UNIVERSIDAD CENTROCCIDENTAL "LISANDRO ALVARADO"
DECANATO DE CIENCIAS Y TECNOLOGIA

APLICACIÓN DE MINERIA DE DATOS PARA LA PREDICCION DEL RENDIMIENTO DE LA CAÑA DE AZUCAR

MARIA AUXILIADORA PEREZ

Barquisimeto, Febrero de 2005

UNIVERSIDAD CENTROCCIDENTAL "LISANDRO ALVARADO"
DECANATO DE CIENCIAS Y TECNOLOGÍA
POSTGRADO EN CIENCIAS DE LA COMPUTACION

APLICACIÓN DE MINERIA DE DATOS PARA LA PREDICCION DEL RENDIMIENTO DE LA CAÑA DE AZUCAR

Trabajo de grado para optar al grado de

Magíster Scientiarum en Inteligencia Artificial

Por: MARIA AUXILIADORA PEREZ

Barquisimeto, Febrero de 2005

Al Padre Todopoderoso, fuente de todo conocimiento
A mis padres, guías y pilares de mi vida
A mi abuela Chica, mi eterna inspiración
A Neymar y Nelson por su amor

AGRADECIMIENTO

A la Universidad Centroccidental "Lisandro Alvarado" por proporcionarme las herramientas necesarias para la formación y conocimiento durante estos años de estudio.

A la profesora Maritza Bracho de Rodríguez por su asesoría, supervisión y amable colaboración en la elaboración de este trabajo.

A la profesora Gladys Marante por ser una fuente de motivación y modelo a seguir.

Al los profesores Carlos Lameda y Sonia Córdoba por colaborar en todo momento y ser un punto de apoyo.

A mis familiares, amigos y compañeros de trabajo por su paciencia y colaboración.

A mis compañeros de estudio, con quienes tuve la oportunidad de compartir conocimientos a lo largo de la Maestría en Ciencias de la Computación Mención Inteligencia Artificial.

Al personal de la Gerencia de Gestión de Sistemas de la empresa Azucarera Río Turbio por brindarme un ambiente propicio para aplicar todos los conocimientos adquiridos en esta Maestría.

A todos aquellos que extendieron su mano amiga....A todos Muchas Gracias.

INDICE

INDICE DE ILUSTRACIONES

Figura **Página**

Cuadros **Página**

Tablas **Página**

APLICACIÓN DE MINERIA DE DATOS PARA LA PREDECCION DEL RENDIMIENTO DE LA CAÑA DE AZUCAR

Autor: María Auxiliadora Pérez
Tutor: Maritza Bracho de Rodríguez
Año : 2005

RESUMEN

En la actualidad existe un gran interés comercial e investigativo en el desarrollo de nuevas tecnologías en el área del Descubrimiento de Conocimiento en Bases de Datos y más específicamente en el área de Minería de Datos, en virtud del innegable valor táctico y estratégico de la información almacenada en una Base de Datos. El descubrir patrones relevantes en los datos almacenados en las bases de datos y convertirlos en conocimiento útil para la toma de decisiones es sin lugar a dudas una meta tras la cual se han reunido esfuerzos significativos. El reto de encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes representa una línea de investigación en constante crecimiento. Este trabajo presenta la aplicación de un proceso de Minería de Datos en el contexto de los datos agrícolas y más específicamente para la predicción del rendimiento de la Caña de Azúcar, mediante el diseño y desarrollo de modelos de Aprendizaje Automático. Los modelos obtenidos mediante las técnicas de Arboles de Decisión utilizando el algoritmo C4.5 y las redes neuronales multicapas con retropropagación, demuestran ser una solución eficiente al problema planteado. La identificación de las características que describen las unidades de producción mayores rendimientos de la empresa y la relación existente entre las variables objeto de estudio, constituye el aporte de nuevo conocimiento mediante el estudio de los datos históricos almacenados en la Base de Datos.

Palabras Clave: Descubrimiento de Conocimiento, Minería de Datos, Predicción, Aprendizaje Automático.

INTRODUCCION

Cada día se genera una gran cantidad de información en el mundo, con el aumento considerable del tamaño de las Bases de Datos y los requerimientos de información del usuario. Básicamente, la generación masiva de datos surge y se ve notablemente beneficiada por los avances tecnológicos de los sistemas de almacenamiento y las mejoras sustanciales a nivel de costo de los mismos.

Sin lugar a dudas, el valor táctico o estratégico de los grandes almacenes de datos está en proporción directa con la capacidad de analizarlos. De esta forma surge la interrogante respecto a cómo los sistemas de información pueden constituir un elemento vital del negocio y como los datos almacenados en los mismos, representan un elemento de inestimable importancia para la toma de decisiones de la empresa.

La mayoría de las empresas multinacionales generan más información en una semana que la que cualquier persona podría leer en toda su vida, e incluso las pequeñas empresas generan un volumen de datos que no son capaces de manejar. Contrariamente a lo que pudiera esperarse, esta explosión de datos no supone un aumento de nuestro conocimiento, puesto que resulta imposible procesarlos con los métodos clásicos. Técnicas tradicionales de análisis de información como lo son: análisis de regresión, de clusters, de discriminantes, de factores, entre otros, (Michalsky y otros (1983)), ya han sido rebasadas tanto por la cantidad, como por la velocidad con la que crecen los datos.

Para superar este problema, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos y permiten realizar un análisis con detenimiento de los mismos de forma automática. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

Así nace el concepto de minería de datos, cuyo objetivo fundamental es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante

algoritmos eficientes. Dada la gran gama de hipótesis plausibles que se ajustan a los datos, el problema computacional representa un reto hasta ahora poco enfrentado.

La búsqueda de información en grandes volúmenes de datos, surgidos por la acumulación a lo largo de cierto tiempo, ha sido objeto de especial interés del ámbito académico, en el área de la estadística y más recientemente en Inteligencia Artificial. Es en esta área donde se estudian y desarrollan algoritmos que implementan los distintos modelos de aprendizaje y su aplicación a la resolución de problemas prácticos.

Aplicaciones de minería de datos han sido desarrolladas en diversas áreas, entre las que se pueden mencionar: detección de fraudes, optimización de campañas de mercadeo, análisis de riesgos en créditos, descripción y segmentación de clientes, clasificación de cuerpos celestes, minería de texto, minería en Internet, entre otras.

Un caso particular lo constituye la industria azucarera nacional, la cual posee un interés comercial por explotar la información almacenada en sus Bases de Datos con la finalidad de convertir dicha información en conocimiento apropiado para satisfacer las necesidades de los usuarios, de forma tal que este nuevo conocimiento sea estratégico en un mundo globalizado como el nuestro.

Es importante señalar que los métodos de aprendizaje automático, empleados en el proceso de minería de datos no corresponden a un estándar genérico que resuelve todo tipo de problemas, sino que consisten en una metodología dinámica e iterativa que depende en gran parte del problema planteado, de la disponibilidad de la fuente de datos, del conocimiento de las herramientas necesarias, de la metodología aplicada, de los requerimientos y de los recursos con los que se cuente al momento de implementar dicho proceso.

CAPITULO I

EL PROBLEMA

Planteamiento del Problema

Las técnicas de minería de datos se utilizan para mejorar los procesos de negocio en los que se maneja un gran volumen de datos. La construcción de modelos que aporten soluciones a problemas del mundo real es uno de los retos que actualmente tiene la Minería de Datos como herramienta de descubrimiento de conocimiento inteligente.

Los algoritmos de minería de datos realizan en general tareas de predicción de datos desconocidos. Es así como, predecir el comportamiento de una variable o de un conjunto de variables dado su comportamiento en el pasado, es uno de los problemas más estudiados en este contexto.

Existen implementaciones de procesos de minería de datos en empresas del área agroalimentaria (Díaz y Morillas (2003)), donde se analizan las características contables de estas empresas que definen las de mayor rentabilidad económica, generando un modelo de reglas difusas extraídas de una base de datos.

Sin embargo, la aplicación de procesos de minería de datos inteligente en empresas agrícolas, enfocando el estudio desde el punto de vista del análisis de las características de la producción y manejo del cultivo, con el objeto de mejorar sus índices de rendimiento y rentabilidad, es un área donde este tipo de estudios tiene un carácter inédito.

Uno de los problemas que actualmente está enfrentando la industria azucarera nacional, es la disminución de la productividad, representada por las toneladas de caña de azúcar producida por hectárea cultivada y la disminución del rendimiento de

la misma, constituida por el porcentaje de sacarosa extraída. Aunque se ha experimentado un aumento de la superficie cultivada, la producción de este renglón alimenticio no muestra el mismo comportamiento.

Es así como se propone efectuar un estudio que dadas las características de las unidades de producción de caña de azúcar y los rendimientos obtenidos en el pasado, permita predecir el rendimiento de las mismas. Las características de las unidades de producción de caña de azúcar corresponden a los valores técnicos que describen el comportamiento de la misma y se encuentran registrados en una Base de Datos como resultado del registro de las operaciones de la empresa mediante sus sistemas de información.

El rendimiento es un valor que se obtiene de análisis físico-químicos practicados a muestras, tanto de la caña cultivada como de la sacarosa producto del procesamiento de la caña. El análisis del comportamiento del rendimiento es elaborado utilizando técnicas tradicionales de análisis de datos como lo es el análisis estadístico.

El descubrimiento de nuevo conocimiento en esta área, que pueda servir de apoyo para la toma de decisiones técnicas en el proceso de cosecha y en la determinación de políticas de gestión que contribuyan a mejorar sustancialmente la productividad y los rendimientos de la empresa, representa uno de los propósitos de la presente investigación.

Objetivos de la Investigación

Objetivo General

Desarrollar un modelo de aprendizaje automático que permita predecir el rendimiento de la Caña de Azúcar, mediante la aplicación de minería de datos.

Objetivos Específicos

- Desarrollar un modelo basado en Árboles de Decisión que permita predecir el rendimiento de una unidad de producción de caña de azúcar, dadas sus características.
- Desarrollar un modelo basado en Redes Neuronales Multicapas con retropropagación para predecir el rendimiento de una unidad de producción de caña de azúcar, dadas sus características.
- Comparar los modelos desarrollados para establecer el que aporte una mejor solución al problema planteado en términos de exactitud e interpretabilidad.
- Aplicar la metodología CRISP-DM en el desarrollo del modelo, para lograr una mejor definición y validación del mismo

Justificación e Importancia

El estudio y aplicación de técnicas inteligentes para el análisis de información almacenada en bases de datos es un tema de investigación en el que ya se han involucrado muchas especialidades, pero que sus aplicaciones en agronomía y más específicamente en la producción de azúcar aún no han sido explotadas suficientemente.

La aplicación de un proceso de Minería de Datos en el área agrícola, concretamente en la producción azucarera, apoyará los procesos de toma de decisión de la unidad de Gestión Agrícola de la empresa Azucarera Río Turbio C.A., tanto en el ámbito técnico como el gerencial. La necesidad existente de herramientas inteligentes que apoyen sus procesos de decisión con la subsiguiente mejora en la rentabilidad económica y un incremento de la productividad, que beneficiaria tanto a los productores como al industrial, con la incorporación de nuevo conocimiento, mas allá de lo que en la actualidad están reportando sus sistemas de información, justifica el presente trabajo y los objetivos fundamentales del mismo.

La existencia de grandes volúmenes de información estructurada y almacenada en Bases de Datos de los procesos del negocio y la necesidad de análisis de esta información, posibilita la creación de una nueva generación de técnicas y herramientas computacionales con la capacidad de asistir a usuarios en el análisis automático e inteligentes de datos.

En este contexto, la ausencia de trabajos exhaustivos de minería de datos aplicados al área Agrícola Azucarero, permitirá la aplicación de la experiencia obtenida mediante esta investigación en otras áreas del negocio, generando la posibilidad de nuevos estudios e investigaciones.

Alcance

El proyecto consistirá en la aplicación de un proceso de Minería de Datos, en el cual se obtendrán dos modelos de Aprendizaje Automático para la predicción del rendimiento de la caña de azúcar, para luego comparar cual de ambos modelos resuelve el problema planteado de acuerdo a criterios de exactitud e interpretabilidad.

La aplicación de los algoritmos de minería de datos requiere la realización de una serie de actividades previas orientadas a preparar los datos de entrada y generar el conjunto de datos para el entrenamiento y validación del modelo, según el Proceso Estándar entre Industrias para Minería de Datos *"CRISP-DM",* debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido.

Los métodos a considerar para la implementación de dichos modelos son: Aprendizaje Inductivo por Arboles de Decisión y Aprendizaje en Redes Neuronales Multicapas con Retropropagación. Los aspectos teóricos de éstas técnicas se mostrarán en el Capitulo II. Basándose en los resultados de la experimentación con éstas técnicas, se escogerá el modelo que resuelva el problema planteado de acuerdo a los criterios de exactitud y generalidad.

Limitaciones

Las técnicas de Minería de Datos han surgido a partir de sistemas de aprendizaje inductivo en computadores, siendo la principal diferencia entre ellos los datos sobre los que se realiza la búsqueda de nuevo conocimiento (Holsheimer y Siebes (1994)). En el caso del aprendizaje de máquinas, se usa un conjunto de datos pequeño y cuidadosamente seleccionado para entrenar el sistema. En la Minería de Datos, se parte de una Base de Datos, generalmente grande, en la que los datos han sido generados y almacenados para propósitos diferentes del aprendizaje de los mismos.

La mayoría de los algoritmos de aprendizaje al ser aplicados sobre las Bases de Datos, se encuentran con dificultades no previstas por los sistemas de aprendizaje tradicionales, puesto que en el mundo real, las bases de datos suelen ser dinámicas, incompletas, ruidosas y muy grandes (Frawley y otros (1991)) y gran parte del trabajo que se realiza en la inducción de conocimiento en bases de datos trata de solucionar estos problemas.

Entre los inconvenientes que pudieran presentarse durante el desarrollo de la investigación están:

1. Datos incompletos: El manejo de datos incompletos en una base de datos puede deberse a pérdida de valores de algún atributo, o a la ausencia del mismo en la vista que el sistema posee sobre los datos. El impacto en los resultados dependerá de si el dato incompleto es relevante o no para el objetivo del sistema de aprendizaje.
2. Ruido e incertidumbre: El ruido presente en los datos viene determinado tanto por el tipo de valores de los atributos como por la exactitud en la medida de dichos valores.
3. Tamaño de la base de datos: El tamaño de la base de datos suele ser muy superior al tamaño del conjunto de entrenamiento de muchos sistemas de aprendizaje, es por ello que en las bases de datos muy grandes, un análisis completo de todos los datos es inabordable y deben emplearse técnicas específicas que aceleren el aprendizaje sobre las mismas.

4. Datos Dinámicos: En la mayoría de las bases de datos, los datos son modificados de forma continua. Cuando el valor de los datos almacenados está en función del tiempo, el conocimiento inducido varía según el instante en que se obtenga, y por ello es deseable un sistema que funcione de forma continua, en modo secuencial, para tener siempre actualizado el conocimiento extraído.

CAPITULO II

MARCO TEORICO

Antecedentes

Después de realizadas las búsquedas pertinentes en la Universidad Centroccidental "Lisandro Alvarado", en el Instituto Nacional de Investigaciones Agrícolas y en la empresa Azucarera Río Turbio C.A., se ha determinado que los antecedentes de ésta investigación son escasos, ya que no existen proyectos realizados con anterioridad que enfoquen específicamente la Minería de Datos como una herramienta para la predicción del rendimiento de la caña de azúcar *(Saccharum spp*.), por lo que no se pueden citar otras investigaciones para este fin, a menos que sean simplemente investigaciones basadas en Minería de Datos sin ser aplicadas a la industria azucarera y más específicamente en el desarrollo de un modelo de aprendizaje automático que permita predecir el rendimiento de la caña de azúcar.

Tal es el caso de (Díaz y Morillas (2003)), donde se implementó un proceso de Minería de Datos para caracterizar contablemente las empresas agroalimentarias de mayor rentabilidad económica, generando para ello un modelo de reglas difusas. En dicho trabajo se expone la metodología aplicada para el desarrollo del modelo, así como también las técnicas utilizadas para la construcción del modelo generado.

Así mismo, existen varios ensayos de campo aplicados en el área de influencia de Azucarera Río Turbio C.A. tales como: Mago y otros (1986), Zérega y otros (1991), De Sousa y Rea (1993) que realizan análisis de variables que afectan el rendimiento y la calidad del azúcar producida, utilizando para ello técnicas estadísticas, por lo cual sirven de aporte documental a la presente investigación, sin que de hecho constituyan un antecedente de la misma.

Bases Teóricas

Evolución de la Tecnología de Minería de Datos

En la década de los años sesenta surgen los Sistemas de Gestión de Bases de Datos, cuya función principal es la de proporcionar la infraestructura necesaria para almacenar, recuperar y manipular datos, soportando transacciones y actividades en línea.

En la década de los ochenta nace el concepto de Almacén de Datos. Según Han (2001), un Almacén de Datos es una colección de datos orientada a temas, integrado, no volátil y variante en el tiempo para el soporte del proceso de toma de decisiones.

Así, se denomina Almacenamiento de Datos al proceso de construcción y uso de un Almacén de Datos. La construcción de un Almacén de Datos requiere de la integración, limpieza y consolidación de los datos. La utilización de un Almacén de Datos frecuentemente necesita de una colección de tecnologías de soporte de decisiones (Han (2001)).

El objetivo del Almacén de Datos es agrupar los datos con el propósito de facilitar su posterior análisis, de forma que sean útiles para acceder y analizar información sobre la propia empresa. A este tipo de datos se les conoce como “informativos”.

Dada la complejidad que pudiera tener un Almacén de Datos se ha planteado la necesidad de abordar los proyectos por áreas temáticas de análisis, así en lugar de crear el "gran repositorio" de la empresa, se han desarrollado proyectos más pequeños, creando así cubos de información que respondan las necesidades de un área específica, por ejemplo el Cubo de Datos de mercadeo, o dentro del Cubo de Datos de Mercadeo el Cubo de Datos de Nuevos Negocios, entre otros. Un Cubo de Datos puede verse como una bodega dentro de un gran almacén de datos, que alberga datos para un propósito específico.

Los sistemas que manejan estos datos se denominan Sistemas de Procesamiento Analítico en Línea. Esta tecnología está basada en el concepto de cubo de información. Un cubo de información es una estructura para almacenar información

que permite realizar análisis multidimensional y se basa en métricas y dimensiones. Una métrica es una medición matemática de una variable del negocio, representa lo que se quiere medir. Una dimensión es la variable contra la que se quiere medir.

Estas herramientas ofrecen un gran poderío para revisar, graficar y visualizar información multidimensional, en características temporales, espaciales o propias. Requieren de una alta participación de un usuario humano, pues son interactivas y requieren de la guía de un experto.

Es así como, la evolución de estas tecnologías nos ha llevado a lo que conocemos con el nombre de minería de datos, tal como se muestra en la figura 1, en respuesta a la automatización de las tareas de análisis de información y a la necesidad de obtener el conocimiento oculto que guardan los datos.

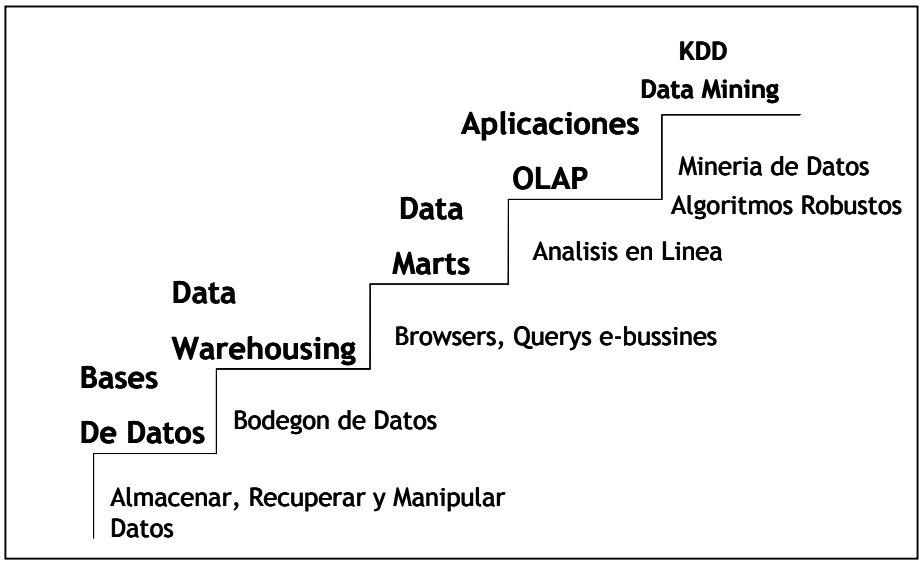

Figura 1. Evolución de la Minería de Datos. (El Autor)

Una definición de Minería de Datos es según Bigus (1996), el descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos.

En la definición anterior el término "*información valiosa*" se entiende como información que ayuda al proceso de toma de decisiones ó representa una ventaja competitiva para el negocio. El término "*gran colección de datos*" se considera

como una cantidad de información almacenada que va desde un gigabyte hasta cientos de terabytes.

Según Fayyad y otros (1996) el proceso de Descubrimiento de Conocimiento en Bases de Datos se define como el proceso de identificación no trivial de patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles en los datos. Aquí el término "*datos*" representa un conjunto de hechos, por ejemplo, casos en una Base de Datos.

Etapas del Proceso de Minería de Datos

Usualmente, un estudio de Descubrimiento de Conocimiento comprende de la aplicación iterativa e interactiva de los siguientes pasos: (a) preparación de datos, (b) selección de características, (c) aplicación de un algoritmo de extracción de conocimiento, (d) evaluación e interpretación del modelo resultante para tomar la decisión de qué constituye conocimiento y qué no lo es. Es oportuno mencionar que varios autores se refieren al proceso de minería como la aplicación de un algoritmo para extraer patrones de datos y a Descubrimiento de Conocimiento al proceso completo (pre-procesamiento, minería, post-procesamiento).

En la figura 2 se ilustran los pasos del proceso de Descubrimiento de Conocimiento en Bases de Datos.

Así, la preparación de los datos se refiere al proceso de filtrar los datos originales contenidos en la fuente de datos ya sea una base de datos ó un almacén de datos, ya que la mayoría de las veces no es posible utilizar un algoritmo de minería sobre estos datos. Este proceso permite eliminar valores incorrectos, no válidos, desconocidos, según las necesidades y el algoritmo a usar, se obtienen muestras de los mismos ó reducen el número de valores posibles.

La etapa de *selección de características* reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería. Los métodos para la selección de características son básicamente dos: aquellos basados en la elección de los mejores

atributos del problema, y aquellos que buscan variables independientes mediante pruebas de sensibilidad, algoritmos de distancia o heurísticos.

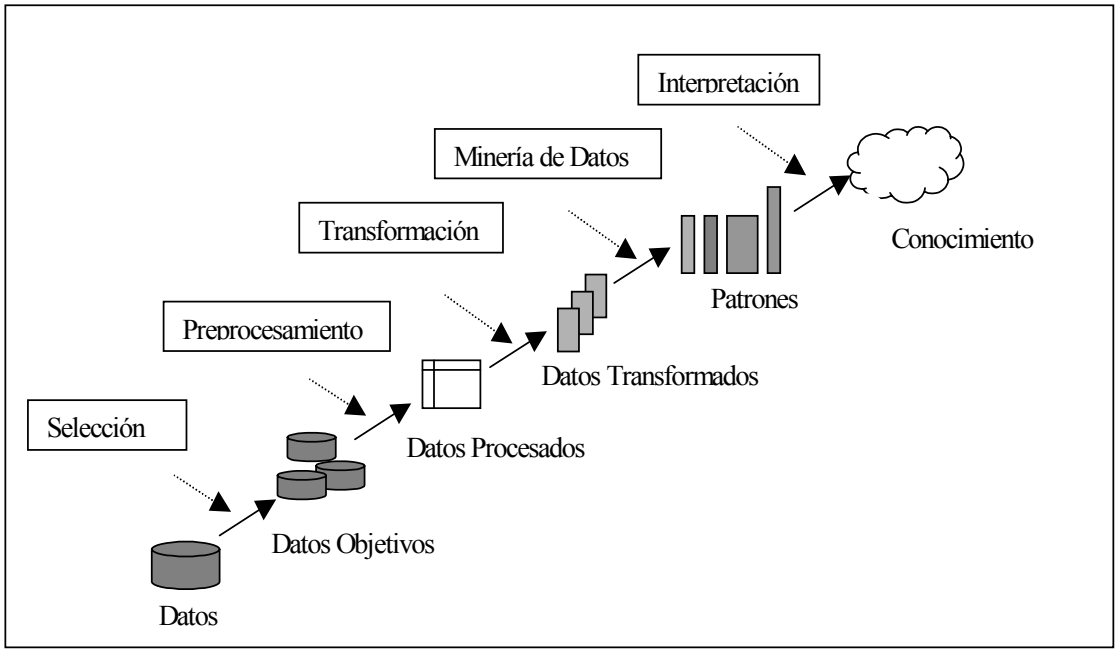

Figura 2. Proceso de Descubrimiento de Conocimiento. (Ramos y otros (2004))

La aplicación de una técnica de aprendizaje, compete a la utilización de un algoritmo de minería, del cual se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.

Fayyad y otros (1996) definen esta fase como un paso en el proceso de descubrimiento de conocimiento que consiste en la aplicación de un algoritmo particular de minería de datos que, bajo algunas limitaciones de eficiencia computacional, produce una enumeración particular de patrones de los datos.

Una vez obtenido el modelo, se debe proceder a su *validación*, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.

Técnicas de Minería de Datos

Es importante mencionar las principales técnicas utilizadas en minería de datos, las cuales fundamentalmente se dividen en:

1. Clasificación: Partiendo de una base de datos de observaciones, se buscan leyes o patrones que nos indiquen el comportamiento de una variable respecto a otras. Ejemplos de esta técnica son: generadores de árboles de decisión, generadores de reglas de asociación, redes neuronales, algoritmos genéticos, clasificadores bayesianos, modelos de regresión, clasificadores difusos, entre otros.
2. Agrupamiento: Dado un conjunto de casos en una Base de Datos, se busca agruparlas dentro de un número de clases preestablecidas, de acuerdo a criterios de distancia o similitud. Algunas técnicas muy utilizadas son: K Medias, Redes Autoorganizativas, Sistemas de Clasificación Automática Bayesiana, Teoría de Resonancia Adaptativa, entre otros.
3. Técnicas de Reducción de Dimensión y Visualización de la Información: Su objetivo es reducir al mínimo el número de variables y visualizar los puntos N dimensionales para detectar estructuras o características de forma visual. Algunos ejemplos son: Análisis de Componentes Principales, Gráficos de Coordenadas Paralelas, entre otros.

Según Kdnuggets (2004) las técnicas de minería de datos mas utilizadas por los análisis en sus proyectos, aparecen en el gráfico reflejado en la figura 3.

Los problemas a resolver en minería de datos se dividen según Weiss y otros (1998) en dos categorías generales: (a) supervisados o predictivos y (b) no supervisados o de descubrimiento de conocimiento. En el Cuadro 1 se muestran algunos problemas típicos de minería de datos.

Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de 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. Esta forma de trabajar se conoce como aprendizaje supervisado.

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en este caso hay que recurrir a los métodos no supervisados o de descubrimiento de conocimiento que revelan patrones y tendencias en los datos actuales, (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio científico o de negocio.

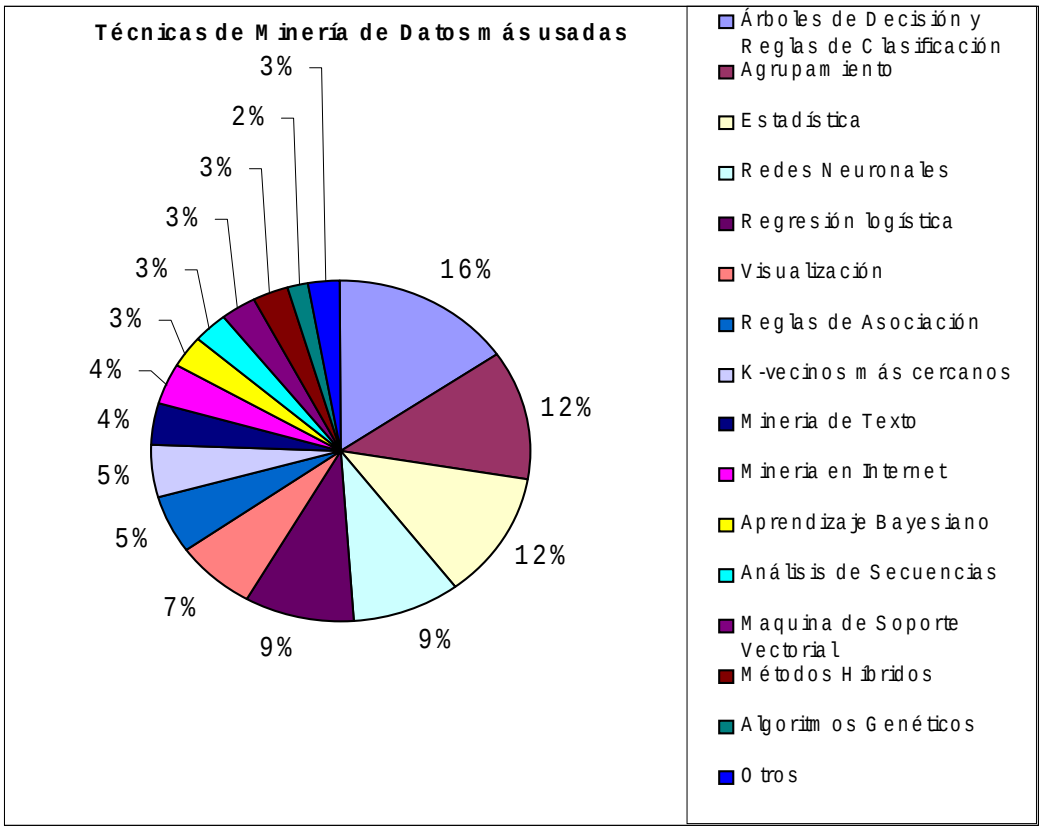

Figura 3. Técnicas de Minería de Datos más usadas. KDNUGGETS(2004)

Cuadro 1

Tipos de Problemas de Minería de Datos

Predicción	Descubrimiento de Conocimiento
Clasificación	Detección de Desviaciones
Regresión	Segmentación de Bases de Datos
Series de Tiempo	Agrupamiento
	Reglas de Asociación
	Sumarización
	Visualización
	Minería de Textos

Fuente: Weiss y otros (1998)

Aprendizaje Automático

Al hablar de minería de datos inteligente, (Michalsky y otros (1998)), se hace referencia específicamente a la aplicación de métodos de aprendizaje automático u otros métodos similares, para descubrir y enumerar patrones presentes en los datos. Por esto, este apartado abordará los mecanismos que hacen posible que las máquinas aprendan, los diferentes enfoques, paradigmas y la taxonomía del aprendizaje automático.

Una de las definiciones más controversiales en el campo del Aprendizaje Automático es la correspondiente a que se entiende por *aprender.* Según Michalsky y otros (1983), aprender denota cambios en el sistema que son adaptativos en el sentido de que le capacitan para realizar en ocasiones posteriores la misma tarea a partir de la misma población, más eficaz y eficientemente.

Según Mitchell (1997), una definición de *aprendizaje automático* es la siguiente: Un programa de computación se dice que aprende de la experiencia ***E*** con respecto a alguna clase de tareas ***T*** y rendimiento ***P***, si su rendimiento en las tareas ***T***, medido por ***P***, se incrementa con la experiencia ***E***.

Existen dos grandes tipos de aprendizaje, el inductivo y el deductivo. En el aprendizaje inductivo el sistema de aprendizaje aplica la inducción a los hechos u observaciones suministradas, para obtener nuevo conocimiento. Básicamente, las funciones que se aprenden son hipótesis acerca de una función desconocida que esta subyacente en los datos. En el aprendizaje por deducción, partiendo de los hechos de un dominio, conocidos previamente, se deduce nuevo conocimiento

Existen tres enfoques principales en el estudio del aprendizaje automático: Uno que se refiere al análisis teórico y al desarrollo de algoritmos de aprendizaje general; un segundo que estudia el desarrollo de modelos computacionales del proceso de aprendizaje humano, lo que se denomina modelos cognitivos y el tercero que se dedica a la construcción de sistemas de aprendizaje para aplicaciones específicas.

Entre las principales técnicas de aprendizaje automático se encuentran: (a) Aprendizaje Inductivo por Árboles de Decisión, (b) Aprendizaje por Agrupamiento, (c) Aprendizaje en Redes Neuronales Multicapas con Retropropagación, (d) Aprendizaje Probabilístico o Bayesiano, (e) Aprendizaje basado en instancias, (f) Aprendizaje evolutivo, (g) Aprendizaje lógico inductivo, (h) Aprendizaje por Refuerzo.

Aprendizaje Inductivo por Árboles de Decisión

Consiste en la creación de un modelo de clasificación a partir de un conjunto de entrenamiento y de un inductor. Los registros del conjunto de entrenamiento deben pertenecer a un pequeño grupo de clases predefinidas, cada clase corresponde a un valor de la etiqueta. El modelo inducido (clasificador) consiste en una serie de patrones que son útiles para distinguir las clases.

Una vez que se ha inducido el modelo se puede utilizar para predecir automáticamente la clase de otros registros no clasificados (de etiqueta desconocida). Es un método para la aproximación de funciones de valores discretos, robusto frente a datos con ruido y capaz de aprender expresiones disyuntivas. Existe una familia de

algoritmos de árboles de decisión que incluye los ampliamente utilizados: ID3, C4.5, y ASSISTANT.

Aprendizaje por Agrupamiento

Los algoritmos de agrupamiento encuentran grupos de datos que son similares. Se divide un conjunto de datos de modo que los registros con contenido similar estén en el mismo grupo, y los grupos sean tan diferentes entre sí como sea posible. Puesto que las categorías no son especificadas a priori, el agrupamiento es comúnmente referenciado como aprendizaje no supervisado.

Aprendizaje en Redes Neuronales Multicapas con Retropropagación

Son modelos predictivos no lineales que aprenden directamente del entrenamiento y reensamblan redes de neuronas biológicas en su estructura. Las redes neuronales incluidas dentro de los modelos conexionistas, son sistemas formados por un conjunto de sencillos elementos llamadas neuronas artificiales. Estas neuronas están interconectadas a través de unas conexiones con unos pesos asociados, que representan el conocimiento en la red.

Proveen un método práctico y general para el aprendizaje a partir de ejemplos de funciones reales, discretas, entres otras. Cada neurona calcula la suma de sus entradas, ponderadas por los pesos de las conexiones, le resta un valor de umbral y le aplica una función no lineal; el resultado sirve de entrada a las neuronas de la capa siguiente.

Uno de los algoritmos más usado para entrenar redes neuronales es le de Retropropagación, que utiliza el gradiente descendente para ajustar los parámetros de la red de forma que se ajusten mejor a los datos de entrenamiento de entrada-salida. Es un método iterativo para propagar los términos de error (diferencia entre los valores obtenidos y los valores deseados), necesarios para modificar los pesos de las conexiones interneuronales.

Este aprendizaje es robusto frente a la aparición de errores en los datos de entrenamiento. La red neuronal, completamente "ignorante" al principio, efectúa un aprendizaje partiendo de los ejemplos, para luego transformarse, a través de modificaciones sucesivas, en un modelo susceptible de rendir cuenta del comportamiento observado en función de las variables descriptivas. La construcción del modelo es automática y directa desde los datos.

Las redes neuronales han sido utilizadas con éxito en diferentes tipos de problemas entre los que se pueden mencionar: Auto-asociación, Clasificación de patrones, Detección de regularidades. Las principales desventajas para usar redes neuronales en minería de datos son: el aprendizaje es bastante más lento que en un sistema de aprendizaje simbólico, el conocimiento obtenido por las mismas no es representable en forma de reglas inteligibles, es difícil incorporar conocimiento de base o interacción del usuario en el proceso de aprendizaje de una red neuronal.

Aprendizaje Probabilístico o Bayesiano

El razonamiento bayesiano provee un acercamiento probabilístico a la inferencia. Está basado en asumir que las cantidades de interés están gobernadas por distribuciones de probabilidad y que las decisiones óptimas pueden ser realizadas por medio de razonamientos sobre estas probabilidades y datos observables. Provee una visión cuantitativa para pesar la evidencia que soporta distintas hipótesis.

El razonamiento bayesiano provee las bases para el aprendizaje de algoritmos que manipulan directamente probabilidades, y un ámbito para analizar cómo operan otros algoritmos que las manipulan explícitamente.

Aprendizaje Basado en Instancias

A diferencia de aquellos métodos de aprendizaje que construyen una descripción general, y explícita de la función objetivo a partir de los datos de entrenamiento, estos métodos simplemente guardan dichos datos. La generalización sobre estos ejemplos

se pospone hasta que una nueva instancia debe ser clasificada. Cada vez que una nueva instancia es encontrada, se calcula su relación con los ejemplos previamente guardados con el propósito de asignar un valor de la función objetivo para la nueva instancia.

El aprendizaje basado en instancias incluye el vecino más cercano y método de regresión pesado localmente que asumen que las instancias pueden ser representadas como puntos en el espacio euclideo. Incluyen también a los métodos de razonamiento basado en casos, que utilizan una representación más compleja y simbólica de los datos. Los métodos de aprendizaje basados en instancias son denominados "perezosos" pues dilatan el procesamiento hasta que una nueva instancia deba ser clasificada. Una ventaja de este retraso es que no se estima la función objetivo una vez para todo el espacio de instancias, sino que se hace en forma local y diferente para cada nueva instancia a clasificar.

Aprendizaje Evolutivo

Uno de los representantes de esta clase de técnica son los algoritmos genéticos. El aprendizaje de estos algoritmos está basado en la simulación de la evolución. Las hipótesis que se aprenden originalmente fueron representadas como cadenas de bits, cuya interpretación depende del tipo de aplicación. Dichas hipótesis pueden llegar a ser descritas por expresiones simbólicas o aún por programas de computación. La búsqueda de una hipótesis apropiada comienza con una población, o colección, de hipótesis iniciales.

Los miembros de la actual población pasan a la próxima generación, por medio de operaciones tales como mutación aleatoria o cruce, asociados a procesos de evolución biológica. En cada paso, la hipótesis es evaluada en base a una medida de aptitud, y las mejores son seleccionadas en forma probabilística para pasar a la próxima generación. Estos algoritmos han sido aplicados en forma exitosa a una variada gama de tareas de aprendizaje y a otros problemas de optimización. Por

ejemplo, han sido utilizados para aprender una colección de reglas de control de un robot y para optimizar la topología y los parámetros de una red neuronal.

Aprendizaje Lógico Inductivo

Una de las más expresivas y humanamente legibles representaciones para las hipótesis aprendidas son las reglas Si-Luego-Entonces. Existen varios algoritmos que utilizan este tipo de representación. En particular, uno de ellos que emplea reglas que contienen variables llamadas cláusulas de Horn de primer orden. Debido a que un conjunto de estas cláusulas puede ser considerado un programa en el lenguaje lógico de programación Prolog, su aprendizaje es generalmente denominado programación lógica inductiva.

Aprendizaje por Refuerzo

Este tipo de aprendizaje se refiere a cómo un agente autónomo que actúa en un entorno, puede aprender a elegir acciones óptimas que lo conduzcan a alcanzar objetivos. Este problema genérico cubre tareas tales como el aprendizaje del control de un robot móvil, aprendizaje de cómo optimizar operaciones en una factoría, y aprendizaje de cómo realizar jugadas en juegos de tableros.

Cada vez que un agente realiza una acción en su entorno, un entrenador provee un premio o penalización que indica la bondad del estado resultante. Por ejemplo, cuando se entrena a un agente para jugar un juego, el entrenador debe proveer una recompensa positiva cuando el juego es ganado, negativa si se pierde y cero en los otros estados. La tarea del agente es la de aprender a partir de esta recompensa indirecta y retrasada, a elegir secuencias de acciones que produzcan la mayor acumulación de recompensas. Un ejemplo de estos tipos de algoritmos es el Q-learning, que permite adquirir estrategias de control óptimas a partir de recompensas retrasadas, aun cuando el agente no posee un conocimiento inicial del efecto de las

acciones en el entorno. Estos algoritmos están relacionados con la programación dinámica, frecuentemente empleada en la resolución de problemas de optimización.

Taxonomía de los Sistemas de Aprendizaje

Michalsky y otros (1983), propone una taxonomía de los sistemas de aprendizaje que supone tres grupos, cada uno originado por una clasificación según un criterio diferente, tal como se muestra en el Cuadro 2.

Cuadro 2

Taxonomía de los Sistemas de Aprendizaje

Predicción	Descubrimiento de Conocimiento
Por la estrategia utilizada en el proceso de aprendizaje	Aprendizaje por rutina. Aprendizaje por instrucciones. Aprendizaje por analogía. Aprendizaje basado en ejemplos. Aprendizaje por observación y descubrimiento.
Por la representación del conocimiento extraído	Parámetros en expresiones algebraicas. Árboles de decisión. Gramáticas formales. Reglas de producción. Expresiones basadas en la lógica formal y formalismos afines. Grafos y Redes. Marcos y esquemas. Programas de computación y otros procedimientos codificados. Taxonomías. Representaciones múltiples.
Por el dominio de aplicación del sistema que aprende	Medicina, Química, Matemática La industria en general

Fuente: Milchasky *y otros* (1983)

Metodología CRISP-DM

Existen varias metodologías en el mercado para la implantación de Minería de Datos. Una de ellas es CRISP-DM, Proceso Estándar entre Industrias para Minería de Datos, definida por un grupo de compañías con amplia trayectoria en el uso de la Minería de Datos. Según varios autores, entre los que se destacan: Gamberger y Otros (2001), Ramos y Giménez (2004), esta metodología consta de seis fases: (a) Comprensión del Problema, (b) Comprensión de los Datos, (c) Preparación de los Datos, (d) Modelación, (e) Evaluación de los Resultados y (f) Despliegue de los Resultados. La figura 4 ilustra las fases de ésta metodología

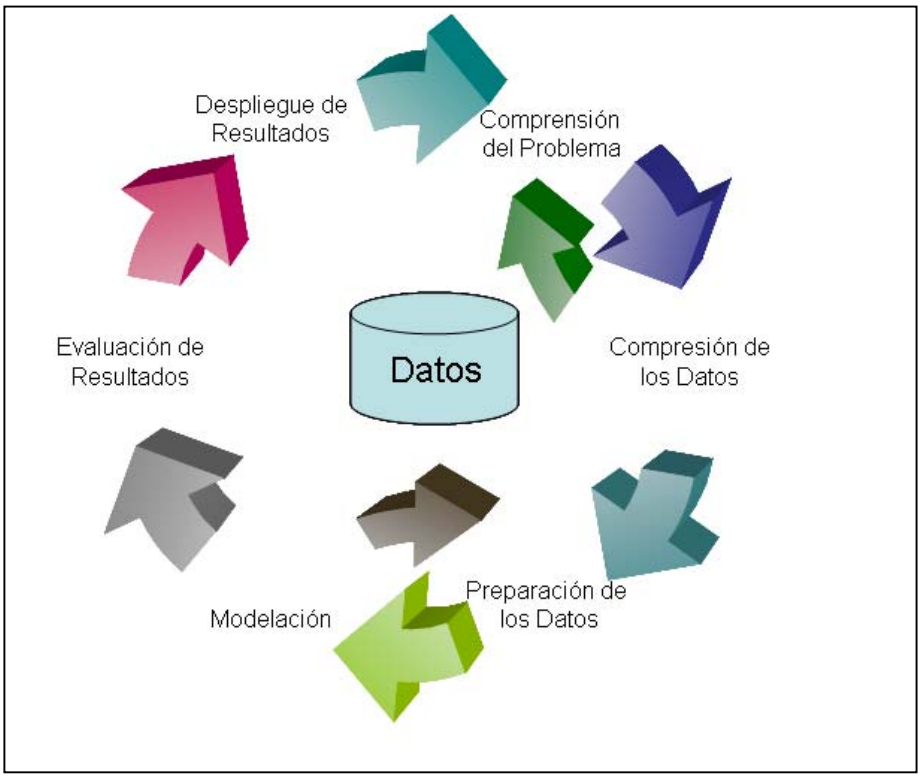

Figura 4. Fases de la Metodología CRISP-DM. (El Autor)

Esta metodología, junto con la metodología SEMMA, son las principales metodologías utilizadas por los analistas en los proyectos de minería. El gráfico

mostrado en la Figura 5, evidencia los resultados de la encuesta realizada por el portal de Minería de Datos Kdnuggets, en la que se confirma dicho uso.

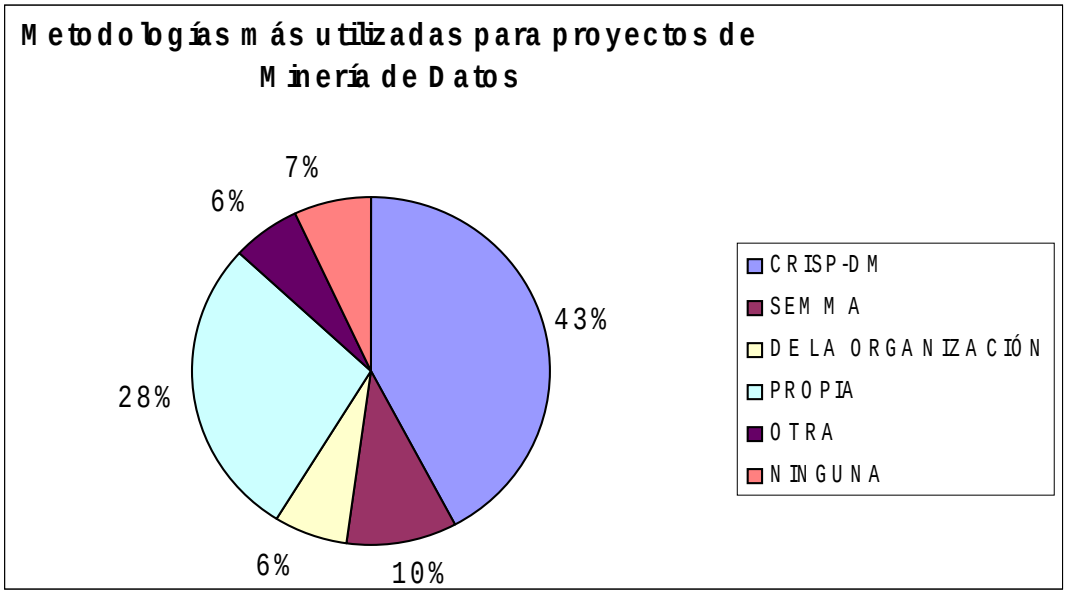

Figura 5. Metodologías para proyectos de Minería de Datos. KDNUGGETS(2004)

Fases del Estudio

Compresión del Problema.

Esta fase abarca en líneas generales, las siguientes actividades:

1. *Determinación de los objetivos*: El primer paso y el más importante es entender la necesidad de hacer minería de datos, determinando cual es el problema que se desea resolver, para que se convierta en el *objetivo* del proceso de minería de datos.
2. *Definición de Criterios de Éxito*: Una vez definido el problema, es necesario disponer de criterios de éxito para el proceso de minería de datos. Esos criterios pueden ser objetivos (cuantitativos), o pueden ser subjetivos o de naturaleza cualitativa. Los resultados deben contener algunas nuevas percepciones acerca de las relaciones entre las variables del dominio del problema.

3. *Calificación de la Situación*: Una vez definido el problema y sus criterios de solución, hay que tomar en cuenta los aspectos relacionados al problema, tales como: conocimiento experto o previo disponible acerca del problema, existencia de datos suficientes para intentar resolver el problema, etc.
4. *Determinación de las metas de la Minería de datos*: Consiste en una traducción de los objetivos del proyecto en términos de tecnología de minería de datos.
5. *Producción de un Plan del Proyecto*: Finalmente, se crea un plan para el proyecto que describa los pasos a seguir y las técnicas empleadas en cada paso.

Compresión de los datos.

El aspecto principal de la minería de datos está dado por los datos. Las actividades a desarrollar en esta fase son:

1. *Recolectar los datos iniciales*: El primer paso es la adquisición de los datos iniciales y su preparación para futuro procesamiento. El proceso de adquisición de datos puede producir las siguientes salidas: listas de datos adquiridos, localización de datos y métodos a usar para su adquisición y problemas y soluciones relacionados a la adquisición de datos.
2. *Descripción de los datos*: Luego de adquiridos, estos deben ser descritos, lo cual significa principalmente establecer el volumen de los datos (número de registros y campos por registro), identificación y significado de cada campo y la descripción del formato inicial de los datos.
3. *Exploración de los Datos*: Este paso no es obligatorio, pero si útil en mucho aspectos. El rol principal de la exploración de datos en esta fase es encontrar una estructura general para los datos. La exploración no está directamente relacionada con la solución al problema (esa es una tarea para las técnicas de modelación de minería de datos), sino que envuelve la aplicación de pruebas estadísticas básicas que revelen propiedades en los datos recién adquiridos: Si tiene campos nominales, se crean tablas de frecuencia y para los campos numéricos, se grafica su distribución y se buscan dependencias.

4. *Verificación de la Calidad de los Datos*: Aquí se realizan chequeos sobre los datos para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los datos faltantes, encontrar valores fuera de rango (que pueden representar ruido o un nuevo e interesante fenómeno). La idea en este punto es asegurar la completitud y correctitud de los datos. Completitud se refiere a la proporcionalidad y regularidad de los valores faltantes y correctitud se refiere al descubrimiento de valores erróneos en los datos y su posible solución.

Preparación de los Datos.

Aunque el núcleo del proceso es la aplicación de las técnicas de modelación de minería de datos y la evaluación de los modelos resultantes basándose en sus valores predictivos o descriptivos, no debe disminuirse la importancia que tienen los esfuerzos en la preparación de los datos. La fase de preparación de los datos está dividida en:

1. *Selección de Datos*: Un subconjunto de los datos adquiridos en las fases previas es seleccionado, basado en criterios también establecidos en fases anteriores: calidad de los datos (completitud y correctitud), limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de minería de datos preseleccionadas.
2. *Limpieza de los Datos*: Este paso complementa al anterior, también es uno de los que más tiempo consumen, debido a la enorme cantidad de técnicas que pueden aplicarse para optimizar la calidad de los datos con vistas a la fase de modelación. Algunas técnicas son: normalización de los datos (por ejemplo, de una escala decimal al rango [0,1]), discretización de campos numéricos, tratamiento de valores ausentes (hay una gran cantidad de técnicas para realizar esta tarea: reemplazo el valor faltante con una constante global, reemplazo del valor faltante con la media, con la media de la clase e incluso técnicas más complejas que pretenden *predecir* el valor), reducción del volumen de datos (por ejemplo, eliminando campos con bajo potencial de predicción o redundantes).

3. *Construcción de Nuevos Datos*: Aquí se crean nuevas estructuras a partir de los datos seleccionados, por ejemplo: generación de nuevos campos a partir de dos o más ya existentes, creación de nuevos registros (muestras), fusión de dos tablas que contengan atributos diferentes para el mismo objeto, agregación de nuevos campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.
4. *Formateo de los Datos*: Este paso en la preparación de los datos, implica transformaciones sintácticas de los datos sin modificar su significado, esto con la idea de permitir o facilitar el empleo de alguna técnica de minería de datos en particular. Algunas ejemplos son: reordenación de los campos y/o registros de la tabla (algunas herramientas de modelación requieren que los campos estén en cierto orden, las redes neuronales requieren que los registros estén ubicados aleatoriamente), ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (remover comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.)

Modelación.

Lo novedoso y abundante de las técnicas disponibles y de los algoritmos involucrados en la fase de modelación hace de ésta, la fase más interesante del proceso de minería de datos. Los pasos importantes en la fase de modelación son:

1. *Selección de la Técnica de Modelación*: Al principio del proceso de minería de datos se establece el problema a resolver y la meta de minería de datos implicada, ahora es el momento de seleccionar una técnica de minería de datos en concreto. Cuando se escoge una técnica apropiada entre numerosas técnicas de modelación disponibles en minería de datos se debe tener en cuenta el objetivo principal del proyecto y su relación con la principal división de las herramientas de minería de datos de acuerdo al tipo de problema. La primera división de las técnicas de modelación de minería de datos está hecha sobre la base del tipo de tarea de descubrimiento de conocimiento que se desea: *Predicción* o *Descripción*.

2. *Generación de Pruebas para el Modelo*: Luego de construido un modelo, se debe generar un procedimiento o mecanismo para probar la calidad y validez del modelo. Por ejemplo, en una tarea supervisada de la minería de datos como la clasificación, es común usar la rata de error como medida de la calidad. En consecuencia, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.
3. *Construcción del Modelo*: Una vez que la técnica de modelación ha sido seleccionada, se procede a ejecutarla sobre los datos previamente preparados para generar un modelo. Todas las técnicas de modelación tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los parámetros óptimos para la técnica de modelación es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.
4. *Calificación del Modelo*: Una vez que los modelos son generados, estos son interpretados de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc.)

Evaluación de los resultados.

En las fases previas (sobre todo en la de modelación), la evaluación se refería a la exactitud y generalidad del modelo generado, mientras que en esta fase involucra la evaluación del modelo con respecto a los objetivos del proyecto. En esta fase se debe decidir si hay o no razones para construir un modelo deficiente (relación costo - beneficio), si es aconsejable probar el modelo en un problema real. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable

calificar el modelo con relación a otros objetivos diferentes a los originales?, esto podría revelar información adicional.

Otro paso de esta fase es la Revisión del Proceso, que se refiere a calificar al proceso entero de minería de datos con la idea de identificar elementos que pudieran ser mejorados. Por último, en esta fase se toma una decisión acerca de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría decidirse pasar a la fase de despliegue de resultados, sino, podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de minería de datos.

Despliegue de los resultados.

En esta fase se define una estrategia para desplegar los resultados de la minería de datos.

1. *Monitoreo y Mantenimiento*: Si los modelos resultantes del proceso de minería de datos son desplegados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitoreo y mantenimiento para ser construidas sobre los modelos. La retroalimentación generado por el monitoreo y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.
2. *Reporte Final:* Es la conclusión del proyecto de minería de datos. Resume los puntos importantes del proyecto, la experiencia ganada y explica los resultados producidos.

Sistema WEKA

El sistema WEKA (*Waikato Environment for Knowledge Analysis*) fue desarrollado en la Universidad de Waikato en Nueva Zelanda. Está implementado en el lenguaje de programación Java y ha sido probado en los ambientes operativos Windows, Linux y Macintosh. Implementa algoritmos de minería de datos que

pueden aplicarse a bases de datos desde su línea de comando o bien desde su interfaz gráfica.

Este sistema incluye una variedad de herramientas para transformar conjuntos de datos. Permite realizar preprocesamientos de datos para transformarlos en un esquema de aprendizaje, a fin de que sus resultados puedan ser analizados. Una manera de usar WEKA es aplicar un método de aprendizaje a conjuntos de datos y analizar los resultados para extraer información. Otra es aplicar varios métodos de aprendizaje y comparar sus resultados en orden de escoger una predicción. La atención de WEKA se centra en los algoritmos de clasificación y filtro, sin embargo, también incluye la implementación de algoritmos para el aprendizaje de reglas de asociación y el agrupamiento de datos (*clustering*).

Definición de Términos Básicos

Agrupamiento: Búsqueda de grupos de datos que sean similares. Conjunto de datos cuyos registros con contenido similar están en el mismo grupo, y los grupos sean tan diferentes entre sí como sea posible.

Almacén de Datos: Colección de datos orientada a temas, integrado, no volátil y variante en el tiempo para el soporte del proceso de toma de decisiones.

Aprender: Denota cambios en el sistema que son adaptativos en el sentido de que le capacitan para realizar en ocasiones posteriores la misma tarea a partir de la misma población, más eficaz y eficientemente.

Aprendizaje: Según la psicología conductista, el aprendizaje es la capacidad de experimentar cambios adaptativos para mejorar el rendimiento. Según el enfoque cognoscitivo de la psicología, el aprendizaje consiste en la construcción y modificación de la representación del conocimiento.

Aprendizaje Automático: Un programa de computación se dice que aprende de la experiencia ***E*** con respecto a alguna clase de tareas ***T*** y rendimiento ***P***, si su rendimiento en las tareas ***T***, medido por ***P***, se incrementa con la experiencia ***E***.

Aprendizaje por Deducción: Partiendo del conocimiento suministrado y/o poseído, se deduce el nuevo conocimiento.

Aprendizaje por Inducción: El sistema de aprendizaje aplica la inducción a los hechos u observaciones suministradas, para obtener nuevo conocimiento.

Árboles de Decisión: Forma de representación utilizada en los sistemas de aprendizaje supervisado, para clasificar ejemplos en un numero finito de clases.

Cubo de Datos: Bodega dentro de un gran almacén de datos que alberga data para un propósito específico.

Descubrimiento de Conocimiento en Bases de Datos: Proceso global de identificación no trivial de patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles en los datos.

Minería de Datos: Descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos.

Redes Neuronales: Modelos no lineales conexionistas formados por un conjunto de elementos llamados neuronas artificiales. Estas neuronas están interconectadas a través de unas conexiones con sus pesos asociados, que representan el conocimiento de la red.

Sistemas de Procesamiento Analítico en Línea: Herramientas que manejan el Almacén de Datos o los Cubos de Datos, permiten revisar, graficar y visualizar información multidimensional, en características temporales, espaciales o propias.

CAPITULO III

MARCO METODOLOGICO

Tipo de Investigación

Según el problema planteado referido a la aplicación de minería de datos para predecir el rendimiento de la caña de azúcar en la empresa Azucarera Río Turbio C.A., ubicada en la Hacienda La Unión, Sector Chorobobo, Estado Yaracuy y en función de sus objetivos, el tipo de investigación esta enmarcado en la modalidad de proyecto factible, dado que corresponde a una propuesta sustentable en un modelo realizable, que satisface una necesidad, referida a tecnología, métodos y procesos.

El proyecto se sustenta sobre la base de investigación documental, ya que se han realizado revisiones de los aspectos teóricos-prácticos relacionados con el Aprendizaje Automático para realizar Minería de Datos.

El trabajo a desarrollar se enmarca dentro de la Línea de Investigación de Inteligencia Artificial de la Maestría en Ciencia de la Computación, y más específicamente en el área de Aprendizaje Automático, aplicada a la Minería de Datos.

Fases del Estudio

En atención a la modalidad de investigación de Proyecto Factible, se seguirá una serie de fases contempladas dentro del Manual para la presentación de Trabajos de Especialización, Maestría y Doctorado (UCLA 2002), Estas fases son: Diagnóstico, Factibilidad y Diseño de la propuesta.

Fase de Diagnóstico

Diseño de la Investigación o Procedimiento

Con el objeto de dar cumplimiento a los objetivos de la investigación, el procedimiento a seguir contempla la aplicación de la metodología CRISP-DM, cuyas fases son las siguientes:

1. Compresión del Problema.
2. Compresión de los Datos.
3. Preparación de los Datos.
4. Modelación.
5. Evaluación de Resultados.
6. Despliegue de Resultados.

En la fase de modelación se considerarán el método de aprendizaje inductivo por Arboles de Decisión y más específicamente el algoritmo C4.5 propuesto por Quinlan en 1993 y el aprendizaje por Redes Neuronales Multicapas con Retropropagación.

Población o Universo de Estudio

El universo de estudio está constituido por el conjunto de registros almacenados en una Base de Datos Oracle versión 9i, provenientes de las operaciones diarias de la empresa mediante la utilización de dos (02) sistemas desarrollados por la Gerencia de Gestión de Recursos Humanos y Sistemas como lo son: el Sistema SAGAZ (Sistema de Administración Agrícola de Azucarera) y el Sistema ROMANA (Sistema utilizado para el registro, control y pesaje de todos los productos que entran y salen de la empresa). Mediante ambos sistemas se procesa la información proveniente de cada unidad de producción y cada registro contiene los atributos a ser evaluados en la presente investigación.

Técnicas e Instrumentos de Recolección de Información

A partir de los datos almacenados en la Base de Datos Oracle 9i se realizarán consultas a la misma mediante el estándar SQL (Lenguaje Estructurado de Consultas) con la finalidad de realizar la extracción de los registros candidatos para la construcción de los conjuntos de datos. Estos conjuntos de datos se generarán en archivos planos, para facilitar su acceso desde las herramientas de software que se utilizarán en el desarrollo de los modelos.

Técnicas de Análisis de los Datos

Las técnicas de aprendizaje automático a considerar para el tratamiento de los datos son: Aprendizaje Inductivo por Árboles de Decisión y Aprendizaje en Redes Neuronales Multicapas con Retropropagación, dada la naturaleza de los datos.

Fase de Estudio de Factibilidad

Se ha analizado la factibilidad técnica, operativa y económica del proyecto, las cuales se detallan a continuación:

Factibilidad Técnica

La factibilidad técnica en estudios de Minería de Datos se refiere a la existencia de suficientes datos, presencia de datos que contengan rasgos relevantes al dominio del problema planteado, existencia de poco ruido en los datos y el dominio que se tenga para la aplicación de los métodos de minería de datos.

Respecto a la existencia de suficientes datos, se cuenta con información histórica de los últimos nueve (09) años de operación de la empresa, registrados a partir de sistemas de información desarrollados por la misma, realizándose mediante estos

sistemas un conjunto de validaciones que permiten asegurar que los datos almacenados poseen un nivel de ruido aceptable para realizar el proceso de minería.

Así mismo se cuenta con el dominio de los métodos de minería de datos que permiten realizar la predicción objeto de la presente investigación, así como la disponibilidad de expertos en el área de Inteligencia Artificial y en el área de producción de caña de azúcar.

Factibilidad Operativa

El proyecto se considera factible operativamente puesto que existe el compromiso formal de las unidades organizativas de la empresa involucradas en el proyecto de forma tal que se garantice que el resultado del presente trabajo de investigación represente una solución a un problema real para el usuario final y se constituya en una herramienta para el soporte de toma de decisiones.

Factibilidad Económica

La factibilidad económica de este proyecto de investigación se ha determinado en función de la inversión económica a realizar, para lo que se ha realizado una estimación de gastos, los cuales se reflejan en la tabla 1. Dado que se cuenta con los recursos económicos, recursos de hardware y software necesarios, esta investigación se considera factible económicamente.

Tabla 1.

Estimación de gastos

Descripción	Costo Estimado (Bs.).
Adquisición de Bibliografía	800.000,00
Materiales y Suministros	800.000,00
Aranceles y otros gastos	800.000,00
TOTAL	2.400.000,00

Fuente: El Autor

CAPITULO IV

PROPUESTA DE ESTUDIO

Justificación

El proceso de Minería de Datos consiste en el descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos. Uno de sus principales objetivos es extraer información útil a partir de grandes cantidades de datos. Es así como el objeto del presente trabajo se centra en el estudio de las características que definen a una unidad de producción de Caña de Azúcar como una unidad productora de Azúcar de máximo rendimiento.

El azúcar se obtiene de la planta de la caña por la reacción de fotosíntesis debiendo separarse en el proceso de fabricación otros componentes como son la fibra, las sales minerales, ácidos orgánicos e inorgánicos, y obteniéndose una sacarosa de alta pureza en forma de cristal. Es producida por los cañeros en época de zafra, de manera natural, semi mecanizada y mecanizada transportándose a la factoría mediante camiones y/o chatas tiradas por tractores.

En la recepción de la caña se realizan dos operaciones fundamentales: por un lado el Control de Peso, realizado en balanzas electrónicas computarizadas, por diferencia de pesaje entre el transporte con carga y el transporte vacío. De esta forma se logra obtener el peso de la materia prima ingresada. Por otra parte, en el laboratorio, se realiza un análisis individual de la Caña de Azúcar ingresada mediante toma de muestras de las cuales, se obtienen valores como el Brix, el Pol, entre otros.

El registro de la información proveniente de estos procesos se realiza mediante sistemas de información desarrollados por personal de la empresa, en una Base de Datos Oracle 9i. Este conjunto de datos representa el punto de partida para el proceso de minería de datos. Así mismo la aplicación de una Metodología que guíe de forma

estructurada el proceso es un factor determinante en el éxito del mismo. La meta es inducir un modelo para poder predecir a que clase pertenece cada una de las unidades de producción de la empresa, dados los valores de los atributos.

Descripción de la Propuesta

Se aplicará la metodología CRISP-DM para el desarrollo y seguimiento del proyecto de Minería de Datos. Partiendo del conjunto de datos registrados en la Base de Datos, se procederá a la selección de los atributos más relevantes que describen el problema planteado con el soporte del experto del Área Agrícola. Utilizando consultas vía SQL (Lenguaje de Consulta Estructurado) y las herramientas del Ambiente WEKA para el Análisis de Conocimiento, se generará un archivo plano que contendrá el conjunto de casos de estudio.

A partir del archivo plano generado se crearán varios archivos con formato ".arff" como conjuntos de entrenamiento y un archivo para realizar la prueba del modelo creado.

El sistema WEKA (Ambiente Waikato para el Análisis de Conocimiento) de la Universidad de Waikato en Nueva Zelanda, es un conjunto de herramientas que contiene la implementación de algoritmos de minería de datos, tales como, Aprendizaje inductivo por Árboles de Decisión y más específicamente el algoritmo C4.5 y el Aprendizaje en Redes Neuronales Multicapas con Retropropagación. Tal y como se mencionó en el Capitulo I en su aparte Alcance, estos métodos son los seleccionados para realizar la fase de modelado.

A partir de los resultados que se obtengan de la aplicación de ambos algoritmos, se procederá a determinar si el nuevo conocimiento aporta una solución al problema planteado. Seguidamente se realizaran las observaciones y recomendaciones a que hubiera lugar, las cuales servirán de base para la integración del conocimiento descubierto con el sistema de información de Azucarera Río Turbio C.A.

Es significativo destacar la importancia que el nuevo conocimiento descubierto por este estudio de minería de datos representará a nivel de la toma de decisiones gerenciales en el área agrícola y por ende de la empresa.

CAPITULO V

EJECUCION DE LA PROPUESTA

En este capitulo se mostrará paso a paso la ejecución del procesos de Minería de Datos, desde la Fase I Compresión del Problema hasta la Fase VI Despliegue de Resultados. El ambiente de trabajo lo constituye un equipo IBM Modelo A50 (Pentium IV, 2.8 GHz, 256 MB RAM), Sistema Operativo Windows XP Profesional, en la estación de trabajo, Servidor de Base de Datos Oracle 9i para el acceso a los datos vía SQL (Lenguaje de Consultas Estructurado), Servidor de Aplicaciones para el acceso a los Sistemas de información de la empresa, WEKA (Ambiente Waikato para el Análisis de Conocimiento).

La primera fase de la Metodología CRISP-DM Compresión del Problema ha sido abordada ampliamente en el Capitulo I Planteamiento del Problema y en Capitulo IV Propuesta de estudio, es por ello que a continuación se aborda la ejecución de la propuesta partiendo de la fase Compresión de los Datos.

Aplicación de la Fase II Comprensión de los Datos

Recolección de los Datos Iniciales

Se dispone de datos registrados en un conjunto de tablas de una Base de Datos Oracle, correspondientes a ocho (08) períodos de Zafra. Una Zafra, define el período o época de corte en los campos de cultivos y recepción de la caña de azúcar, para ser molida y/o convertida en Azúcar o sus derivados. A los efectos de este estudio, la zafra comprende desde el día que la empresa comienza el período de recepción de

caña de azúcar y molienda hasta el día en que se cierra la recepción de caña de azúcar y concluye el proceso de molienda.

Se contemplaron de esta forma los períodos de Zafra desde 1998-1999 hasta la Zafra 2002-2003. Debido a que el período de Zafra 2003-2004 presentó inconvenientes como cañas no fertilizadas, cortes a destiempo, entre otros, la Gerencia de Gestión Agrícola solicitó la exclusión de los registros correspondientes a este período de zafra 2003-2004 dado que esta información no aporta datos confiables para el estudio de Minería de Datos.

La Base de Datos esta compuesta por un conjunto de cincuenta y dos (52) estructuras de datos llamadas *Tablas*, donde se registran las operaciones diarias de la empresa. Mediante un examen exhaustivo de la información registrada en cada una de las Tablas del sistema de información, se determinó que dos (02) estructuras de datos contienen la información requerida en el presente estudio. Una sección del Modelo Entidad-Relación se muestra en la figura 6.

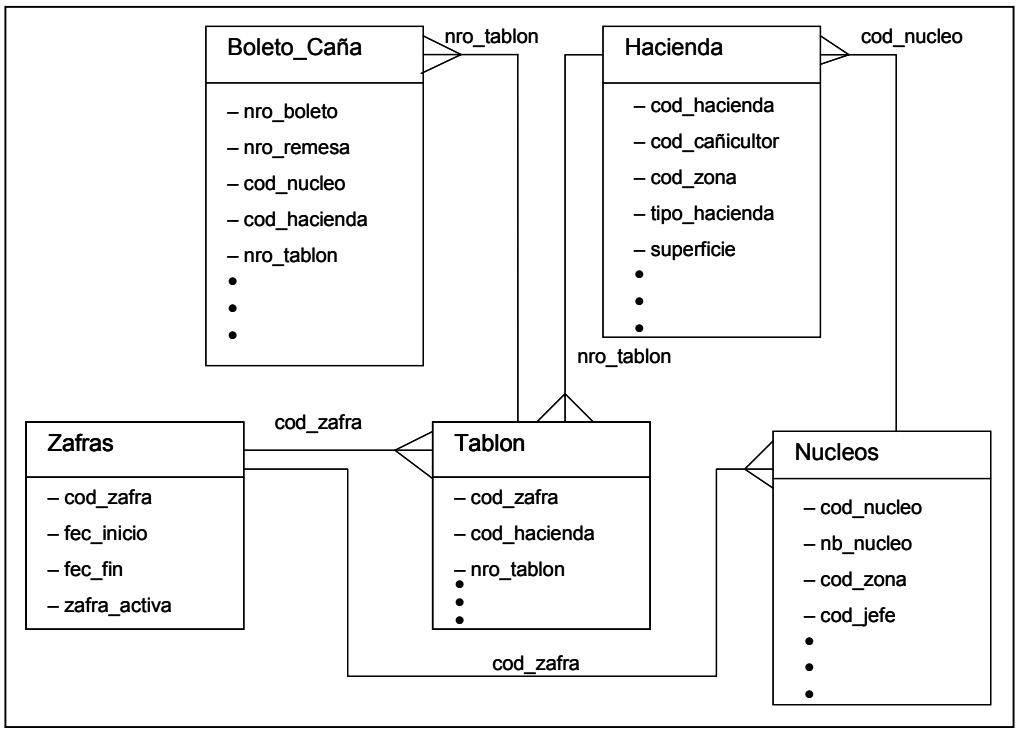

Figura 6. Modelo Entidad-Relación de la Base de Datos

La primera de estas estructuras es la Tabla de nombre ***Boleto_Caña***, la cual comprende los registros de cada una de las recepciones de Caña de Azúcar, así como toda la información relacionada al proceso productivo que sigue la Caña de Azúcar desde su recepción hasta la obtención de azúcar refinada y la segunda, es la Tabla de nombre ***Tablón***, en la que se registra información relacionada a cada Unidad de Producción de Caña de Azúcar. La descripción detallada de Tablas Boleto_Caña y Tablón se muestra en las Tablas 2 y 3 respectivamente.

Descripción de los Datos

En esta sección se muestra una descripción detallada de los datos, número de registros o casos de la Base de Datos, identificación y la descripción del formato inicial de los datos. La Tabla ***Tablón*** contiene 55.735 registros, en la Tabla 2 se muestra una descripción de cada uno de los campos de la estructura de datos, el formato y si admite o no valores nulos. La Tabla ***Boleto_Caña*** contiene 236.838 registros, en la Tabla 3 se muestra una descripción ampliada de la misma.

Tabla 2

Descripción Tabla Tablón

Variable	Descripción	Tipo	Valores Nulos
Cod_Zafra	Código del periodo de cosecha	Caracter(9)	No
Cod_Hacienda	Código de la hacienda	Caracter(6)	No
Nro_Tablon	Identificador de tablones de hacienda	Caracter(3)	Si
Superficie_Tablon	Cantidad de hectáreas de un tablón	Numérico(6,2)	Si
Superficie_Cosechada	Cantidad de hectáreas sembradas	Numérico(6,2)	Si
Fecha_Corte_Anterior	Fecha Anterior en que se realizó el corte	Fecha	No
Fecha_Corte_Actual	Fecha Actual en que se realizó el corte	Fecha	No
Caña_Estimada	Cantidad de caña que se estima cosechar	Numérico(9,3)	Si
Tipo_Corte	Tipo de Corte (Manual , Mecánico)	Caracter(1)	Si
Fecha_Quema	Fecha en que se realizó la quema	Fecha	Si
Caña_Arrimada	Cantidad real de caña cosechada	Numérico(12,3)	Si
Azucar_Producida	Cantidad real de azúcar extraída	Numérico(8,3)	Si
Cod_Variedad	Código de la variedad de la caña	Caracter(4)	No
Cod_Nucleo	Código que identifica al grupo de cosecha	Caracter(6)	Si
Fe_Liquidacion	Fecha en que se liquidó el tablón	Fecha	Si
Superficie_Semilla	Cantidad de hectáreas sembradas de semilla	Numérico(6,2)	Si

Fuente: Base de Datos Sistema SAGAZ

Tabla 3

Descripción Tabla Boleto_Caña

Variable	Descripción	Tipo	Valores Nulos
Nro_Boleto	Consecutivo de entrada a romana	Numérico(6)	No
Nro_Remesa	Número de documento de salida de la hacienda	Numérico(6)	No
Cod_Nucleo	Código que identifica al grupo de cosecha	Caracter(6)	No
Cod_Hacienda	Código de la hacienda	Caracter(6)	No
Nro_Tablon	Código de tablón	Caracter(3)	Si
Placa	Número de la placa del transporte	Caracter(7)	No
Peso_Bruto	Toneladas del transporte con carga	Numérico(6)	No
Peso_Tara	Toneladas de la tara	Numérico(6)	No
Peso_Neto	Diferencia entre el peso bruto y el peso tara	Numérico(6)	No
Status_Boleto	Estado del boleto	Caracter(1)	No
Número_Pesadas	Número de veces que se realiza la pesada	Numérico(1)	Si
Brix	Valor de Brix obtenido en análisis químico	Numérico(5,2)	Si
Pol	Valor de Pol obtenido en análisis químico	Numérico(5,2)	Si
Extracción	Valor de Extracción obtenido en análisis químico	Numérico(5,2)	Si
Pureza	Valor de Pureza obtenido en análisis químico	Numérico(5,2)	Si
Rendimiento	Valor del Rendimiento obtenido en análisis químico	Numérico(5,2)	Si
Cod_Analista	Código del analista de laboratorio	Caracter(5)	Si
Fec_Bruto	Fecha del pesaje del transporte con carga	Fecha	Si
Fec_Tara	Fecha del pesaje de la tara	Fecha	Si
Fec_Analisis	Fecha de realización análisis químicos	Fecha	Si
Fec_Quema	Fecha en que se realizó la quema	Fecha	Si
Fec_Quema_Perito	Fecha de revisión del perito de la quema	Fecha	Si
Fec_Salida	Fecha de salida del transporte	Fecha	Si
Ml_Gastados	Valor de MI obtenido en análisis químico	Numérico(4,2)	Si
Reductores	Valor de Reductores obtenido en análisis químico	Numérico(4,2)	Si
Ph	Valor de Ph obtenido en análisis químico	Numérico(3,2)	Si
Acidez	Valor de Acidez obtenido en análisis químico	Numérico(3,2)	Si
Observación	Observación del boleto	Caracter(40)	Si
Cod_Cañicultor	Código del Cañicultor	Caracter(5)	No
Romana	Código de la romana donde se realizo el pesaje	Caracter(1)	Si
Zafra	Código del periodo de cosecha	Caracter(9)	Si
Dia_Zafra	Fecha correspondiente al día de zafra	Fecha	Si
Factor_Liquidacion	Factor de rendimiento para liquidación de caña	Numérico(7,6)	Si
Azúcar	Cantidad de azúcar obtenida	Numérico(6,3)	Si
Fec_Llegada	Fecha de llegada del transporte a la empresa	Fecha	Si
Ml_Sedimento	Valor de MI_Sedimento en análisis químico	Numérico(4,2)	Si
Cod_Vehiculo	Código del Transporte	Caracter(7)	Si
Cedula_Cond	Cédula del conductor del transporte	Caracter(8)	Si
Nro_Orden_Cosecha	Número de documento que autoriza la cosecha	Caracter(6)	Si
Nro_Analisis	Número consecutivo de análisis químico	Caracter(10)	Si
Peaje	Peaje por donde pasó el transporte	Caracter(30)	Si

Fuente: Base de Datos Sistema SAGAZ

En este punto es importante destacar, que en ambas tablas el porcentaje de atributos que admiten valores nulos esta en el orden de entre el 80% y 81%, con solo alrededor de un 20% de garantía de que los campos contengan alguna información, vía el modelo de datos. La Base de Datos cuenta con un conjunto de procedimientos almacenados, llamados paquetes o procedimientos almacenados (bloques de programas que realizan una tarea específica), cuya función es validar la entrada de datos y asegurar que los datos de cada transacción sean correctos y completos.

Mediante la aplicación de sentencias SQL se pudo constatar que, menos del 1% de los campos contienen información nula al finalizar una operación completa de entrada y procesamiento de Caña de Azúcar, por lo que se concluye que la validación de entrada de datos en el Sistema de Información es robusta, facilitando la tarea de limpieza de datos del proyecto de minería de datos.

El archivo de datos se obtuvo mediante la selección de las características más relevantes registradas en las tablas, de acuerdo al criterio del experto del área y la aplicación iterativa de sentencias SQL hasta lograr la creación de un archivo plano. En la Figura 7 se visualiza la sentencia SQL definitiva.

```
select a.ZAFRA,a.COD_NUCLEO,a.COD_HACIENDA,a.NRO_TABLON
,(b.fecha_corte_actual-b.fecha_corte_anterior) edad
,avg(a.PESO_NETO) tn_ca#a,avg(a.BRIX) brix,avg(a.POL) pol
,avg(a.EXTRACCION) extraccion,avg(a.PUREZA) pureza
,avg(a.RENDIMIENTO) rendimiento,avg(a.AZUCAR) azucar
from boleto_ca#a a
  , tablon b
where a.nro_tablon   = b.nro_tablon
 and a.zafra         = b.cod_zafra
 and a.cod_hacienda = b.cod_hacienda
 and nvl((b.fecha_corte_actual-b.fecha_corte_anterior) ,0) between 250 and 720
 and a.zafra < '2003-2004'
group by a.ZAFRA, a.COD_NUCLEO, a.COD_HACIENDA, a.NRO_TABLON
     ,(b.fecha_corte_actual-b.fecha_corte_anterior)
order by 1,12,11
```

Figura 7. Sentencia SQL para la generación del archivo de datos.

El archivo de datos esta formado por un total de doce (12) atributos o variables y un total de 18809 casos o renglones de datos. La tabla 4 que se muestra seguidamente contiene una descripción de cada uno de los atributos del archivo de datos, el significado de cada variable objeto de estudio y el tipo de dato. Del conjunto de atributos, los primeros cuatro (04) corresponden a atributos identificadores de cada unidad de producción y los siguientes ocho (08) a atributos clasificadores de las unidades de producción.

En la sentencia SQL, el atributo ***Edad*** es el resultado de la diferencia de las variables *fecha_corte_actual* menos *fecha_corte_anterior,* ambas variables son de tipo fecha. Se excluyeron edades menores a 250 días y edades mayores a 720 días, como consecuencia de verificar que tales valores corresponden a errores en la trascripción de las fechas en el sistema.

Tabla 4

Descripción de los atributos del archivo de datos

Variable	Descripción	Tipo
Zafra	Periodo en que se Cosecha la Caña de Azúcar	Carácter
Cod_Un	Código que identifica el grupo que cosecha la Caña de Azúcar	Carácter
Cod_Ha	Código de la Hacienda a la cual se le cosecha la Caña de Azúcar	Carácter
Nro	Unidad llamada Tablón que corresponde a una superficie física delimitada de una Hacienda	Carácter
Edad	Días transcurridos desde la ultima cosecha a la fecha	Numérico
Tn_Caña	Cantidad de Caña Cosechada expresada en toneladas	Numérico
Brix	El Brix de una solución es la concentración (expresada en g de concentrado en 100 g de solución) de una solución de sacarosa pura en agua. Representa la cantidad de sólidos no-sacarosa que contiene el azúcar	Numérico
Pol	Porcentaje de sacarosa aparente que contiene el azúcar	Numérico
Extracción	Porcentaje de jugo extraído de la Caña de Azúcar	Numérico
Pureza	Mide el porcentaje de pureza de una solución y es el producto de dividir los grados Brix entre los grados Pol	Numérico
Rendimiento	Porcentaje de azúcar teórica a extraer de la Caña de Azúcar	Numérico
Azúcar	Cantidad de azúcar real extraída de la Caña de Azúcar	Numérico
Clase	Etiqueta de Clase	Carácter

Fuente: El Autor

La cláusula ***AVG*** (Promedio) en la sentencia SQL, se utiliza para promediar la

entrada de caña por tablón, toda vez que una cosecha de un tablón corresponde a un número ***n*** de transportes con materia prima.

Exploración de los Datos

Para la exploración de los datos se utilizaron dos herramientas. La primera XLSTAT 7.5 con el objeto de obtener las estadísticas descriptivas de las variables. La segunda, las opciones de preprocesamiento y filtrado de datos de WEKA. Las estadísticas descriptivas que se muestran en la figura 8 detallan el comportamiento estadístico de las variables del archivo de datos. Seguidamente se exponen los resultados estadísticos obtenidos de cada una de las variables numéricas.

	EDAD	TN_CA#A	BRIX	POL	EXTRACCION	PUREZA	RENDIMIENTO	AZUCAR
Núm. de valores utilizados	18809	18809	18795	18795	18795	18795	18795	18807
Núm. de valores ignorados	0	0	14	14	14	14	14	2
Núm. de val. min.	7	2	1	1	1	1	1	1
% de val. min.	0.037	0.011	0.005	0.005	0.005	0.005	0.005	0.005
Mínimo	250.000	0.000	11.833	8.220	29.010	58.330	2.370	0.056
Primer cuartil	343.000	16561.418	17.864	14.385	65.466	79.503	7.632	1.333
Mediana	368.000	19930.000	19.025	15.473	67.461	81.180	8.333	1.622
Tercer cuartil	397.000	23513.333	20.198	16.508	69.057	82.635	8.930	1.955
Máximo	720.000	39150.000	26.900	22.120	75.608	90.510	12.200	3.854
Rango	470.000	39150.000	15.067	13.900	46.598	32.180	9.830	3.798
Media	377.615	20273.697	19.025	15.418	66.822	80.908	8.233	1.667
Media geométrica	373.335		18.947	15.331	66.714	80.863	8.163	1.602
Media armónica	369.567		18.868	15.240	66.593	80.816	8.086	1.533
Curtosis (Pearson)	7.185	-0.312	0.102	0.314	9.659	3.011	1.016	0.007
Asimetría (Pearson)	2.110	0.292	-0.022	-0.206	-2.117	-0.970	-0.596	0.453
Curtosis	7.188	-0.311	0.102	0.314	9.663	3.012	1.017	0.007
Asimetría	2.111	0.292	-0.022	-0.206	-2.117	-0.970	-0.596	0.453
CV (desviación típica/media)	0.163	0.252	0.090	0.105	0.054	0.033	0.125	0.278
Varianza de muestra	3776.146	26121861.619	2.930	2.605	13.053	7.211	1.052	0.214
Varianza estimada	3776.347	26123250.489	2.930	2.605	13.054	7.211	1.052	0.214
Desviación típica de muestra	61.450	5110.955	1.712	1.614	3.613	2.685	1.026	0.463
Desviación típica estimada	61.452	5111.091	1.712	1.614	3.613	2.685	1.026	0.463
Desviación típica media	40.867	4117.534	1.367	1.275	2.527	2.015	0.799	0.371
Desviación absoluta mediana	26.000	3464.286	1.165	1.062	1.753	1.552	0.646	0.306
Desviación típica de la media	0.448	37.268	0.012	0.012	0.026	0.020	0.007	0.003
Límite inf. IC de la media	376.736	20200.649	19.001	15.395	66.770	80.870	8.218	1.660
Límite sup. IC de la media	378.493	20346.744	19.050	15.441	66.873	80.947	8.248	1.673

Figura 8. Estadísticas Descriptivas (XLSTAT 7.5)

.

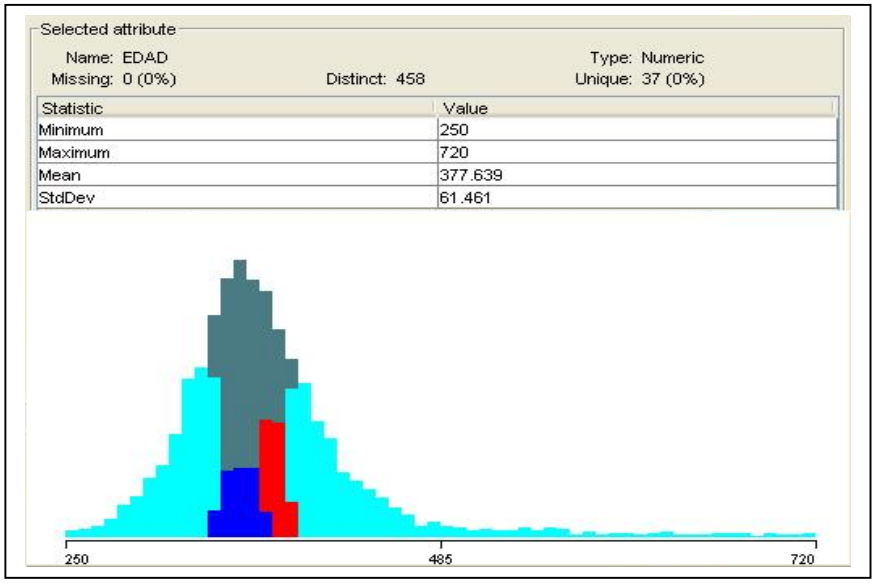

Figura 9. Valores Estadísticos y Distribución Variable Edad (XLSTAT 7.5)

En la figura 8, se muestra gráficamente la distribución de la Variable Edad y los Valores Estadísticos tales como Valor Mínimo, Valor Máximo, Media y Desviación Estándar. Así mismo, se muestra la cantidad de valores perdidos, valores únicos y valores distintos presentes en los datos. Los colores de la gráfica representan las clases del conjunto de datos.

El proceso de aprendizaje utilizado tanto por Arboles de Decisión Algoritmo C4.5 como por las Redes Neuronales Multicapas con Retropropagación se circunscribe dentro del paradigma de Aprendizaje Supervisado, debido a que el mecanismo para lograr el aprendizaje se basa en exponer al sistema un conjunto de ejemplos de los que se conoce una etiqueta o clase, dicho conjunto se denominará conjunto de entrenamiento.

En color Azul Rey se presentan casos de la clase 1, en color Rojo se muestran los casos de la clase 2. La clase 3 aparece con el color Azul Claro y por ultimo la clase 4 en color verde.

Tabla 5

Descripción de las clases

Clase	Edad	Rendimiento	Azúcar
1	> 345 y < 375 días	>= 8,24 y < 8,64	> 1,74
2	> 375 y < 390 días	>=7,23 y < 8,23	>= 1,54 y < 1,73
3	< 345 y > 390 días	< 7,23	< 1,53
4	Elementos que no clasifican en las clases anteriores		

La Tabla 5 muestra un resumen de las clases utilizadas. La clase 1 corresponde a registros cuya edad esta entre 345 y 375 días, el rendimiento entre 8.24 y 8.64 y el azúcar producida por cada 20 toneladas de caña es mayor a 1.74. La clase 2 agrupa las entidades con edades entre 375 y 390 días, un rendimiento entre 7.23 y 8.23, el azúcar producida varia entre 1.54 y 1.73. La clase 3 se caracteriza por presentar edades menores 345 o mayores a 390, con rendimientos menores a 7.0 y azúcar producida menor a 1.53. Por último, la clase 4 corresponde a entidades que no se clasifican en ninguna de las categorías antes mencionadas.

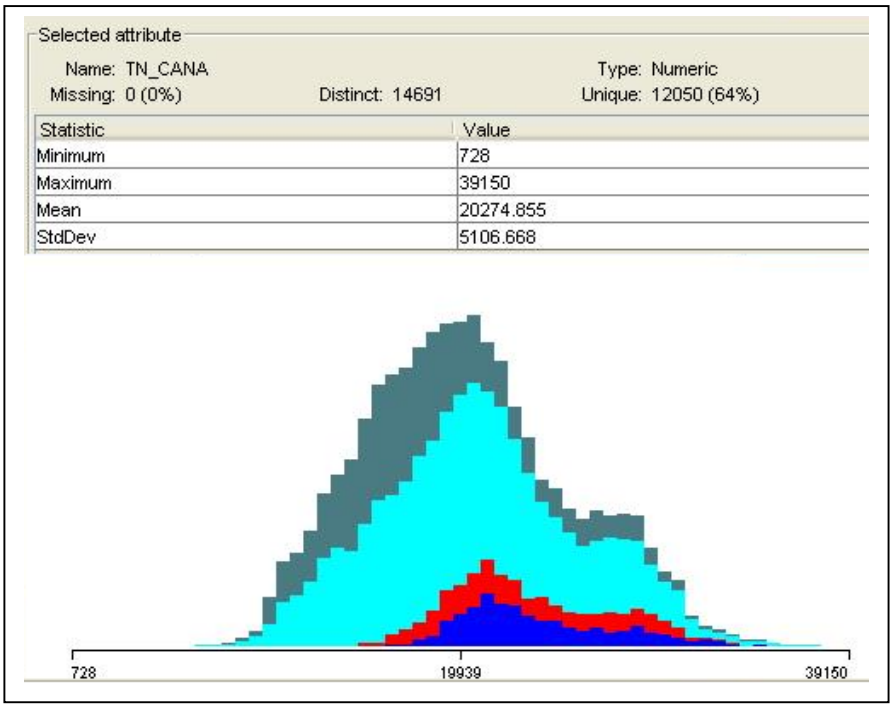

Figura 10. Valores Estadísticos y Distribución Variable Tn_Caña (XLSTAT 7.5)

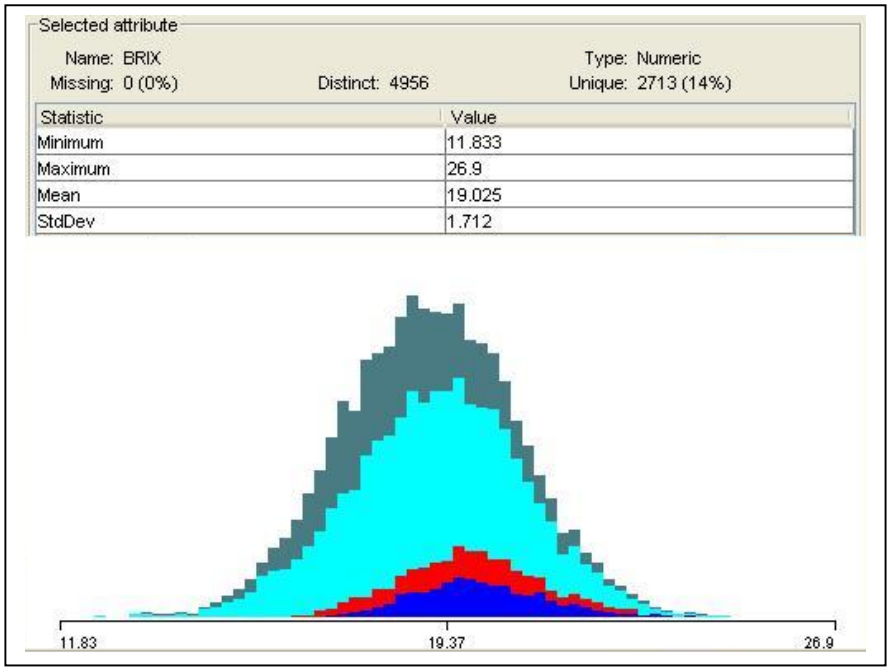

Figura 11. Valores Estadísticos y Distribución Variable Brix (XLSTAT 7.5)

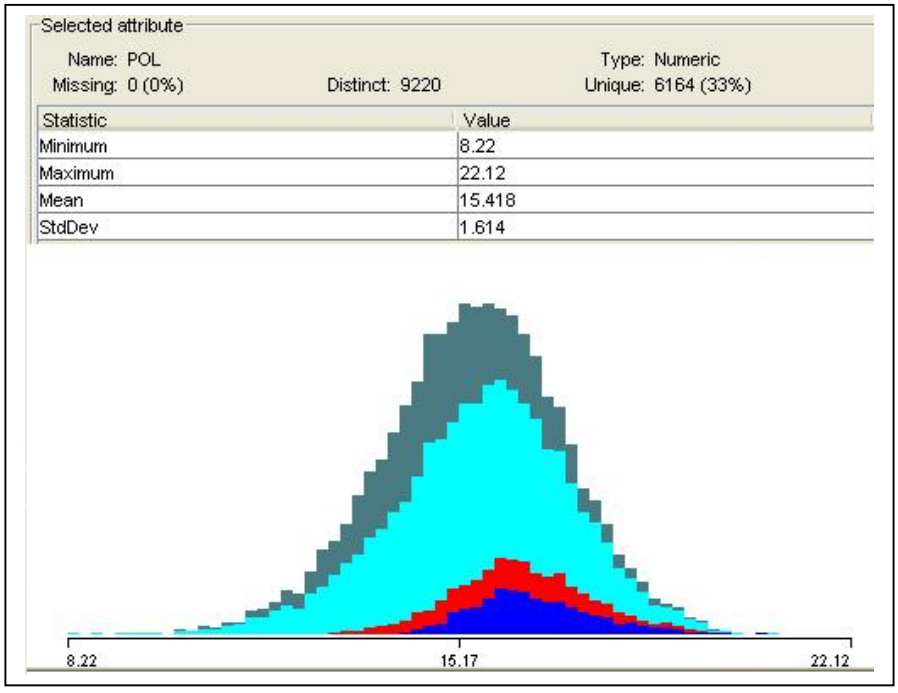

Figura 12. Valores Estadísticos y Distribución Variable Pol (XLSTAT 7.5)

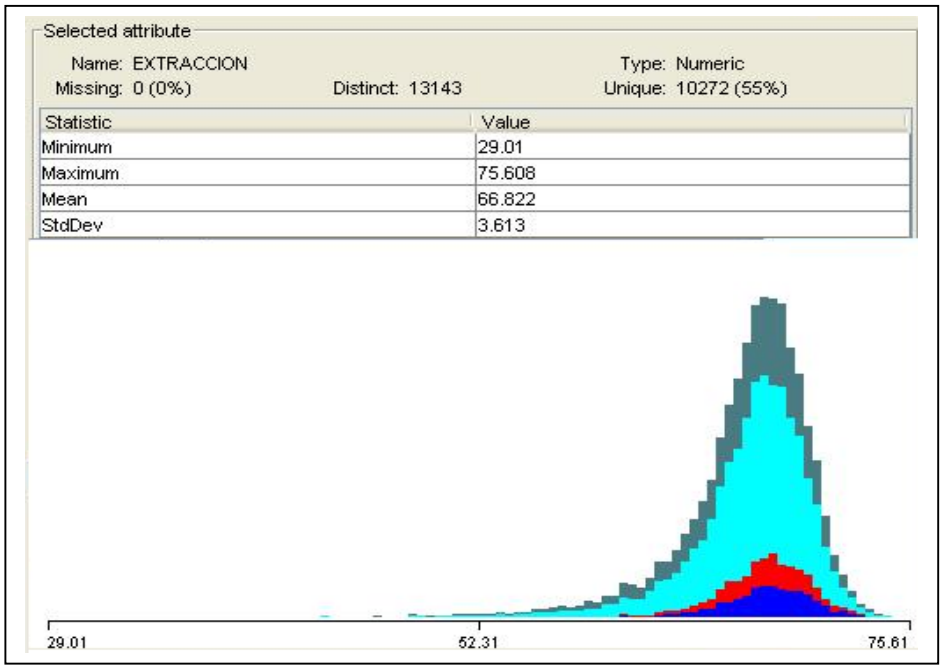

Figura 13. Valores Estadísticos y Distribución Variable Extracción (XLSTAT 7.5)

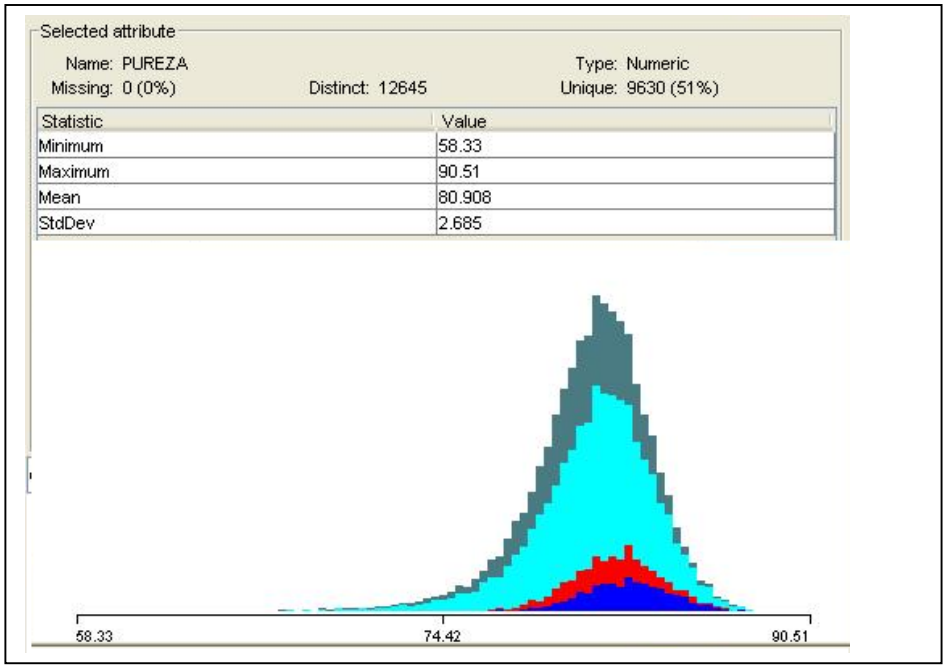

Figura 14. Valores Estadísticos y Distribución Variable Pureza (XLSTAT 7.5)

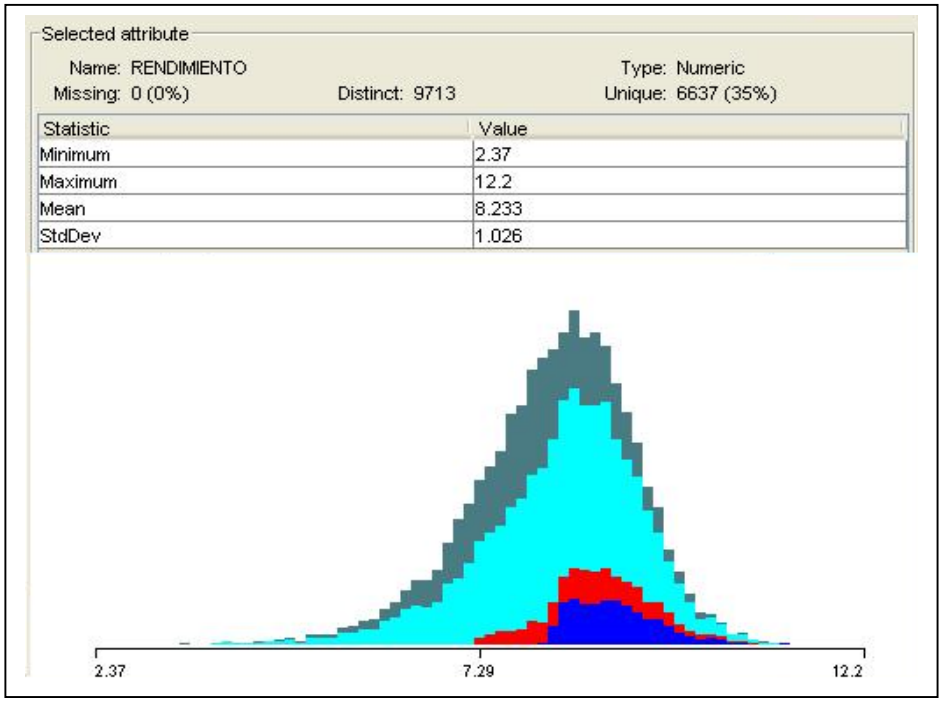

Figura 15. Valores Estadísticos y Distribución Variable Rendimiento (XLSTAT 7.5)

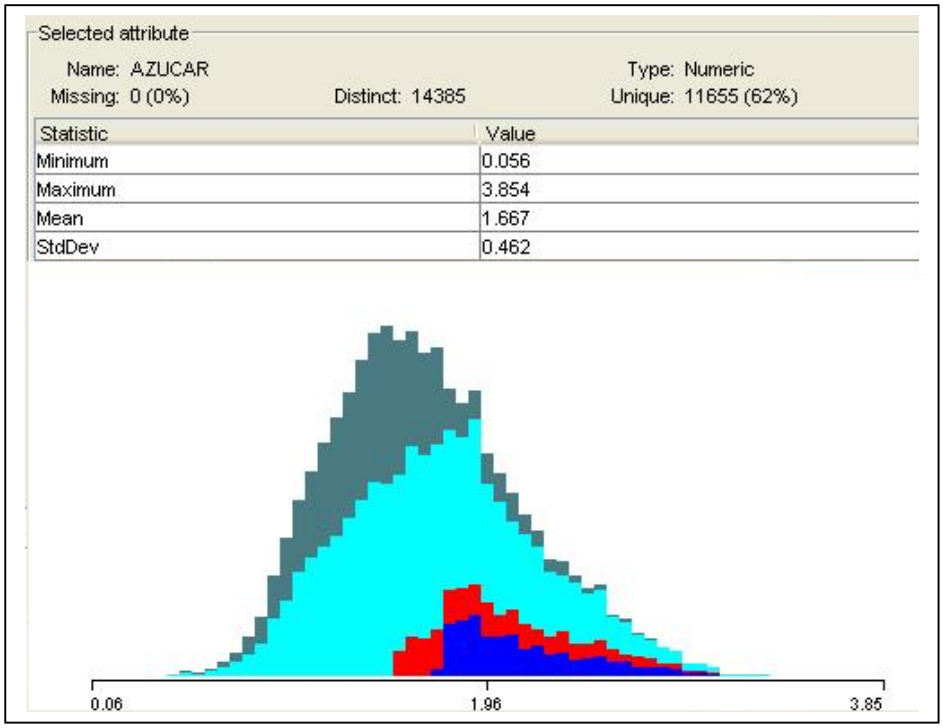

Figura 16. Valores Estadísticos y Distribución Variable Azúcar (XLSTAT 7.5)

Las clases utilizadas fueron aportadas por el experto del área agrícola de acuerdo a análisis estadísticos previos y estándares del negocio azucarero. Respecto al comportamiento estadístico de las variables tenemos que, sobre la base de los valores de las medias y los rangos, el promedio de la edad de la caña de azúcar está por encima del valor medio, tal como se muestra en la figura 9.

La variable Tn_Caña de la figura 10, se ubicó por encima de los valores permitidos. Tanto la variable Brix como la variable Pol, que aparecen en las figuras 11 y 12 respectivamente, muestran un comportamiento entre los valores normales.

La variable Extracción que se visualiza en la figura 13, exhibe un comportamiento entre los valores medios aceptables. La figura 14 muestra el comportamiento de la variable Pureza, la cual se ubicó entre los valores normales. En el caso de la variable Rendimiento que aparece en la figura 15, está por encima de los valores permitidos. Para los valores de la variable Azúcar, la media está en el valor mínimo por debajo del punto central, tal como se aprecia en la figura 16.

Verificación de la Calidad de los Datos.

La etapa de verificación de la calidad se realizó de forma muy estrecha con las tareas de selección y limpieza de los datos de la Fase III Preparación de los Datos, dado que las actividades ejecutadas en esta etapa, garantizan la consistencia de los datos individuales de los campos, así como el tratamiento de los datos faltantes.

Aplicación de la Fase III Preparación de los Datos

Selección y Limpieza de Datos.

La selección y limpieza de los datos se garantizó mediante la inclusión de restricciones en el ámbito de la cláusula ***Select*** del SQL. Estas validaciones incluyen el filtrado de edades en el rango comprendido entre 250 y 720 días, dado que este valor proviene de la diferencia de dos campos tipo fecha. Registros fuera de este

rango se consideran errores de usuario al procesar la información.

La sentencia ***NVL*** se utilizó para asignar el valor cero a los campos de tipo fecha cuyo resultado era un valor nulo, por lo que registros que no cumplen esta condición fueron excluidos del archivo de datos objeto de estudio. La carga del archivo de datos con la herramienta WEKA, realiza automáticamente la verificación de que cada fila del archivo contiene un valor en cada atributo. De no cumplirse esta condición la carga finaliza con un error que indica la fila que presenta problemas, para su respectiva eliminación. De esta forma, se efectuó el preprocesamiento del archivo de datos, descartando las filas que presentaron al menos un atributo con valores nulos.

Construcción de Nuevos Datos.

En el presente proyecto se contó con cantidad suficiente de datos reales por lo que no fue necesaria la construcción de nuevas estructuras de datos. Para el archivo de datos los campos se construyeron los nuevos datos ***Tn_Caña, Brix, Pol, Extracción, Pureza, Rendimiento, Azúcar,*** mediante la aplicación del promedio de los boletos de entrada de materia prima agrupándolos por ***zafra, cod_nucleo, cod_hacienda, nro_tablon***. El calculo del atributo ***Edad*** mostrado con anterioridad, corresponde igualmente a la construcción de un nuevo dato para el proyecto de minería.

Formateo de los Datos.

Los datos iniciales fueron modificados a los fines de permitir la aplicación de la técnica de modelo propuesta, sin modificar su significado. Los datos transformados fueron los siguientes: El atributo Zafra posee un formato inicial formado por cuatro dígitos seguidos de un guión más cuatro dígitos. Este guión fue sustituido por el número cero. Similar sucede con el atributo Cod_Ha. Este atributo posee en su contenido un carácter guión que fue sustituido por un número cero.

Aplicación de la Fase IV Modelado

Selección de la Técnica de Modelación

En esta etapa del proceso se seleccionaron dos (02) técnicas clasificadas dentro de la categoría de supervisados o predictivos (Weiss y otros (1998)), como lo son los Árboles de Decisión y las Redes Neuronales Multicapas con Retropropagación, con el propósito de cumplir los objetivos de esta investigación. Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos.

En la presente investigación se utilizó el enfoque planteado por Han y otros 2001, donde la predicción puede ser vista como la construcción y uso de un modelo para evaluar la clase de un ejemplo no etiquetado o para evaluar el valor o el rango de valores de un atributo que es probable que una muestra dada tenga. En Minería de Datos es comúnmente aceptado referirse al término predicción, para la predicción de etiquetas de clase como clasificación.

Los árboles de decisión son una forma de representación sencilla, muy usada entre los sistemas de aprendizaje supervisado, para clasificar ejemplos en un número finito de clases. Se basan en la partición del conjunto de ejemplos según ciertas condiciones que se aplican a los valores de los atributos. Su potencia descriptiva viene limitada por las condiciones o reglas con las que se divide el conjunto de entrenamiento. Los sistemas basados en árboles de decisión forman una familia llamada TDIDT (Top-Down Induction of Decision Trees), cuyo representante más conocido es ID3.

ID3 (Interactive Dichotomizer) se basa en la reducción de la entropía media para seleccionar el atributo que genera cada partición (cada nodo del árbol), seleccionando aquél con el que la reducción es máxima. Los nodos del árbol están etiquetados con nombres de atributos, las ramas con los posibles valores del atributo, y las hojas con las diferentes clases. C4.5 es una variante de ID3, que permite clasificar ejemplos con

atributos que toman valores continuos.

Las redes neuronales, incluidas dentro de los modelos conexionistas, son sistemas formados por un conjunto de sencillos elementos de computación llamados neuronas artificiales. Estas neuronas están interconectadas a través de unas conexiones con unos pesos asociados, que representan el conocimiento en la red. Cada neurona calcula la suma de sus entradas, ponderadas por los pesos de las conexiones, le resta un valor umbral y le aplica una función no lineal (por ej. sigmoidal), el resultado sirve de entrada a las neuronas de la capa siguiente.

El algoritmo más usado para entrenar redes neuronales es el retropropagación, el cual utiliza un método iterativo para propagar los términos de error (diferencia entre valores obtenidos y valores deseados), necesarios para modificar los pesos de las conexiones interneuronales. Aplica el método de descenso de gradiente en el espacio de parámetros (pesos), para encontrar mínimos locales en la función de error.

Arboles de Decisión Algoritmo C4.5

- ***Generación de Pruebas para el Modelo***

Inicialmente se abordará el tema sobre la obtención tanto del conjunto de entrenamiento, como el conjunto de prueba. Para la construcción de los conjuntos de entrenamiento se utilizó la herramienta de filtros no supervisados aplicados a conjuntos de datos de WEKA. Al conjunto de datos iniciales, se le aplicó un filtro de instancias no supervisado, cuyo formato es el siguiente:

Formato: weka.filters.unsupervised.instance.Randomize

Este filtro tiene por objetivo, cambiar el orden en el conjunto de entrada al azar, mediante un generador de números aleatorios. El parámetro que requiere para la ejecución es la semilla para el generador de números aleatorios. Posteriormente se aplicó el filtro de instancias no supervisado:

Formato: weka.filters.unsupervised.instance.Remove Percentage

El resultado que se obtiene en este caso es eliminar del conjunto de datos, el número de casos correspondiente el porcentaje suministrado como parámetro. Una vez aplicados los filtros en cada caso, el resultado fue almacenado en un archivo con un nombre identificador y extensión .arff. De esta forma, se obtuvieron los conjuntos de entrenamiento.

La Tabla 6 muestra un resumen de las pruebas realizadas, el conjunto de entrenamiento utilizado y el conjunto de validación con el que fue probado el modelo de Árboles de Decisión con el algoritmo C4.5.

Tabla 6.

Diseño de Pruebas para el Modelo Arboles de Decisión Algoritmo C4.5

Prueba	Conjunto de Validación: Prueba.arff 18793 Instancias		
	Conjunto de Entrenamiento		
	Tamaño	Nombre del Archivo	Factor de Confianza
1	9396	Caña50%Aleatorio.arff	0.25
2	5637	Caña30%Aleatorio.arff	0.25
3	4698	Caña25%Aleatorio.arff	0.25
Prueba	**Conjunto de Validación: Prueba.arff 18793 Instancias**		
	Conjunto de Entrenamiento		
	Tamaño	**Nombre del Archivo**	**Factor de Confianza**
4	4698	Caña25%Aleatorio.arff Prueba: Caña50%Aleatorio.arff	0.25
5	9396	Caña50%Aleatorio.arff	0.25
6	9396	Caña50%Aleatorio.arff	0.10
7	7517	Caña40%Aleatorio.arff	0.10
8	5637	Caña30%Aleatorio.arff	0.10

Fuente: El Autor

Red Neuronal Multicapa con Retropropagación.

- ***Generación de Pruebas para el Modelo.***

La Tabla 7 muestra el resumen de las pruebas realizadas, el conjunto de entrenamiento utilizado y el conjunto de validación con el que fue probado el modelo de Redes Neuronales Multicapa con Retropropagación.

Tabla 7.

Diseño de Pruebas para el Modelo Redes Neuronales Multicapa con Retropropagación

Prueba	Conjunto de Validación: Prueba.arff 18793 Instancias		
	Conjunto de Entrenamiento		
	Tamaño	Nombre del Archivo	Nro de Epocas y Factor de Aprendizaje
9	9396	Caña50%Aleatorio.arff	E=500 LR=0.3
10	9396	Caña50%Aleatorio.arff	E=500 LR=0.2
11	4698	Caña25%Aleatorio.arff	E=1000 LR=0.2
12	4698	Caña25%Aleatorio.arff	E=2000 LR=0.2
13	4698	Caña25%Aleatorio.arff	E=5000 LR=0.2
14	5637	Caña30%Aleatorio.arff	E=5000 LR=0.1

Fuente: El Autor

Construcción del Modelo

La organización general de esta sección, los resultados obtenidos en la construcción de los modelos, siguiendo el orden de los casos de prueba elaborados. Primeramente, se trataran los casos de Árboles de Decisión y posteriormente los de Redes Neuronales Multicapas con Retropropagación.

Árboles de Decisión Algoritmo C4.5

El formato de aplicación del algoritmo para el caso de Prueba 1 es el siguiente:

Formato:	***weka.classifiers.trees.J48 -C 0.25 -M 2***

Los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 9396 casos, con un factor de confianza de 0.25. El árbol de decisión generado, muestra en el nodo raíz el atributo *Edad*, el cual determina la primera decisión. Los números entre paréntesis al final de cada hoja corresponden al número de ejemplos en esa hoja. Si en la hoja no todos los elementos son de la misma clase, se muestra el número de ejemplos no clasificados. El tamaño del árbol es igual al número de nodos del árbol en este caso es 37 nodos y el tiempo de ejecución es 4.17 segundos.

El árbol de decisión generado se muestra gráficamente en la figura 17. El error que presenta el clasificador, para el conjunto de entrenamiento es de 0,4417% de instancias no clasificadas en forma correcta, clasificando correctamente el 99.55% de las instancias.

El algoritmo implementado por WEKA utiliza el estadístico Kappa (Cohen (1980)) para medir la coincidencia de la predicción con la clase real. Un valor de 0 indica, que la posibilidad de coincidencia era debida al azar y un valor de 1 indica una coincidencia perfecta en las predicciones. Las valores entre 0 y 0,2 se consideran muy malos, entre 0,2 y 0,4 malos, entre 0,4 y 0,6 regulares, entre 0,6 y 0,8 buenas y entre 0,8 y 1 muy buenas o excelentes. En este caso de prueba el valor obtenido es 0.9926.

Instancias Correctamente Clasificadas	18710	99.5583 %
Instancias Incorrectamente Clasificadas	83	0.4417 %
Estadístico Kappa	0.9926	
Error Absoluto Medio	0.0031	
Número Total de Instancias	18793	

A continuación se presenta la precisión del modelo en forma detallada por clases. La columna PV (Positivo Verdadero), es la proporción de elementos que están

clasificados dentro de una clase, de entre todos los elementos que realmente son de la clase. Constituye la parte de la clase que ha sido capturada por el clasificador. La columna FP (Falso Positivo), representa la proporción de ejemplos que han sido clasificados dentro de una clase, pero pertenecen a una clase diferente. La columna Precisión muestra la proporción de ejemplos que realmente son de una clase de entre todos los elementos que han sido clasificados dentro de la misma. Estas medidas se utilizaron para comparar los clasificadores generados.

Precisión del Modelo

PV	FP	Precisión	Clase
0.992	0.001	0.985	1
0.985	0.001	0.984	2
1	0	1	3
0.991	0.003	0.993	4

La matriz de confusión o matriz de contingencia es una matriz de tamaño *n*x*n*, donde *n* es el número de clases. El número de instancias clasificadas correctamente es la suma de la diagonal de la matriz, las demás instancias han sido clasificadas incorrectamente. En este caso la matriz de confusión generada es la siguiente:

Matriz de Confusión

a	b	c	d	<--	Clase
1525	0	0	13	|	a = 1
0	1431	0	22	|	b = 2
0	0	10411	1	|	c = 3
23	24	0	5343	|	d = 4

Para el caso de *prueba 2*, los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 5637 casos, con un factor de confianza de 0.25. El tamaño del árbol generado es de 21 nodos y el tiempo de ejecución del algoritmo 1.92 segundos. Al igual que el caso de prueba 1 el atributo raíz del árbol es el atributo *Edad.* En la figura 18 se observa gráficamente el árbol obtenido.

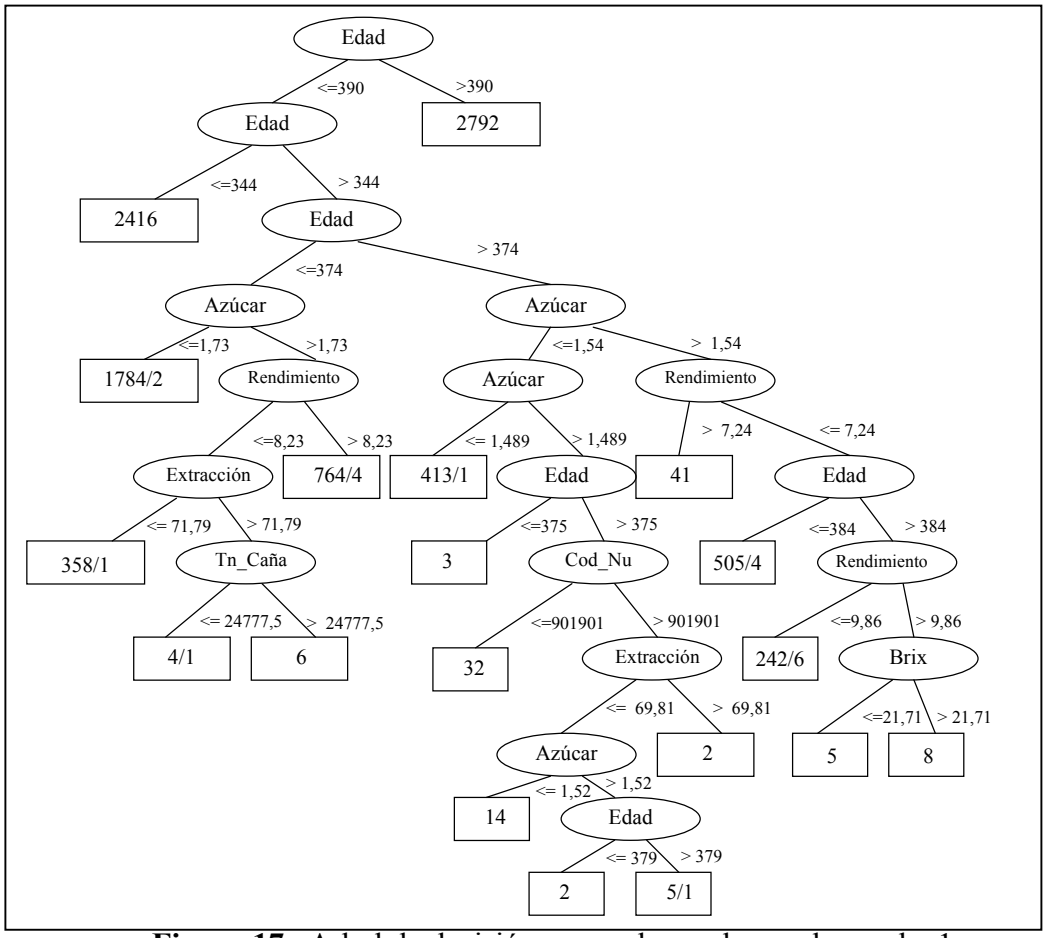

Figura 17. Arbol de decisión generado en el caso de prueba 1.

En esta prueba clasifica correctamente el 99.54% de las instancias, con una reducción del 20% del número de instancias de entrenamiento, obteniéndose una mejora del tiempo de ejecución de 2.25 segundos respecto al caso de prueba 1 y con la disminución del tamaño del árbol de 16 nodos. El error que presenta el clasificador para el conjunto de entrenamiento en este caso de prueba es de 0,4523% de instancias no clasificadas en forma correcta. El valor del estadístico Kappa en esta prueba es 0.9924.

Instancias Correctamente Clasificadas	18708	99.5477 %
Instancias Incorrectamente Clasificadas	85	0.4523 %
Estadístico Kappa	0.9924	
Error Absoluto Medio	0.0038	
Número Total de Instancias	18793	

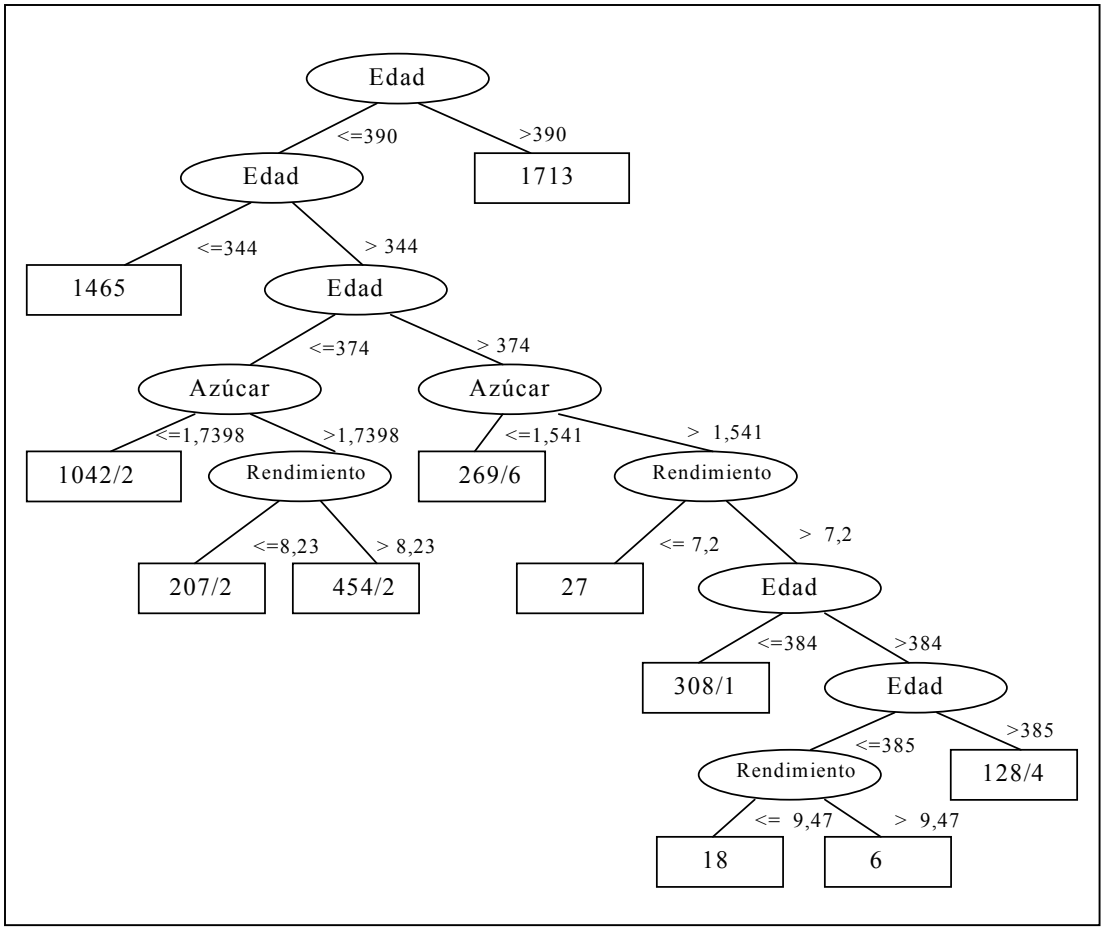

Figura 18. Árbol de decisión generado en el caso de prueba 2.

La precisión del modelo y la matriz de confusión obtenidas en esta prueba son las siguientes:

Precisión del Modelo

PV	PF	Precisión	Clase
0.991	0.001	0.991	1
0.975	0.001	0.986	2
1	0	1	3
0.994	0.004	0.991	4

Matriz de Confusión

a	b	c	d	<-- Clase
1524	0	0	14	a = 1
0	1417	0	36	b = 2
0	0	1041	1	c = 3
14	20	0	5356	d = 4

Para el caso de *prueba 3*, los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 4698 casos, con un conjunto de validación de 18793 instancias, con un factor de confianza de 0.25. El resultado de la ejecución en este caso fue Memoria Insuficiente.

Para verificar que en el caso de prueba anterior el error de memoria se debe al tamaño del conjunto de validación, en el caso de *prueba 4*, los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 4698 casos, conjunto de validación 9396 instancias, con un factor de confianza de 0.25. El tamaño del árbol generado es de 21 nodos y el tiempo de ejecución del algoritmo 1.7 segundos. Al igual que en los casos anteriores, el atributo raíz del árbol es el atributo *Edad.*

En esta prueba clasifica correctamente el 99.68% de las instancias, con un error de 0,3193% de instancias no clasificadas correctamente, manteniéndose el tamaño del árbol en 21 nodos, como en el caso de prueba 2. El tiempo de ejecución disminuye como consecuencia de procesar menor número de instancias sin afectar las medidas de eficiencia. El valor del estadístico Kappa en esta prueba es 0.9947.

Instancias Correctamente Clasificadas	9366	99.6807 %
Instancias Incorrectamente Clasificadas	30	0.3193 %
Estadístico Kappa	0.9947	
Error Absoluto Medio	0.0029	
Número Total de Instancias	9396	

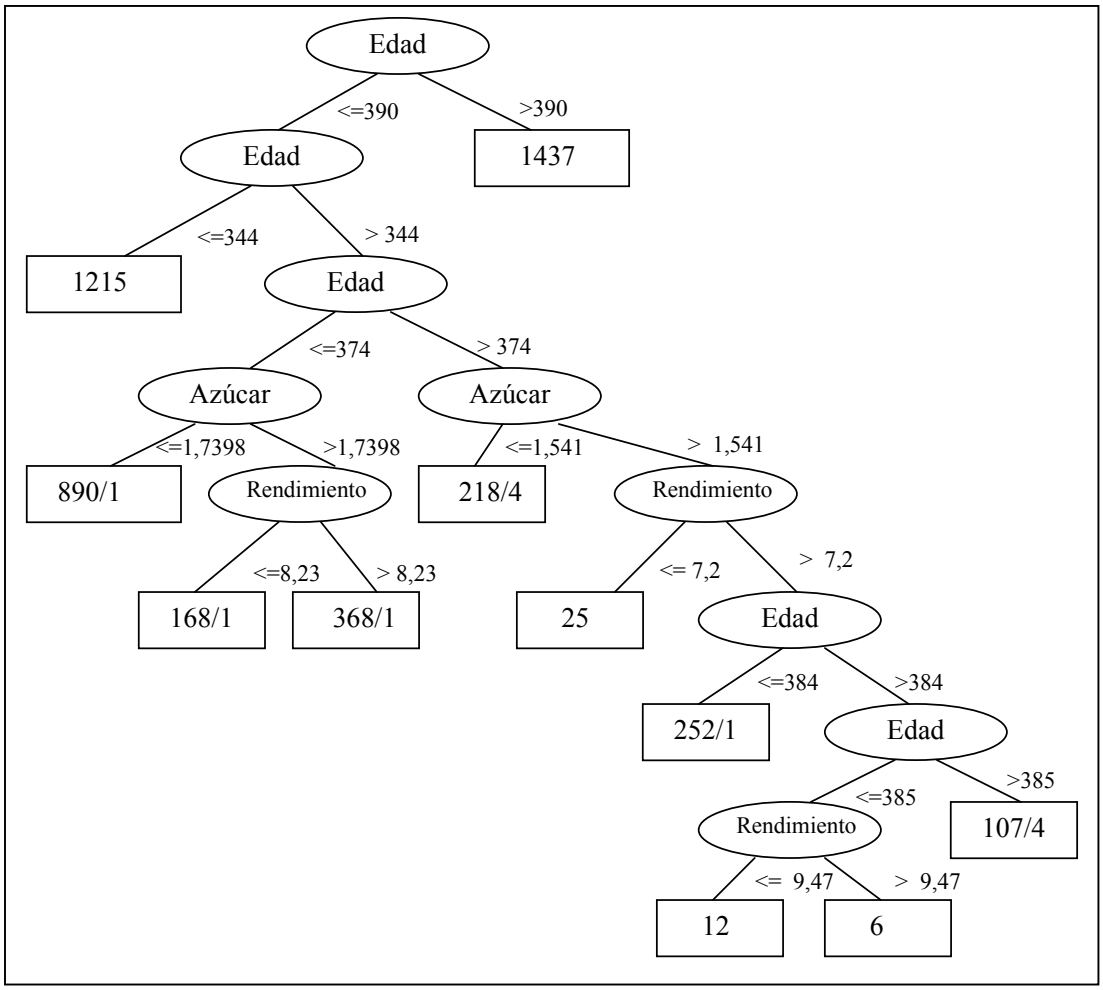

Figura 19. Árbol de decisión generado en el caso de prueba 4.

La figura 19 muestra el árbol obtenido. La precisión del modelo y la matriz de confusión obtenidas en esta prueba son las siguientes:

Precisión del Modelo

PV	PF	Precisión	Clase
0.992	0	0.995	1
0.987	0.001	0.987	2
1	0	1	3
0.995	0.002	0.994	4

Matriz de Confusión

a	b	c	d	<-- Clase
760	0	0	6	a = 1
0	744	0	10	b = 2
0	0	5208	0	c = 3
4	10	0	2654	d = 4

El formato de aplicación del algoritmo para el caso de Prueba 5 es el siguiente:

Formato:	***weka.classifiers.trees.DecisionStump***

Este caso de prueba se seleccionó para verificar el comportamiento de los datos al aplicar otro algoritmo de generación de arboles que proporciona la herramienta WEKA. El algoritmo "***Decision Stump"*** genera un árbol de decisión con una única división. Dado que en los casos de prueba anteriores el nodo raíz es el atributo Edad y en aras de verificar si un algoritmo más sencillo provee una buena solución al problema planteado, se realizó el caso de prueba 5. El conjunto de entrenamiento utilizado consta de 4698 instancias. En esta prueba clasifica correctamente el 58.16% de las instancias, el tiempo de ejecución es 0.99 segundos. El valor del estadístico Kappa en esta prueba es 0.3406. Los resultados obtenidos en este caso, descartan la aplicación del algoritmo "***Decision Stump***" como opción para la solución del problema planteado.

En los casos de *prueba 6 al 8*, los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 9396,7517 y 5637 instancias respectivamente y un factor de confianza de 0.10. El tamaño de los árboles generados fueron 15,15 y 21 nodos y los tiempos de ejecución de 4.17, 2.75 y 2.57 segundos. Al igual que en los casos anteriores, el atributo raíz del árbol es el atributo *Edad.*

En prueba 6 se clasificaron correctamente el 99.55% de las instancias, y el valor del estadístico Kappa en esta prueba es 0.9926. Para el caso de prueba 7 se clasificaron correctamente el mismo número de instancias del caso de prueba 6 con

un conjunto de entrenamiento de menor tamaño, generando el mismo árbol del caso de prueba 6.

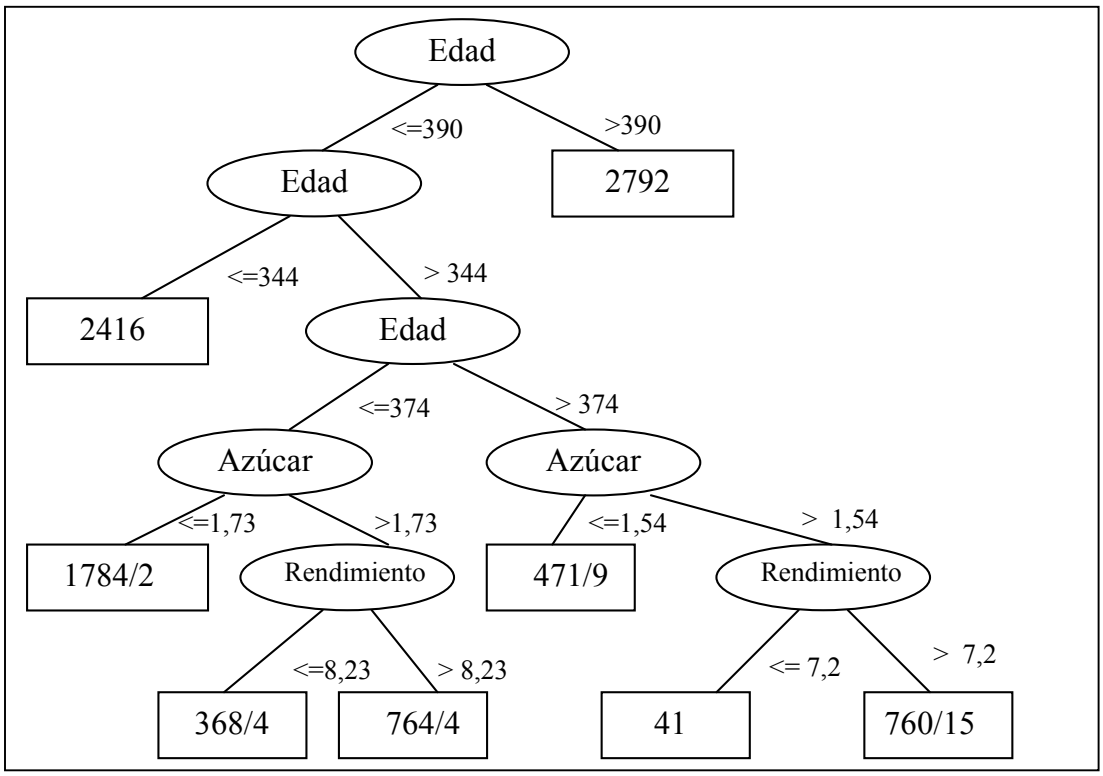

Figura 20. Árbol de decisión generado en el caso de prueba 6.

La figura 20 presenta gráficamente el árbol generado por el caso de prueba 6. Las figuras 21 y 22 muestran los árboles obtenidos para los casos de prueba 7 y 8 respectivamente. Las medidas de eficiencia obtenidas en el caso de prueba 8 son las siguientes: 99.54% de las instancias clasificadas correctamente, valor del estadístico Kappa 0.9926.

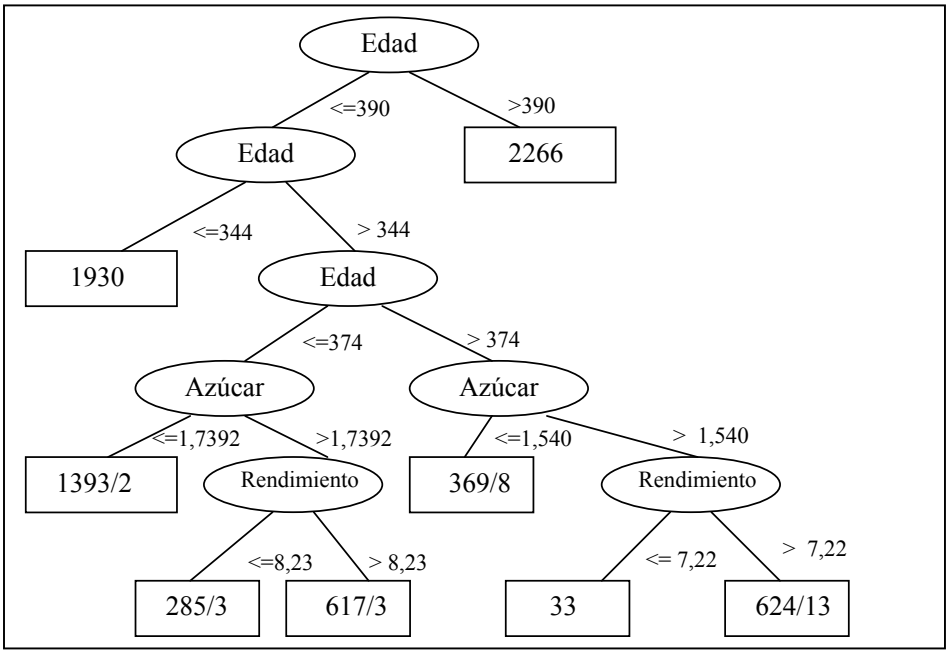

Figura 21. Árbol de decisión generado en el caso de prueba 7.

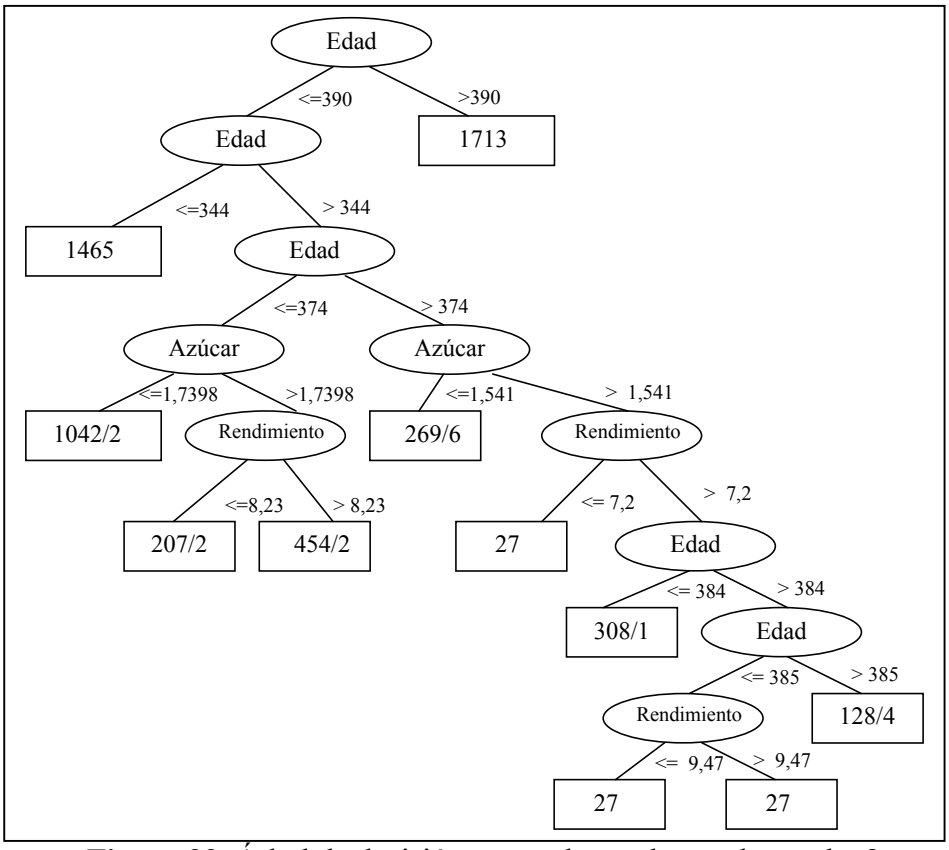

Figura 22. Árbol de decisión generado en el caso de prueba 8.

Seguidamente las figuras 23 y 24, muestran las gráficas comparativas de tiempos de ejecución y tamaños de árboles generados en cada una de las pruebas realizadas para el algoritmo de árboles de decisión C4.5.

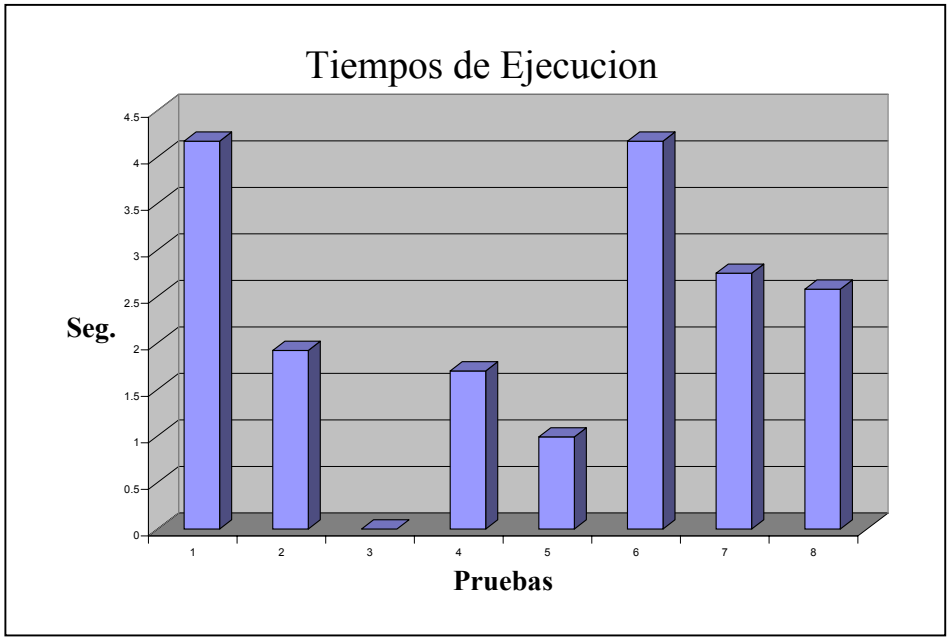

Figura 23. Tiempos de ejecución para las ocho pruebas del algoritmo C4.5.

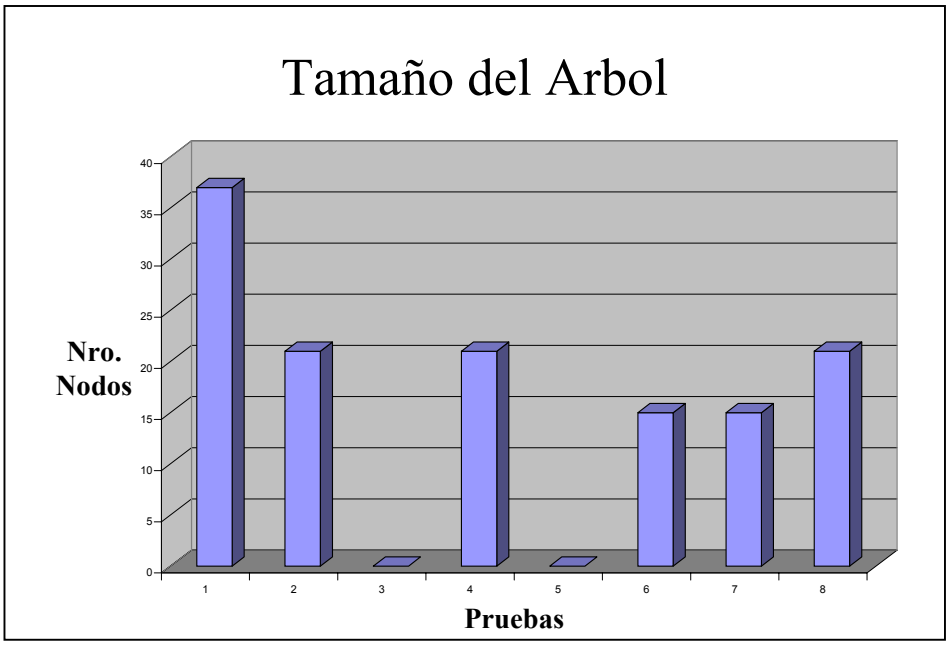

Figura 24. Tamaños de los árboles para las ocho pruebas del algoritmo C4.5.

Tabla 8.

Resultados Pruebas Arboles de Decisión Algoritmo C4.5

Exp	Número de Hojas	Tamaño del Arbol	Tiempo para la construcción del modelo	Instancias correctamente clasificadas	Instancias incorrectamente clasificadas	Error Absoluto Medio
1	19	37	4.17 seg.	99.55%	0.44%	0.0031
2	11	21	1.92 seg.	99.54%	0.45%	0.0038
3	Memoria Insuficiente					
4	11	21	1.7 seg.	99.68%	0.31%	0.0029
5			0.99 seg.	58.16%	41.83%	0.2381
6	8	15	4.17 seg.	99.55%	0.44%	0.004
7	8	15	2.75 seg.	99.55%	0.44%	0.0041
8	11	21	2.58 seg.	99.54%	0.45%	0.0038

Fuente: El Autor

La tabla 8, presenta un resumen de los resultados obtenidos en las pruebas para la obtención del modelo basado en árboles de decisión.

Para la generación del modelo utilizando la técnica de Redes Neuronales Multicapas con Retropropagación, la arquitectura de la red utilizada para la clasificación del modelo a capas es 12 unidades de entrada, 8 unidades de la capa oculta y 4 unidades de salida, tal como se muestra en la figura 25 que aparece en la siguiente pagina.

El formato de aplicación del algoritmo para el caso de Prueba 9 es el siguiente:

Formato: weka.classifiers.functions.MultilayerPerceptron -L 0.3 –M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G –R

Los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 9396 casos, número de épocas 500 y tasa de aprendizaje de 0.3. El tiempo de ejecución para la construcción del modelo fue 1000.25 seg.

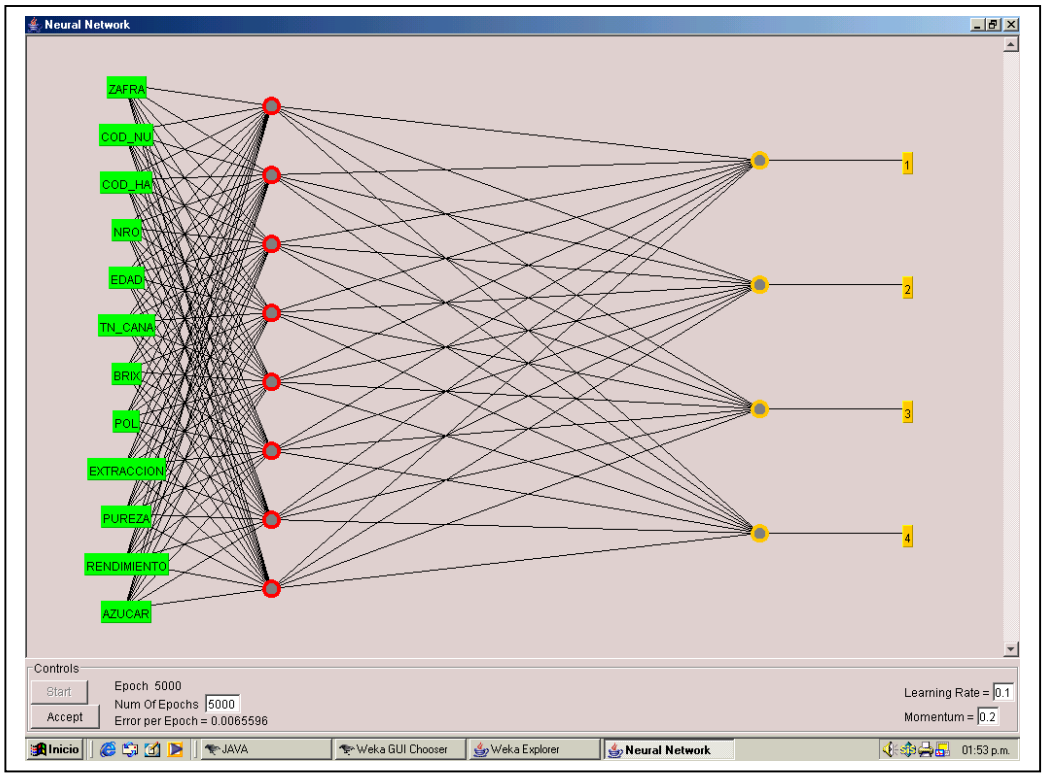

Figura 25. Arquitectura de la red neuronal usada (WEKA Explorer)

El error que presenta el clasificador, para el conjunto de entrenamiento es de 3.2352% de instancias no clasificadas en forma correcta, el número de instancias clasificadas correctamente es del 96.76%. El estadístico Kappa mide la coincidencia de la predicción con la clase real, en este caso su valor es 0.9455.

Instancias Correctamente Clasificadas	18185	96.7648 %
Instancias Incorrectamente Clasificadas	6080	3.2352 %
Estadístico Kappa	0.9455	
Error Absoluto Medio	0.0239	
Número Total de Instancias	18793	

Precisión del Modelo

PV	PF	Precisión	Clase
0.873	0.002	0.92	1
0.952	0.009	0.923	2
0.996	0.033	0.985	3
0.944	0.01	0.959	4

Matriz de Confusión

a	b	c	d	<-- Clase
1343	50	87	58 |	a = 1
0	1383	26	44 |	b = 2
3	7	10372	30 |	c = 3
35	105	163	5087 |	d = 4

Para el caso de prueba 10, los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 9396 casos, numero de épocas 500 y tasa de aprendizaje de 0.2. El tiempo de ejecución para la construcción del modelo fue 967.46 seg., clasificó correctamente el 96.94% de las instancias. El valor del estadístico Kappa en este caso fue 0.9487.

Instancias Correctamente Clasificadas	18218	96.9404 %
Instancias Incorrectamente Clasificadas	575	3.0596 %
Estadístico Kappa	0.9487	
Error Absoluto Medio	0.0247	
Número Total de Instancias	18793	

Precisión del Modelo

PV	PF	Precisión	Clase
0.886	0.004	0.953	1
0.955	0.011	0.88	2
0.998	0.023	0.982	3
0.942	0.01	0.975	4

Matriz de Confusión

a	b	c	d	<-- Clase
1363	60	42	73 |	a = 1
2	1387	17	47 |	b = 2
0	10	10393	9 |	c = 3
65	20	130	5075 |	d = 4

Para el caso de prueba 11, los parámetros de ejecución del algoritmo fueron un conjunto de entrenamiento de 4698 instancias, con un número de épocas de 1000 y una tasa de aprendizaje de 0.2. El tiempo de ejecución para la construcción del

modelo fue 894.9 seg., clasificando correctamente el 96.42% de las instancias. El valor del estadístico Kappa en este caso fue 0.9407.

En los casos de prueba del 12 al 14, se modificaron los parámetros de ejecución del algoritmo tales como el conjunto de entrenamiento, el numero de épocas y la tasa de aprendizaje obteniéndose el mayor porcentaje de instancias clasificadas correctamente en el caso de prueba 13, para un porcentaje de 97.49% de instancias clasificadas correctamente.

A continuación se muestra la figura 26 con una gráfica comparativa del tiempo de ejecución de cada una de las pruebas realizadas para la red neuronal multicapa propuesta.

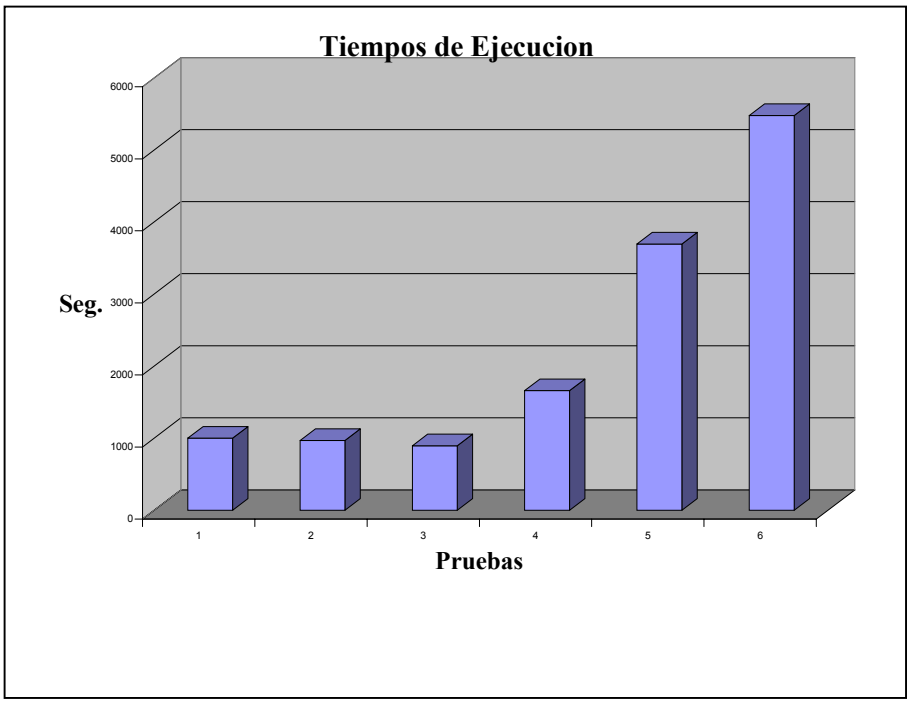

Figura 26. Tiempos de Ejecución pruebas red neuronal multicapas

La Tabla 9 presenta un resumen de los resultados obtenidos en la construcción de los modelos mediante la técnica de redes Neuronales Multicapas con Retropropagación.

Tabla 9.

Resultados Experimentos Redes Neuronales Multicapa con Retropropagación

Exp	Tiempo para la construcción del modelo	Instancias correctamente clasificadas	Instancias incorrectamente clasificadas	Error Absoluto Medio
9	1025 seg.	96.76%	3.23%	0.0239
10	967.46 seg.	96.94%	3.05%	0.0247
11	894.9 seg.	96.42%	3.57%	0.0243
12	1659.9 seg.	97.04%	2.95%	0.0196
13	3695.45 seg.	97.49%	2.50%	0.0159
14	5479.53	13.85%	86.14%	0.3757

Fuente: El Autor

Calificación del Modelo.

Los resultados obtenidos en la generación del modelo utilizando la técnica de árboles de decisión, exhiben una tasa de error de 0.4% de instancias no clasificadas correctamente, lo que permite calificar a estos modelos como exactos. El error obtenido en la construcción del modelo de red neuronal multicapa con retropropagación, el error de instancias no clasificadas correctamente está entre el 2% y 3%, siendo esta cifra mayor que en el modelo de árboles de decisión y en consecuencia estos modelos muestran menor exactitud de clasificación. En líneas generales ambos esquemas de aprendizaje mostraron una solución con un alto nivel de exactitud.

Un análisis de los tiempos de ejecución de ambos modelos permite calificar el modelo de árbol de decisión como de mayor eficiencia respecto al tiempo, dado que el promedio estuvo alrededor de los 4 segundos, mientras que el rango del tiempo de ejecución de la red neuronal estuvo entre 800 segundos y 5000 segundos.

Asi mismo, respecto al uso de otro recurso computacional como lo es la memoria del equipo, ambos modelos exhiben un comportamiento similar en el uso intensivo

de la memoria de la maquina, presentando en algunas pruebas el error de memoria insuficiente.

Aplicación de la Fase V Evaluación de los Resultados

Esta fase está destinada a la evaluación de los resultados obtenidos desde el punto de vista de los objetivos del proyecto de minería de datos. Desde el inicio del proyecto de minería se manejó la hipótesis de que uno de los factores que mayor incidencia tiene tanto en la productividad como en la rentabilidad del cultivo de la Caña de Azúcar, es el referente a la edad de la planta al momento de cosecharla.

Esto se evidencia en las características que describen cada una de las clases, las cuales fueron suministradas por el experto del área agrícola. Los resultados obtenidos mediante la construcción del modelo de árbol de decisión, muestra explícitamente que éste es el factor decisivo en el éxito de todo el proceso productivo.

Los modelos generados confirman las apreciaciones empíricas existentes y permiten establecer nuevas estrategias a la luz del nuevo conocimiento descubierto.

El clasificador generado mediante la técnica de árbol de decisión aporta información valiosa, en cuanto descubre relaciones entre siete de las doce variables del conjunto de datos, tal como lo muestra el árbol generado en el caso de prueba 1. Para conjuntos de entrenamientos mas pequeños, prueba la veracidad de las clasificación suministrada, dado que genera árboles que confirman la fuerte relación existente entre las variables Edad, Azúcar y Rendimiento.

El clasificador obtenido a través de la red neuronal muticapas con retropropagación, aunque muestra altos valores de exactitud no facilita la compresión los resultados obtenidos por parte del experto del área agrícola, por lo que su baja interpretabilidad conlleva a que el usuario final del proyecto no utilice el modelo obtenido.

Técnicamente, fue posible constatar que el comportamiento de los datos de la empresa respecto a los valores de estándares mundiales del negocio azucarero está bastante ajustado. Es así como la aplicación del estándar interno manejado por la

empresa en la determinación de las clases para realizar la clasificación, permitió establecer a través de datos históricos de las unidades de producción de caña de azúcar, la cantidad de unidades de producción que describen un comportamiento de alta rentabilidad y altos rendimientos.

La información aportada por los modelos generados permite a la empresa reorganizar estrategias tanto financieras como técnicas tendientes a mejorar la productividad de cada unidad de producción.

La información valiosa que guardan los datos registrados en el sistema de información permitió al gerente del área agrícola visualizar un nuevo panorama de la problemática que enfrentan las unidades de producción, proporcionado conocimiento que será incorporado a su sistema de información.

Aplicación de la Fase VI Despliegue de Resultados

El reporte final de los resultados obtenidos y la conclusión del proyecto de minería de datos se presenta en el Capitulo VI Conclusiones y Recomendaciones de la presente investigación, todo esto con la finalidad de evitar repeticiones innecesarias de información.

CAPITULO VI

CONCLUSIONES Y RECOMENDACIONES

La presente investigación logró cumplir con todos los objetivos propuestos desarrollando un modelo de aprendizaje automático que permite predecir el rendimiento de la Caña de Azúcar, lo cual fue obtenido aplicando técnicas de minería de datos. Se obtuvo un modelo de aprendizaje automático que resuelve el problema planteado desde el punto de vista del proyecto de minería, con lo cual se cumple el objetivo general del proyecto.

Los modelos generados permiten clasificar las unidades de producción de caña de azúcar con una exactitud de un 99.5%, contribuyendo así a la detección de las unidades más productivas y las que no muestran este comportamiento, detectándose así mismo que patrones comunes presentan dichas unidades. El trabajo de investigación representa un aporte novedoso en cuanto a la aplicación de técnicas de minería de datos en el área agrícola azucarero, pudiendo servir de antecedente a futuras investigaciones que implementen otras técnicas de minería con la finalidad de comparar los resultados.

Para lograr el objetivo general propuesto se cumplieron todos los objetivos específicos planteados. Se desarrollo un modelo basado en arboles de decisión mediante el algoritmo C4.5, tomando en cuenta los atributos más relevantes que describen el proceso en estudio. Los resultados obtenidos desde el punto de vista de la exactitud reflejan un excelente desempeño, dado que el error de instancias mal clasificadas esta en el orden de menos del 0.5%. Los tiempos de ejecución para la generación del modelo estuvieron en el orden de 4 segundos.

Se desarrollo un modelo basado en la red neuronal multicapas con retropropagación, que igualmente clasifica las unidades de producción de acuerdo a la criterio de clasificación suministrado por la Gerencia de Gestión Agrícola. Frente al

conjunto de datos, los modelos de experimentación de la red reflejan un desempeño de alrededor del 97% de instancias correctamente clasificadas, utilizando un considerable aumento del tiempo de ejecución. Los tiempos de ejecución para la generación este modelo estuvo en el orden de 5000 segundos en el peor de los casos y 800 segundos en el mejor de los casos.

Se realizó la comparación de los modelos de desarrollados de acuerdo a los resultados anteriores estableciendo que desde el punto de vista técnico el modelo que exhibe mejor comportamiento en cuanto a tiempo de ejecución y exactitud en la clasificación es el modelo basado en arboles de decisión utilizando el algoritmo C4.5. Además, desde el punto de vista del usuario final, el conocimiento representado en el árbol de decisión es más explícito y de mayor comprensión que los modelos de red presentados.

Se aplico la Metodología CRISP-DM en el desarrollo del proyecto de minería facilitando el la planificación, ejecución y seguimiento de cada una de las etapas del proceso, demostrándose una vez mas que la aplicación de la misma contribuye a la obtención de resultados confiables, por lo que es ampliamente utilizada por los profesionales del área.

Se recomienda realizar estudios adicionales en esta área utilizando técnicas de aprendizaje no supervisado para realizar comparaciones y establecer mejoras en cuanto a la calidad de los modelos obtenidos.

Así mismo, se le recomienda a la Gerencia de Gestión Agrícola de Azucarera Río Turbio C.A y a la Gerencia de Gestión de Sistemas, iniciar a la brevedad posible la puesta en marcha del Sistema de Control de Hacienda en su totalidad, a los fines de registrar la información relacionada al manejo de las haciendas y la implantación de un sistema para el registro de las precipitaciones y oscilación térmica, con la finalidad de incorporar estas variables a un futuro estudio de minería de datos mas completo.

Por otra parte, seria interesante el desarrollo de un ambiente con fines educativos para la extracción de conocimiento similar al de WEKA, que ayuden a los estudiantes noveles de Inteligencia Artificial en la compresión del proceso de minería de datos y

las técnicas involucradas; donde participen profesionales de diferentes disciplinas, como Ingeniería de Software, Informática, Estadística, entre otros.

REFERENCIAS BIBLIOGRAFICAS

Bigus, J. 1996. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. The McGraw Hill Companies Incorporated Press, USA.

Camacho, S. De Niño, T. Pire, R. Rodríguez, Ayolaida. 2002. Manual para la Presentación del Trabajo Conducente al Grado Académico de Especialización, Maestría, Doctorado. Universidad Centroccidental "Lisandro Alvarado" Vice-Rectorado Académico. Dirección de Postgrado. Venezuela.

Cohen, J. 1980. A coefficient of agreement for nominal scales. Education and Psychological Measurement. USA.

Díaz, B. Morillas, A. 2003. Minería de Datos y Lógica Difusa. Una aplicación al estudio de la Rentabilidad Económica de las Empresas Agroalimentarias en Andalucía. Universidad de Málaga. URL: campusvirtual.uma.es/morillas/Dataminingylogicadifusa.pdf. (Consulta: Marzo 15, 2004)

De Sousa, O. Rea, R. 1993. Correlación entre los componentes de rendimiento y calidad en cinco cultivares híbridos de Caña de Azúcar. Revista Científica Caña de Azúcar. Vol. 11(01): 45-52. URL: http://www.ceniap.gov.ve/bdigital/cana/cana1101/texto/correlacion.htm (Consulta: Diciembre 12, 2003)

Fayyad, U. Piatetsky-Shapiro, G. Smyth, P. Uthurusamy, R. 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, USA.

Frawley, G. Piatetsky-Shapiro, G. Matheus, C. 1991. Knowledge Discovery in Databases: An Overview. AAAI Press/The MIT Press, USA.

Gamberger D, Smuc Tomislay and Mari Ivan. 2001. Data Mining Server. Laboratory Information System - Rudjer Boskovic Institute. URL: http://dms.irb.hr/tutorial/tut_intro.php. 2001.(Consulta: Marzo 15, 2004)

Groth, R. 1998. Data Mining. Building Competitive Advantage. Prentice Hall. USA.

Han, J. Kamber, M. 2001. Data Mining. Concepts and Techniques. Morgan Kaufmann Publishers, USA.

Houtsma, M. Swami, A. 1995. Set Oriented Mining for Association Rules in Relational Databases. In Proceeding of the 11th IEEE International Conference on Data Engineering. Taiwan.

Holsheimer, M. Siebes, A. 1994. Data Mining: The Search for Knowledge in Databases. URL: ftp://ftp.cwi.nl/pub/CWIreports/AA/CS-R9406.ps.Z, (Consulta: Diciembre 01, 2003)

John. Y, Langari, R 1999. Fuzzy Logic Intelligence, Control and Information. Printice Hall, USA.

Kdnuggets. 2004. Portal de Data Mining, Web Mining & Knowledge Discovery. URL: http://www.kdnuggets.com/ (Consulta: Mayo 08, 2004)

Kennedy, R. Lee, Y. Van Roy, B. Reed, C. Lippman, R. 1995-1997. Solving Data Mining Problem through Pattern Recognition. Printice Hall, USA.

Mago, Pedro. Galíndez, O. 1986. Epoca de Siembra y Cosecha en dieciocho (18) variedades comerciales de Caña de Azúcar en Río Turbio Venezuela. Revista Científica Caña de Azúcar Vol. 4 (1): 27-63 URL: http://www.ceniap.gov.ve/bdigital/cana/cana0401/texto/epoca.htm (Consulta: Diciembre 09, 2003)

Mannila, H. Toivonen, H. Verkamo, I. 1994. Efficient Algorithms for Discovering Association Rules. Proceedings of the AAAI Workshop on Knowledge Discovery in Databases (KDD-94). USA.

Masters, T. 1993. Practical Neural Network Recipes in C++. Morgan Kaufmann Publishers, USA.

Mendel, J. 2001. Uncertain Rule-Based Fuzzy Logic Sistems: Introduction and New Directions. Printice Hall, USA.

Metodología CRISP-DM 1996. URL: http://www.crisp-dm.org/ (Consulta: Febrero 23, 2004)

Michalsky, R. Bratko, I. Kubat, M. 1998. Machine Learning and Data Mining, Methods and Applications, John Wiley & Sons Ltd. England.

Michalsky, R. Carbonell, J. Mitchell, T. 1983. Machine Learning: An Artificial Intelligence Approach. Morgan-Kauffman Publishers, USA.

Mitchell, T. 1997. Machine Learning. WCB/McGraw-Hill. USA.

Ramos, E., Giménez, C. 2004. Proceso de Desarrollo de Minería de Datos. Universidad Central de Venezuela, Facultad de Ciencias, Escuela de Computación, Inteligencia Artificial. URL: http://strix.ciens.ucv.ve/~iartific/Material/MetodosMineria.pdf (Consulta: Mayo 02, 2004).

Rodríguez, M. Alvarez, J. Mesa, J. González, A. 2002. Metodologías para la Realización de Proyectos de Data Mining. URL: http://www.aeipro.com/congreso_03/pdf/mayte@api.uniovi.es_dc2336ab68ff252c5840828af4bc7999.pdf (Consulta: Enero 12, 2004)

Savasere, A. Omiecinski, E. Navathe, S. 1995. An Efficient Algorithm for Mining Association Rules in Large Databases. In Proceeding of the 21 nd International Conference on Very Large Databases. Suiza.

Swingler, K. 1996. Applying Neural Networks a Practical Guide. Morgan Kaufmann Publishers, USA.

Weiss, S. Indurkhya, N. 1998. Predictive Data Mining a Practical Guide. Morgan Kaufmann Publishers, USA.

Witten, I. Frank, E. 1999. Data Mining. Practical Machine Learning Tools and Techniques with Java Implementacions.. Morgan Kaufmann Publishers, USA.

Zérega, L. Hernández, T. Valladares, J. 1991. Caracterización de suelos y aguas afectadas por sales en zonas cañameleras de Azucarera Río Turbio. Revista Científica Caña de Azúcar Vol. 09 (1): 5-52 URL: http://www.ceniap.gov.ve/bdigital/cana/cana0901/texto/caracterizacion.htm (Consulta: Diciembre 11, 2003)