vista de la Minería de Datos es la capacidad de inducir estos árboles automáticamente de grandes volúmenes de datos (Marante Jacas & Marante Jacas, 2009).

Dentro de sus ventajas se encuentran que son muy fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes, son computacionalmente rápidos. Además representan las relaciones mediante reglas que son fáciles de interpretar, permitiendo representar los resultados en lenguajes naturales, de allí que los conocimientos obtenidos se puedan representar en cualquier lenguaje de bases de datos como SQL. A esto se le suma que brindan una gran facilidad de interpretación y el ser robusto frente a datos atípicos u observaciones mal etiquetadas.

Los árboles de decisión tienen la capacidad de utilizar datos binarios y no categóricos sin necesidad de transfórmalos, esta es una característica muy importante en la medida que la mayoría de las restantes aplicaciones tienen problemas para tratar con este tipo de variables. Además son insensibles a variables de escala, por lo que pueden ser usadas sin ningún tipo de normalización.

Aunque presentan algunas pequeñas desventajas:

- Las reglas de asignación son bastantes sensibles a pequeñas perturbaciones en los datos (inestabilidad).
- Dificultad para elegir el árbol óptimo.
- Ausencia de una función global de las variables y como consecuencia pérdida de la representación geométrica.
- Requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hojas sea significativa.

Algunos algoritmos que pertenecen a esta clasificación son:

- Algoritmo ID3
- Algoritmo C4.5

Algoritmo ID3

Este algoritmo es usado en la búsqueda de hipótesis o reglas, dado un conjunto de ejemplos. El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, el atributo a clasificar es el objetivo, el cual es de tipo binario -positivo o negativo, sí o no, válido o inválido-.

Así el algoritmo puede conseguir las hipótesis que clasifiquen ante nuevas instancias, es decir, conocer si dicho ejemplo va a ser positivo o negativo.

ID3 realiza toda esta labor descrita a través de la construcción de un árbol de decisión.

Pseudocódigo del Algoritmo

```
Id3 (Ejemplos, Atributo-objetivo, Atributos)
  Si todos los ejemplos son positivos devolver un nodo positivo
  Si todos los ejemplos son negativos devolver un nodo negativo
    Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo en
       Ejemplos.
  En otro caso
     Sea A Atributo el MEJOR de atributos
     Para cada v valor del atributo hacer
        Sea Ejemplos (v) el subconjunto de ejemplos cuyo valor de atributo A es v
                Si Ejemplos (v) está vacío devolver un nodo con el voto mayoritario del Atributo
                   objetivo de Ejemplos
        Sino Devolver Id3 (Ejemplos (v), Atributo-objetivo, Atributos/ {A})
```

Obsérvese que la construcción del árbol se hace de forma recursiva, siendo las tres primeras líneas y la penúltima, los casos bases que construyen los nodos hojas.

Algoritmo C4.5

C4.5 es un algoritmo utilizado para generar un árbol de decisiones desarrollado por Ross Quinlan.

Este algoritmo construye sus árboles de decisión a partir de un conjunto de datos de entrenamiento de la misma forma que ID3, aplicando el concepto de entropía[4] de la información.

En cada nodo del árbol, C4.5 elige uno de los atributos de los datos que mayor efectividad tiene en dividir su conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio, la ganancia de la información normalizada, diferencia de entropía, es el resultado de la elección de un atributo para separar los datos. El atributo que obtiene la mayor información normalizada es elegido para tomar la decisión. Después el algoritmo se ejecuta de forma recursiva en las sublistas.

Los casos bases son:

- Todas las muestras en la lista pertenecen a la misma clase. Cuando esto sucede, simplemente se crea un nodo hoja para indicar al árbol la selección de esa clase.
- Ninguna de las características proporciona ningún beneficio en la información. En este caso, C4.5 crea un nodo de decisión encima usando el valor esperado de la clase.
- Instancia de clase previamente-no vista encontrada. C4.5 nuevamente crea un nodo de decisión más arriba en el árbol usando el valor esperado.

C4.5 hizo una serie de mejoras al algoritmo ID3:

- La manipulación de atributos continuos y discretos. Con la finalidad de manejar los atributos continuos, se crea un umbral y luego se divide la lista en aquellos cuyo valor es mayor que el umbral y en los que son iguales o están por debajo del mismo.
- Gestionar los datos de formación con atributos cuyo valor falta. C4.5 permite atributos cuyo valor no existe, simplemente estos no son usados luego para el cálculo de la ganancia y entropía.
- Manipulación de Atributos con diferentes costos.

[4] **Entropía:** ***es una medida de la incertidumbre asociada a una variable aleatoria.***

- Poda de árboles después de ser creados. El algoritmo va hacia arriba en el árbol y elimina las ramas que han sido creadas y que ya no son de utilidad, reemplazándolas por nodos hojas.

Redes Neuronales

Las redes neuronales son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales, específicamente el funcionamiento del cerebro humano. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

El principal desafío que afrontó esta técnica fue la forma de representar computacionalmente el proceso de aprendizaje de un ser humano, para lograr el objetivo de crear fuertes aplicaciones "inteligentes", basadas en la combinación de neuronas interconectadas en una red, capaces de resolver problemas relacionados con reconocimientos de formas, patrones y predicciones, entre muchos otros.

Esta técnica de inteligencia artificial, en los últimos años se ha convertido en uno de los instrumentos de uso frecuente para detectar categorías comunes en los datos, esta técnica se asemeja a los árboles de decisión, en cuanto a que resuelve problemas de clasificación y regresión también, debido a que son capaces de detectar y aprender complejos patrones no lineales, difícilmente descriptibles por medios de reglas y características de los datos. Son excelentes clasificadores superiores a otros como: los mencionados árboles de decisión, algoritmos basados en ejemplares y redes de cuantización vectorial.

Una de las principales características de las redes neuronales, es que son capaces de trabajar con datos incompletos e incluso paradójicos, que dependiendo del problema puede resultar una ventaja o un inconveniente. Además pueden aprender de la experiencia, generalizar de casos anteriores a nuevos casos y abstraer características esenciales a partir de entradas que representan información irrelevante.

El hecho de que las redes neuronales estén basadas en la estructura del sistema nervioso, hace que presenten las mismas características, por lo que la unidad central de la red es la neurona y esta tiene la obligación de comunicarse. Para lograr la analogía sináptica y analógica, las señales llegadas a la sinapsis son las entradas de las neuronas, estas son ponderadas –atenuadas o simplificadas- a través del parámetro denominado peso, asociado a esa sinapsis. La señal tiene la habilidad de excitar a la neurona –peso positivo- o de inhibirla –peso negativo-. Si la sumatoria de todas las entradas ponderadas es mayor que el umbral entonces la neurona se activa –dando salida-, en el caso contrario se cierra.

La sinapsis es susceptible a diversos eventos como la fatiga, deficiencia de oxígeno, la falta de uso, entre otras, esta habilidad de ajustar la señal es un mecanismo de aprendizaje.

Esta técnica posee dos formas de aprendizaje: supervisado y no supervisado.

Debido a su utilidad existen en la actualidad decenas de miles de estas, pero todas mantienen los elementos básicos, ya que estos son los que le permiten reproducir un comportamiento parecido al cerebro humano.

Algunos ejemplos de redes neuronales son:

- El Perceptrón – Perceptrón Simple-.
- El Perceptrón Multicapa.
- **Los Mapas auto organizados (SOM)**, también conocidos como redes de Kohonen.

2.2.1.2 Regresión

La regresión consta esencialmente en obtener un modelo que permita predecir el valor numérico de alguna variable. El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo.

A la combinación de los datos de origen y los datos de la predicción se le denomina serie.

Un ejemplo de algoritmo de regresión es el Algoritmo de serie temporal de Microsoft.

Algoritmo serie temporal:

El algoritmo de serie temporal de Microsoft proporciona los algoritmos de regresión que se optimizan para la previsión en el tiempo de valores, una característica importante de este algoritmo es su capacidad para llevar a cabo predicciones cruzadas -la predicción cruzada también es útil para crear un modelo general que se puede aplicar a múltiples series-.

Los modelos de serie temporal no requieren columnas adicionales de nueva información como entrada para predecir una tendencia. Un modelo de serie temporal puede predecir tendencias basadas únicamente en el conjunto de datos original utilizado para crear el modelo. Es posible también agregar

nuevos datos al modelo al realizar una predicción e incorporar automáticamente los nuevos datos en el análisis de tendencias.

2.2.1.3 Predicción

Consiste en intentar conocer resultados futuros a partir de la modelación de los datos actuales. Por ejemplo si se crea un modelo de variables para saber si el cliente compra o no compra, al aplicar dicho modelo a un futuro cliente, se puede predecir si comprará o no.

Se refiere tanto a la predicción de valores en los datos como a la predicción de clases utilizando la identificación de distribuciones en los datos disponibles.

2.2.2 Técnicas de Aprendizaje no Supervisado.

Las principales técnicas de aprendizaje no supervisado son agrupamiento y asociación, aunque también se pueden encontrar otras mucho menos conocidas como los patrones secuenciales, extracción de características y detección anómala, producto de no ser cubiertas de una manera oficial todavía por la Minería de Datos.

2.2.2.1 Agrupamiento

Agrupamiento es una técnica más de aprendizaje automático, en la que el aprendizaje realizado es no supervisado, puesto que busca encontrar relaciones entre variables descriptivas pero no la que guardan con respecto a la variable objetivo. Es una técnica bastante específica y utilizada mayormente para detectar secuencias típicas dentro de un conjunto de eventos.

Fundamentalmente consiste en agrupar datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, aunque existen otras más robustas o que permiten extenderla a variables discretas, de manera que las clases sean similares entre sí y distintas con las otras clases. Estas clases se obtienen directamente de los datos de entrada usando medidas de similaridad, creando grupos a su vez lo más parecido posible y a su vez lo más distinto posible a otros grupos –por ejemplo clientes más rentables/clientes menos rentables -.

El representar los datos por una serie de grupos, representa la pérdida de detalles, pero consigue la simplificación de los mismos.

Una característica que lo diferencia de la clasificación es que no se conocen ni las clases ni su número.

Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas. Este método debido a su naturaleza flexible se puede combinar fácilmente con otro tipo de técnica de Minería de Datos, dando como resultado un sistema híbrido.

Desde un punto de vista práctico, el agrupamiento juega un papel muy importante en aplicaciones de Minería de Datos, tales como exploración de datos científicos, recuperación de la información y minería de texto. Además en aplicaciones sobre bases de datos espaciales -tales como GIS o datos procedentes de astronomía-, aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, en teoría de la señal -para eliminar ruidos-.

Un problema relacionado con el análisis de los grupos es la selección de los factores en las tareas de clasificación, debido a que no todas las variables tienen la misma importancia a la hora de agrupar los objetos.

Distancias

Cuando se van a agrupar los datos en los distintos grupos, esto se realiza teniendo en cuenta los criterios de distancia y similitud. Dígase que similitud es un concepto matemático inverso de “distancia”. Es la cualidad, condición o circunstancia de tener una persona o un objeto, rasgos, elementos o propiedades semejantes a los de otra; aspecto, característica o situación que tienen en común o en el que se parecen dos o más personas o cosas.

Si se desea conocer la similitud entre dos instancias o individuos, es necesario elegir una función de distancia y calcular con ella la distancia entre estos individuos.

Por lo que surge la interrogante ¿Qué funciones de distancia podemos utilizar para dicha tarea? Existen varias funciones de distancia como la Euclidiana, Manhattan o Minkowski, pero también se pueden encontrar otras métricas que cumplen con todos los requisitos de una función de distancia y que pueden llegar a superarlas en ciertos contextos.

Distancia Euclidiana:

La distancia Euclidiana se calcula entre dos elementos i, j y viene dada por la raíz cuadrada de la sumatoria de los cuadrados de la diferencia entre los valores i, j de todas las variables.

$$d_{ij}\sqrt{\sum_{v=1}^{p}(X_{iv}-X_{jv})^2}$$

Distancia de Manhattan o distancia por cuadras:

Esta es una de las medidas de distancia más utilizadas y hace referencia a recorrer un camino no en diagonal, usando el camino más corto sino zigzagueando, como se haría en la ciudad de Manhattan.

$$d_{ij}=\sum_{v=1}^{p}|X_{iv}-X_{jv}|$$

Algoritmos de Agrupamiento

A continuación se muestra un esquema simplificado de un algoritmo de agrupamiento:

Figura 5 Esquema general de un algoritmo de agrupamiento

Las principales características deseables por estos algoritmos son (Henandez Orallo, Ramírez Quintana, & Ferri, 2001):

- **Escalabilidad:** La escalabilidad de un algoritmo viene dada por su capacidad para realizar un agrupamiento de datos en función de un elevado número de observaciones. Se dice de un

algoritmo con gran escalabilidad cuando este es capaz de agrupar datos en bases de millones de observaciones. También dentro de este concepto se incluye la habilidad para trabajar con distintos tipos de atributos: numéricos, binarios, discretos y alfanuméricos.

- **Descubrimiento de grupos con formas arbitrarias:** El algoritmo debe poder establecer grupos de formas arbitrarias, pues como la mayoría de los algoritmos se basan en la distancia Euclidiana, los grupos tienden a tener forma (circular) y densidad similares.
- **Requerimientos mínimos en el conocimiento del dominio para determinar los parámetros de entrada:** Las herramientas no deben solicitarle al usuario que introduzca la cantidad de clases que quiere considerar, ya que dichos parámetros en muchas ocasiones no son fáciles de determinar, y esto hace que sea difícil controlar la calidad del algoritmo.
- **Habilidad para tratar con datos ruidosos:** La mayoría de las Bases de Datos contienen datos con comportamiento extraño, datos faltantes, desconocidos o erróneos. Muchos de estos algoritmos son sensibles a tales datos y pueden derivarlos a grupos de baja calidad.
- **Insensibilidad al orden de las observaciones de entrada:** Algunos algoritmos para un mismo conjunto de datos, dependiendo del orden en que se analicen, los grupos devueltos pueden ser diferentes. Es importante entonces que el algoritmo sea insensible al orden de los datos, y que el conjunto de grupos devuelto sea siempre el mismo.
- **Alta dimensionalidad:** Una Base de Datos o un Almacén de Datos puede contener varias dimensiones o atributos, por lo que es bueno que el algoritmo trabaje de manera eficiente y correcta no sólo en repositorios con pocos atributos, sino también en repositorios con un alto espacio dimensional y/o gran cantidad de atributos.
- **Agrupamiento basado en restricciones:** Es un gran desafío el agrupar los datos teniendo en cuenta no sólo el comportamiento, sino también que satisfagan ciertas restricciones.
- **Interpretación y uso:** Los usuarios esperan que los resultados del agrupamiento sean comprensibles, fáciles de interpretar y de utilizar.

Estos algoritmos pueden presentar diferencias entre sí por las reglas heurísticas que utilizan y el tipo de aplicación para el cual fueron diseñados.

Normalmente se clasifican por:

- El tipo de dato que manejan -numérico, categórico y/o mixto-.

- El criterio utilizado para medir la similitud entre los puntos.
- Los conceptos y técnicas de agrupamiento empleadas -lógica difusa y estadísticas-.

Existen varias técnicas de agrupamiento que varían de acuerdo a la arquitectura que utilizan. Una clasificación general divide los algoritmos en: agrupamiento particional, agrupamiento jerárquico, agrupamiento basado en densidad y agrupamiento basado en GRID.

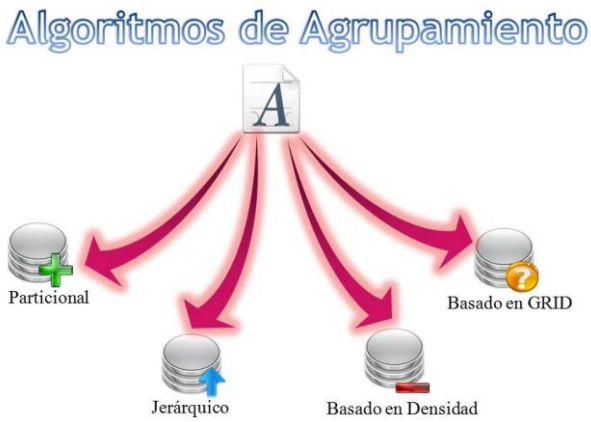


Figura 6 Algoritmos de Agrupamiento

Alguno de los algoritmos que emplean esta técnica más utilizados son:

- Algoritmo K Medias (K-means, en inglés).
- Algoritmo K-medoids.
- Fuzzy c-means.
- Grupos próximos a un entorno (nearest neighborhood clustering).
- Redes de aprendizaje competitivo.
- Agrupamiento Conceptual (COBWEB).
- Reconocimiento de Voz.
- Agrupamiento Demográfico.

Reconocimiento de Voz

El objetivo de la construcción de este algoritmo es que permite la identificación de los fonemas de personas a partir de variables que identifican los caracteres del sonido para un flujo continuo de ondas sonoras y silencios.

El reconocimiento de voz generalmente se utiliza como una interfaz entre el ser humano y la computadora, para algún software.

El mismo debe cumplir 3 tareas:

- Preprocesamiento: convierte la entrada de voz a una forma que el reconocedor pueda procesar.
- Reconocimiento: identifica lo que se dijo -traducción de señal a texto-.
- Comunicación: envía lo reconocido al sistema software o hardware que lo requiere.

Figura 7 Componentes de una aplicación

K-Medias (K-Means)

Es uno de los algoritmos más utilizados para hacer agrupamiento por su sencillez. En primer lugar se debe especificar por adelantado cuantos grupos se van a crear, este es el parámetro K, para lo cual se seleccionan k elementos aleatoriamente, que representarán el centro o media de cada grupo.

A continuación cada una de las instancias, ejemplos, es asignada al centro del grupo más cercano de acuerdo con la distancia Euclidiana que le separa de él. Para cada uno de los grupos así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos grupos.

Finalmente se repite el proceso completo con los nuevos centros de los grupos. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos grupos, ya que los puntos centrales de los grupos se han estabilizado y permanecerán invariables después de cada iteración (Silveira Martineaux & Fernández Pérez, 2008).

Esta peculiaridad de trabajo del algoritmo garantiza una elevada semejanza intra-grupos y desemejanza inter-grupos. La similitud entre los grupos se mide desde el punto medio de los mismos, lo cual puede ser visto como su centro de gravedad.

El objetivo de este método es crear grupos homogéneos en su interior y heterogéneos entre sí. Un criterio para evaluar la homogeneidad-heterogeneidad entre objetos es por la proximidad media de cada individuo del grupo. Esta puede ser determinada por la suma de los cuadrados de la diferencia de cada objeto con la media de cada grupo j. Esta función es conocida como la función objetivo (Marante Jacas & Marante Jacas, 2009).

$$\sum_{i=1}^{n_j} \left(X_{ij} - \text{Promedio}\, X_j\right)^2$$

Donde:

Xij es el valor de la variable para cada individuo del grupo j (1, 2, 3…) del grupo j.

n_j Es la dimensión del grupo j.

Pseudocódigo:

```
Iniciar
Seleccionar k objetos (cada uno es inicialmente el centro de cada grupo)
Hacer
        Asignar restantes objetos al grupo más similar (grupo con centroide más cercano)
        Calcular nuevamente los  centroides de los grupos
Mientras: No Criterio de Convergencia → Grupos no estén suficientemente compactos y separados del resto (Función Objetivo)
```

El algoritmo está presente en muchas herramientas de Minería de Datos ya que resulta relativamente sencillo su desarrollo computacional.

Sus principales ventajas son su velocidad, aspecto considerable cuando se tienen grandes volúmenes de datos y la posibilidad de cambiar los puntos iniciales para obtener resultados diferentes.

Mientras que tiene como principal desventaja su dificultad para lidiar con grupos que no tengan forma convexa, pues fue diseñado sobre la distancia óptima entre todos los puntos o centroide, por lo que promueve la construcción de grupos convexos aunque estos no sean la solución más adecuada para el problema en cuestión.

Agrupamiento Conceptual (COBWEB)

El agrupamiento conceptual es una solución ante la ineficacia del K-Medias cuando los atributos no son numéricos, pues se basa en la vecindad entre los elementos de la población para justificar la necesidad de un agrupamiento cualitativo frente al agrupamiento cuantitativo. En este tipo de agrupamiento una partición de los datos es buena si cada clase tiene una buena interpretación conceptual.

A semejanza de los humanos, COBWEB forma los conceptos por agrupación de ejemplos con atributos similares. Representa los grupos como una distribución de probabilidad sobre el espacio de los valores de los atributos, generando un árbol de clasificación jerárquica en el que los nodos intermedios definen subconceptos.

2.2.2.2 Asociación

La técnica de asociación se basa en buscar correlaciones entre diferentes atributos de un conjunto de datos. Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente, es relativamente alta. Las asociaciones se expresan como condiciones atributo-valor y deben estar presentes varias veces en los datos.

La aplicación más común de esta clase de técnica es la creación de reglas de asociación, que pueden utilizarse en un análisis de la cesta de compra.

Reglas de Asociación

Los modelos de asociación se generan basándose en conjuntos de datos que contienen identificadores para casos individuales y para los elementos que contienen los casos. Un grupo de elementos de un caso se denomina un conjunto de elementos.

Un modelo de asociación se compone de una serie de conjuntos de elementos y de las reglas que describen cómo estos elementos se agrupan dentro de los casos. Son un conjunto de transacciones, donde cada transacción es un conjunto de ítems, una regla de asociación es una expresión de la forma XY, donde X e Y son conjuntos de ítems.

Un ejemplo de regla de asociación sería: “el 30% de las transacciones que contienen niños, también contienen pañales; y el 2% de las transacciones contienen ambas cosas”. En este caso el 30% es el nivel de confianza de la regla y 2% es la cantidad de casos que respaldan la regla.

2.3 HERRAMIENTAS

Los procesos de Minería de Datos son un conjunto de tareas o procesos en los cuales se involucran una serie de herramientas, estas herramientas suelen ser incompletas, ya que no es posible crear una herramienta genérica para las tareas de inteligencia artificial.

Las herramientas de Minería de Datos suelen dividirse en dos grandes grupos:

- Técnicas de verificación: el sistema se limita a comprobar hipótesis suministradas por el usuario.
- Métodos de descubrimiento: Se han de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción.

Generalmente las aplicaciones actuales que se utilizan para los procesos de Minería de Datos tienen implementadas funcionalidades que las ubican en ambos grupos (Marante Jacas & Marante Jacas, 2009).

Es muy importante cuando se va a seleccionar una herramienta de Minería de Datos, tener en cuenta las características de la tecnología existente en el centro donde se utilizará, aspectos tanto de software como de hardware, entre los que se encuentran:

- Escalabilidad: Se evalúa si la herramienta permite aprovechar operaciones con bases de datos en paralelo, con procesadores adicionales, de esta forma se podría trabajar con un set de datos más robusto y se podrían construir más modelos.

- Capacidad para manejar datos: Aspecto muy importante, porque permite realizar alguna limpieza autónoma de los datos, como el tratamiento de los valores perdidos, permitiendo descartarlos, promediar, alertar y excluir.
- Velocidad y exactitud: Estas características contribuyen a la evaluación del rendimiento global de la herramienta.

A continuación se expondrán algunas de las herramientas más completas para la aplicación de la Minería de Datos.

2.3.1 Orange

Orange es un software desarrollado en la facultad de informática de la Universidad de Ljubljana.

Orange consta de una serie de componentes y algoritmos desarrollados en C++ interconectados con Phython[5]. Además posee un entorno gráfico bastante cómodo y tiene gran facilidad para la visualización de datos y análisis para los novatos y expertos.

El software está operado bajo la licencia de **GPL (Licencia Pública General, acrónimo en inglés)**, existen actualmente distribuciones para Windows, Linux, y Macintosh, es flexible y rápido (KDnuggets, 2010).

Esta herramienta tiene implementados los siguientes métodos de aprendizaje:

- Método Bayesiano
- Árboles de decisión
- Árboles de regresión
- K-Vecinos más próximos
- Reglas de Asociación.

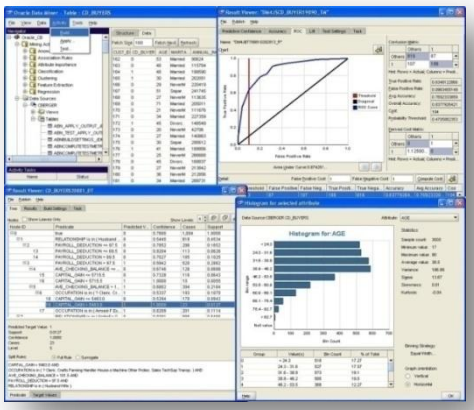
[5] **Phyton** ***es un lenguaje de script moderno y avanzado, es perfectamente programable y extensible, implementa algoritmos de MD, así como operaciones de preprocesamiento y representación gráfica de datos.***

Además incluye algoritmos para realizar las siguientes tareas (Gil Bellosta, 2010):

- Preprocesamiento: selección de variables y discretización.
- Modelización predictiva: árboles de clasificación, k-vecinos más próximos, regresión logística y clasificadores basados en reglas.
- Métodos de agrupación de modelos: boosting, bagging y bosques aleatorios.
- Visualización: SOM (self-organization maps), agrupamiento jerárquico, k-medias y escalado multidimensional.
- Validación: remuestreo, validación cruzada y medidas estadísticas del ajuste.

2.3.2 Oracle Data Mining

Oracle Data Mining es un potente software incluido en Oracle Database.



Los algoritmos de aprendizaje supervisado que Oracle Data Mining incluye son:

- Redes Bayesianas.
- Árboles de Decisión.
- Modelos Lineales Generalizados.
- Máquinas de Vectores Soporte.

Y los no supervisado:

- k-Means mejorado.
- Agrupamiento de Partición Ortogonal.
- Reglas de Asociación.
- Factorización de Matrices no Negativas.

Oracle Data Mining incluye Oracle Data Miner el cual está integrado a la Base de Datos, diseñado para funcionar con conjuntos inmensos de datos, tiene una interfaz gráfica de usuarios orientada a tareas para el análisis de datos que tiene el fin de crear y aplicar modelos de Minería de Datos .

Oracle Data Miner guía al analista de datos a través del proceso de Minería de Datos con total flexibilidad y presenta los resultados en formatos gráficos y tabulares. Realiza exitosamente las tareas más difíciles como lo son la preparación de los datos y el ajuste de parámetros.

Oracle Data Miner está orientado a los siguientes propósitos:

- Explicar: La importancia de los atributos.
- Predecir - Clasificación / Regresión-.
- Agrupar -Clustering / Segmentación-.
- Detectar: Detección de anomalías y "outliers"[6].
- Mapear: Proyectar datos en menos dimensiones.

Los beneficios que brinda Oracle Data Miner son:

- Al tener algoritmos de Minería de Datos en la Base de Datos elimina movimiento y exposición de los datos.
- Presenta un amplio rango de algoritmos que pueden aplicarse a la mayoría de los problemas de Minería de Datos.
- Trabaja sobre diferentes plataformas: Las aplicaciones pueden ser desarrolladas y luego instaladas en otra plataforma.
- Es parte de la Tecnología Oracle (GRID, RAC, JAVA y PL/SQL).

[6] ***Datos con características considerablemente diferentes al resto de los pertenecientes al conjunto.***

2.3.3 SPSS Clementine

SPSS Clementine[7] es una herramienta de SPSS Inc., la cual fue comprada en julio del 2009 por la compañía IBM, por lo que pasó a conocerse como IBM SPSS Modeller.

Es un potente software que combina modernas técnicas de modelamiento con poderosas herramientas de acceso, manipulación y exploración de datos en una interfaz simple e intuitiva.

SPSS Clementine tiene una serie de características propias que lo hacen unos de los mejores sistemas de Minería de Datos que existen actualmente en el mercado:

- Visualización Interactiva: Fácil entendimiento de los datos (Información Creativa, 2010).
- Muy buena preparación de los datos: Posibilidad de acceso y combinación de datos de múltiples fuentes, especifica los valores perdidos, deriva nuevas variables de trabajo y produce información resumida.
- Incrementa la productividad con su enfoque visual de la manipulación de los datos.
- Soporta la metodología estándar CRISP-DM: Los proyectos de esta solución y cada una de sus etapas pueden organizarse eficientemente utilizando el administrador de Proyectos CRIP-DM.

SPSS Clementine utiliza las técnicas de evaluación tablas estadísticas y gráficos de ganancias, y las técnicas de publicación de modelos punto o scoring de bases de datos y scoring en tiempo real. Este software tiene una capacidad extendida por lo que cubre todos los aspectos de las interacciones de los clientes ya sea **Minería de Textos (Text Mining)** y/o **Minería Web (Web Mining)**.

Con este software se puede contar con tres muestras en vez de dos, entrenamiento, prueba y evaluación. Pues tiene un nodo que automáticamente crea las particiones que se necesitan para el análisis.

También permite analizar, grandes volúmenes de datos en crudo, almacenados en grandes bases de datos transaccionales y/o registros de programas y cuenta con los métodos de Redes Neuronales de mayor uso: Kohonen, Prune y Radial Basis.

El sistema además cuenta con los modelos **GRI (Generalizad Rule Induction)** el cual los tiene incluido y permite generar reglas que sintetizan patrones en los datos.

[7] **Sitio Web Oficial *http://www.spss.com***

Diversas fases del proceso de Minería de Datos utilizan gráficos y diagramas para explorar los datos almacenados de una forma u otra. SPSS Clementine cuenta con un sistema de visualización muy avanzado que se pueden clasificar en cuatro tipos.

- Gráficos para comprender de una forma mucho mejor los datos y las distribuciones.
- Gráficos para manipular los registros y campos previos a las operaciones.
- Gráficos para comprobar la distribución y las relaciones entre campos recién derivados.
- Gráficos de apoyo al modelado.

Una de las características destacables de SPSS Clementine es haber acercado el análisis de datos a un público más global y mejorado la productividad, debido a su amigable interfaz y el poder realizar en un mismo entorno el análisis de datos, el proceso ETL[8] y la modelización. Se podría decir de ella que ha sido el Windows[9] de la Minería de Datos.

SPSS Clementine es un software privativo con grandes requerimientos de hardware, clasificándose estas como sus principales desventajas.

Total de técnicas de modelado que implementa:

- Técnicas Supervisadas:
  - ✓ Árboles de Clasificación y Regresión
  - ✓ Redes Neuronales
  - ✓ C 5.0
  - ✓ Quest
  - ✓ CHAID (Chi-square Automatic Interaction Detector)
  - ✓ Regresión Lineal
  - ✓ Regresión Logística
  - ✓ K vecinos más próximos (k-Nearest Neighbor)

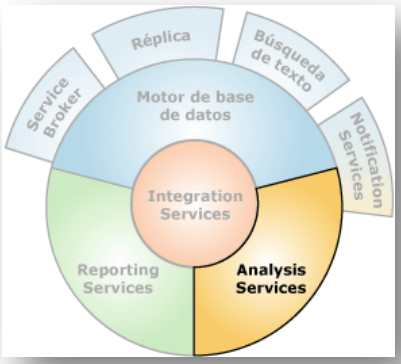
- Técnicas no Supervisadas:
  - ✓ K- Medias
  - ✓ Mapas Auto organizados o redes de Kohonen

[8] **Primera fase de un proyecto de MD:** ***extracción, transformación y carga de la información de los datos necesarios.***
[9] **Windows:** ***Sistema operativo muy conocido y extendido por sus características y funcionalidades.***

- ✓ Bi-etápico
- ✓ A priori
- ✓ GRI (General Rules of Interpretation)
- ✓ Sequence
- ✓ Carma
- ✓ Detección de Anomalías

2.3.4 Microsoft SQL Server 2005

Microsoft SQL Server 2005 es una plataforma de bases de datos que se utiliza en el procesamiento de transacciones en línea (OLTP) a gran escala, el almacenamiento de datos y las aplicaciones de comercio electrónico; es también una plataforma de inteligencia de negocios para soluciones de integración, análisis y creación de informes de datos.



Microsoft SQL Server 2005 Analysis Services (SSAS) permite diseñar, crear y administrar estructuras multidimensionales con datos de detalle y agregados de diversos orígenes de datos, como bases de datos relacionales, en un solo modelo lógico unificado y compatible con los cálculos integrados.

Analysis Services facilita el análisis rápido e intuitivo de grandes cantidades de datos creados a partir de este modelo de datos unificado, que se puede poner a disposición de los usuarios en varios idiomas. Además permite trabajar con almacenes de datos, puestos de datos, bases de datos de producción y almacenes de datos operativos, y admite el análisis de datos históricos y en tiempo real.

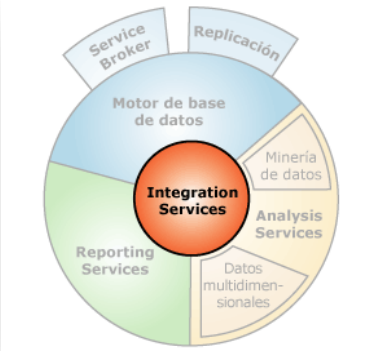
Analysis Services contiene las características y herramientas necesarias para crear complejas soluciones de Minería de Datos.

- ➢ Un conjunto de algoritmos de Minería de Datos estándar del sector.
- ➢ Las **Extensiones de Minería de Datos (DMX)**: es un lenguaje eficaz para crear la estructura de modelos de Minería de Datos nuevos, para entrenar esos modelos y para explorar, administrar y

realizar predicciones con ellos, eficaz para administrar modelos de Minería de Datos y crear complejas consultas predictivas.

Analysis Services incluye los siguientes tipos de algoritmos:

- Algoritmos de clasificación.
- Algoritmos de regresión.
- Algoritmos de agrupamiento.
- Algoritmos de asociación.
- Algoritmos de análisis de secuencias.



SQL Server 2005 Integration Services (SSIS) es el componente de extracción, transformación y carga (ETL) de SQL Server 2005. Sustituye al componente de ETL de SQL Server anterior, Servicios de Transformación de Datos (DTS).

Integration Services es una plataforma para la creación de soluciones de integración y transformaciones de datos de alto rendimiento, puede resolver complejos problemas empresariales mediante la copia o descarga de archivos, el envío de mensajes de correo electrónico como respuesta a eventos, la actualización de almacenes de datos, la limpieza y Minería de Datos, y la administración de objetos y datos de SQL Server.

Integration Services puede extraer y transformar datos de diferentes orígenes, como archivos de datos XML, archivos planos y orígenes de datos relacionales, y posteriormente, cargarlos en uno o varios destinos.

Las herramientas gráficas de Integration Services se pueden usar para crear soluciones sin escribir una sola línea de código.

2.3.5 Knime

KNIME (Konstanz Information Miner) es una herramienta desarrollada por la Cátedra de Bioinformática y Minería de la Información en la Universidad de Konstanz, Alemania.

El grupo encabezado por Michael Berthold la utiliza para la enseñanza y la investigación en la Universidad. Un buen número de nuevos métodos de análisis de datos desarrollados en la cátedra se integran en esta herramienta, pero no todos los módulos, son parte de la versión estándar KNIME todavía.

Es una herramienta basada en eclipse, que permite la ejecución tanto de rutinas de WEKA como de R dando una gran versatilidad a la misma. Así mismo incluye el proyecto **Inteligencia de Negocios y Herramientas de Reportes (BIRT, siglas en inglés)** como motor de reportes (Curto Díaz, 2010).

La versión base de KNIME ya incorpora más de 100 nodos de procesamiento de datos de entrada/salida, preprocesamiento y limpieza, modelado, análisis y Minería de Datos, así como diversos puntos de vista interactivo, tales como diagramas de dispersión y coordenadas paralelas.

Por diseño, KNIME fue construido para ser extensible con nuevas funcionalidades en forma de nuevos nodos, los cuales pueden ser implementados en sólo cuestión de horas para así ampliar su capacidad de comprender y prestar apoyo de primer nivel para datos específicos de dominio. Esta modularidad y extensibilidad le permite ser empleada en entornos de producción comercial, así como en la enseñanza e investigaciones.

KNIME es liberada bajo un esquema de doble licenciamiento, donde la licencia de código abierto -GPL- le permite ser descargado, distribuido y utilizado libremente (Dill & Domko, 2010).

2.3.7 WEKA

La weka es un ave endémica de Nueva Zelanda, cuyo nombre científico es *Gallirallus australis*. Es una gallinácea en peligro de extinción, famosa por su curiosidad y agresividad, de aspecto pardo y muy similar a las gallinas (García Morate, 2001).

Figura 8 Imagen de una Weka

Su nombre fue seleccionado por la Universidad de Waikato para su proyecto de desarrollo de Análisis del Conocimiento, basado en técnicas de Minería de Datos. El proyecto tuvo sus inicios en el año 1993.

WEKA (Waikato Environment for Knowledge Analysis) es un software desarrollado en lenguaje de programación Java, lo que lo hace independiente de la arquitectura, pues funciona sobre cualquier plataforma en la que se encuentre corriendo una máquina virtual Java. Este opera bajo la licencia GPL[10], lo que significa que este programa es de libre distribución y difusión. Contiene una extensa colección de algoritmos de aprendizaje por computadora o ML (del inglés, Machine Learning) para realizar tareas de Minería de Datos, en forma de un conjunto de librerías.

Para usar estas librerías existen dos formas de hacerlo: se pueden llamar desde su interfaz o desde las propias clases Java. De necesitarse por parte del usuario es posible agregar códigos para extender la funcionalidad del paquete, pues los desarrolladores de esta herramienta proporcionan el código fuente completo (Witten, Frank, Trigg, Geoffrey Holmes, & Jo Cunningham, 1999).

La versión original de WEKA fue un front-end en **TCL/TK (Tool Command Language/ToolKit)** para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para preprocesamiento de datos desarrolladas en C para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura, pero la versión más reciente basada en Java -WEKA 3-, empezó a desarrollarse en 1997 y tiene aplicaciones en diferentes áreas.

[10] **GNU Public License.** ***http://www.gnu.org/copyleft/gpl.html***

WEKA posee una serie de herramientas muy importantes para la extracción de conocimiento y soporta varias tareas de Minería de Datos.

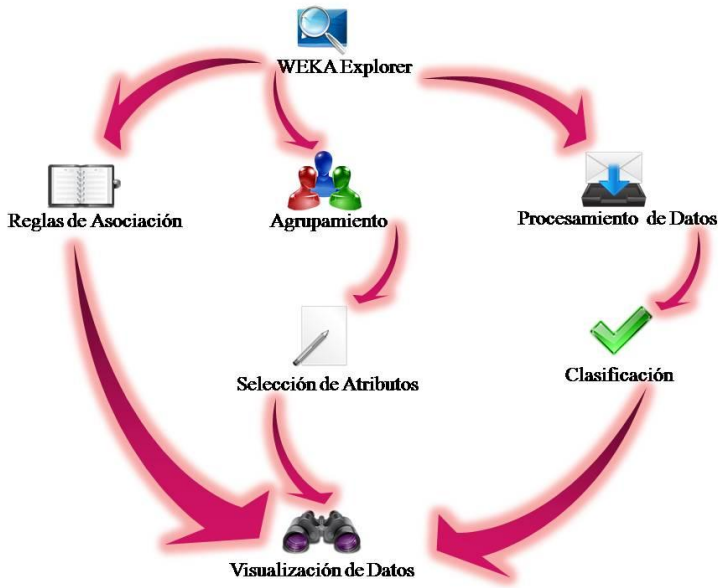


Figura 9 Algoritmos de WEKA

Las técnicas de WEKA se fundamentan en la asunción de que los datos están disponibles en un fichero plano -flat file- o una relación, en la que cada registro de datos está descrito por un número fijo de atributos, normalmente numéricos o nominales, aunque también se soportan otros tipos. Además proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity, sigla en inglés) y puede procesar el resultado devuelto por una consulta hecha a la base de datos.

No puede realizar Minería de Datos multirelacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con este software.

WEKA contiene herramientas para diferentes tareas básicas:

- Preprocesamiento: Multitud de herramientas para el preprocesamiento de los datos, por ejemplo discretización de variables.
- Clasificación: Algoritmos de clasificación, distribuidos por paquetes, por ejemplo ID3 y C4.5.
- Agrupamiento: Diferentes algoritmos de segmentación como el simple k-means.

- Asociación: Algoritmos para encontrar relaciones de asociación entre variables.
- Selección de atributos: Una vez cargados los datos, la herramienta es capaz de buscar las mejores variables del modelo.
- Visualización: Herramienta de visualización de datos en los ejes cartesianos, con muchas posibilidades.

Entonces se pueden deducir como las principales ventajas de WEKA que:

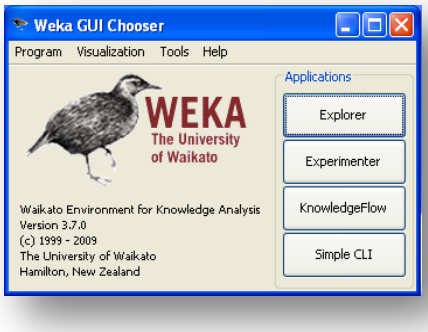
- Está disponible libremente bajo la licencia pública general de GNU.
- Es portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado.
- Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

2.4 ANÁLISIS DE ALGORITMOS EN WEKA

La herramienta WEKA propuesta anteriormente presenta una colección de herramientas de visualización y algoritmos que permitirá el análisis de los datos y modelado predictivo del SIGDC. La misma cuenta con una interfaz gráfica (Weka GUI Chooser) que posibilita el acceso a las funcionalidades de una forma mucho más fácil.

El sistema presenta cuatro interfaces principales:

- Explorer
- Experimenter
- Knowledge Flow
- Simple CLI



Simple CLI es la abreviatura de **Interfaz Simple de Línea de Comandos** (Simple Command-Line Interface), la cual básicamente es una consola que permite acceder a todas las opciones de WEKA desde línea de comandos. Su apariencia es muy simple pero es extremadamente potente pues permite realizar cualquier operación soportada por WEKA de forma directa, aunque es difícil de operar pues se necesita de un conocimiento completo de la aplicación.

Explorer (Explorador) es una interfaz que dispone de varios paneles que brindan el acceso a los componentes principales, permitiendo realizar operaciones sobre un sólo archivo de datos. En estos momentos es el modo más usado y descriptivo de WEKA. Permite tareas de:

- **Preprocess (Preprocesamiento):** Este panel dispone de diferentes opciones que posibilitan la preparación de los datos, ya sea importando los mismos desde una base de datos o ficheros CSV, arff u otras extensiones. Los algoritmos de filtrado de atributos contenidos en esta sección son los encargados del preprocesamiento de los datos luego de cargados los mismos. Estos filtros posibilitan el proceso ETL, pudiéndose transformar por ejemplo datos numéricos en valores discretos y eliminar registros o atributos siguiendo criterios previamente especificados.
- **Classify (Clasificación):** Esta sección posibilita la aplicación de los algoritmos de clasificación.
- **Associate (Asociación):** Proporciona acceso a las reglas de asociación.
- **Cluster[11]:** Permite el acceso a las técnicas de agrupamiento.
- **Selected Attributes (Selección de Atributos):** Proporciona algoritmos para identificar los atributos más significativos en un conjunto de datos.
- **Visualize (Visualizar):** Muestra al usuario una matriz de puntos dispersos, donde cada punto puede ser seleccionado para su análisis en detalle.

Experimenter (Experimentador) es una interfaz encargada de la automatización de tareas facilitando la realización de experimentos a gran escala.

Knowledge Flow (Flujo de Conocimiento) es una interfaz que ofrece soporte para el aprendizaje incremental y permite generar proyectos de Minería de Datos mediante la generación de flujos de información (Hernández Orallo & Ferri Ramírez, Marzo 2006).

[11] **Cluster (Término en inglés):** ***Conjunto o racimo de objetos, que tienen características comunes.***

En WEKA los modelos pueden ser validados de las siguientes formas:

- Use training set: Entrena el método con todos los datos disponibles y después realiza la evaluación sobre los mismos datos.
- Supplied test set: Carga un conjunto de datos, regularmente diferentes a los de aprendizaje, con los cuales se realizará la evaluación
- Cross-validation: Realiza la evaluación mediante la técnica de validación cruzada. En este caso se establece el número de pliegues a utilizar.
- Percentage Split: Define un porcentaje con el que aprende el modelo y la evaluación se realiza con los datos restantes.

2.4.1 Técnica Clasificación. Árboles de Decisión. Algoritmo C4.5

Una de las técnicas más eficaces y relevantes, los árboles de decisión, están representados dentro de la interfaz Classify, los cuales se encuentran ejemplificados a través de su algoritmo **C4.5 –J48** nombre que recibe en WEKA-. Este es uno de los algoritmos más estandarizados y de amplio uso en la práctica del aprendizaje supervisado.

La interfaz brinda al usuario la posibilidad de modificar los parámetros específicos para el algoritmo, los cuales en caso de no ser modificados se tomarían con sus valores por defecto predefinidos y luego se debe seleccionar la opción deseada para validar el modelo.

Al terminarse el modelo predictivo, WEKA brinda información acerca de algunos parámetros del conjunto de datos, así como el modelo aprendido representado textualmente. Además se incluye como resultado el error de la evaluación del modelo, importante elemento de comparación con otros algoritmos de clasificación, en el cual el C4.5 ofrece ser una de las mejores opciones.

El árbol de decisión puede ser visualizado a través de la herramienta, que muestra en cada una de sus hojas los ejemplos de la muestra que son cubiertos (García Morate, 2001).

2.4.2 Técnica Agrupamiento. K-Medias

El algoritmo K-Medias está localizado dentro de la interfaz Cluster, donde aparece con el nombre de Simple Kmeans. Se encuentra implementado dentro de la clase "*weka.clusterers.SimpleKMeans.java".*

Las principales opciones de configuración que Weka ofrece para este algoritmo son:

- El número de grupos o clases que se deben crear.
- El seed -semilla- a partir del que se genera el número aleatorio para inicializar los centros de los grupos.
- La función de la distancia a utilizar para la comparación de los casos, es por defecto la distancia Euclidiana.
- El número de iteraciones máximas a realizar.

Al pulsar el botón derecho sobre "Simple Kmeans" de la parte de "Result List", se puede ver en qué grupo es ubicado cada clase, así como estudiar la forma de asignación de las clases a los grupos, dependiendo de los atributos que hayan sido seleccionados para dibujar la gráfica (Hernández Orallo & Ferri Ramírez, Marzo 2006).

La implementación está concebida para admitir atributos del tipo simbólico y numérico, el resto del funcionamiento del algoritmo es similar al descrito al principio del capítulo. Se utiliza el número aleatorio obtenido a partir de la semilla empleada para calcular los centroides iniciales, luego los k ejemplos correspondientes a los k números enteros siguientes al número aleatorio obtenido serán los que conformen dichos centroides.

2.4.3 Técnica Asociación. Reglas de Asociación. Apriori

Para encontrar reglas de asociación entre los atributos se selecciona la interfaz Associate. El sistema provee el paquete "WEKA.associations.Apriori" que posee la implementación del algoritmo de aprendizaje de reglas de asociación Apriori.

El algoritmo puede ser configurado con varias opciones:

- UpperBoundMinSupport: indica el límite superior de cobertura requerido para aceptar un conjunto de ítems. De no ser suficientes los conjuntos de ítems para generar las reglas requeridas se comienza a disminuir el límite hasta el límite inferior "LowerBoundMinSupport".
- MinMetric: indica la confianza mínima, u otras métricas dependiendo del criterio de ordenación, para mostrar una regla de asociación.
- NumRules: indica el número de reglas que se quiere aparezcan en pantalla. La ordenación de estas reglas en pantalla puede configurarse mediante la opción "MetricType".

- Lift: confianza de la regla dividida por el número de ejemplos cubiertos por la parte derecha de la regla.

Es importante destacar que las reglas de asociación aprendidas en muchas ocasiones se encuentran lastradas por la presencia de atributos que están fuertemente descompensados, pues al estos tener baja cobertura las reglas son filtradas, lo cual puede ser mitigado parcialmente al cambiar el método de selección de la regla. Además, estos métodos sólo funcionan con datos nominales.

2.5 DISEÑO DEL SIGDC

El diseño de este Sistema tiene como propósito final obtener, almacenar, manipular, administrar e intercambiar y hacer uso de la información provista por los datos, de forma que fluya como un todo, existiendo una alta cohesión entre todos los procesos y sistemas que lo componen:

- **Sistema Gestor de Base de Datos (SGBD)**
- **Sistema Gestor de Base de Modelos (SGBM).**
- **Sistema de Generación y Gestión del Diálogo (SGGD).**

Los usuarios tendrán acceso al SIGDC mediante la interfaz de usuario o SGGD, a través de la cual podrán manipular los elementos necesarios de los componentes SGBD y SGBM de acuerdo a sus necesidades.

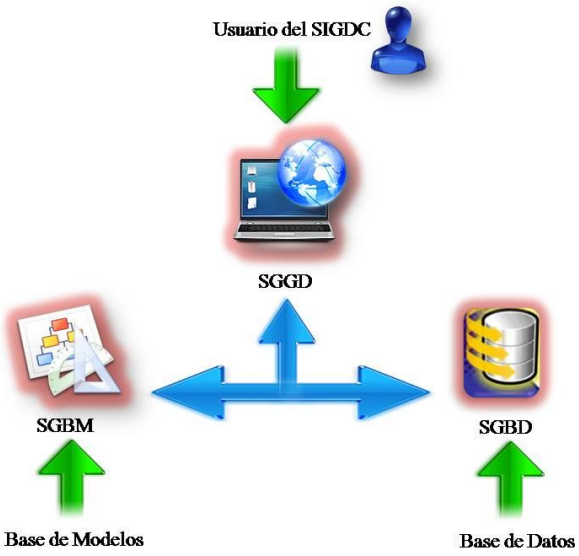


Figura 10 Diseño del SIGDC

2.5.1 Sistema Gestor de Base de Datos

El Sistema Gestor de Base de Datos está integrado por un conjunto de programas o software que permiten al usuario crear una base de datos, para mantener y asegurar la integridad total de sus datos, así como su seguridad. Además provee al sistema de un conjunto de facilidades para controlar el acceso a los datos, gestionar la concurrencia y la restauración de la base de datos desde copias de seguridad.

Para la creación de la Base de Datos el primer paso a tener en cuenta es definirla, lo que incluye especificar estructura, tipo y restricciones de datos; luego viene la construcción, que consiste en guardar los datos en algún medio o recurso controlado por el mismo SGBD; y por último se debe poder manipular los datos guardados, es decir, realizar consultas, generar informes y actualizarla.

Este componente proporciona al sistema una abstracción respecto a la forma de almacenamiento y procesamiento de los datos, separando a los usuarios de los aspectos físicos de la estructura, contenido y del análisis de la Base de Datos (Lichilín Ríos & Rodríguez Betancourt, Junio, 2008).

EL SGBD quedaría conformado esencialmente por el SADC del Centro de Clima del INSMET con las adaptaciones necesarias para cumplimentar los requisitos necesarios. El diseño distribuido del mismo permite su independencia de otros subsistemas pertenecientes a otros departamentos – Pronóstico

Agrometeorología, Contaminación Atmosférica- por el momento, los cuales pasarían a formar parte del SGBD una vez que los recursos tecnológicos y el capital humano permitan y garanticen su unificación satisfactoria.

2.5.1.1 Sistema de Administración de Datos Climáticos

El Sistema de Administración de Datos Climáticos es una herramienta creada con el objetivo de conservar, gestionar y consultar la información medida u observada en las estaciones meteorológicas convencionales, automáticas y de aire superior del país, el cual contribuye a la conservación de los documentos originales de las observaciones que anteriormente sólo se encontraban en formato duro y que poseían información de alta demanda para los especialistas e investigadores del centro.

El SADC tuvo en cuenta para su confección el registro climático asentado y validado por décadas, así como las condiciones en que fueron medidas las variables climatológicas para su confección. El mismo se encuentra en funcionamiento en el país desde el año 2006.

Este sistema consta de dos formas de asimilación de los datos: en tiempo real y tiempo diferido.

- Los primeros son aquellos datos que se recogen diariamente en todas las estaciones del territorio nacional y que se incorporan automáticamente a la Base de Datos a través del sistema Ciber – Operador Meteorológico.

  La herramienta **Ciber – Operador Meteorológico (C.O.M)** garantiza todo el proceso de captación y chequeo de la información de las observaciones realizadas en la estación, la actualización de metadatos y la confección del telegrama sinóptico, el cual es recibido en la sede del INSMET e incorporado a la Base de Datos Meteorológica.

- Los segundos son los datos que se encuentran archivados en formato duro y que de forma gradual se están incorporando al sistema.

El SADC está configurado por tres niveles para el procesamiento de las observaciones y mediciones realizadas:

- Nivel de Estación
- Nivel de Centro Meteorológico Provincial
- Sistema de Administración de Datos Climáticos a nivel nacional

Los servicios brindados por el SADC en estos momentos sólo se encuentran disponibles en la intranet de la institución. Los mismos se clasifican en dos tipos:

- Especializados
- A petición del usuario

Los servicios especializados son aquellos que están estandarizados y que son calculados con periodicidad para obtener valores meteorológicos – promedio de la temperatura, rosa de los vientos, valores máximos y mínimos de precipitaciones, temperaturas– de interés para los especialistas, investigadores e instituciones interesadas.

Los servicios a petición de usuarios se clasifican en dos formas manuales o automatizados. Los automatizados son aquellos que son realizados mediante la Base de Datos a través de los DTS (Data Transformation Server, encargado de transformar los datos del modelo almacenado al solicitado por los usuarios), o JOB (se encarga de la ejecución de tareas programadas), ambas potentes herramientas provistas por SQL Server. Mientras que los manuales son aquellos que no se encuentran implementados por no ser frecuentemente solicitados y que se realizan de forma independiente por los informáticos encargados de estas cuestiones en el centro.

Todos estos servicios encuentran como usuarios finales a Investigadores, Departamentos –como por ejemplo la defensa civil -, Empresas - como la Eléctrica-, Ministerios –Agricultura, Salud, FAR - y entidades que los usan para sus investigaciones dentro de sus materias específicas.

Su confección permitió la creación de un estándar de datos climáticos, tanto a nivel nacional, provincial como de la estación en sí, pues permite compatibilizar criterios, herramientas de cálculo y análisis así como la generalización de metodologías en todas las provincias del país, esto convierte al sistema en el rector del dato climático en Cuba.

Este sistema tiene como unos de sus grandes logros el permitir incorporar a Cuba al grupo de naciones que han desarrollado herramientas de esta categoría, elevando considerablemente el valor y la calidad de la información climatológica nacional. Además brinda la posibilidad de intercambio de información con otras naciones, incluso hasta dar asesoría técnica a otras que no posean un sistema de administración para sus datos climáticos.

2.5.2 Sistema Gestor de Base de Modelos

El SGBM tiene entre sus principales objetivos transformar los datos en bruto en información nueva y útil a través de los modelos matemáticos, algoritmos y técnicas computacionales provistas por la Minería de Datos y la Visualización.

Una de las principales utilidades del SGBM es que brinda soporte e independencia a los modelos provistos por los mismos. Además puede asistir en la construcción de los modelos, especialmente cuando son muy complejos (Lichilín Ríos & Rodríguez Betancourt, Junio, 2008).

En las diferentes herramientas de modelado desarrolladas, el conocimiento acerca del sistema es representado por ecuaciones o reglas lógicas, mejoradas con una representación explícita de la incertidumbre. Una vez el modelo ha sido formulado, pueden usarse una variedad de métodos matemáticos para analizarlo.

2.5.2.1 Visualización científica

La Visualización Científica de los Datos es una novedosa técnica usada en la actualidad para contribuir a la presentación de la información hallada por otras técnicas y procesos, en lo que concierne a este trabajo la Minería de Datos.

Numerosas herramientas de esta variedad están sometidas a la aplicación de un campo en específico, como la medicina, la química molecular y la meteorología.

El objetivo principal de los métodos de visualización incluidos en las herramientas de Minería de Datos es la utilización de datos multiparamétricos, ya que estos permiten encontrar representaciones visuales aplicadas para un conjunto de datos previamente entregados, o sea, conseguir que dichas representaciones visuales permitan a los investigadores y especialistas descubrir nuevas relaciones y/o comportamientos de patrones entre los datos, dando a conocer lo que realmente transcurre con ellos.

Esta técnica no solamente da la posibilidad al usuario de visualizar los datos, sino también que le permite reconocer, comprender y evaluar los mismos de una forma muchos más representativa y comprensible, favoreciéndole una mejor interpretación del resultado alcanzado, algo difícil de lograr a través de la forma numérica clásica, aún siendo analizados rigurosamente.

La visualización es el proceso de generación de imágenes mediante el filtrado, mapeo y renderizado de datos. Existen técnicas de visualización para diferentes clases de datos, por ejemplo, datos de fluidos, volumen y datos multivariados (Romel, González Herrera, & Pérez Risquet, 2008).

Este proceso distingue principalmente los tipos de datos escalares, vectoriales y tensoriales. Los tipos de datos escalares se dividen en subgrupos de una dimensión, dos dimensiones, tres o "n" dimensiones, las que pueden variar con el tiempo cuando se tenga una acumulación histórica de observaciones. Los datos que pertenecen al subgrupo de la tercera o "n" dimensión como las bases de datos estadísticas y relacionales, se asocian a los datos escalares multidimensionales para lograr una visualización con mayores probabilidades de éxito.

Además algunas de sus características la convierten en una herramienta muy usada en el manejo y la interpretación de la información, el análisis de los datos, la comunicación de modelos y conceptos, los permite realizar de forma más fácil, pues una imagen o animación logra más claridad en la comprensión y aprovechamiento de los datos. Otra de sus características es que simplifica la comunicación entre los investigadores que hacen uso de la misma, independientemente de su lenguaje.

2.5.3 Sistema de Generación y Gestión del Diálogo

El SGGD es el sistema encargado de gestionar cómo el usuario puede interactuar con el sistema de la manera más intuitiva y fácil de comprender para él. Esta necesidad viene dada debido a que a menudo los usuarios son gestores sin formación informática y las interfaces con estas características presentan una mayor aplicabilidad ayudando en la construcción del modelo y en la interacción con el mismo (Lichilín Ríos & Rodríguez Betancourt, Junio, 2008).

Este SGGD estaría conformado además por una red de conocimiento lo que permitiría un espacio de colaboración, investigación e intercambio en el centro, lo cual estaría posibilitado a través del portal de la Biblioteca, el cual cuenta con toda la ayuda y soporte por parte de la entidad, garantizando un fuerte espacio colaborativo y de esta forma un aprendizaje colaborativo, como actividad social que es, ya que estaría implicado el intercambio de conocimiento, dígase compartir o adquirir el mismo, proceso conocido como "construcción del conocimiento" (Jonassen, Mayes, & McAlesee, 1992).

La importancia de incluir a dicho sistema una red de conocimientos está dado con el fin de lograr el objetivo de poder gestionar el conocimiento, al este encontrarse provisto por un grupo de utilidades que

facilitarían el trabajo colaborativo entre los especialistas e investigadores, interesados e involucrados con el proceso y a su vez facilitar la creación y administración de una estructura robusta para dicho conocimiento.

2.5.3.1 Redes de Conocimiento

En la sociedad globalizada del siglo XXI, las redes de conocimiento constituyen una de las máximas expresiones del hombre como productor de conocimientos y su necesidad de intercambiar y transferir lo que aprende y crea, a partir de la interacción social dentro de una plataforma tecnológica y un contexto muy particular.

La producción de conocimiento científico está estrechamente relacionada con las organizaciones formales que se crean para ello, los procesos de investigación que se desarrollan en ellas son el resultado de una integración tanto de recursos intelectuales como financieros, de allí la importancia de capacitar a los recursos humanos de la misma con la conciencia de hacer de ese espacio colaborativo todo un éxito.

Existen varias concepciones y definiciones en diferentes bibliografías que abordan el tema, por ejemplo para Beltrán y Castellanos (Moreno Beltrán & Castellanos Galvin, 2003) una red de conocimiento se define "como una comunidad de personas que, de modo formal o informal, ocasionalmente, a tiempo parcial o de forma dedicada, trabajan con un interés común y basan sus acciones en la construcción, el desarrollo y la compartición mutua de conocimientos". Por otro lado, para Seufert (Rodríguez, Araujo, & Yulianov, 2003) las redes de conocimiento son las "redes que se establecen ente los individuos, los grupos y las organizaciones donde no solamente son importantes las relaciones bilaterales, sino la integridad de las actividades desempeñadas por la propia red de conocimientos."

Naturalmente se pueden asociar ciertas características en común que abordan un concepto en sí definiéndolo como las relaciones humanas en la producción, almacenamiento, distribución, transferencia, acceso y análisis de los conocimientos producidos por el hombre de manera sistemática por la investigación o por el interés personal o grupal por compartir datos a través de cualquier medio, generalmente electrónicos en la actualidad.

Al hablar de los modelos adecuados para su confección pueden también aparecer disímiles ideas, pero sin dudas dos componentes fundamentales y comunes en todas ellas lo son: un grupo de personas que

conviven en sociedad y una plataforma tecnológica que optimice la producción y transferencia del conocimiento científico producido por ellas (Martín-Moreno Cerrillo, diciembre, 2004).

Son algunas de sus características más básicas:

- Ser expresiones de la interacción humana en un contexto social propio e íntimamente ligado al desarrollo de las civilizaciones.
- Producir, almacenar y distribuir conocimiento científico por medio de cualquier método de transmisión tecnológica.
- Transformar el entorno en la búsqueda constante del enriquecimiento intelectual del ser humano en su quehacer innovativo y creativo a través del estudio sistemático que ofrece la investigación científica pluridisciplinaria (Royero, 2004).
- Ser un medio de cooperación para intercambiar información, fomentar los valores cooperativos, compartir y reconocer el valor estratégico del conocimiento y promover la creación del mismo.

El objetivo principal perseguido con estas redes de conocimiento en las organizaciones es el avance simétrico del conocimiento, de forma que ayudando a otro grupo a avanzar, el conocimiento del propio grupo avance también en relación con los objetivos que persigue.

En el seno de estas redes es posible además llevar a cabo discusiones científicas, relativas a la creación del conocimiento, al desarrollo de las actividades en la organización, a la aplicación de los avances tecnológicos, a los procesos de innovación y a la organización del aprendizaje. Todo esto tiene como pilares principales la información provista por el proceso de MD que permitiría todo el intercambio mencionado anteriormente.

CAPÍTULO 3: VALIDACIÓN DE LA PROPUESTA

En este capítulo se analiza la factibilidad del diseño propuesto, desde el punto de vista de su futura implementación así como de la satisfacción de los usuarios finales mediante algunos ejemplos o casos de estudio.

Estos experimentos serán desarrollados con la herramienta WEKA para revelar la información capaz de extraerse a partir de simples datos en bruto a través de las técnicas y algoritmos provistos por la Minería de Datos.

3.1 CASOS DE ESTUDIO

3.1.1 Caso de estudio de entrenamiento: Jugar Tenis.

El primer experimento constituye un caso de entrenamiento al ser un ejemplo clásico dentro de la bibliografía. El mismo expone la relación existente entre un juego de tenis y las condiciones meteorológicas que posibiliten o no el mismo.

Relation: weather

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	**play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Figura 11 Datos del caso de estudio.

La figura 11 muestra la funcionalidad “Viewer” de Weka, como se puede apreciar son 5 atributos con 14 instancias.

En la pantalla principal en la pestaña de preprocesamiento aparece información de utilidad sobre el atributo que se selecciona -en el ejemplo temperatura- , dígase tipo de atributo -numérico-, ocurrencia de ausencia de este valor en los registros -0- , valores distintos -12-, valores únicos -10-, valor máximo -85-, valor mínimo -64-y promedio -73.5-.

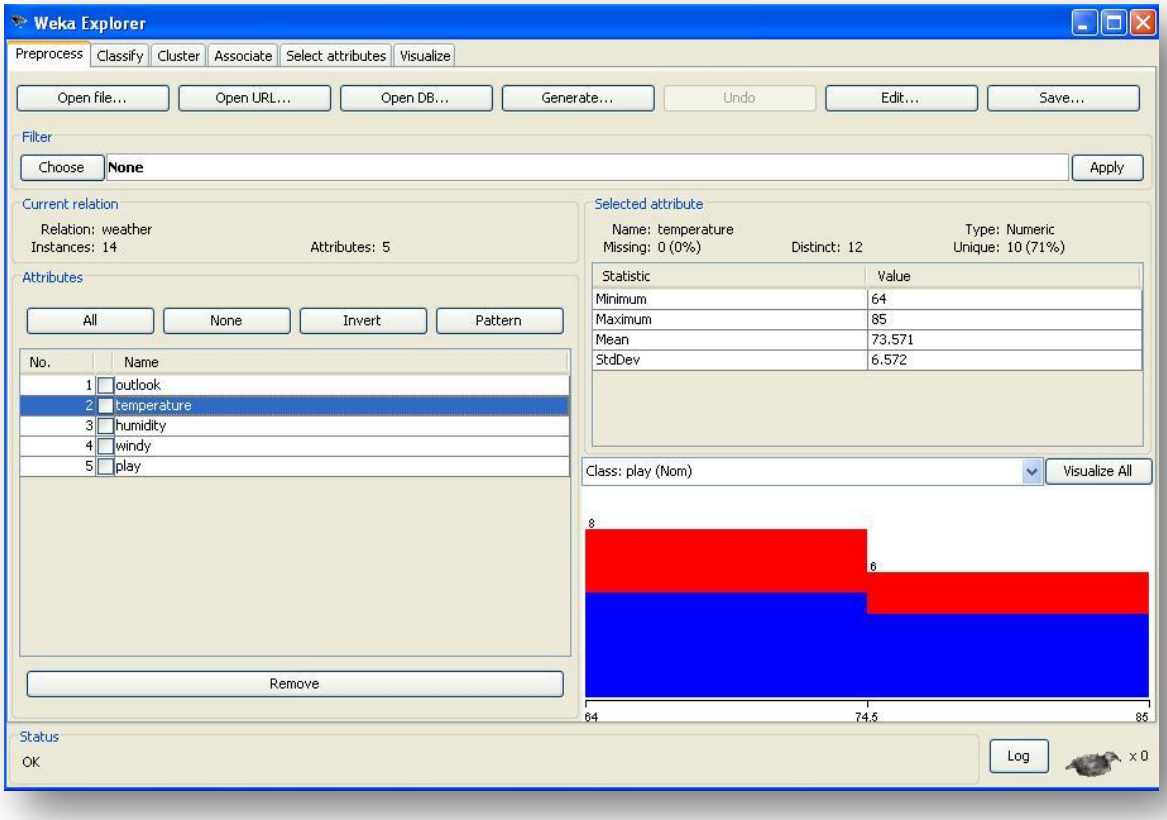


Figura 12 Explorador de WEKA. Preprocesamiento

Otra de las utilidades brindadas durante el preprocesamiento de los datos es la visualización de los atributos en función de una clase. Donde en el histograma obtenido se aprecia la distribución de los valores en columnas y con colores se identifica su distribución en la clase especificada –color rojo: no, color azul: sí-.

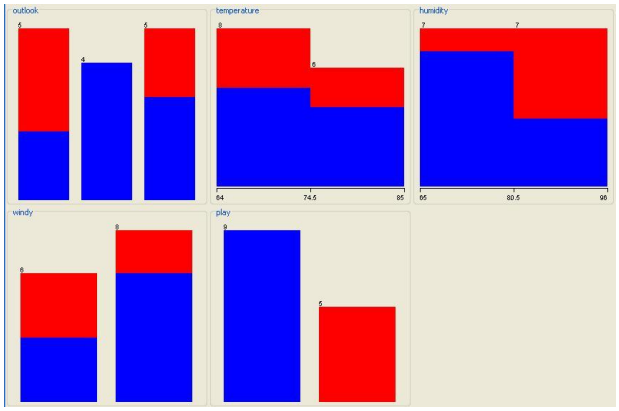


Figura 13 Histograma de los atributos en función de la clase "play".

A continuación se pasará a obtener un modelo aplicando la técnica clasificación a través de un árbol de decisión. El algoritmo que se selecciona es el J48 en WEKA equivalente al C4.5.

El ejemplo se realiza utilizando como validación del modelo la opción "use training set", la cual entrena el método con todos los datos disponibles y posteriormente realiza la evaluación sobre los mismos datos.

Se puede apreciar en la parte derecha de la figura 14 el modelo aprendido de manera textual, el cual muestra información acerca de los parámetros y de la precisión con que se ha realizado.

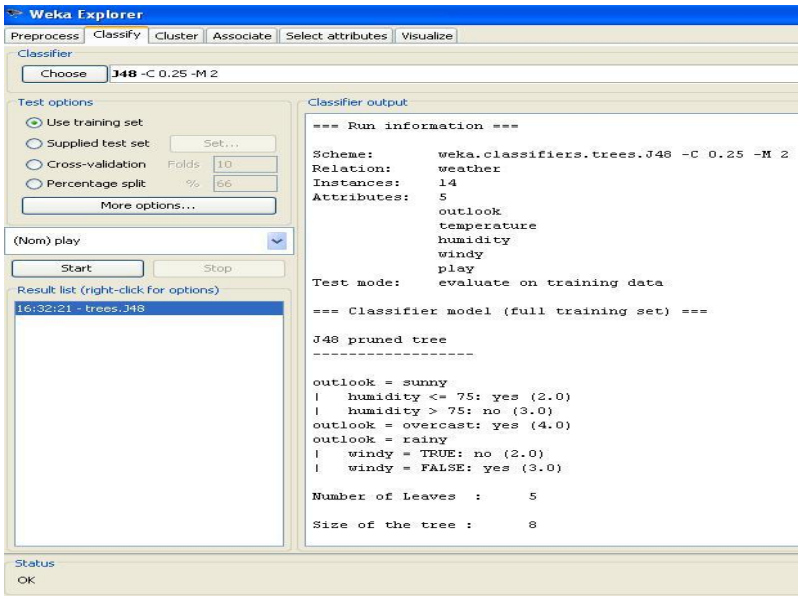


Figura 14 Técnica de Clasificación. Árbol de Decisión. Algoritmo C4.5. Resultados.

Debajo se visualiza el árbol para una mejor comprensión de los datos.

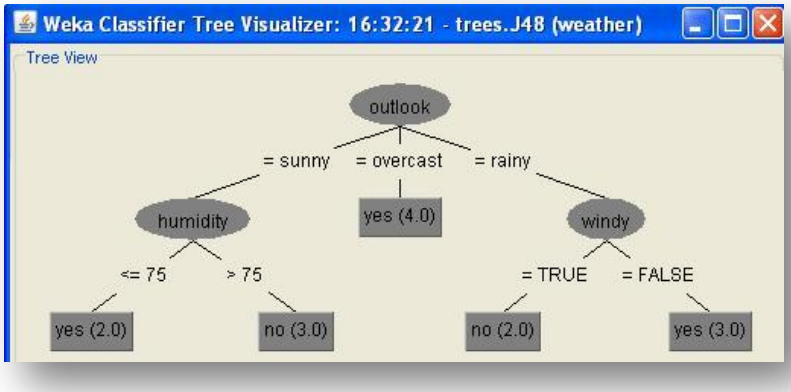


Figura 15 Árbol de Decisión. Resultado obtenido de aplicar el algoritmo C4.5.

Con este árbol se puede llegar a la conclusión de que siempre que esté nublado sin precipitación se podrá jugar tenis, existiendo una proporción directa entre ambos atributos. Además se puede deducir de una forma fácil y rápida que las condiciones para jugar tenis dependen además, de que si está soleado, la humedad sea menor de un 75% y que si está lluvioso, no exista viento.

Dicho experimento nos ha posibilitado predecir en cuestiones de minutos cuáles son las condiciones meteorológicas necesarias con una certeza elevada para poder efectuar un juego de tenis.

3.1.2 CASO DE ESTUDIO: CÁLCULO DE LA NUBOSIDAD.

El segundo experimento se ha realizado a partir de los siguientes datos reales tomados en la estación meteorológica de Casablanca, un total de 8082 registros distribuidos en 17 atributos:

- Visibilidad Horizontal -Km- (VV).
- Nubosidad -octavos de cielo cubierto- (N).
- Temperatura del bulbo seco -grados Celsius- (Ts).
- Temperatura del bulbo húmedo -grados Celsius- (Th).
- Humedad Relativa -%- (HR).
- Temperatura del Punto de Rocío -grados Celsius- (Td).
- Nubosidad de nubes bajas -octavos de cielo cubierto- (Nh).
- Cantidad de Precipitación -milímetros- (RR).

- Cumulonimbo Apilatus (CbCap).
- Cumulonimbo Calvus (CbCal).
- Cúmulo congestus (Cucon).[12]
- Altostratus traslucidus (As1).
- Altostratus opacus (As2)[13].

El objetivo del mismo es determinar los **modelos lineales (LM)** provistos por WEKA para calcular la nubosidad al aplicarse la técnica clasificación, con el algoritmo M5 –M5P, en WEKA-, el cual pertenece a la categoría de los árboles de decisión. Con los modelos obtenidos se hace posible validar las observaciones meteorológicas realizadas por los especialistas, permitiendo detectar posibles errores cometidos durante las mismas.

```
=== Classifier model (full training set) ===

Nh <= 3.5 :
|   Nh <= 1.5 :
|   |   Th <= 22.05 :
|   |   |   VV <= 3.5 :
|   |   |   |   P <= 1017.85 : LM1 (407/69.147%)
|   |   |   |   P >  1017.85 :
|   |   |   |   |   HR <= 82.5 :
|   |   |   |   |   |   Td <= 17.15 : LM2 (126/41.075%)
|   |   |   |   |   |   Td >  17.15 : LM3 (48/68.197%)
|   |   |   |   |   HR >  82.5 : LM4 (110/24.803%)
|   |   |   VV >  3.5 : LM5 (556/90.219%)
|   |   Th >  22.05 :
|   |   |   VV <= 3.5 : LM6 (815/83.545%)
|   |   |   VV >  3.5 :
|   |   |   |   HR <= 65.5 : LM7 (261/88.012%)
|   |   |   |   HR >  65.5 : LM8 (494/100.151%)
|   Nh >  1.5 :
|   |   Td <= 19.85 : LM9 (839/62.895%)
|   |   Td >  19.85 : LM10 (1847/75.129%)
Nh >  3.5 : LM11 (2579/36.576%)
```

```
LM num: 1
N =
        -0.8211 * VV
        + 0 * dd
        - 0.0102 * Ts
        + 0.0226 * Th
        - 0.0006 * HR
        - 0.0072 * Td
        - 0.1194 * Po
        + 0.0002 * P
        + 0.0061 * Nh
        + 0.001 * RR
        + 0.0025 * CbCap
        + 0.0087 * CuCon
        + 0.0774 * As2
        + 124.352
```

[12] **CbCap, CbCal y CuCon** ***son tipos de nubes bajas, las cuales producen precipitación de tipo chubasco.***
[13] **As1 y As2** ***son tipos de nubes medias las cuales producen precipitación del tipo lluvia.***

El algoritmo muestra un resultado de 11 modelos lineales, con un error medio absoluto de 1.2125, la opción de prueba seleccionada para su ejecución fue una validación cruzada de 10 pliegues. Los modelos lineales constituyen las hojas del árbol mostrado a continuación.

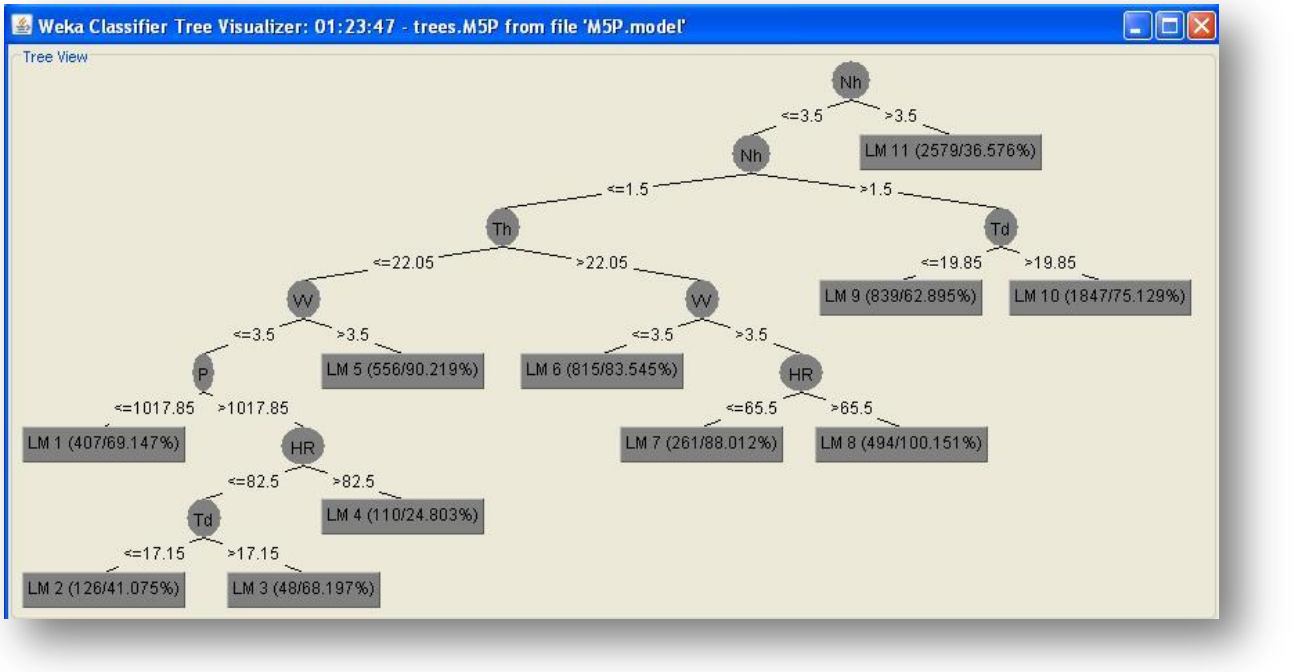


Figura 16 Resultado del algoritmo M5P.

3.1.3 CASO DE ESTUDIO: PREDICCIÓN DE LA TEMPERATURA.

El tercer experimento tiene como objetivo realizar la predicción de la temperatura, lo cual constituiría un servicio de gran utilidad para el Centro de Clima.

Para obtener los modelos de predicción se han tomado las temperaturas de las 10:00 (Temp10) pasado meridiano, la 1:00 (Temp1) y 4:00 (Temp4) antes meridiano con el propósito de predecir la temperatura de las 7:00 (Temp7) antes meridiano.

Para la realización del mismo se han utilizado dos técnicas con resultados de precisión bastante similares, la regresión lineal y el algoritmo aplicado en el caso de estudio anterior, M5, ambos son técnicas de aprendizaje supervisado perteneciente a la categoría de regresión y clasificación respectivamente. A continuación se aprecia sus resultados así como las respectivas fórmulas para realizar dicha predicción.

=== Classifier model (full training set) ===
"Linear Regression Model"

Temp7 =

0.1015 * Temp10 +
0.9362 * Temp4 +
-1.0612

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.9694
Mean absolute error	0.4857
Root mean squared error	0.6552
Relative absolute error	22.2456 %
Root relative squared error %	24.4998
Total Number of Instances	2653

=== Classifier model (full training set) ===
"M5 pruned model tree"
Temp4 <= 23.15: LM1 (1203/26.013%)
Temp4 > 23.15: LM2 (1450/22.325%)

LM num: 1
Temp7 =
0.091 * Temp10
+ 0.935 * Temp4
- 0.8514

LM num: 2
Temp7 =
0.109 * Temp10
+ 0.7849 * Temp4
+ 2.5559
Number of Rules: 2

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.9703
Mean absolute error	0.4778
Root mean squared error	0.646
Relative absolute error	21.8838 %
Root relative squared error	24.1552 %
Total Number of Instances	2653

Se hace el ejercicio en tiempo real en el mes en curso obteniéndose una predicción significativa con una diferencia de pocas décimas con respecto al valor real.

Día	Hora 10:00 pm	Hora 1:00 am	Hora 4:00 am	Hora 7:00 am	Predicción
1	28	27,6	27,5	27	**27,5**
2	26,7	26,4	26,3	26	**26,3**
3	26	25,8	25,2	26,2	**25,2**
4	28,9	27,8	26,8	26,9	**27,0**
5	29,1	28,2	27,4	27,5	**27,5**
6	29,1	27,9	26,1	27,5	**26,3**
7	28,3	27,8	26,3	26,9	**26,4**
8	28,8	27,4	27,2	27,3	**27,3**
9	27,8	26,9	26,2	25,8	**26,3**

10	27,7	26,5	26,7	26,8	**26,7**
11	24,4	24,8	24	25,5	**23,9**
12	28,4	27,5	27	27,5	**27,1**
13	28,1	27,7	25,4	25,8	**25,6**
14	27,8	25,8	26,2	26,3	**26,3**
15	28,4	27,7	26,5	26,8	**26,6**
16	28,2	27,4	26,6	26,6	**26,7**
17	28,6	27,5	26,5	27,2	**26,7**
18	28,1	27,2	26,4	27	**26,5**
19	28,8	28,6	26	27	**26,2**
20	28,7	28	26,2	26,2	**26,4**

Se obtiene como resultado que la temperatura de la 1 de la mañana no es relevante para la estimación de la temperatura de las 7 de la mañana, esto se corrobora con el gráfico de las temperaturas registradas en el termograma de la estación meteorológica Casablanca porque la marcha de la temperatura en el intervalo de 10 pm a 7 am, es prácticamente una recta, donde el valor de la 1 de la mañana es muy próximo a ella.

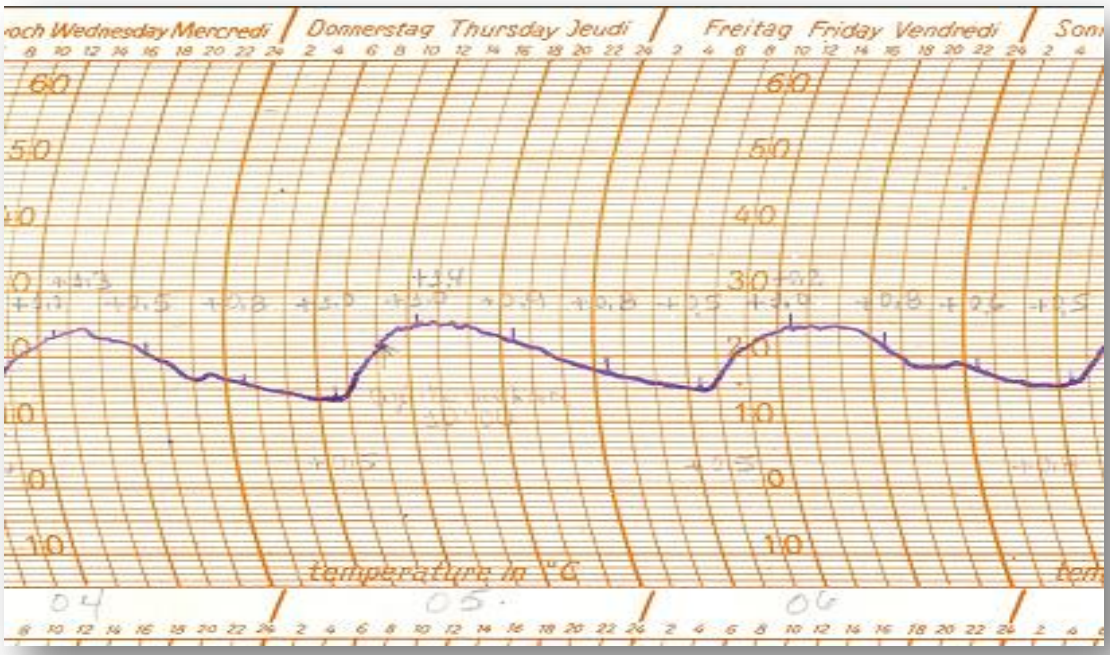


Figura 17 Termograma de la estación meteorológica de Casablanca.

3.2 ANÁLISIS INTEGRADO DE LOS RESULTADOS.

La realización de estos casos de estudios con la herramienta WEKA obtuvo los siguientes resultados:

- Su relevancia para el control de la calidad de los datos, a partir de los ejercicios resueltos que han permitido la validación de diferentes elementos climáticos aplicando las técnicas de Minería de Datos.
- El poder calcular (o estimar) los valores faltantes en la base de datos de las distintas variables.
- Modelos matemáticos para predecir elementos climáticos.
- Todas estas técnicas con sus modelos se incorporarán como nuevas herramientas al SADC para aumentar la calidad de los servicios provistos por el mismo en la actualidad.

Estos resultados son de interés para las funciones realizadas en el Centro de Clima, ver anexo 1.

CONCLUSIONES

Con la presente investigación se dan por cumplidos los objetivos propuestos en el trabajo de diploma:

- Se realizó un estudio de los principales conceptos y características de temas de interés como: Bases de Datos, Descubrimiento de Conocimientos en Bases de Datos, Procesamiento en Paralelo y Sistemas de Información.
- Se profundizó en el estudio de las principales técnicas supervisadas y no supervisadas, algoritmos y herramientas provistos por la Minería de Datos.
- Se seleccionó e instaló en el INSMET la herramienta WEKA.
- Se efectuó un conjunto de pruebas para demostrar la pertinencia y factibilidad de la propuesta.
- Se desarrolló la propuesta de un Sistema de Información para la Gestión de Datos Climáticos teniendo en cuenta la tecnología existente en el Centro de Clima del INSMET y sus requisitos, el cual permitirá ampliar y brindar nuevos servicios de información con los resultados obtenidos y generar un nuevo conocimiento para los investigadores del centro, que a su vez tendrán a su disposición un espacio de colaboración e intercambio a través de la Red de Conocimientos.

BIBLIOGRAFÍA

Altamirano, R. B., Pupo, A. R., Milanés, N. M., Ramírez, M. S., & Suárez, R. P. (2006). *SISTEMA DE ADMINISTRACION DE DATOS CLIMATICOS.* Informe Científico de Resultado, Instituto de Meteorología, La Habana.

AstroSeti. (2003). *Seti@Home Detección de Señales de otras civilizaciones*. Recuperado el 7 de Abril de 2010, de http://seti.astroseti.org/setiathome/

Aular, M., Josefina, Y., & Talavera Pereira, R. (2007). Minería de Datos como soporte a la toma de decisiones empresariales. *23* (52), 104-118.

Berthold, M., & Hands, D. (2003). *Intelligent Data Analysis. An Introduction* (Second Edition ed.). Springer.

Blanco, L. (2004). Complejidad, caos y administración de empresas. Un acercamiento desde los sistemas de información y conocimiento. La Habana.

Bressán, G. E. (2003). *Almacenes de datos y Minería de Datos.* Trabajo monográfico de adscripción para la Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura, Argentina.

Curto Díaz, J. (2010). *Recursos globales para expertos de Business Intelligence y Data Warehousing*. Recuperado el 16 de Marzo de 2010, de BeyeNetwork: http://www.beyenetwork.es/blogs/curtodiaz/archives/data-mining/

Date, C. J. (2003). *Introducción a los Sistemas de Bases de Datos.* La Habana, Cuba: Félix Varela.

Dill, F., & Domko, M. (2010). *KNIME*. Recuperado el 21 de Marzo de 2010, de http://www.knime.org/

Fayyad, U. M., Piatetsky-Shapiro, G., Smith, P., & R., U. (1996). *Advances in Knowledge Discovery and Data Mining.* AAAI Press / The MIT Press.

Febles Rodríguez, J. P., & González Pérez, A. (2002). Aplicación de la minería de datos en la bioinformática. *ACIMED , 10* (2).

Frank, E., & Witten, I. H. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (Segunda ed.). San Francisco, California, Estados Unidos: Morgan Kaufmann.

García Morate, D. (2001). *Manual de Weka.*

Gil Bellosta, C. J. (2010). *Datanalytics*. Recuperado el 14 de Marzo de 2010, de http://www.datanalytics.com/

Gobierno de Australia. Departamento de Meteorología. (20 de Octubre de 2005). (Commonwealth) Recuperado el 10 de febrero del 2010 de febrero de 2010, de http://www.bom.gov.au/wmo/climate/ccl/opag1.shtml

Haag, S. (2007). *Management Information Systems for the Information Age* (7ma ed.). México: McGraw-Hill.

Henandez Orallo, J., Ramírez Quintana, J., & Ferri, C. (2001). *Introducción a la Minería de Datos.* Madrid: Pearson.

Hernández Orallo, J., & Ferri Ramírez, C. (Marzo 2006). *Práctica de Minería de Datos. Introducción al WEKA.* Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software, Universidad Politécnica de Valencia, Valencia.

Hornick, M. F., Marcadé, E., & Venkayala, S. (2007). *Java Data Mining: Strategy, Standard and Practice.* San Francisco, California, Estados Unidos: Morgan Kaufmann.

Información Creativa. (17 de Marzo de 2010). *Catalogo de Soiftware.com*. Recuperado el 20 de Marzo de 2010, de Portal Especializado de Software y Servicios Relacionados para el Sector Empresarial: http://www.catalogodesoftware.com/

Jesús, E. D., & Zaiane. (30 de Noviembre de 2009). *MATI: Sobre la letra digital.* Recuperado el 8 de febrero de 2010, de MATI: Sobre la letra digital: http://www.mati.unam.mx

Jonassen, D., Mayes, T., & McAlesee, R. (1992). *Designing Environments for Constructive Learning.* Berlín: Springer-Verlag.

KDnuggets. (2010). *Data Mining Community's Top Resource since 1997*. Recuperado el 15 de Marzo de 2010, de http://www.kdnuggets.com/software/suites.html

Kioskea.net. (24 de Abril de 2009). *Kioskea.net*. Recuperado el 5 de Abril de 2010, de Kioskea.net: http://es.kioskea.net/contents/pc/processeur.php3

Lichilín Ríos, Y., & Rodríguez Betancourt, Y. L. (Junio, 2008). *Propuesta de un Sistema de Información para el Control Interno.* Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas, Ciudad de La Habana.

Marante Jacas, D., & Marante Jacas, D. (2009). *Aplicación de la minería de datos para la exploración y detección de patrones delictivos.* Trabajo de Diploma para optar por el título de Ingeniero Informático, Universidad de las Ciencias Informáticas, La Habana, Cuba.

Martín-Moreno Cerrillo, Q. (diciembre, 2004). *Aprendizaje colaborativo y redes de conocimiento.* Granada: Grupo Editorial Universitario.

Microsoft Corporation. (2010). *Microsoft|Technet*. Recuperado el 7 de Marzo de 2010, de Recursos para Profesionales de TI: http://technet.microsoft.com/es-es/library/

Molina Félix, L. C. (Noviembre de 2002). *Universidad Abierta de Catalunya(UOC).* Recuperado el 5 de marzo de 2010, de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

Molina López, J. M., & García Herrero, J. (2004). *Técnicas de Análisis de Datos: Aplicaciones prácticas utlizando Microsoft Excel y WEKA.* Madrid.

Moreno Beltrán, R., & Castellanos Galvin, S. (2003). Definición de un modelo de redes de conocimiento como soporte a la transferencia del conocimiento generado en clusters de investigación. *GTI (Gerencia,Tecnológica, Informática) , 2* (2).

Moreno García, M. N., Miguel Quintales, L. A., García Peñalvo, F. J., & Polo Martín, M. J. *Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software.* Universidad de Salamanca, Departamento de Informática y Automática.

Nicholson, S. (9 de Enero de 2006). *Bibliomining Retrospective.* Recuperado el 3 de Marzo de 2010, de Bibliomining Retrospective: http://www.bibliomining.com/

Núñez. E., M. A., & Cubas. D., L. O. (2005). Programa SAROM.

Orallo Hernández, J., Quintana Ramírez, M. J., & Ramírez Ferri, C. (2004). *Introducción a la Minería de Datos.* Prentice Hall.

Rodríguez, A., Araujo, A., & Yulianov, E. (2003). Redes virtuales para la gestión del conocimiento: El caso de las universidades. *Centro para la gestión del conocimiento en la universidad* , 427-439.

Rojas Figueroa, E. (6 de Mayo de 2009). *Edays: A los 21, tecnologias y herramientas en tus manos.* Recuperado el 1 de Marzo de 2010, de http://edays.netau.net/2009/05/06/inteligencia-artificial-algoritmos-geneticos/

Romel, V. R., González Herrera, I. Y., & Pérez Risquet, C. (2008). *Estudio de la factibilidad de la aplicación de técnicas de visualización de datos multiparamétricos para el análisis visual de datos meteorológicos.* Universidad Central de Las Villas Marta Abreu (UCLV), Villa Clara.

Royero, J. (2004). *Las redes sociales de conocimiento: El nuevo reto de las organizaciones de investigación científica y tecnológica.* Anaco, Venezuela.

Silveira Martineaux, K., & Fernández Pérez, R. (2008). *Comparación de algoritmos de clasificación y agrupamiento aplicando técnicas de Minería de Datos.* Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas, Universidad de las Ciencias Informáticas, Ciudad de La Habana.

Vallejos, S. J. (2006). *Minería de Datos.* Trabajo de Adscripción , Universidad Nacional del Nordeste, Argentina.

Witten, I. H., Frank, E., Trigg, L., Geoffrey Holmes, M. H., & Jo Cunningham, S. (1999). *Weka: Practical machine learning tools and techniques with java implementations.* Universidad de Waikato, Departamento de Ciencias de la Computación, Nueva Zelanda.

Zhu, X., & Davidson, I. (2007). *Knowledge discovery and data mining; challenges and realities.* Farmington Hills, Michigan: Book News, Inc.

ANEXO 1

AVAL

Por este medio se avala que el trabajo de diploma **"Propuesta de un Sistema de Información para la Gestión de Datos Climáticos"**, de los alumnos José Barlia Bernal y Jorge Fonseca Córdova, tiene una aplicación práctica directa en las funciones que realiza el Centro del Clima (CenClim), del Instituto de Meteorología. Como principales resultados se han obtenido los siguientes:

1. Ejecución de ejercicios de validación de diferentes variables climáticas, en el control de calidad de los datos, con la aplicación de reglas de asociación entre ellas.
2. Cálculo (o estimación) de los valores faltantes de distintas variables en el Sistema de Administración de Datos Climáticos.
3. Determinación de fórmulas para predecir variables climáticas como la temperatura.

Aunque son preliminares, todos ellos fueron implementados por el Grupo de Datos del CenClim en este período. Sin embargo, lo más importante en esta dirección es que en el futuro será posible la introducción a fondo de tales técnicas en la Climatología General y Aplicada, a fin de encontrar relaciones no visibles entre los diferentes elementos climáticos, en un tiempo considerablemente menor que de costumbre.

Es de nuestro interés la continuación de trabajos de esta índole, que impulsan el desarrollo de la actividad climatológica en nuestro país.

A. Vladimir Guevara
Jefe Centro del Clima – Instituto de Meteorología

Dado en Casablanca, Ciudad de La Habana, a los 18 dias del mes de junio del año 2010, "Año 52 de la Revolución".