IMPLEMENTACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DE INFORMACIÓN DEL ATLAS SÍSMICO COLOMBIANO

JOHANN MANUEL JAIMES RINCÓN
FABIO ANDRÉS VERGEL CRIADO

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
FACULTAD DE INGENIERÍA DE SISTEMAS
SISTEMAS DE INFORMACIÓN E INGENIERÍA DE SOFTWARE
BUCARAMANGA
2008

IMPLEMENTACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DE INFORMACIÓN DEL ATLAS SÍSMICO COLOMBIANO

JOHANN MANUEL JAIMES RINCÓN
FABIO ANDRÉS VERGEL CRIADO

Trabajo de grado presentado para optar el título de: Ingeniero de Sistemas

Director
Ms.C Javier Hernández Cáceres

Asesor
PHD. Juan Carlos García Díaz

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
FACULTAD DE INGENIERÍA DE SISTEMAS
SISTEMAS DE INFORMACIÓN E INGENIERÍA DE SOFTWARE
BUCARAMANGA
2008

Nota de aceptación:

_____

_____

_____

_____

_____

_____
Firma del Jurado

_____
Firma del Jurado

_____
Firma del Director

Bucaramanga, Agosto 12 de 2008

TABLA DE CONTENIDO

pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

pág.

LISTA DE ANEXOS

pág.

GLOSARIO

AC: análisis de Clúster.

AD: análisis Discriminante.

CP: componentes principales.

RESUMEN

Hoy en día, se estima que el crecimiento de información digital es del 60% anual, lo cual indica, que hay gran cantidad de información almacenada en las bases de datos de diferentes entidades y organismos. Sin embargo, permanece oculta mucha información valiosa que al ser identificada y explotada permite establecer patrones o normas a las empresas, permitiendo reducir costos ó maximizar ganancias y, así mismo, generan un nuevo conocimiento.

A nivel nacional, las empresas explotadoras de hidrocarburos son de las más interesadas en extraer nuevo conocimiento de sus bases de datos, como es el caso del ICP (Instituto Colombiano del Petróleo), organismo que en el año 2005, desarrolló el Atlas sísmico colombiano, toda vez que le permitió, almacenar la información sísmica colombiana. En este atlas sísmico reposa gran cantidad de información útil que por medio de las técnicas de minería de datos le proporcionan al ICP poder establece estrategias, normas para minimizar costos y maximizar ganancias y productividad.

En este proyecto se implementara un prototipo computacional desarrollado con el cual, se pretenden analizar los datos obtenidos en el atlas sísmicos colombianos que se encuentran almacenados en la base de datos ICP, y poder facilitar el trabajo de análisis de las técnicas sobre grandes volúmenes de datos, en razón de obtener resultados rápidos y confiables.

Palabras Clave: Atlas sísmico de Colombia, Minería de Datos, WEKA, KDD, CRISP-DM, ICP.

INTRODUCCIÓN

Actualmente se cuenta con diferentes opciones al momento de elegir una herramienta computacional que permita aplicar técnicas de minería de datos a información que está almacenada en las bases de datos; estas herramientas se encuentran bajo licencias libres o propietarias, sin embargo, la complejidad de unas y alto costo de otras evitan que se pueda acceder a estas herramientas evitando así mismo que sea extraída información valiosa de diferentes bases de datos. Como objetivo principal de este proyecto está el diseño e implementación de un prototipo computacional que, así como las herramientas actuales, cuenta con algoritmos de técnicas de minería de datos pero con la diferencia que las técnicas implementadas en este prototipo son de fácil acceso y entendimiento, permitiendo al usuario aplicarlas a la información deseada.

Durante el año 2005, el ICP creó el Atlas sísmico colombiano, el cual cuenta con información sísmica colombiana de forma detallada[1]. En este atlas se incluyó una técnica de minería de datos conocida como redes bayesianas, la cual produjo resultados satisfactorios, razón por la cual, nace el interés del ICP en implementar técnicas de minería de datos adicionales las cuales ayuden a complementar la técnica actual y, con la aplicación de las técnicas de minería de datos en conjunto, se puede tomar decisiones que permitan optimizar los procesos de exploración, explotación y producción de hidrocarburos en Colombia.

[1] Bienvenidos: ECOPETROL: Disponible en internet. [Online]. [Visitada Agosto 21 de 2007]. Disponible en Internet: <http://www.ecopetrol.com.co/Informe_Anual/....htm >.

Este proyecto contiene un alto componente de investigación debido a que la información sísmica tiene características especiales que requieren una análisis sobre cuáles son los algoritmos que mejor se adaptan a ese tipo de información.

1. MARCO TEÓRICO

1.1 PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

Ciertos sistemas producen una cantidad enorme de datos; estos datos contienen valiosa información que puede resultar muy útil. Estos inmensos volúmenes de bases de datos hacen que para el usuario un análisis sea muy difícil extraer información importante para una organización o para el interesado.

Es conocido que los datos por si solos no producen beneficio directo. Su verdadero valor radica en extraer información útil para la toma de decisiones o la exploración y comprensión del fenómeno que produjo los datos. Tradicionalmente en la mayoría de los campos este análisis de datos se hacía mediante procesos manuales o semiautomáticos: uno o más analistas con conocimientos de los datos y con la ayuda de técnicas estadísticas proporcionaban resúmenes y generaban informes, o validaban modelos sugeridos manualmente por los expertos. Sin embargo, este proceso, es irrealizable conforme aumenta el tamaño de los datos y el número de dimensiones o parámetros se incrementa.

Por todo lo anterior, surge la necesidad de metodologías para el análisis inteligente de datos, que permitan descubrir un conocimiento útil a partir de los datos. Este es el proceso de KDD (Knowledge Discovery In Databases) [2]

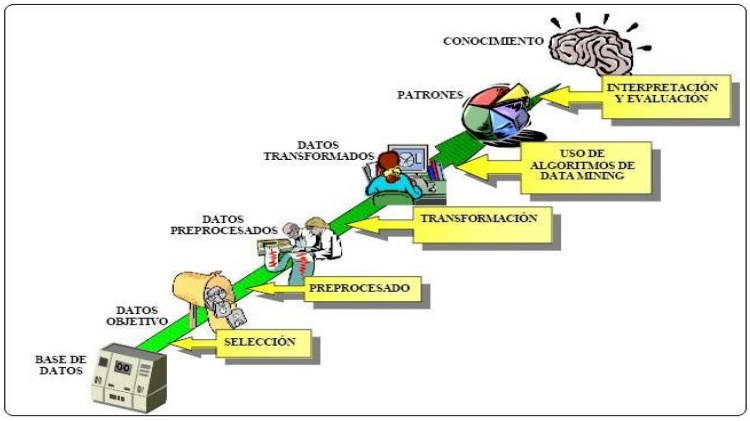
[2] Red Española de Minería de Datos y Aprendizaje. [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: <http://www.lsi.us.es/redmidas/>

KDD puede ser definido como “El proceso no trivial de identificar patrones válidos, nuevos (originales), potencialmente útiles y deseablemente entendibles en datos. Un proceso para encontrar conocimiento desconocido o no reconocido previamente, potencialmente útil, de datos estructurados almacenados en BD” [3].

Analizando un poco mas se puede decir en general que este proceso de descubrimiento requiere de ciertas etapas esenciales y primordiales para la obtención de información útil, las cuales se describen a continuación:

1.1.1 Proceso de KDD. Las diferentes etapas que se aplican en este proceso se pueden observar de una manera tangible en la (Figura 1).

Figura 1. Fases de un proceso de minería de datos



Fuente: Molina, Luis Carlos. (2000). Torturando los Datos Hasta que Confiesen. Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña. Barcelona, España.

[3]OHRN, A. 1999. “Discernibility and Rough Sets in Medicine: Tools and Applications”, Department of Computer and Information Science, Norwegian University of Science and Technology, N-7941 Trondheim, Norway.

Este proceso es iniciado con datos simples, que al realizar sus diferentes etapas de proceso, se termina con la obtención de conocimiento útil. El diseño metodológico adoptado para el desarrollo de esta tesis esta basado en la metodología de aplicaciones de KDD (Descubrimiento de Conocimiento en BD).

1.1.1.1 Bases de datos. Los datos usualmente se encuentran almacenados en una tabla ASCII, mediante bodegas de bases de datos que están optimizadas para el proceso analítico utilizando las herramientas de KDD las cuales pueden incluir mecanismos para almacenar y acceder a datos.

1.1.1.2 Selección. En el proceso de selección se escoge el conjunto o subconjunto de datos de la base de datos, y la selección de muestras de datos las cuales son llamadas instancias.

1.1.1.3 Pre procesado. Cuando se llega al pre-procesamiento, se hace la limpieza de los datos y la eliminación de ruidos, datos atípicos, casos extremos, y el manejo de los datos faltantes y desconocidos.

1.1.1.4 Transformación. Al llegar al proceso de transformación al formato requerido por el algoritmo especifico de minería de datos.

1.1.1.5 Minería de Datos. Cuando se llega a la minería de datos comienza la búsqueda de patrones de interés en una forma particular de representación, la búsqueda se puede expresar como un modelo o como un patrón que expresa cierta dependencia entre datos, el modelo tiene como función los análisis

estudiados anteriormente y la forma de representarlos (función lineal, conjunto de reglas), en esta etapa también se encuentra el criterio de preferencia y la estrategia de búsqueda y antes de terminar se va a pasar por una evaluación, esa evaluación es importante, pues la interpretación de los patrones descubiertos, pueden beneficiarse usando la visualización y también se puede borrar patrones redundantes o irrelevantes[4], los patrones pueden compararse con conocimiento previamente almacenado, y por ultimo se obtendrá el la incorporación de conocimiento descubierto en un sistema de desempeño y se entregara documentado y reportados a las personas interesadas en el desarrollo del proyecto.

1.1.1.6 Interpretación y Evaluación. Finalmente, se obtiene nueva información valiosa, y hasta ahora desconocida, respecto a la base de datos inicial, es momento de realizar una interpretación de los resultados obtenidos, los cuales pueden ser nuevas reglas o patrones de comportamiento, generalidades, entre otras dependiendo de las técnicas de usadas en la etapa de minería de datos.

Para el desarrollo del prototipo que servirá de soporte a la aplicación de la metodología de KDD se utilizará una metodología convencional para el desarrollo del software el cual se va a aplicar el ciclo de vida lineal.

[4] Descubrimiento de Conocimiento en Bases de Datos [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonografiaMD.PDF >.

1.2 MINERÍA DE DATOS

La minería de datos es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Aunque se basa en varias disciplinas, algunas de ellas más tradicionales, se distingue de ellas en la orientación más hacia el fin que hacia el medio, hecho que permite nutrirse de todas ellas sin prejuicios. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido a la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permita a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que se debe actuar y tomar decisiones[5]. La minería de datos es solo una etapa, sin embargo es la más importante de lo que se ha venido llamando el proceso de extracción de conocimiento a partir de datos. Este proceso consta de varias fases e incorpora muy diferentes técnicas de los campos del aprendizaje automático, la estadística, la base de datos, los sistemas de toma de decisión, la inteligencia artificial y otra área de la informática y la gestión de información[6].

La Minería de Datos surgió como una integración de múltiples disciplinas y tecnologías tales como: la estadística, el soporte a la toma de decisiones, el aprendizaje computacional, la gestión y almacenamiento de bases de datos, inteligencia artificial, el apoyo a la toma de decisiones, entre otros muchos campos. Para la realización de estos procesos se aplican técnicas procedentes de muy diversas áreas, como pueden ser los algoritmos genéticos, las redes neuronales, los árboles de decisión, etc. [7]

---

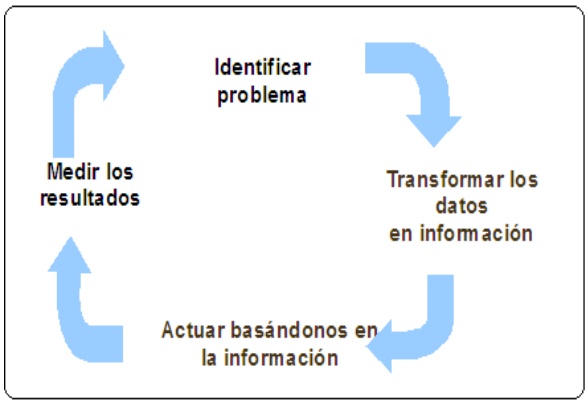
[5] Minería de datos inteligente [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < http://www.itba.edu.ar...datos.htm >.

[6] HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 15

[7] Informática y Salud.BIOCHIPS Y BIOINFORMÁTICA [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < http://www.conganat.org/SEIS/i_s/i_s25/i_s25_1.htm >.

1.2.1 Proceso de Minería de Datos. El proceso de minería de datos está compuesto por diferentes etapas, entre los cuales se encuentran (ver figura 2):

Figura 2. Etapas del proceso de minería de datos



Fuente: Informática y Salud.BIOCHIPS Y BIOINFORMÁTICA [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < http://www.conganat.org/SEIS/i_s/i_s25/i_s25_1.htm >.

A.) Identificar el problema: Delimitar los objetivos que el cliente desea bajo la orientación del especialista en Minería de datos.

B.) Transformar los datos en información: Es la selección, limpieza, transformación y reducción de los datos en las bases de datos. Esta es la epata más extensa en el proceso de minería de datos.

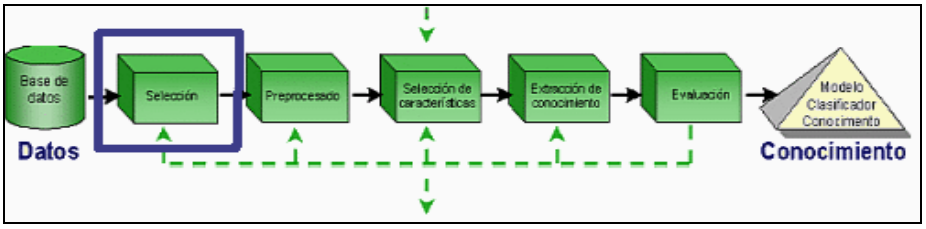
C.) Actuar basándonos en la información: Se inicia realizando un análisis estadístico de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una aproximación. Dependiendo de los objetivos planteados en la etapa inicial se puede usar una o más técnicas de minería de datos.

D.) Medir los resultados: Verifica si los resultados obtenidos son coherentes con los objetivos planteados. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones[8].

1.2.2 Fases de la Minería de datos. Al igual que el KDD, la minería de datos consta de una serie de etapas lineales que se deben seguir sin importar la técnica de minería de datos a usar, estas etapas se nombran a continuación:

1.2.3. Selección

Figura 3. Selección de datos.



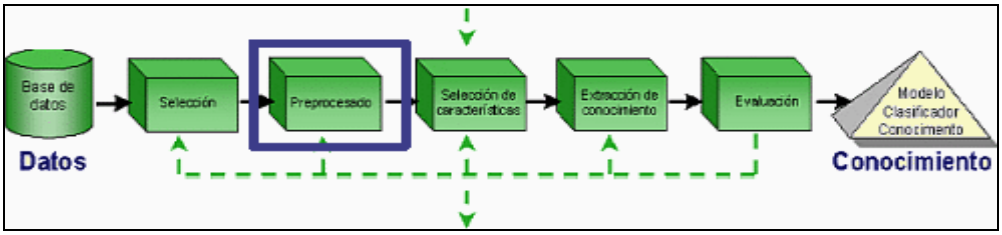
Fuente: DAEDALUS - DATA, Decisions and Language, S. A. Minería de Datos [online, Articulo] 2006. [Citado el 24 de febrero 2066]. Disponible en Internet: http://www.daedalus.es/mineria-de-datos/proceso-de-mineria-de-datos

En la etapa de selección se extrae una pequeña cantidad de datos de la base de datos, conocida como muestra, para aplicar la técnica de minería de datos deseada, esto debido a que casi en ninguna situación se obtienen buenos resultados aplicando las técnicas de minería de datos a la totalidad de los datos de las bases de datos[9].

[8] Mati – Minería de datos. ¿Para qué nos sirve en una publicación digital? [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet:< http://www.mati.unam.mx/...51>.
[9]DAEDALUS - Proceso de minería de datos .[Online]. [Visitada Septiembre de 2007]. Disponible en Internet: URL: < http://www.daedalus.es.../datos >.

1.2.2.1 Pre procesado

Figura 4. Preprocesado de datos



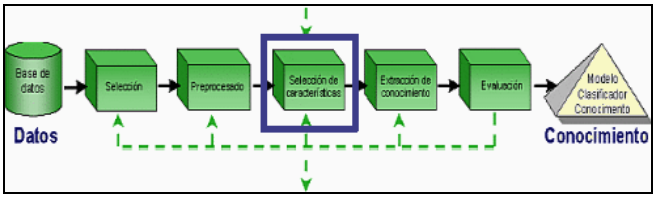
Fuente: Ibíd., p. 17

En esta etapa se eliminan los datos no válidos, como datos atípicos, nulos o desconocidos según la técnica de minería de datos a usar. Además se obtiene una muestra de los datos o se reduce en la cantidad de datos a usar[10].

[10] DAEDALUS - Proceso de minería de datos .[Online]. [Visitada Septiembre de 2007]. Disponible en Internet: URL: < http://www.daedalus.es.../datos >.

1.2.2.2 Selección de características

Figura 5. Selección de características.

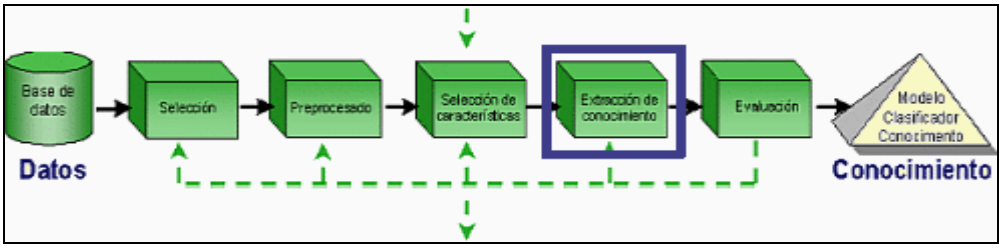


Fuente: Ibíd., p. 17

Después de las etapas de preprocesado y selección de características se tiene una gran cantidad de datos los cuales se debe procurar reducir, aun más, de acuerdo a las principales características entre sí, tales como: selección de atributos más comunes en los datos y atributos acordes con la técnica de minería de datos a usar[11].

1.2.2.3 Extracción de conocimiento

Figura 6. Extracción de conocimiento.



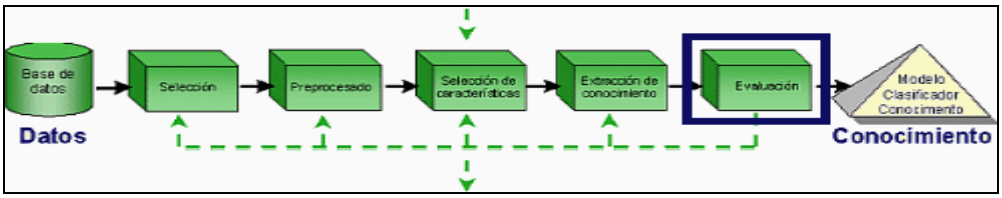
Fuente: Ibíd., p. 17

[11] DAEDALUS - Proceso de minería de datos .[Online]. [Visitada Septiembre de 2007]. Disponible en Internet: URL: < http://www.daedalus.es.../datos >.

En la etapa de extracción de conocimiento se obtienen modelos, reglas ó patrones de acuerdo a la técnica de minería de datos usada. En algunas ocasiones se puede usar más de una técnica de minería de datos simultáneamente para aplicársela a los datos, sin embargo, esta posibilidad es muy escaza debido a que cada técnica de minería de datos requiere un preprocesamiento diferente[12].

1.2.2.4 Evaluación

Figura 7. Evaluación



Fuente: Ibíd., p. 17

Una vez obtenido el modelo generado por la técnica de minería de datos se debe proceder a su evaluación, comprobando que las conclusiones generadas por este sean validas, en el caso de obtener más de un modelo se debe seleccionar el que más se ajuste a los objetivos inicialmente planteados[13].

***1.2.3* Aplicaciones de la minería de datos.** Muchas son las disciplinas relacionadas con la minería de datos, entre ellas se encuentra: Bases de datos;

[12] DAEDALUS - Proceso de minería de datos .[Online]. [Visitada Septiembre de 2007]. Disponible en Internet: URL: < http://www.daedalus.es.../datos >.
[13] DAEDALUS - Proceso de minería de datos .[Online]. [Visitada Septiembre de 2007]. Disponible en Internet: URL: < http://www.daedalus.es.../datos >.

Visualización, comprensión y evaluación de los resultados; estadística; computación y diseño de herramientas computacionales; Recuperación y análisis de información, inteligencia artificial, entre otras. Es por esto que la minería de datos tiene un gran campo de acción, algunos de estos son nombrados a continuación:

- **Bancos y entidades financieras:** Obtención de patrones que permitan establecer que personas solicitan un crédito lo devuelven o no.

- **Análisis de mercados:** Análisis de la frecuencia de compra de los clientes; aprovechamiento de ofertas; Fidelidad de los clientes.

- **Ventas de servicios:** Análisis e identificación de potenciales clientes; identificación de clientes que deseen renovar el servicio; identificación de clientes potencialmente costosos (seguros); identificación de fraudes;

- **Telecomunicaciones:** Identificación de patrones de llamadas; identificación de daños en las redes.

- **Medicina:** Diagnostico de enfermedades; identificación de pacientes con riesgo de enfermedades; Diagnostico de medicina; evaluación de respuesta del paciente a la medicina suministrada.

- ***Ventajas de la minería de datos.*** Permite sacar el máximo provecho a las bases de datos.
  - Ahorrar sumas considerables de dineros a las empresas.
  - Maximizar ganancias y reducir costos.
  - Relación de los investigadores con el personal ejecutivo de la empresa para ejecutar la toma de decisiones.

- Contribuye a la toma de decisiones estratégicas y financieras en una empresa.
- Dependiendo de las técnicas de minería de datos, genera modelos descriptivos o predictivos los cuales permiten identificar patrones de comportamiento o reglas ocultas en los datos.

1.2.4 Ejemplo de Minería de datos. Uno de los ejemplos típicos de aplicación de minería de datos es el análisis de las compras en un supermercado. En determinado supermercado se quiere establecer la relación de la compra entre diferentes productos. Las compras de cierta cantidad de clientes son mostradas a continuación.

Tabla 1. Compras en un supermercado

Clientes	Huevos	Aceite	Pañales	Vinos	Leche	Man/quilla	Carn
1	Si	No	No	Si	No	Si	Si
2	No	Si	No	No	Si	No	Si
3	No	No	Si	No	Si	No	No
4	No	Si	Si	No	Si	No	No
5	Si	Si	No	No	No	Si	No
6	Si	No	No	Si	Si	Si	Si

Fuente: HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 7

Analizando los datos de compras de clientes, se puede establecer, por ejemplo, que el 100 por cien de las veces que se compra Pañales también se compra leche; que si se compra huevos y vino entonces se compra carne. Mediante el análisis de

estos resultados se podría tomar decisiones como ubicación de productos, ofertas y promociones, entre otras. [14]

1.3 TÉCNICAS DE MINERÍA DE DATOS

Cabe mencionar que las técnicas de Minería de datos han sido usadas para una variedad de tareas o aplicaciones, sin embargo, desde un punto de vista global, se propone dos categorías de problemas generales, la *predicción* y la *descripción*[15] [16]

Minería de datos Predictiva (MDP): La Predicción se basa en algunas variables o campos de la Base de Datos para predecir valores desconocidos o futuros de otras variables de interés. Usa primordialmente técnicas estadísticas.

Minería de Datos para descubrimiento de conocimiento (MDDC): La Descripción, en cambio, se centra en encontrar patrones interpretables por el ser humano, a partir de la descripción de los datos. Usa principalmente técnicas de inteligencia artificial. [17]

[14] HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 7

[15] Data Mining–Enfoque [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < http://datamining.iespana.es/dataminig_enfoque.htm >.

[16] eGobierno e Innovación Gubernamental. [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet:< www.tidap.gob.mx/enc_dataw/JulioVargasdwh.ppt >.

[17] Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. [Online]. [Visitada en Octubre de 2007] .Disponible en Internet:< www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf >.

Tabla 2. Clasificación de técnicas de minería de datos

Predictivos	Descubrimiento de Información
Arboles de decisión	Reglas de asociación
Inducción temporal	Análisis de componentes principales
Regresión	Redes Neuronales
Series Temporales	Análisis de Clúster

Fuente: Aplicación de técnicas de Minería de Datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software [online, Articulo]. [Visitada 17 de octubre 2007]. Disponible en Internet: <http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf> P.3

1.3.1 **Análisis Factorial.** El análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables. Esos grupos homogéneos se forman con variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros.

Cuando recogemos un gran número de variables de forma simultánea, como por ejemplo en un cuestionario de satisfacción laboral, podemos estar interesados en averiguar si las preguntas del cuestionario se agrupan de alguna forma característica. Aplicando un análisis factorial a las respuestas de los sujetos podemos encontrar grupos de variables con significado común y conseguir reducir el número de dimensiones para explicar las respuestas de los sujetos.

El análisis factorial es, por tanto, una técnica de reducción de la dimensionalidad de los datos. Su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos en una forma concisa y segura como una ayuda a la construcción de nuevos conceptos y teorías[18].

- El Análisis Factorial puede ser exploratorio o confirmatorio.

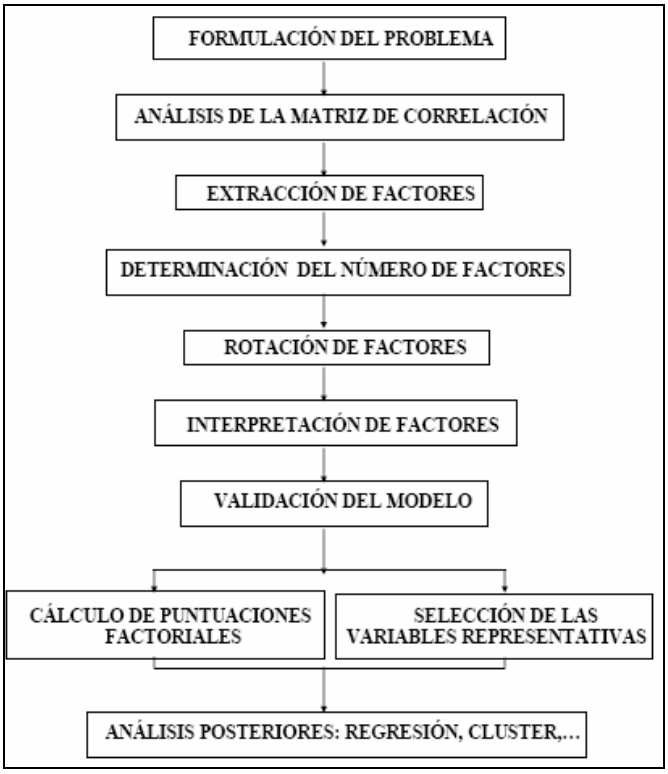
- El análisis exploratorio se caracteriza porque no se conocen a priori el número de factores y es en la aplicación empírica donde se determina este número.

- Por el contrario, en el análisis de tipo confirmatorio los factores están fijados a priori, utilizándose contrastes de hipótesis para su corroboración[19].

A diferencia de lo que ocurre en otras técnicas como el análisis de varianza o el de regresión, en el análisis factorial todas las variables del análisis cumplen el mismo papel: todas ellas son independientes en el sentido de que no existe a priori una dependencia conceptual de unas variables sobre otras. En la figura 8 se ilustran los pasos necesarios para la realización de un Análisis Factorial.

[18] Formación y prevención de riesgos laborales: Bases para la adquisición de una cultura preventiva en los centros educativos. [Online]. [Visitada en Diciembre de 2007] Disponible en Internet:< http://hera.ugr.es/tesisugr/17116788.pdf >.
[19] El Análisis Factorial. [Online]. [Visitada en Diciembre de 2007] Disponible en Internet: URL: < www.ciberconta.unizar.es/.../.pdf >.

Figura 8. Pasos para la realización del análisis Factorial



Fuente: El Análisis Factorial. [Online]. [Visitada en Diciembre de 2007] Disponible en Internet: < www.ciberconta.unizar.es/.../.pdf >.

Una breve descripción de los pasos para la realizar el análisis factorial, mostrado en la figura 8, son 4:

- **Objetivos del Análisis Factorial:** Se basa en establecer los objetivos que se quieren alcanzar con el análisis factorial, identificación del tipo de datos mediante la reducción de datos, aplicación de otras técnicas de Análisis Factorial para finalmente hacer la selección de la nueva cantidad de variables.

- **Análisis de los factores:** Análisis de las relaciones entre los factores, y análisis del tamaño muestral.

- **Análisis de resultados:** Se procede a analizar los resultados evaluando las matrices tales como la matriz anti-imagen y contraste de esfericidad de Bartlett.

- **Análisis Posteriores:** Implementación de otras técnicas, como: Análisis de Regresión, análisis de clúster, análisis de componentes principales.

1.3.1.1 Análisis de Componentes Principales. El Análisis de Componentes Principales (ACP) es una etapa del Análisis Factorial, su objetivo, al igual que Análisis Factorial, es la síntesis de información ó reducción de la dimensión de los datos (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible[20].

Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iníciales (habrá que estudiar tanto el signo como la magnitud de las correlaciones). Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre el objeto de estudio.

ACP trata de hallar componentes (factores) que sucesivamente expliquen la mayor parte de la varianza total. En el Análisis de Componentes Principales, el primer factor o componente sería aquel que explica una mayor parte de la varianza total, el segundo factor sería aquel que explica la mayor parte de la varianza restante, es decir, de la que no explicaba el primero y así sucesivamente hasta encontrar

[20] Análisis Multivariados, Dr. Roberto Mercado Hernández. [Online]. [Visitada en Octubre de 2007] Disponible en Internet < www.feoc.ugto.mx/super/obtenga/D0003.ppt >.

una cantidad mínima de componentes que expliquen la mayor parte de la varianza total. De este modo sería posible obtener tantos componentes como variables originales, para obtener una explicación del cien por cien de la varianza total, aunque esto, en la práctica, no tiene sentido.

Por ejemplo, el investigador tiene una gran cantidad de datos sobre numerosas variables diferentes correlación entre ellas, este análisis nos ayudara a comprimir un poco los datos. Con la ayuda del análisis factorial, las variables de combinación pueden detectarse y especificarse, y el análisis también dice lo estrechamente que estos factores están vinculados con las variables originalmente verificadas. A veces se sitúa también una condición suplementaria sobre los factores, concretamente que no deben tener correlación alguna entre ellos y estén por lo tanto en "ángulo recto" uno con respecto a otro.

Objetivos del Análisis de Componentes Principales

- Reducir la mayor cantidad de datos posible de los datos originales

- Mediante la reducción de datos, reducir la magnitud del problema que se está estudiando.

- Descartar variables que aportan poca información.

Modelo Matemático

“Al estudiar un conjunto de n individuos mediante p-variables es posible encontrar nuevas variables denominadas Y(k), k=1,...,p que sean combinaciones lineales de

las variables originales X(j), e imponer a este sistema ciertas condiciones que permitan satisfacer los objetivos del análisis por componentes principales.”[21]

$$y(k) = \sum_{j=1}^{p} l(jk)\, x(j), \qquad k = 1,\ldots, p$$

Ecuación 1

1.3.1.1.1 Obtención de los componentes principales. Se debe tener en cuenta que con un numero $r < p$ de combinaciones lineales se intentara explicar la mayor parte de la varianza total.

- La primera componente principal será la combinación lineal **z**1 = **Xa**1 que tenga varianza máxima.

- Las siguientes componentes se definen de la misma forma, es decir, se intenta la máxima varianza con combinaciones lineales que no sean correlacionadas con las combinaciones previamente calculadas.

En este punto es donde se aprecia que para poder explicar el 100% de la varianza es necesario obtener una cantidad de CP[22] igual a la cantidad de variables originales.

[21] Uso del método de análisis de componentes principales para la caracterización de fincas agropecuarias. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet: < www.ceniap.gov.ve/.../htm >.

[22] CP, Componentes principales

1.3.2 Análisis de clúster. El análisis clústers es una técnica analítica para desarrollar subgrupos significativos de individuos u objetos. El objetivo es clasificar una muestra de entidades (personas u objetos) en un número pequeño de grupos donde sus entidades son homogéneas entre si pero, así mismo, los grupos son heterogéneos entre sí. En el análisis de clústers, a diferencia del análisis discriminante, los grupos no están predefinidos. Por consiguiente, se usa la técnica para identificar los grupos[23].

Habitualmente, el análisis de clústers implica al menos dos etapas:

- La primera es la medida de alguna forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad en la muestra.

- La segunda etapa es describir las personas o variables para determinar su composición. Este paso puede llevarse a cabo aplicando el análisis discriminante a los grupos identificados por la técnica de clústers.

Este análisis se subdivide en dos técnicas: (1) como técnica de *agrupación de variables* es similar al análisis factorial, la diferencia es que el A.C. [24] Es menos restrictiva, es decir, no exige linealidad, ni simetría, etc.) y admite varios métodos de estimación de la matriz de distancias, (2) como técnica de *agrupación de casos* es similar al A.D. [25], la diferencia es que el A.C detecta el numero optimo de grupos y su composición únicamente a partir de la similaridad existente entre los casos y no asume ninguna distribución especificas para las variables.

Es decir, la técnica de Clústers se encarga de identificar tipologías o grupos donde los elementos guardan una gran similitud entre sí y muchas diferencias con los de

[23] Principal – Tipos de técnicas Multivariantes. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet: < http://www.virtual.unal.edu.co/.../tiposmulti.htm >
[24] AC, Análisis de Clústers
[25] AD, Análisis Discriminante

otros grupos, esta basada en los atributos que describen los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar Minería de datos sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización.

Principalmente el Clustering o la segmentación es utilizada para hallar agrupaciones de elementos de acuerdo a características comunes, dividiendo los datos en grupos similares de objetos, donde Cada grupo es llamado cluster y contiene todos los objetos que son similares entre si y distintos a los objetos de otros grupos.

Su objetivo fundamental es segmentar a un grupo diverso en un conjunto de subgrupos o “cluster”. A diferencia de clasificación, Clustering no depende de clases predefinidas. Y es el primer paso en segmentación de mercado. Por ejemplo: un cluster particular de síntomas puede indicar una enfermedad particular.

Como se había dicho antes, el análisis de clústers es útil especialmente para realizar tareas de agrupación de datos tales como segmentación de datos con características comunes, donde la idea es el poder agrupar casos de acuerdo a atributos comunes, es decir, agrupación de datos sin tener una clasificación, ya que consiste en particionar un conjunto heterogéneo en subconjuntos homogéneos.

A continuación se presenta dos de los métodos básicos de Clustering, los cuales son: Clustering Particional y Clustering Ascendente Jerárquico.

- **Clústers Particional:** Su objetivo es obtener una partición de los objetos en grupos o clusters de tal forma que todos los objetos pertenezcan a alguno de los k clusters posibles y que por otra parte los clusters sean disjuntos.

- **Clústers Ascendente Jerárquico:** En el Clustering ascendente jerárquico se pretende ir agrupando en cada paso aquellos 2 objetos (o conglomerados) más cercanos, para de esta forma ir construyendo una estructura conocida como dendograma, la cual parte en su base de tantos conglomerados como objetos a clasificar, los cuales son agrupados finalmente en un único grupo conteniendo todos los objetos.

Si bien el costo computacional asociado a un Clustering ascendente jerárquico es superior al que se relaciona con un Clústers Particional, el dendograma que se obtiene con el primer método es mas rico que una simple partición, ya que posibilita la obtención de distintas particiones, simplemente variando el nivel de corte de dicha estructura.

En este tipo de análisis, a diferencia de otros, sólo de dispone de los valores de p variables X explicativas, para N sujetos, y el objetivo es agruparlos en K grupos ($K \leq N$), de tal manera que los individuos que pertenecen a un grupo se parezcan lo más posible entre sí con respecto a esas variables, y a su vez difieran lo máximo posible de los individuos de otros grupos. Este planteamiento es completamente diferente de la metodología estadística habitual ya que aquí no hay una hipótesis previa. Un posible ejemplo puede ser el buscar grupos de procesos médicos para valoración de costes, de tal manera que los grupos sean lo más homogéneos en cuanto a los recursos empleados.

Cuando se analizan sólo dos variables los datos son representables en unos ejes XY y de forma visual se puede intentar determinar una posible formación de grupos, por lo que una sencilla técnica a emplear es buscar, mediante algún

método de reducción de variables (por ejemplo análisis de componentes principales), obtener dos nuevas variables, función de las originales, que conserven una gran parte de la variabilidad original, y representarlas gráficamente para una inspección visual[26].

Algunas de las Aplicación de la técnica análisis de Clústers es en el área de la Medicina con la Identificación de enfermedades mediante determinados síntomas, Educación con la Identificación de áreas donde los estudiantes son destacados (Matemáticas, Lenguas, Artes) y en el Mercadeo: Identificación de tendencia de compras por parte de los clientes.

1.3.3 Análisis de regresión. El análisis de regresión es una de las técnicas mas usadas en el área de mercadeo. Establece la relación entre una única variable dependiente y varias variables independientes. Su formulación básica es la siguiente:

$$Y_1 = X_1 + X_2 + \ldots + X_n$$

Ecuación 2

El objetivo principal de esta técnica es usar las variables independientes, cuyos valores se conocen, para predecir una variable dependiente, cuyo valor se desconoce.

Para la aplicación de esta técnica se debe tener en cuenta aspectos como:

- Adecuación al problema que se pretende resolver
- Establecimiento de una relación estadística entre las variables independientes y las variables dependientes.

[26] Métodos Estadísticos de Clasificación. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet:< http://www.seh-lelha.org/clasifica.htm >.

- Selección de las variables dependientes e independientes a usar de acuerdo al objetivo inicial.
- Tamaño de la muestra seleccionada

El tamaño de la muestra seleccionada debe ser del tamaño adecuado dado que para una muestra pequeña (menor de 20) se llegará a una conclusión certera si hay una variable dependiente y, aun así, solo las relaciones profundamente marcadas serán detectadas; por el contrario, si se cuenta con una muestra grande (mayor a 1000) se podrá tener resultados mucho más acertados dado que el modelo se vuelve sumamente sensible provocando que muchas de las relaciones sean considerablemente acertadas[27].

1.3.4 Árboles de decisión. Representan reglas donde atributos independientes determinan los valores finales. En estos árboles cada nodo representa una propiedad que puede tomar diversos valores (atributos), cada uno de los cuales genera una rama (posibles valores del atributo). Los nodos hojas representan las clasificaciones finales (diferentes clases). Usadas donde se deben tomar decisiones a partir de varias alternativas combinadas y con pesos diferentes. Son útiles en problemas de alta dimensionalidad y pequeño número de valores para cada atributo.

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; por ejemplo, estas reglas pueden ser simplemente relaciones de igualdad entre un atributo y un valor, o relaciones de comparación.

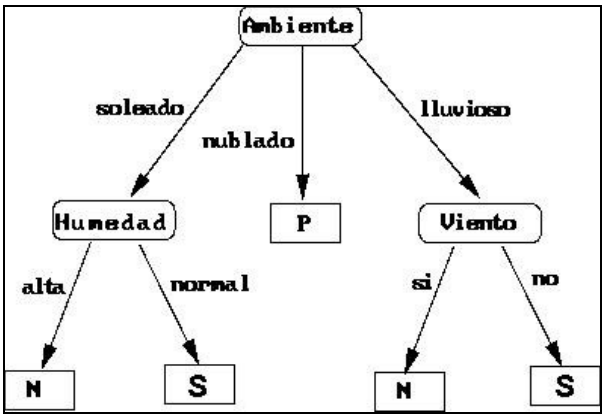
[27] El análisis de Regresión – Joaquín Aldás Manzano.

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 (o medida de desorden) para seleccionar el atributo que genera cada partición (cada nodo del árbol), seleccionando aquél con el que la reducción es máxima[28].

Lo que se quiere es que los árboles tengan la menor cantidad posible de desorden o incertidumbre para que se facilite, aun más, su manejo y entendimiento.

Uno de los ejemplos clásicos de arboles de decisión es la estimación de jugar golf dependiendo de las condiciones climatológicas del día (Ver figura 9).

Figura 9. Árbol de decisión para jugar golf



Fuente: Inducción de Árboles de Decisión. [Online]. [Visitada en Octubre de 2007].Disponible en Internet: < http://ccc.inaoep.mx/~emorales/Cursos/KDD03/node16.html >.

[28] Arboles de clasificación. [Online]. [Visitada en Octubre de 2007].Disponible en Internet: < http://math.uprm.edu/~edgar/treeDaza.html >.

1.3.4.1 Árbol C4.5. El Árbol C4.5, también conocido como J48, es el más usado y así mismo el estándar más reciente (1993). Éste realiza una búsqueda basado en reglas proporcionadas mediante un conjunto de ejemplos.

El conjunto de ejemplo debe estar conformado por tuplas o campos de valores, el objetivo a clasificar debe ser de tipo binario (positivo ó negativo, Si ó No, Valido ó Invalido, etc)

La elección del mejor atributo de los atributos proporcionados se realiza mediante la entropía.

1.3.4.2 Entropía. La entropía es una forma de establecer cuál es el atributo que más aporta al contexto a evaluar. En general, la ecuación final puede variar, sin embargo, en general va a ser como la *Ecuación:*

Si se tienen *p* ejemplos positivos y *n* ejemplos negativos, entonces la entropía es:

$$\mathrm{I}\left(\frac{p}{p+n},\frac{n}{p+n}\right)=-\frac{p}{p+n}\log_2\frac{p}{p+n}-\frac{n}{p+n}\log_2\frac{n}{p+n}$$

Ecuación 3

Donde *p* es el conjunto de los resultados positivos, *n* el de resultados negativos y *d* el total de ejemplos.

1.3.5 Reglas de asociación. Las reglas de asociación que, gracias a su facilidad de uso y sencilla interpretación, es una de las técnicas de minería de datos mas comúnmente utilizadas, incluso sin tener conocimiento previo de minería de datos;

Las reglas de asociación surgieron inicialmente para analizar las costumbres de compras de los clientes de los supermercados permitiendo establecer criterios de ordenamiento de la mercancía en la estantería.

Otro ejemplo práctico es que el 85% de las personas que realizan mantenimiento al automóvil también llenan el tanque de gasolina. Mediante una regla de asociación se puede leer de la forma: Mantenimiento → Gasolina (Mantenimiento entonces gasolina), Donde la parte de izquierda de la flecha se denomina antecedente y la parte derecha consecuente.

Uno de los principales inconvenientes que se tiene al momento de emplear las reglas de asociación es de acuerdo al tamaño de las bases de datos, se debe tener en cuenta que si existe una cantidad establecida de *m* ítem se debe tener una cantidad de 2^m posibles antecedentes; es por esto que es conveniente delimitar las bases de datos a tres restricciones principales:

- **Restricciones sintácticas:** Se puede limitar a que el ítem sea antecedente ó consecuente en determinada situación.

- **Restricciones de soporte:** Se puede limitar a que el ítem este en cierta cantidad de veces en una porción de la base de datos, es decir, que por cierta cantidad de datos cierto porcentaje sea de un ítem especifico.

- **Restricciones de cumplimiento:** Limitar a la posibilidad de ocurrencia de cierto ítem como antecedente o consecuente.

A pesar de las complicaciones que se puede encontrar haciendo uso de este modelo, es bastante practico hacer uso del mismo, debido a los excelentes resultados que arroja, siempre y cuando el ítem u objeto de estudio sea claro y se

cuente con algoritmos los cuales permitan hacer un buen análisis a bases de datos con gran cantidad de datos[29].

1.3.5.1 **Reglas de asociación A priori.** Un algoritmo de aprendizaje de reglas de asociación muy simple y popular es el algoritmo de ***A priori.*** Este algoritmo se basa en la búsqueda de conjuntos de ítems con determinada cobertura. Para lograr esto, se construyen conjuntos formados con un solo ítem que superan la cobertura mínima. Este conjunto de conjuntos se utiliza para construir el conjunto de conjuntos de dos ítems, y asi sucesivamente hasta que llegue a un tamaño en el que no existan conjuntos con la cobertura requerida. A continuación se muestra el algoritmo A priori[30]

Figura 10. Algoritmo a priori

```
Algoritmo A priori (D:Datos, MinC: Cobertura minima)
   i=0
   MIENTRAS Ci != 0
        PARA CADA x=elemento de Ci
        Ci+1 = Selecciona_Candidatos (Li)
        i=i+1
   FIN
   RETORNA C
FIN
```

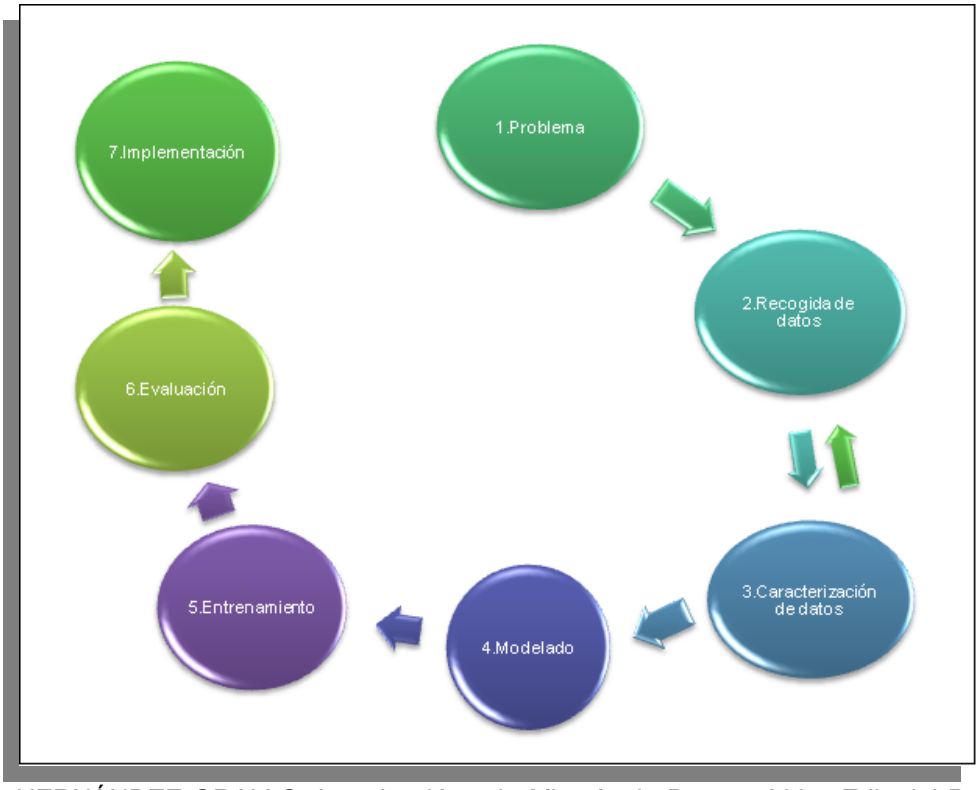
Fuente: Minería de uso de web para predicción de usuarios en la universidad [Online]. [Visitada en Octubre de 2007].Disponible en Internet: <sisbib.unmsm.edu.pe/BibVirtualData/publicaciones/risi/n3_2005/a01.pdf

[29] Reglas de asociación aplicadas a la detección de fraudes con tarjetas de crédito. [Online]. [Visitada en Octubre de 2007].Disponible en Internet:<www.ucu.edu.uy/.../.pdf >.

[30] HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson, p.

2. METODOLOGÍA

Figura 11. Metodologia CRISP DM



Fuente: HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 581

Los datos son el centro de este proceso ya que son la unidad fundamental en la aplicación de Minería de Datos. Las flechas indican las secuencias a seguir, en algunas etapas existe una continua iteración que permitirá encontrar la mejor solución (Ver Figura 11). La intervención del usuario es primordial ya que él será

el encargado de interpretar los resultados en cada una de las etapas, así como de analizar, concluir y hallar una solución con éxito.

2.1 PROBLEMA

En esta fase se plantea el problema el problema, entender los requerimientos y los objetivos de este proyecto.

En este caso el problema consiste en la implementación de un prototipo computacional que mediante una serie de pasos para la aplicación de técnicas de Minería de Datos en el análisis de información del atlas sísmico de Colombia.

2.2 COMPRESIÓN DE LOS DATOS

Esta fase inicia con la recolección y familiarización de los datos. Se debe identificar problemas de calidad y descubrir las primeras potencialidades en los datos o detectar subconjuntos de estos datos. Esta fase se divide en 4 subfases:

- Se deben recopilar los datos.
- Descripción de los datos.
- Exploración de los datos.
- Verificación de calidad de los datos. [31]

El usuario debe extraer de los datos que se van a usar en todo el proceso, esto se denomina datawarehouse[32] (DW). El DW se puede ver como una bodega donde

[31] HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 581

están almacenados todos los datos necesarios para realizar la predicción de exploración y explotación de hidrocarburos. Esta extracción puede ser apoyada por un ingeniero de petróleo y un geólogo que estén encargados la base de datos. Para este proyecto, se determino trabajar con el siguiente datawarehouse (Ver Tabla 3)

Tabla 3. Descripción de las variables

VARIABLE	DESCRIPCIÓN	DEPENDIENTE INDEPENDIENTE
EF	Electrofase	Independiente
LT	Litofase	Independiente
HFU	Unidad de flujo Hidráulico	Independiente
GR	Gamma Ray	Independiente
Log10(LLD)	Profundidad de perfil lateral	Independiente
Log10(MSFL)	Microesférico	Independiente
DT	Tiempo de transito	Independiente
RHOB	Densidad	Independiente
NPHI	Porosidad entre los pozos	Independiente
PEF	Potencial Fotoeléctrico	Independiente
KMAX	Permeabilidad	Independiente

[32] WOLFF Carmen Gloria. La Tecnología Datawarehousing. 1999 [online, Articulo]. [Citado el 27 de agosto 2006]. Disponible en Internet: <http://www.inf.udec.cl/revista/ediciones/edicion3/cwolff.PDF> p. 2.

	Máxima	
K90	Permeabilidad de 90%	Independiente
KVERT	Permeabilidad vertical	Independiente
POR	Porosidad de laboratorio	Dependiente

F**uente:** Ing. Hermes Martínez

2.3 CARACTERIZACIÓN DE LOS DATOS

Esta fase inicia con la Selección de un conjunto de rasgos y valores que caracterizan los datos según su integración, selección, limpieza y transformación. Se debe tener en cuenta la normalidad, homocedasticidad (se refiere al supuesto de que las variables dependientes exhiban iguales niveles de varianza a lo largo del rango de predicción de las variables)

Linealidad (la ausencia de linealidad provoca que el coeficiente de correlación no mida adecuadamente la relación entre los pares de variables) de los datos. Selección de datos (razones de inclusión / exclusión), Limpieza de datos Construcción de datos (atributos derivados, registros generados), Integración de datos (datos mezclados) y formateo de datos. Los datos puede ser clasificados en:

Escalas nominales o categóricas. En este caso, los números se comportan como etiquetas. En esta escala la variable Porosidad (POR) ya que su valor está dado por (Aceptable, Bueno, Excelente).

Escalas ordinales o numéricas. No sólo consiguen distinguir entre valores, sino que además establece un orden entre ellas. Clasifican las variables independientes que tengan valor numérico. Estas clasificaciones se pueden agrupar de la siguiente manera **escalas métricas** (escalas de intervalo y razón)**, escalas no métricas** (escalas nominales y ordinales). La caracterización de las variables se puede observar en la Tabla 4.

Tabla 4. Caracterización de las variables

VARIABLE	CLASIFICACIÓN	ESCALA
EF	Independiente	Métrica
@LT	Independiente	Métrica
HFU	Independiente	Métrica
GR	Independiente	Métrica
Log10(LLD)	Independiente	Métrica
Log10(MSFL)	Independiente	Métrica
DT	Independiente	Métrica
RHOB	Independiente	Métrica
NPHI	Independiente	Métrica
PEF	Independiente	Métrica
KMAX	Independiente	Métrica
K90	Independiente	Métrica
KVERT	Independiente	Métrica
POR	Dependiente	Métrica / Nominal

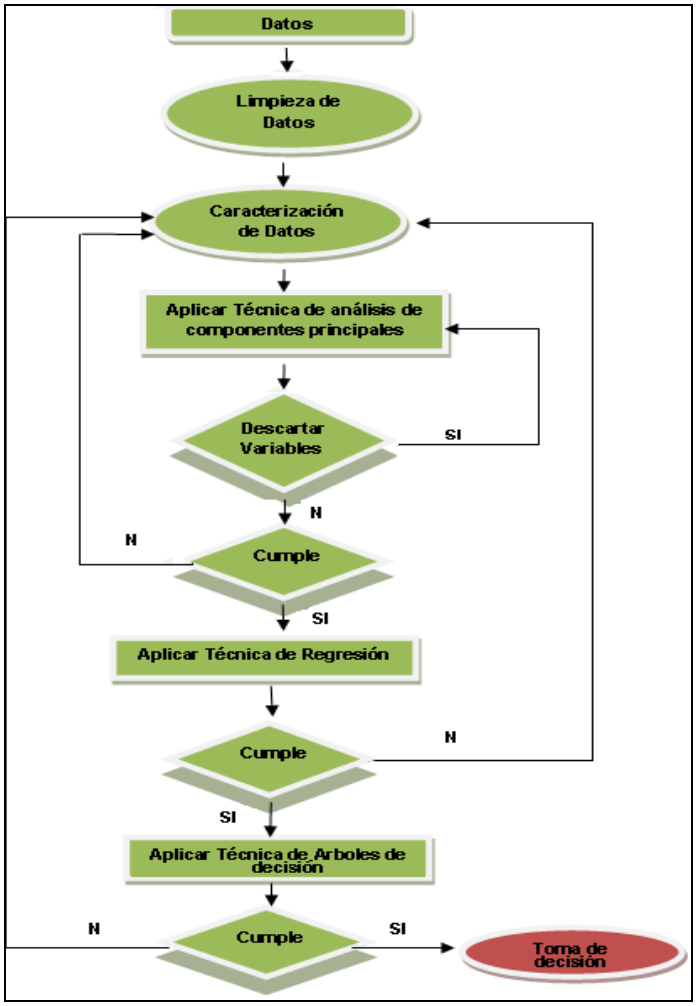
Fuente: Experto Ingeniero de Petróleo

2.4 MODELADO

Esta Fase se Inicia con la extracción de datos desde la base de datos con las variables independientes (GR ,Log10(LLD) ,Log10(MSFL) ,DT , RHOB , NPHI ,PEF KMAX, K90, KVERT, EF, LT, HFU) y dependiente (POR). De las variables dependientes se busca predecir el mejor rendimiento, a partir de los atributos de las variables independientes a través de los modelos que se presentan a continuación y su respectiva evaluación. Se a ensayado el orden de aplicación de las técnicas de minería de datos llegando a la conclusión de los modelo A (Ver figura 12) y del modelo B (Ver figura 13).

2.4.1 ***Modelo A.*** La primera técnica a aplicar es el Análisis de Componentes Principales, porque ayuda a reducir variables; sólo se podrá continuar cuando se cumplan todos los criterios de las técnicas. La segunda técnica es la de Análisis de Regresión para conocer el valor de la variable dependiente con respecto a las independientes. Tercero, el Árbol de Decisión para saber en qué condiciones la variable Porosidad (POR) está en el rango de Aceptable (A), Buena (B) y Excelente (C). Finalmente se toma una decisión. (Ver Figura 12)

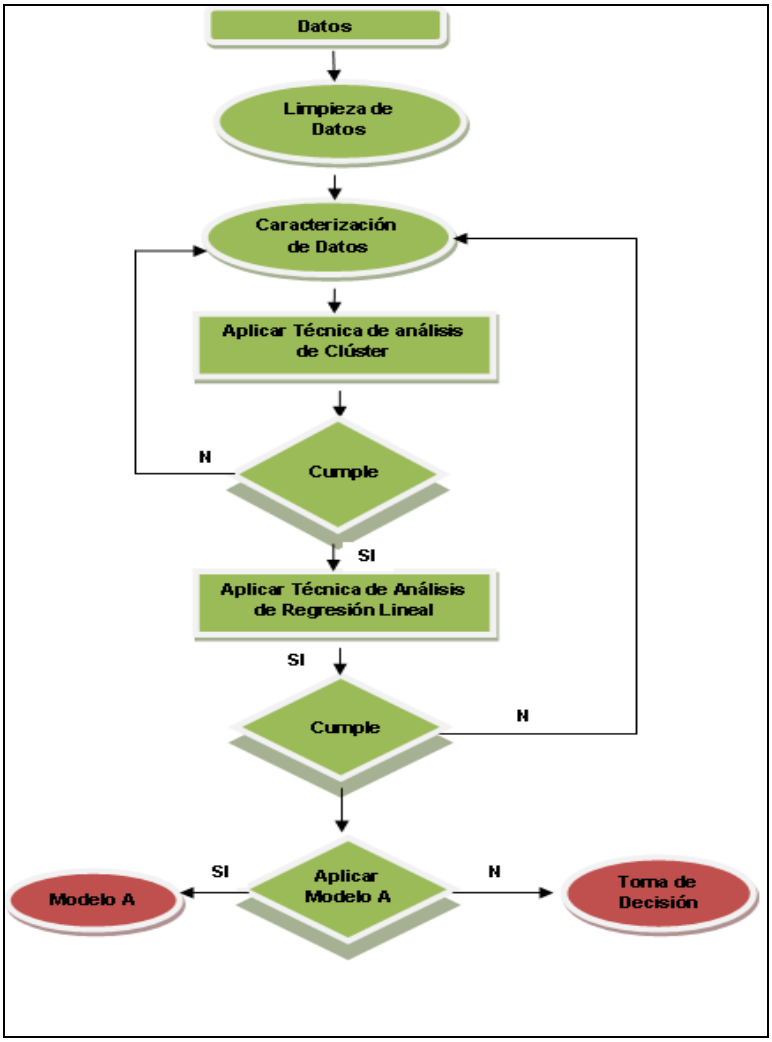
Figura 12. Modelo A



Fuente: Autores de Proyecto.

2.4.2 *Modelo B.*

Figura 13. Modelo B



Fuente: Autores de Proyecto.

La primera técnica a aplicar es Análisis de Clusters porqué indica como se deben agrupar las variables por su homogeneidad dentro del grupo, y por heterogeneidad

con los demás Clusters. Posteriormente se aplica la técnica de Análisis de Regresión, el usuario puede concluir o continuar con el modelo A de la figura 12.

2.5 ENTRENAMIENTO

2.5.1 ***Entrenamiento de los Datos.*** En esta etapa se explica los resultados obtenidos durante la aplicación de las técnicas de minería de datos para que el usuario interprete esos resultados y pueda tomar sus decisiones.

Tabla 5. Descripción de los datos

Variable	Descripción	Característica
EF	Electrofase	Independiente
LT	Litofase	Independiente
HFU	Unidad de flujo Hidráulico	Independiente
GR	Gamma Ray	Independiente
Log10(LLD)	Profundidad lateral	Independiente
Log10(MSFL)	Microesférico	Independiente
DT	Tiempo de transito	Independiente
RHOB	Densidad	Independiente
NPHI	Porosidad entre los pozos	Independiente
PEF	Potencial Fotoeléctrico	Independiente
KMAX	Permeabilidad Máxima	Independiente
K90	Permeabilidad de 90%	Independiente
KVERT	Permeabilidad vertical	Independiente

POR	Porosidad de laboratorio	Dependiente

Fuente: Ing. Hermes Martínez

2.5.2 Resultados de la técnica análisis de regresión. Esta técnica estadística es utilizada para estudiar la relación entre una sola variable dependiente y varias variables independientes. EL principal objetivo de esta técnica es usar las variables independientes, cuyos valores se conocen, para predecir el de la variable dependiente. Las variables que se usaron para explicar los resultados son

Figura 14. Resultados del análisis de regresión.

Linear Regression Model

POR = -0.1113 * EF + 0.1099 * @LT + -0.0389 * GR + 0.1806 * DT + -37.7053 * RHOB + 4.4266 * NPHI + 0.0004 * KMAX + 0.0065 * K90 + -0.0022 * KVERT + 94.7525

"POR" = 94.752

=== Summary ===

Correlation coefficient	**0.9098**
Mean absolute error	**1.3247**
Root mean squared error	**1.8293**
Relative absolute error	**37.6687 %**
Root relative squared error	**41.4964 %**

Total Number of Instances	**904**

Fuente: SATMD

Como se puede observar el coeficiente de correlación (Correlation coefficient) de 0.9098, este resultado es muy bueno ya que indica la fuerza de la relación entre las variables independientes y la variable dependiente. El valor del coeficiente de correlación varía en el intervalo [-1, +1]:

Si coeficiente de correlación = 0, no existe ninguna correlación. El coeficiente indica, por tanto, una independencia total entre las variables, es decir, que la variación de una de ellas no influye en absoluto en el valor que pueda tomar las otras.

Si coeficiente de correlación = 1, existe una correlación positiva perfecta. El coeficiente indica una dependencia total entre las variables denominada relación directa: cuando una de ellas aumenta, las otras también lo hace en idéntica proporción.

Si coeficiente de correlación > 0 y < 1, existe una correlación positiva.

Si coeficiente de correlación = -1, existe una correlación negativa perfecta. El coeficiente indica una dependencia total entre las variables llamada relación inversa: cuando una de ellas aumenta, las otras disminuye en idéntica proporción.

Si coeficiente de correlación >0 y <-1, existe una correlación negativa.

La medida del error absoluto (Mean absolute error) es el promedio de la magnitud de los errores individuales sin tomar en cuenta el signo. O es igual a la imprecisión que acompaña a la medida.

(Root mean squared error) es la medida de las diferencias en promedio de los valores pronosticados y los observados.

2.5.3 Resultados de la técnica de Componentes Principales. Esta técnica es utilizada para reducir la cantidad de variables sin perder la mayor cantidad de información posible.

Figura 15. Matriz de correlación

Correlation matrix

1	0.27	0.02	-0.3	-0.34	-0.46	0.52	-0.53	0.58	-0.17	-0.03	0.12	0.05
0.27	1	0	-0.06	-0.25	-0.27	0.39	-0.52	0.42	-0.02	-0.01	0.2	0.03
0.02	0	1	-0.07	0.04	-0.01	-0.01	-0.01	0.02	0.06	-0.01	-0.01	-0.04
-0.3	-0.06	-0.07	1	0	0.12	-0.24	0.2	-0.23	-0.26	0.08	-0.16	-0.01
-0.34	-0.25	0.04	0	1	0.69	-0.58	0.54	-0.57	0.28	0.05	-0.01	-0.01
-0.46	-0.27	-0.01	0.12	0.69	1	-0.64	0.59	-0.65	0.2	-0.02	-0.13	-0.03
0.52	0.39	-0.01	-0.24	-0.58	-0.64	1	-0.86	0.87	-0.01	-0.02	0.34	0.04
-0.53	-0.52	-0.01	0.2	0.54	0.59	-0.86	1	-0.89	0.09	0.01	-0.33	-0.06

	0.58	0.42	0.02	-0.23	-0.57	-0.65	0.87	-0.89	1	-0.11	0
0.29	0.06										
	-0.17	-0.02	0.06	-0.26	0.28	0.2	-0.01	0.09	-0.11	1	-0.01
0.23	-0.01										
	0.03	-0.01	-0.01	0.08	0.05	-0.02	-0.02	0.01	0	-0.01	1
0.05	0.24										
	0.12	0.2	-0.01	-0.16	-0.01	-0.13	0.34	-0.33	0.29	0.23	0.05	1
0.04											

Fuente: SATMD

Matriz de correlaciones: Es la matriz que indica las correlaciones entre las variables, por ejemplo hay correlación entre dos variables cuando éstas cambian de tal modo que los valores que toma una de ellas son, hasta cierto punto, predecibles a partir de los que toma la otra. Los números sombreados expresan correlación significativa entre las variables.

Tabla 6. Eigenvalores

Eigenvalue	Proportion	Cumulative	
4.55614	**0.35047**	**0.35047**	0.313EF+0.247@LT+0.003HFU-0.127GR-0.328log10LLD-0.365log10MSFL+0.426DT-0.429RHOB+0.434NPHI-0.068PEF-0.008KMAX+0.158K90+0.033KVERT
1.57199	**0.12092**	**0.47139**	-0.004EF+0.081@LT+0.137HFU-0.461GR+0.334log10LLD+0.195log10MSFL+0.083DT-0.057RHOB+0.03 NPHI+0.626PEF-0.009KMAX+0.456K90+0.02 KVERT
1.27043	**0.09773**	**0.56912**	-0.068EF+0.054@LT-0.184HFU+0.17 GR+0.06 log10LLD+0.02 log10MSFL-0.002DT-0.03RHOB+0.006NPHI-0.012PEF+0.685KMAX+0.161K90+0.655KVERT
1.02617	**0.07894**	**0.64806**	-0.269EF+0.266@LT-0.703HFU+0.356GR-0.001log10LLD+0.084log10MSFL+0.057DT-0.081RHOB-0.009NPHI+0.094PEF-0.178KMAX+0.346K90-0.242KVERT
0.94373	0.07259	0.72065	0.279EF-0.351@LT-0.653HFU-0.506GR-0.017log10LLD-0.037log10MSFL-0.012DT+0.105RHOB-0.017NPHI-0.007PEF-0.134KMAX-0.208K90+0.205KVERT
0.79876	0.06144	0.78209	-0.224EF-0.646@LT-0.01HFU+0.034GR-0.25log10LLD-0.283log10MSFL+0.136DT+0.062RHOB+0.034NPHI+0.189PEF+0.375KMAX+0.2 K90-0.385KVERT
0.72154	0.0555	0.8376	0.34 EF+0.257@LT-0.13HFU-0.176GR+0.267log10LLD+0.136log10MSFL-0.086DT-0.01RHOB+0.001NPHI-0.216PEF+0.549KMAX-0.089K90-0.563KVERT

0.67182	0.05168	0.88927	-0.303EF+0.352@LT-0.082HFU-0.109GR-0.245log10LLD-0.142log10MSFL+0.019DT-0.005RHOB-0.028NPHI+0.518PEF+0.176KMAX-0.618K90-0.052KVERT
0.49719	0.03825	0.92752	-0.364EF+0.295@LT+0.032HFU-0.498GR-0.294log10LLD-0.258log10MSFL-0.23DT+0.194RHOB-0.246NPHI-0.351PEF+0.03 KMAX+0.322K90+0.005KVERT
0.4323	0.03325	0.96077	-0.567EF-0.145@LT-0.003HFU-0.258GR+0.168log10LLD+0.388log10MSFL+0.272DT-0.334RHOB+0.289NPHI-0.329PEF+0.024KMAX-0.185K90-0.011KVERT

Fuente: SATMD

Eigenvalores: Son los valores explicados de los factores. Estos determinan cuales factores se deben tomar para el análisis de componentes, dependiendo de la varianza acumulada, se debe tomar en consideración los que sean mayores que 1 para el ejemplo se toman los primeros 4, y con estos eigenvalores se explica el 64,806% de la varianza.

El criterio utilizado para seleccionar las variables de los componentes según los vectores característicos es el que aparece en la tabla 7 de carga factorial.

Tabla 7. Carga Factorial según tamaño de la Muestra

Tamaño muestral	Carga factorial
350	0.30
250	**0.35**
200	0.40
150	0.45
120	0.50
100	0.55
85	0.60
70	0.65
60	0.70
50	0.75

Fuente: Autores del Proyecto

Para nuestro caso empleamos un tamaño muestral superior a 300, por lo que tomamos la carga factorial de 0.30 como criterio de selección.

Tabla 8. Eigenvectores

V1	V2	V3	V4	V5	V6	V7	V8	V9	V
0.3132	-0.0036	-0.0678	-0.2691	0.279	-0.2237	0.3402	-0.3026	-0.3636	-0.5
0.247	0.0813	0.054	0.266	-0.3506	-0.6463	0.2566	0.3525	0.295	-0.1
0.0031	0.1375	-0.1835	-0.703	-0.6532	-0.0101	-0.1297	-0.0815	0.0322	-0.0
-0.127	-0.4613	0.1698	0.3565	-0.5061	0.034	-0.1758	-0.1088	-0.4977	-0.2
-0.328	0.3344	0.0597	-0.0005	-0.0167	-0.2497	0.2675	-0.2451	-0.2942	0.1
-0.3646	0.1946	0.0205	0.0839	-0.0366	-0.283	0.1361	-0.1424	-0.2581	0.3
0.4256	0.0826	-0.0022	0.0571	-0.0118	0.1361	-0.086	0.0185	-0.2295	0.2
-0.4287	-0.057	-0.0301	-0.0812	0.105	0.0623	-0.0096	-0.0047	0.194	-0.3
0.4343	0.0304	0.0065	-0.0094	-0.0171	0.0343	0.001	-0.028	-0.2459	0.2
-0.0682	0.6265	-0.0116	0.0942	-0.007	0.1889	-0.2162	0.518	-0.3512	-0.3
-0.0081	-0.0087	0.6855	-0.1776	-0.1345	0.3752	0.5493	0.1763	0.0301	0.0
0.1576	0.4562	0.1606	0.3456	-0.2079	0.2003	-0.0894	-0.6181	0.3223	-0.1
0.0332	0.0202	0.6552	-0.2424	0.2051	-0.3853	-0.5627	-0.0518	0.0049	-0.0

Fuente: SATMD

Los valores que conforman el eigenvector explican a la variable en cada uno de ellos; para el ejemplo anterior los eigenvectores que se tienen en cuenta son los cuatro primeros, y estos contienen una lista de valores para cada variable, donde los valores mayores a 0.3 significa que esta mejor representada en el eigenvector correspondiente, por ejemplo que la variable RHOB esta explicada en V1, porque su valor corresponde a -0.4256, mientras que en el V2 corresponde a 0.0826, en el v3 corresponde a -0.0301, y en el v4 corresponde -0.0812, como se indica a continuación:

Tabla 9. Eigenvectores clasificados

V1	V2	V3	V4	VAR
0.3132	-0.0036	-0.0678	-0.2691	EF
0.247	0.0813	0.054	0.266	@LT
0.0031	0.1375	-0.1835	**-0.703**	HFU
-0.127	**-0.4613**	0.1698	0.3565	GR
-0.328	**0.3344**	0.0597	-0.0005	log10LLD
-0.3646	0.1946	0.0205	0.0839	log10MSFL
0.4256	0.0826	-0.0022	0.0571	DT
-0.4287	-0.057	-0.0301	-0.0812	RHOB
0.4343	0.0304	0.0065	-0.0094	NPHI
-0.0682	**0.6265**	-0.0116	0.0942	PEF
-0.0081	-0.0087	**0.6855**	-0.1776	KMAX
0.1576	**0.4562**	0.1606	0.3456	K90
0.0332	0.0202	**0.6552**	-0.2424	KVERT

Fuente: SATMD

De los resultados arrojados por el prototipo SATMD se puede concluir lo siguiente:

El Componente Principal (V1) tiene la mayor correlación positiva con las variables EF, DT, NPHI, mientras que tiene correlación negativa con RHOB, Log 10MSFL y casi nula con el resto de variables.

El Componente Principal (V2) tiene la mayor correlación positiva con las variables Log10LLD, PEF, K90, mientras que tiene correlación negativa con GR y casi nula con el resto de variables.

El Componente Principal (V3) tiene la mayor correlación positiva con las variables KMAX, KVERT y casi nula con el resto de variables.

El Componente Principal (V4) tiene la mayor correlación negativa con HFU, casi nula con el resto de variables.

El coeficiente de correlación negativa indica una dependencia entre las variables llamada relación inversa; cuando una de ellas aumenta, las otras disminuye en idéntica proporción.

2.5.4 Resultados técnica de Clústers. El algoritmo que se va usar para esta técnica es el simplekmeans, los resultados que arroja son los siguientes (Ver Figura 16): Los centroides de los cluster, las instancias que se usaron en cada uno de ellos y la comparación de las medias así como cuáles variables hacen parte de un clúster

Figura 16. Resultados Técnica de Clústers

```
Means
Number of iterations: 4
Within cluster sum of squared errors: 611.2873733832989
Cluster centroids:
Cluster 0
```

Mean/Mode: 4.1568 6.3464 5.0435 22.3358 19.4626 17.9488
56.5114 2.5870 0.0716 6.0498 33.9157 4.8716 6.0458
Std Devs: 2.0785 1.8957 0.7778 11.4754 4.2018 4.7869
4.5774 0.0761 0.0441 2.0583 469.5473 20.7804 45.0515
Cluster 1
Mean/Mode: 3.7550 5.9932 1.7168 24.3732 20.0607 19.0961
55.1969 2.6108 0.0569 5.9490 34.1968 4.4204 9.7141
Std Devs: 2.0546 1.9982 0.6849 12.3019 4.9943 5.8313
4.9010 0.0764 0.0450 1.9685 483.4010 21.7899 132.3213

Clustered Instances
0 459 (51%)
1 445 (49%)

Fuente: SATMD

De un total de 904 registros, el cluster0 tiene 459 y el cluster1 445. Es decir, se agruparon correctamente 904 registros.

Tabla 11. Comparación de medias de clústers

Variables	Cluster 0	Cluster 1
"EF"	**4,1568**	3,7550
"@LT"	**6,3464**	5,9932
"HFU"	**5,0435**	1,7168
"GR"	22,3358	**24,3732**
"LOG10LLD"	19,4626	**20,0607**
"LOG10MSFL"	17,9488	**19,0961**

"DT"	**56,5114**	55,1969
"RHOB"	2,5870	**2,6108**
"NPHI"	**0,0716**	0,0569
"PEF"	**6,0498**	5,9490
"KMAX"	33,9157	**34,1968**
"K90"	**4,8716**	4,4204
"KVERT"	6,0458	**9,7140**

Fuente: SATMD

Según la Comparación de medias que arrojan los resultados las variables que sea agrupan en el Cluster 0 son EF, @LT, HFU, DT, NPHI, PEF, K90; y las variables que se agrupan en el Cluster 1 son GR, log10LLD, log10MSFL, RHOB, KMAX, KVERT.

2.5.5 Resultados de la técnica árboles de decisión. Esta técnica recibe como entrada una situación descrita por un conjunto de atributos y los clasifica respecto a una clase dependiendo del grado de entropía, que es el nivel de desorden en los valores del atributo y la ganancia, correspondiente a la cantidad de información que se gana al seleccionar un atributo, dando como resultado una estructura de clasificación en forma de árbol, donde las ramas están etiquetadas con los posibles valores de la prueba y las hojas representan los valores de la clase, el sumario del Algoritmo J48 y la Matriz de Confusión. Esto se puede observar en la Figura 17, y a continuación su respectiva explicación.

Figura 17. Resultados Técnica de Árboles de decisión

```
J48 pruned tree
------------------

RHOB <= 2.58
|   RHOB <= 2.47
|   |   NPHI <= 0.14
|   |   |   K90 <= 2.66: A (3.0)
|   |   |   K90 > 2.66
|   |   |   |   DT <= 62.2
|   |   |   |   |   EF <= 4
|   |   |   |   |   |   log10LLD <= 15.804: B (2.0)
|   |   |   |   |   |   log10LLD > 15.804: E (2.0)
|   |   |   |   |   EF > 4: B (5.0)
|   |   |   |   DT > 62.2: E (5.0)
|   |   NPHI > 0.14
|   |   |   KMAX <= 1.82: B (3.0)
|   |   |   KMAX > 1.82
|   |   |   |   RHOB <= 2.44: E (43.0)
|   |   |   |   RHOB > 2.44
|   |   |   |   |   GR <= 10.94: B (3.0)
|   |   |   |   |   GR > 10.94: E (13.0/2.0)
|   RHOB > 2.47
|   |   DT <= 55.5
|   |   |   K90 <= 0.34: A (11.0)
|   |   |   K90 > 0.34
|   |   |   |   @LT <= 5: A (3.0)
|   |   |   |   @LT > 5
```

```
|   |   |   |   PEF <= 5.03
|   |   |   |   |   log10LLD <= 15.946: B (2.0)
|   |   |   |   |   log10LLD > 15.946: A (7.0)
|   |   |   |   PEF > 5.03: B (8.0)
|   |   DT > 55.5
|   |   |   K90 <= 0.15
|   |   |   |   PEF <= 7.23: A (8.0)
|   |   |   |   PEF > 7.23: B (6.0/1.0)
|   |   |   K90 > 0.15
|   |   |   |   KVERT <= 3.24: B (166.0/13.0)
|   |   |   |   KVERT > 3.24
|   |   |   |   |   RHOB <= 2.52
|   |   |   |   |   |   log10MSFL <= 14.349: E (5.0/1.0)
|   |   |   |   |   |   log10MSFL > 14.349: B (5.0/1.0)
|   |   |   |   |   RHOB > 2.52: B (18.0/2.0)
RHOB > 2.58
|   DT <= 56.69: A (526.0/40.0)
|   DT > 56.69
|   |   RHOB <= 2.62
|   |   |   HFU <= 5
|   |   |   |   KMAX <= 1.24: A (24.0/8.0)
|   |   |   |   KMAX > 1.24: B (16.0/3.0)
|   |   |   HFU > 5: B (5.0)
|   |   RHOB > 2.62: A (15.0)

------------------------------------------------------------------------------
```

Fuente: SATMD

Figura 18. Resultados Técnica de Árboles de decisión

```
Number of Leaves  :         25

Size of the tree :      49


=== Summary ===

Correctly Classified Instances         833               92.146  %
Incorrectly Classified Instances        71               7.854  %
Kappa statistic                          0.8436
Mean absolute error                      0.0926
Root mean squared error                  0.2152
Root relative squared error             52.0113 %
Total Number of Instances              904

=== Confusion Matrix ===

   a   b   c   <-- classified as
 549  16   1 |   a = A
  48 219   2 |   b = B
   0   4  65 |   c = E
```

Fuente: SATMD

El número de niveles de hojas (ramas sin hijos) es de 25 y el número de ramas del árbol es de 49.Se muestra el número de instancias clasificadas correctamente, 833 y 71 incorrectamente clasificadas. El Estadístico de Kappa dice

Tabla 12. Estadístico Kappa

Kappa	Grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

Fuente: Índices de concordancia. [Online]. [Visitada en Octubre de 2007].Disponible en Internet: < http://www.hrc.es/bioest/errores_2.html >

En el ejemplo se tienen un grado de aceptación Casi Perfecto ya que fue de 0.8436 y los errores son prácticamente muy bajos.

Finalmente se tiene la Matriz de Confusión que es una matriz que contiene los números que reflejan la capacidad predictiva de la función discriminante. Se crea tabulando de forma cruzada el miembro del grupo concreto con el miembro del grupo predicho, los número de la diagonal representa clasificaciones correctas y los números fuera de la diagonal son clasificados incorrectamente.

Se concluye que para la categoría Aceptable se han clasificado 549 instancias correctamente en Aceptable, 16 instancia en la categoría Bueno y 1 en la categoría Excelente. Para la categoría Bueno se ha clasificado 219 instancias correctamente en la categoría Bueno, 48 en la categoría Aceptable y 2 en la categoría Excelente. Para la categoría Excelente se ha clasificado 65 instancias correctamente en la categoría Excelente, 4 en la categoría Bueno y 0 en la categoría aceptable.

También se puede deducir del árbol de decisión las siguiente 25 reglas de decisión que van ayudar a la toma de decisiones con respecto a predecir la variable Porosidad

2.5.5.1 Reglas de decisión. Las regla de decisión se generan por la interpretación a los resultados de la técnica de arboles de decisión donde las ramas están etiquetadas con los posibles valores de la prueba y las hojas representan los valores de la clase.

Regla 1: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= de 0.14 **Y** K90 <= de 2.66 **ENTONCES** POR (porosidad) = A (aceptable).

Regla 2: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= de 0.14, K90 > de 2.66, DT <= de 62.2, EF<= de 4 **Y** log10LLD <= de 15.804 **ENTONCES** POR (porosidad) = B (Buena).

Regla 3: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= de 0.14, K90 > de 2.66, DT <= de 62.2, EF<= de 4 **Y** log10LLD > de 15.804 **ENTONCES** POR (porosidad) = E (Excelente).

Regla 4: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= de 0.14, K90 >de 2.66, DT <= de 62.2 **Y** EF >de 4 **ENTONCES** POR (porosidad) = B (Buena).

Regla 5: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= 0.14, K90 > 2.66 **Y** DT > de 62.2 **ENTONCES** POR (porosidad) = E (Excelente).

Regla 6: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI > 0.14 **Y** KMAX <= de 1.82 **ENTONCES** POR (porosidad) = B (Buena).

Regla 7: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI > de 0.14, KMAX > de 1.82 **Y** RHBO <= de 2.44 **ENTONCES** POR (porosidad) = E (Excelente).

Regla 8: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI > de 0.14, KMAX > de 1.82, RHBO > de 2.44 **Y** GR<= de 10.94 **ENTONCES** POR (porosidad) = B (Buena).

Regla 9: SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI > de 0.14, KMAX > de 1.82, RHBO > de 2.44 **Y** GR > de 10.94 **ENTONCES** POR (porosidad) = E (Excelente).

Regla 10: SI RHBO <= de 2.58, RHBO > de 2.47, DT <= de 55.5 **Y** K90 <= de 0.34 **ENTONCES** POR (porosidad) = A (aceptable).

Regla 11: SI RHBO <= de 2.58, RHBO > de 2.47, DT <= de 55.5, K90 > de 0.34 **Y** @LT < = de 5 **ENTONCES** POR (porosidad) = A (aceptable).

Regla 12: SI RHBO <= de 2.58, RHBO > de 2.47, DT <= de 55.5, K90 > de 0.34, @LT > de 5, PEF <= de 5.03, log10LLD <= de 15.946 **ENTONCES** POR (porosidad) = B (Buena).

Regla 13: SI RHBO <= de 2.58, RHBO > de 2.47, DT <= de 55.5, K90 > de 0.34, @LT > de 5, PEF <= de 5.03, log10LLD > de 15.946 **ENTONCES** POR (porosidad) = A (aceptable).

Regla 14: SI RHBO <= de 2.58, RHBO > de 2.47, DT <= de 55.5, K90 > de 0.34, @LT > de 5 **Y** PEF > de 5.03 **ENTONCES** POR (porosidad) = B (Buena).

Regla 15: SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 =< de 0.15 **Y** PEF <= de 7.23 **ENTONCES** POR (porosidad) = A (aceptable).

Regla 16: SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 =< de 0.15 **Y** PEF > de 7.23 **ENTONCES** POR (porosidad) = B (Buena).

Regla 17: SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 > de 0.15 **Y** KVERT <= de 3.24 **ENTONCES** POR (porosidad) = B (Buena).

Regla 18: SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 > de 0.15, KVERT > de 3.24, RHOB <= de 2.52 **Y** log10MSFL <= de 14.349 **ENTONCES** POR (porosidad) = E (Excelente).

Regla 19: SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 > de 0.15, KVERT > de 3.24, RHOB <= de 2.52 **Y** log10MSFL > de 14.349 **ENTONCES** POR (porosidad) = B (Buena).

Regla 20: SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 > de 0.15, KVERT > de 3.24 **Y** RHOB > de 2.52 **ENTONCES** POR (porosidad) = B (Buena).

Regla 21: SI RHBO > de 2.58 **Y** DT <= de 56.69 **ENTONCES** POR (porosidad) = A (aceptable).

Regla 22: SI RHBO > de 2.58, DT > de 56.69, RHBO <= de 2.62, HFU <= de 5 **Y** KMAX <= de 1.24 **ENTONCES** POR (porosidad) = A (aceptable).

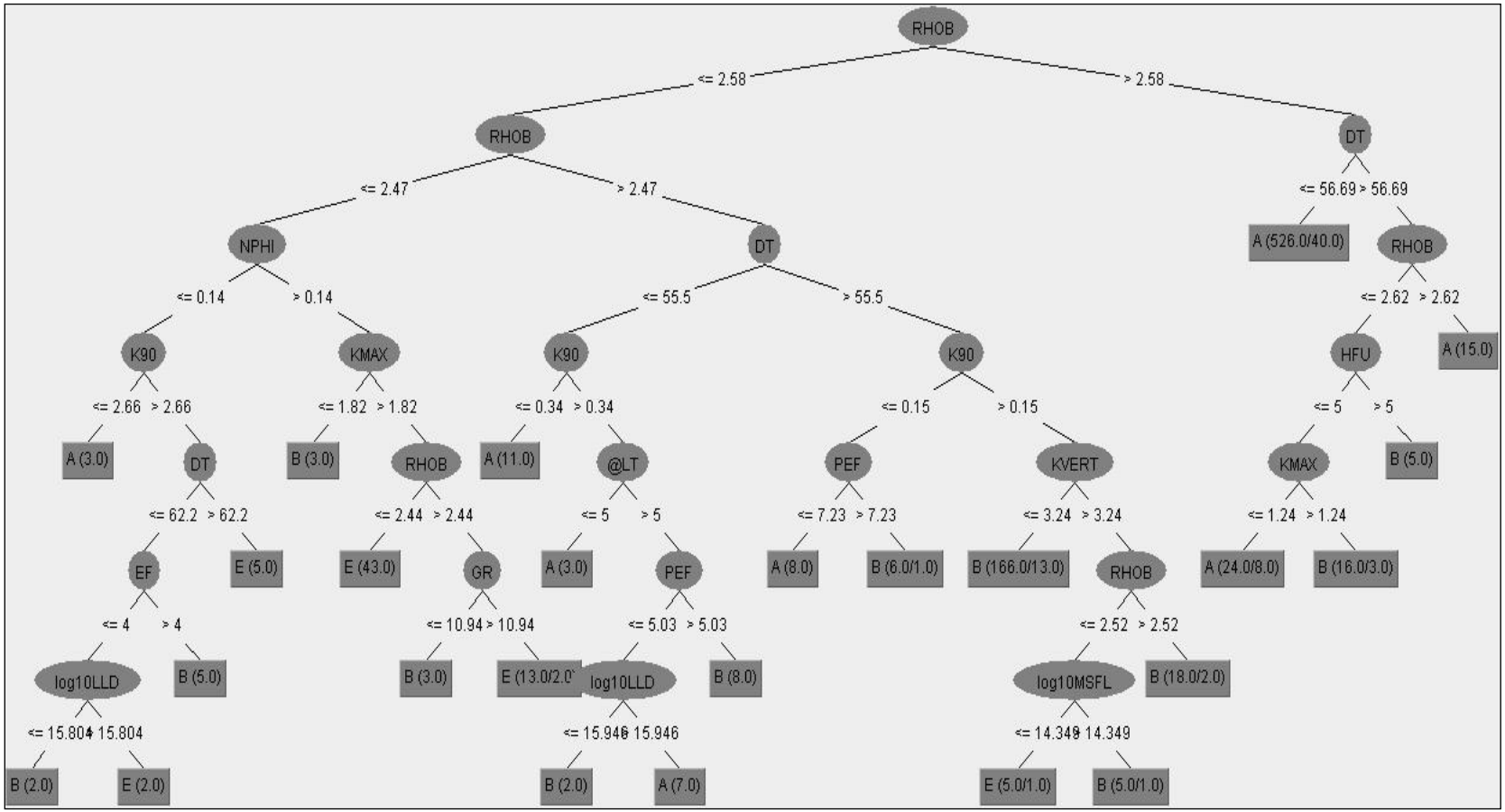
Regla 23: SI RHBO > de 2.58, DT > de 56.69, RHBO <= de 2.62, HFU <= de 5 **Y** KMAX > de 1.24 **ENTONCES** POR (porosidad) = B (Buena).

Regla 24: SI RHBO > de 2.58, DT > de 56.69, RHBO <= de 2.62 **Y** HFU > de 5 **ENTONCES** POR (porosidad) = B (Buena).

Regla 25: SI RHBO > de 2.58, DT > de 56.69 **Y** RHBO > de 2.62 **ENTONCES** POR (porosidad) = A (aceptable).

Según el número de hojas que arroje el árbol de decisión genera el mismo número de reglas de decisión. Para este caso el número de hojas del árbol de decisión es 25 y el número de reglas de decisión es 25. La forma de leer las reglas es la siguiente: Regla 1: **SI** la variable RHBO menor igual de 2.58, la variable RHBO menor igual de 2.47, la variable NPHI menor igual de 0.14 **Y** la variable K90 menor igual de 2.66 **ENTONCES** la predicción de la variable POR será aceptable (A)

Figura 19. Árbol de decisión grafico.



Fuente: SATMD

2.6 EVALUACIÓN

Es necesario evaluar los modelos de la fase anterior, pero ya no sólo desde el punto de vista estadístico respecto a los datos, sino ver si los modelos se ajustan a las necesidades de los objetivos esperados, es decir, si los modelos sirven para responder a los requerimientos del ICP en la predicción de hidrocarburos.

En esta etapa se desarrolla la aplicación de los modelos propuestos iniciando con el modelo B (ver figura 13).

Tabla 13. Clasificación de las Variables

VARIABLE	CLASIFICACIÓN	ESCALA
EF	Independiente	Métrica
@LT	Independiente	Métrica
HFU	Independiente	Métrica
GR	Independiente	Métrica
Log10(LLD)	Independiente	Métrica
Log10(MSFL)	Independiente	Métrica
DT	Independiente	Métrica
RHOB	Independiente	Métrica
NPHI	Independiente	Métrica
PEF	Independiente	Métrica
KMAX	Independiente	Métrica
K90	Independiente	Métrica
KVERT	Independiente	Métrica
POR	Dependiente	Métrica /

		Nominal

Fuente: Experto Ingeniero de Petroleo

El número de registros o la muestra de las variables es de 904. Para esta técnica no se tiene en cuenta la variable dependiente (sea nominal o categórica). Ya se procedió con la fase de recogida y caracterización de datos. A continuación se aplica la técnica de Clusters (SimpleKmeans), con la cual se obtienen los siguientes resultados (Ver figura 20).

Figura 20. Resultados Técnica de Clústers

kMeans

======

Number of iterations: 4

Within cluster sum of squared errors: 611.2873733832989

Cluster centroids:

Cluster 0

Mean/Mode: 4.1568 6.3464 5.0435 22.3358 19.4626 17.9488 56.5114 2.5870 0.0716 6.0498 33.9157 4.8716 6.0458

Std Devs: 2.0785 1.8957 0.7778 11.4754 4.2018 4.7869 4.5774 0.0761 0.0441 2.0583 469.5473 20.7804 45.0515

Cluster 1

Mean/Mode: 3.7550 5.9932 1.7168 24.3732 20.0607 19.0961 55.1969 2.6108 0.0569 5.9490 34.1968 4.4204 9.7141

Std Devs: 2.0546 1.9982 0.6849 12.3019 4.9943 5.8313 4.9010 0.0764 0.0450 1.9685 483.4010 21.7899 132.3213

Clustered Instances

0 459 (51%)

1 445 (49%)

"EF"	**4.1568**	3.7550
"@LT"	**6.3464**	5.9932
"HFU"	**5.0435**	1.7168
"GR"	22.3358	**24.3732**
"log10LLD"	19.4626	**20.0607**
"log10MSFL"	17.9488	**19.0961**
"DT"	**56.5114**	55.1969
"RHOB"	2.5870	**2.6108**
"NPHI"	**0.0716**	0.0569
"PEF"	**6.0498**	5.9490
"KMAX"	33.9157	**34.1968**
"K90"	**4.8716**	4.4204
"KVERT"	6.0458	**9.7141**

Fuente: SATMD

2.6.1 Resultados Técnica de Clústers. Se observa en los resultados la segmentación de las muestras por *clúster*, se han generado dos clústers y en cada uno de ellos se pueden apreciar las medidas de los centroides por cada variable, su media y su respectiva desviación estándar.

En los resultados arrojados por el prototipo se encuentra el número de iteraciones en que fue resuelto éste análisis (4), la suma de cuadrados de los errores entre clústers (611.2873733832989) y se puede observar “Clustered Instances” (instancias agrupadas) que son los porcentajes de las clasificaciones de cada clúster. El clúster 0 tiene 459 instancias agrupadas y el clúster 1 445 instancias agrupadas.

En la comparación de medias se puede ver la clasificación de instancias por clúster, de tal forma de cómo quedan compuestos cada uno de ellos. Las nuevas variables generadas a partir de estos resultados, son colocadas por el prototipo en el Dataset (sabana de datos) para aplicarle un análisis de regresión y comparar el rendimiento que se obtenga en la prueba que se muestra más adelante en componentes principales. El número de instancias para cada clúster queda conformado como se muestra la tabla 14.

Tabla 14. Conformación de Clusters.

Clúster 0	EF	@LT	HFU	DT	NPHI	PEF	K90
Clúster 1	GR	log10LLD	log10MSFL	RHOB	KMAX	KVERT	

Fuente: SATMD

Según la Comparación de medias que arrojan los resultados las variables que sea agrupan en el Cluster 0 son EF, @LT, HFU, DT, NPHI, PEF, K90; y las variables que se agrupan en el Cluster 1 son GR, log10LLD, log10MSFL, RHOB, KMAX, KVERT.

En la tabla 15 se puede observar la distribución de los 3 clusters que se generaron, En la figura 21 se puede ver la distribución de una manera más gráfica.

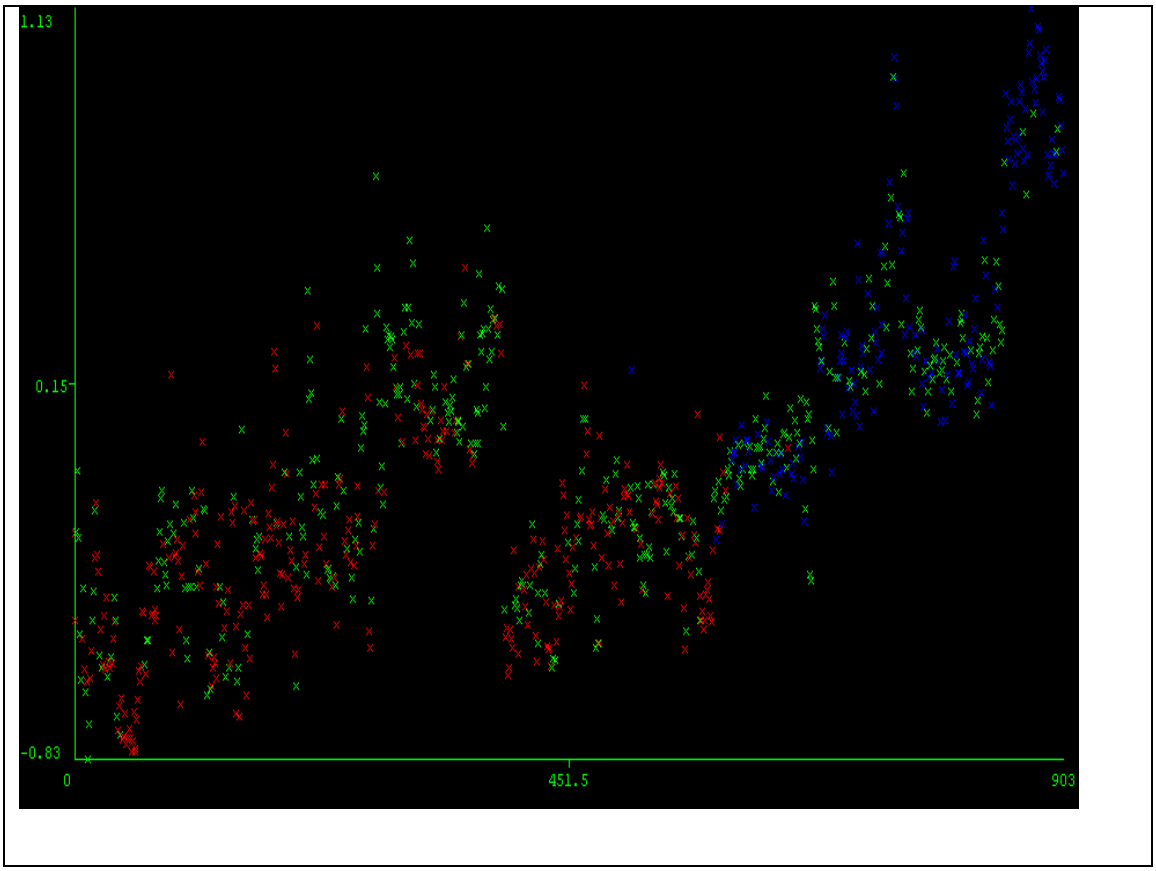
Tabla 15. Distribución de Clústers

Distribución de Clústers			
	cluster2	**cluster1**	**cluster0**
Instancia	310	241	310
V1-PC1	-0.018291339370409773	-0.5167876454506439	0.2704972154070183
V2-PC1	0.27197223877333493	0.3368344622597319	0.840816329675221
V3-PC1	-0.017667027664337857	-0.010260768253196311	-0.046281198238321206
V4-PC1	-0.11203953340656463	-0.17450306934703952	-0.014805511090044415
V5-PC1	0.6784996549206799	0.2721071388943593	0.15904460921603608
V6-PC1	-0.575292555138278	-0.4678011024793972	-0.45834021186668283
V7-PC1	0.22639336396279736	0.36577889155647925	0.25859129073273013
V8-PC1	-0.03442510741408828	-0.35983578693545576	0.35559151806262673
V9-PC1	-0.39294157386147177	-0.5860613669640565	-0.7178349161093657
V10-PC1	0.04411207673650844	-0.11525948274186279	-0.5593468153646047

Fuente: WeKa 3.4.12

El color verde corresponde al cluster 0, El color rojo corresponde al cluster 1 y el color Azul corresponde al cluster 2.

Figura 21. Distribución de los Clústers



Fuente: WeKa 3.4.12

Con los clústers ya conformados, se procede a hacer un análisis de regresión a cada uno de ellos, y así poder validar que estas nuevas variables hacen que el rendimiento sea mayor. Los resultados se pueden observar a continuación (Ver figura 22 y 23).

2.6.2 Resultados Técnica de Análisis de Regresión aplicado a las variables que Pertenecen a cada Clúster.

Figura 22. Resultados Técnica de Análisis de Regresión aplicado al Clúster 0

Linear Regression Model

POR = 0.131 * DT-C0-R1 + 0.8453 * NPHI-C0-R1 + 0.1141 * HFU-C0-R1 + 0.1436 * @LT-C0-R1 + -6.5452

"POR" = -6.545

=== Summary ===

Correlation coefficient	**0.8051**
Mean absolute error	**0.6763**
Root mean squared error	**0.8497**
Relative absolute error	**55.7902 %**
Root relative squared error	**59.315 %**
Total Number of Instances	**459**

Fuente: SATMD

Figura 23. Resultados Técnica de Análisis de Regresión aplicado al Clúster 1

Linear Regression Model

POR = -0.0218 * GR-C1-R1 + -0.017 * log10LLD-C1-R1 + -13.5147 * RHOB-C1-R1 + 0.0002 * KMAX-C1-R1 + -0.0006 * KVERT-C1-R1 + 39.2003

"POR" = 39.2

=== Summary ===	
Correlation coefficient	**0.8355**
Mean absolute error	**0.644**
Root mean squared error	**0.7644**
Relative absolute error	**54.6008 %**
Root relative squared error	**54.943 %**
Total Number of Instances	**445**

Fuente: SATMD

Se puede concluir que el agrupamiento del clúster 0 y clúster 1 es muy bueno para otros análisis con las variables que este contiene, ya que da un rendimiento relativamente alto con una correlación muy buena, para el cluster 0 da un coeficiente de correlación 0.8051 y para el clusters 2 da un coeficiente de correlación 0.8355.Para el modelo A (ver figura 12).

Tabla 16. Clasificación de Variables

VARIABLE	CLASIFICACIÓN	ESCALA
EF	Independiente	Métrica
@LT	Independiente	Métrica
HFU	Independiente	Métrica
GR	Independiente	Métrica
Log10(LLD)	Independiente	Métrica
Log10(MSFL)	Independiente	Métrica
DT	Independiente	Métrica
RHOB	Independiente	Métrica
NPHI	Independiente	Métrica

PEF	Independiente	Métrica
KMAX	Independiente	Métrica
K90	Independiente	Métrica
KVERT	Independiente	Métrica
POR	Dependiente	Métrica / Nominal

Fuente: Experto Ingeniero de Petróleo

2.6.3 ***Resultados Técnica de Componentes Principales.*** Teniendo en cuenta los criterios de esta técnica se observa que solo 21 datos por debajo de la diagonal de 77 son mayores de 0.3 (Ver figura 24).

Figura 24. Resultados Técnica de Componentes Principales

Principal Components Attribute Transformer

Correlation matrix

1 0.27 0.02 -0.3 -0.34 -0.46 0.52 -0.53 0.58 -0.17 -0.03 0.12 0.05

0.27 1 0 -0.06 -0.25 -0.27 0.39 -0.52 0.42 -0.02 -0.01 0.2 0.03

0.02 0 1 -0.07 0.04 -0.01 -0.01 -0.01 0.02 0.06 -0.01 -0.01 -0.04

-0.3 -0.06 -0.07 1 0 0.12 -0.24 0.2 -0.23 -0.26 0.08 -0.16 -0.01

-0.34 -0.25 0.04 0 1 0.69 -0.58 0.54 -0.57 0.28 0.05 -0.01 -0.01

-0.46	-0.27	-0.01	0.12	**0.69**	1	-0.64	0.59	-0.65	0.2	-0.02	-0.13	-0.03
0.52	**0.39**	-0.01	-0.24	**-0.58**	**-0.64**	1	-0.86	0.87	-0.01	-0.02	0.34	0.04
-0.53	**-0.52**	-0.01	0.2	**0.54**	**0.59**	**-0.86**	1	-0.89	0.09	0.01	-0.33	-0.06
0.58	**0.42**	0.02	-0.23	**-0.57**	**-0.65**	**0.87**	**-0.89**	1	-0.11	0	0.29	0.06
-0.17	-0.02	0.06	-0.26	0.28	0.2	-0.01	0.09	-0.11	1	-0.01	0.23	-0.01
-0.03	-0.01	-0.01	0.08	0.05	-0.02	-0.02	0.01	0	-0.01	1	0.05	0.24
0.12	0.2	-0.01	-0.16	-0.01	-0.13	**0.34**	**-0.33**	0.29	0.23	0.05	1	0.04
0.05	0.03	-0.04	-0.01	-0.01	-0.03	0.04	-0.06	0.06	-0.01	0.24	0.04	1

Eigenvalue	Proportion	Cumulative
4.55614	**0.35047**	**0.35047**
1.57199	**0.12092**	**0.47139**
1.27043	**0.09773**	**0.56912**
1.02617	**0.07894**	**0.64806**
0.94373	0.07259	0.72065
0.79876	0.06144	0.78209
0.72154	0.0555	0.8376
0.67182	0.05168	0.88927
0.49719	0.03825	0.92752
0.4323	0.03325	0.96077

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
0.3132	-0.0036	-0.0678	-0.2691	0.279	-0.2237	0.3402	-0.3026	-0.3636	-0.5673	EF
0.247	0.0813	0.054	0.266	-0.3506	-0.6463	0.2566	0.3525	0.295	-0.1449	@LT
0.0031	0.1375	-0.1835	-0.703	-0.6532	-0.0101	-0.1297	-0.0815	0.0322	-0.0029	HFU
-0.127	-0.4613	0.1698	0.3565	-0.5061	0.034	-0.1758	-0.1088	-0.4977	-0.2577	GR
-0.328	0.3344	0.0597	-0.0005	-0.0167	-0.2497	0.2675	-0.2451	-0.2942	0.1684	log10LLD
-0.3646	0.1946	0.0205	0.0839	-0.0366	-0.283	0.1361	-0.1424	-0.2581	0.3875	log10MSFL
0.4256	0.0826	-0.0022	0.0571	-0.0118	0.1361	-0.086	0.0185	-0.2295	0.2722	DT
-0.4287	-0.057	-0.0301	-0.0812	0.105	0.0623	-0.0096	-0.0047	0.194	-0.3343	RHOB
0.4343	0.0304	0.0065	-0.0094	-0.0171	0.0343	0.001	-0.028	-0.2459	0.289	NPHI
-0.0682	0.6265	-0.0116	0.0942	-0.007	0.1889	-0.2162	0.518	-0.3512	-0.3285	PEF
-0.0081	-0.0087	0.6855	-0.1776	-0.1345	0.3752	0.5493	0.1763	0.0301	0.0237	KMAX
0.1576	0.4562	0.1606	0.3456	-0.2079	0.2003	-0.0894	-0.6181	0.3223	-0.1854	K90
0.0332	0.0202	0.6552	-0.2424	0.2051	-0.3853	-0.5627	-0.0518	0.0049	-0.0107	KVERT

Fuente: SATMD

El Componente Principal (V1) tiene la mayor correlación positiva con las variables EF, DT, NPHI, mientras que tiene correlación negativa con RHOB, Log 10MSFL y casi nula con el resto de variables.

El Componente Principal (V2) tiene la mayor correlación positiva con las variables Log10LLD, PEF, K90, mientras que tiene correlación negativa con GR y casi nula con el resto de variables.

El Componente Principal (V3) tiene la mayor correlación positiva con las variables KMAX, KVERT y casi nula con el resto de variables.

El Componente Principal (V4) tiene la mayor correlación negativa con HFU, casi nula con el resto de variables.

El coeficiente de correlación negativa indica una dependencia entre las variables llamada relación inversa; cuando una de ellas aumenta, las otras disminuye en idéntica proporción.

La variable @LT no presentó correlación con ninguno de los 4 componentes porque su variabilidad no fue superior a 0,3 lo que indica que tiene correlación poco significativa.

Tabla 17. Conformación de los Factores

V1	**V2**	**V3**	**V4**
EF	GR	KMAX	HFU
DT	Log10LLD	KVERT	
NPHI	PEF		

RHOB	K90		
Log 10MSFL			

Fuente: Autores del Proyecto

2.6.4 ***Resultados Técnica de Análisis de Regresión.*** La segunda técnica a aplicar según el modelo A es Análisis de Regresión, las variables que se van a usar (GR, Log10 (LLD), Log10 (MSFL), DT, RHOB, NPHI, PEF, KMAX, K90, KVERT, EF, LT, HFU) (Ver figura 25); y los nuevos componentes generados (V1-PC1, V2-PC1, V3-PC1 y V4-PC1) (Ver figura 26).

Figura 25. Resultados Técnica de Análisis de Regresión

Linear Regression Model

POR = (-0.107 * EF) + (-0.0382 * GR) + (0.1716 * DT) + (-39.8059 * RHOB) + (3.9332 * NPHI) + (0.0004 * KMAX) + (0.007 * K90) + (-0.0022 * KVERT) + 101.3891

"POR" = 101.389

=== Summary ===

Correlation coefficient	**0.9089**
Mean absolute error	**1.3315**
Root mean squared error	**1.8381**
Relative absolute error	**37.8605 %**
Root relative squared error	**41.6978 %**

Total Number of Instances	**904**

Fuente: SATMD

Con estos resultados se concluye que POR es de 101.389 y depende en gran medida de la variable RHBO que tiene alta significancia.
El coeficiente de correlación es de 0.9089 indicando una muy buena correlación y los errores son bajos.

2.6.5 Resultados Técnica de Análisis de Regresión a Componentes Principales. Según el modelo A se aplica la técnica de regresión lineal a las variables generadas por la técnica Análisis de Componentes Principales (ver figura 26).

Figura 26. Resultados Análisis de Regresión aplicado a Componentes Principales

Linear Regression Model

POR = 10.0344 * V1-PC1 + 4.9429 * V2-PC1 + 1.9685 * V4-PC1 + 5.786

"POR" = -0.06148178359203982

=== Summary ===

Correlation coefficient	**0.8562**
Mean absolute error	**1.7496**
Root mean squared error	**2.2772**
Relative absolute error	**49.7509 %**

Root relative squared error	**51.6582 %**
Total Number of Instances	**904**

Fuente: SATMD

Se observa que la variable POR es 0.06149 y V1-PC1 determina en gran medida seguida de V2-PC2. El coeficiente de correlación es de 0.8562, indicando una muy buena correlación. Como los resultados cumplen satisfactoriamente los requisitos de la técnica, se aplica Árboles de Decisión.

2.6.6 ***Resultados Técnica de Arboles de Decisión.*** Los resultados aplicados de la técnica de arboles de decisión para las variables independientes (GR, Log10 (LLD), Log10 (MSFL), DT, RHOB, NPHI, PEF, KMAX, K90, KVERT, EF, LT, HFU, @LT) y la variable dependiente (POR (nominal)). (Ver figura 27), (Ver figura 28).

Figura 27. Resultados Técnica de Árboles de decisión.

```
J48 pruned tree
------------------
RHOB <= 2.58
|   RHOB <= 2.47
|   |   NPHI <= 0.14
|   |   |   K90 <= 2.66: A (3.0)
|   |   |   K90 > 2.66
|   |   |   |   DT <= 62.2
|   |   |   |   |   EF <= 4
|   |   |   |   |   |   log10LLD <= 15.804: B (2.0)
|   |   |   |   |   |   log10LLD > 15.804: E (2.0)
```

```
|   |   |   |   EF > 4: B (5.0)
|   |   |   |   DT > 62.2: E (5.0)
|   |   NPHI > 0.14
|   |   |   KMAX <= 1.82: B (3.0)
|   |   |   KMAX > 1.82
|   |   |   |   RHOB <= 2.44: E (43.0)
|   |   |   |   RHOB > 2.44
|   |   |   |   |   GR <= 10.94: B (3.0)
|   |   |   |   |   GR > 10.94: E (13.0/2.0)
|   RHOB > 2.47
|   |   DT <= 55.5
|   |   |   K90 <= 0.34: A (11.0)
|   |   |   K90 > 0.34
|   |   |   |   @LT <= 5: A (3.0)
|   |   |   |   @LT > 5
|   |   |   |   |   PEF <= 5.03
|   |   |   |   |   |   log10LLD <= 15.946: B (2.0)
|   |   |   |   |   |   log10LLD > 15.946: A (7.0)
|   |   |   |   |   PEF > 5.03: B (8.0)
|   |   DT > 55.5
|   |   |   K90 <= 0.15
|   |   |   |   PEF <= 7.23: A (8.0)
|   |   |   |   PEF > 7.23: B (6.0/1.0)
|   |   |   K90 > 0.15
|   |   |   |   KVERT <= 3.24: B (166.0/13.0)
|   |   |   |   KVERT > 3.24
|   |   |   |   |   RHOB <= 2.52
|   |   |   |   |   |   log10MSFL <= 14.349: E (5.0/1.0)
|   |   |   |   |   |   log10MSFL > 14.349: B (5.0/1.0)
```

```
|   |   |   |   RHOB > 2.52: B (18.0/2.0)
RHOB > 2.58
|   DT <= 56.69: A (526.0/40.0)
|   DT > 56.69
|   |   RHOB <= 2.62
|   |   |   HFU <= 5
|   |   |   |   KMAX <= 1.24: A (24.0/8.0)
|   |   |   |   KMAX > 1.24: B (16.0/3.0)
|   |   |   HFU > 5: B (5.0)
|   |   RHOB > 2.62: A (15.0)
```

Fuente: SATMD

Se debe tener en cuenta el Estadístico Kappa en la interpretación de resultados y saber su grado de acuerdo. (Ver Tabla 18).

Tabla 18. Estadístico de de Kappa

Kappa	Grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

Fuente: Índices de concordancia. [Online]. [Visitada en Octubre de 2007].Disponible en Internet: < http://www.hrc.es/bioest/errores_2.html >

Figura 28. Resultados Técnica de Árbol de decisión

```
Number of Leaves  :      25
Size of the tree :    49

=== Summary ===
Correctly Classified Instances         833               92.146  %
Incorrectly Classified Instances        71                7.854  %
Kappa statistic                          0.8436
Mean absolute error                      0.0926
Root mean squared error                  0.2152
Root relative squared error             52.0113 %
Total Number of Instances              904

=== Confusion Matrix ===
   a   b   c   <-- classified as
 549  16   1 |   a = A
  48 219   2 |   b = B
   0   4  65 |   c = E
```

Fuente: SATMD

El número de hojas del árbol es de 25 y el número de nodos 49. Las instancias clasificadas correctamente son de 833 e incorrectas71. Se puede observar con el Estadístico Kappa con un 0.8436 con una aceptación casi perfecta. Los errores son mínimos.

Según la matriz de confusión se han clasificado 549 instancias correctamente en Aceptable, 16 instancia en la categoría Bueno y 1 en la categoría Excelente. Para la categoría Bueno se ha clasificado 219 instancias correctamente en la categoría Bueno, 48 en la categoría Aceptable y 2 en la categoría Excelente. Para la

categoría Excelente se ha clasificado 65 instancias correctamente en la categoría Excelente, 4 en la categoría Bueno y 0 en la categoría aceptable.

Esta técnica genera unas reglas de decisión que han sido interpretadas del árbol de decisión grafico que arroja los resultados Cada Hoja del árbol genera una regla de decisión que para estos resultados generó 25 reglas; pero la idea es optimizar los resultados y buscar una excelencia en la porosidad. Estas son las reglas de decisión que predicen la variable porosidad como Excelente.

Hay varios caminos a seguir para garantizar que la Porosidad sea Excelente, según el Árbol Gráfico (ver Figura 29).

SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= de 0.14, K90 > de 2.66, DT <= de 62.2, EF<= de 4 **Y** log10LLD > de 15.804 **ENTONCES** POR (porosidad) = E (Excelente).

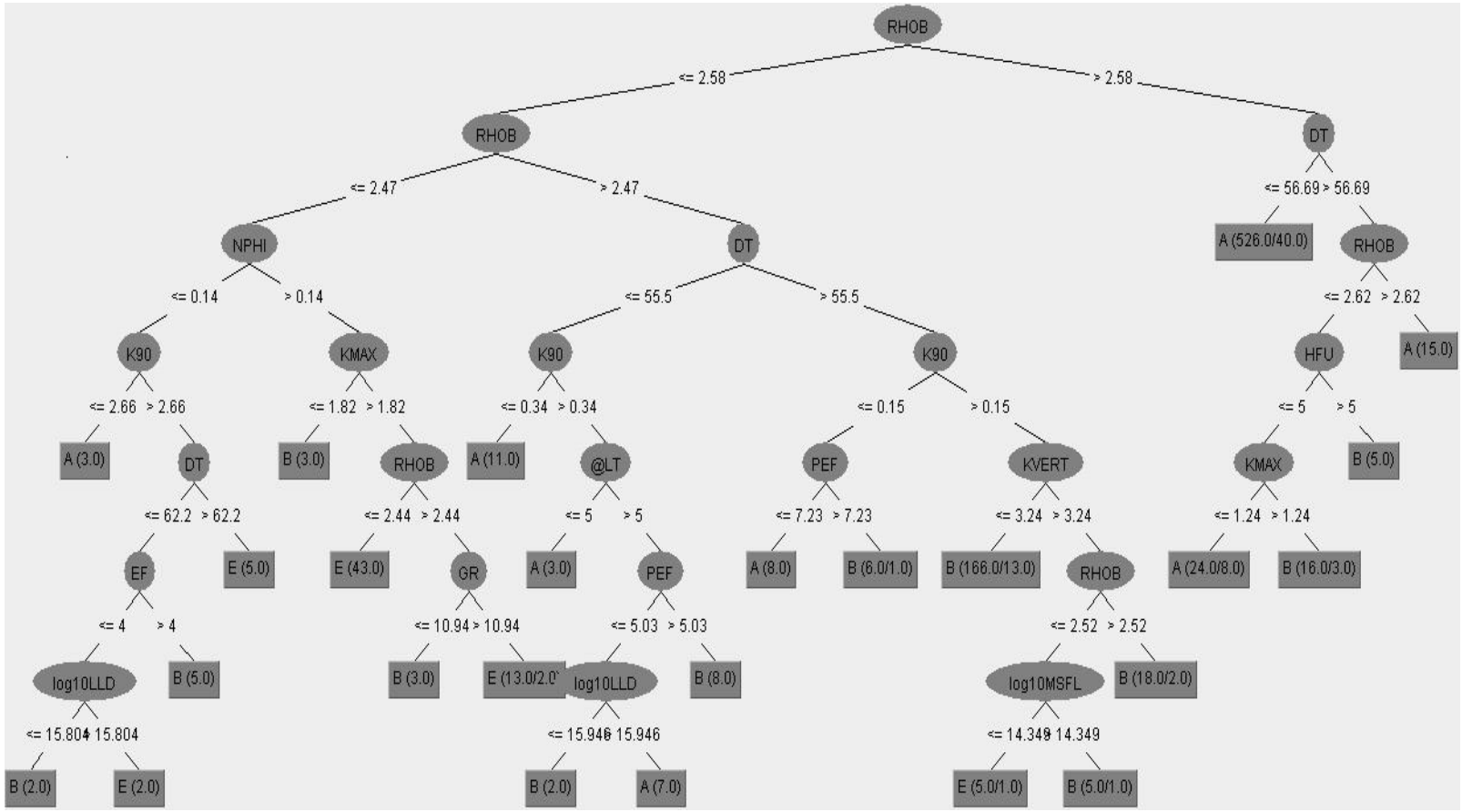
SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI <= 0.14, K90 > 2.66 **Y** DT > de 62.2 **ENTONCES** POR (porosidad) = E (Excelente).

SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI > de 0.14, KMAX > de 1.82 **Y** RHBO <= de 2.44 **ENTONCES** POR (porosidad) = E (Excelente).

SI RHBO <= de 2.58, RHBO <= de 2.47, NPHI > de 0.14, KMAX > de 1.82, RHBO > de 2.44 **Y** GR > de 10.94 **ENTONCES** POR (porosidad) = E (Excelente).

SI RHBO <= de 2.58, RHBO > de 2.47, DT > de 55.5, K90 > de 0.15, KVERT > de 3.24, RHOB <= de 2.52 **Y** log10MSFL <= de 14.349 **ENTONCES** POR (porosidad) = E (Excelente)

Figura 29. Árbol de decisión gráfico



Fuente: SATMD

2.6.7 Resultados aplicando arboles de decisión a los componentes principales.

Siguiendo con la aplicación del Modelo procedemos a aplicar árbol de decisión a los resultados obtenidos por la técnica de componentes principales. Las variables generadas por el análisis de componentes principales (V1-PC1, V2-PC2, V3-PC1, V4-PC1) y los resultados obtenidos son (Ver figura 30).

Figura 30. Resultados Técnica de Árboles de decisión aplicado a componentes principales

```
J48 pruned tree
------------------
V1-PC1 <= 0.077518
|   V1-PC1 <= -0.148091
|   |   V1-PC1 <= -0.321708: A (292.0/10.0)
|   |   V1-PC1 > -0.321708
|   |   |   V2-PC1 <= 0.746961: A (177.0/23.0)
|   |   |   V2-PC1 > 0.746961: B (7.0)
|   V1-PC1 > -0.148091
|   |   V4-PC1 <= 0.182239: A (153.0/55.0)
|   |   V4-PC1 > 0.182239
|   |   |   V1-PC1 <= -0.111005: A (5.0)
|   |   |   V1-PC1 > -0.111005
|   |   |   |   V2-PC1 <= 0.290737
|   |   |   |   |   V3-PC1 <= 0.077048: B (2.0)
|   |   |   |   |   V3-PC1 > 0.077048: A (4.0/1.0)
|   |   |   |   V2-PC1 > 0.290737: B (15.0)
V1-PC1 > 0.077518
|   V1-PC1 <= 0.425499: B (174.0/35.0)
```

```
|   V1-PC1 > 0.425499
|   |   V1-PC1 <= 0.662173
|   |   |   V2-PC1 <= 0.355452: B (5.0)
|   |   |   V2-PC1 > 0.355452
|   |   |   |   V3-PC1 <= -0.072108
|   |   |   |   |   V3-PC1 <= -0.14048: E (3.0/1.0)
|   |   |   |   |   V3-PC1 > -0.14048: B (8.0/1.0)
|   |   |   |   V3-PC1 > -0.072108: E (15.0/2.0)
|   |   V1-PC1 > 0.662173
|   |   |   V2-PC1 <= 0.292433
|   |   |   |   V1-PC1 <= 0.941028
|   |   |   |   |   V1-PC1 <= 0.741592: E (4.0/1.0)
|   |   |   |   |   V1-PC1 > 0.741592: B (4.0)
|   |   |   |   V1-PC1 > 0.941028: E (14.0)
|   |   |   V2-PC1 > 0.292433: E (22.0)

Number of Leaves  :         17
Size of the tree :      33
=== Summary ===
Correctly Classified Instances         775               85.7301 %
Incorrectly Classified Instances       129               14.2699 %
Kappa statistic                          0.7072
Mean absolute error                      0.1469
Root mean squared error                  0.2711
Root relative squared error             65.5092 %
Total Number of Instances              904
=== Confusion Matrix ===
a   b   c   <-- classified as
542  22   2 |   a = A
```

```
88 179   2 |   b = B
1  14  54 |   c = E
```

Fuente: SATMD

Fueron clasificadas 775 instancias correctamente y 129 incorrectamente. Los errores son bajos y Estadístico Kappa da una aceptación Sustancial. El árbol tiene 17 hojas y 33 nodos. Según la matriz de confusión se han clasificado 542 instancias correctamente en Aceptable, 22 instancia en la categoría Bueno y 2 en la categoría Excelente. Para la categoría Bueno se ha clasificado 179 instancias correctamente en la categoría Bueno, 88 en la categoría Aceptable y 2 en la categoría Excelente. Para la categoría Excelente se ha clasificado 54 instancias correctamente en la categoría Excelente, 14 en la categoría Bueno y 1 en la categoría aceptable.

Hay varios caminos a seguir para garantizar que la Porosidad sea Excelente, según el Árbol Gráfico. (Ver Figura 31).

Esta técnica genera unas reglas de decisión que han sido interpretadas del árbol de decisión grafico que arroja los resultados Cada Hoja del árbol genera una regla de decisión que para estos resultados aplicados a los componente principales generó 17 reglas; pero la idea es optimizar los resultados y buscar una excelencia en la porosidad. Estas son las reglas de decisión que predicen la variable porosidad como Excelente.

SI V1-PC1 > 0.077518, V1-PC1 > 0.425499, V1-PC1 <= 0.662173, V2-PC1 > 0.355452, V3-PC1 <= -0.072108 **Y** V3-PC1 <= -0.14048 **ENTONCES** POR (porosidad) = E (Excelente).

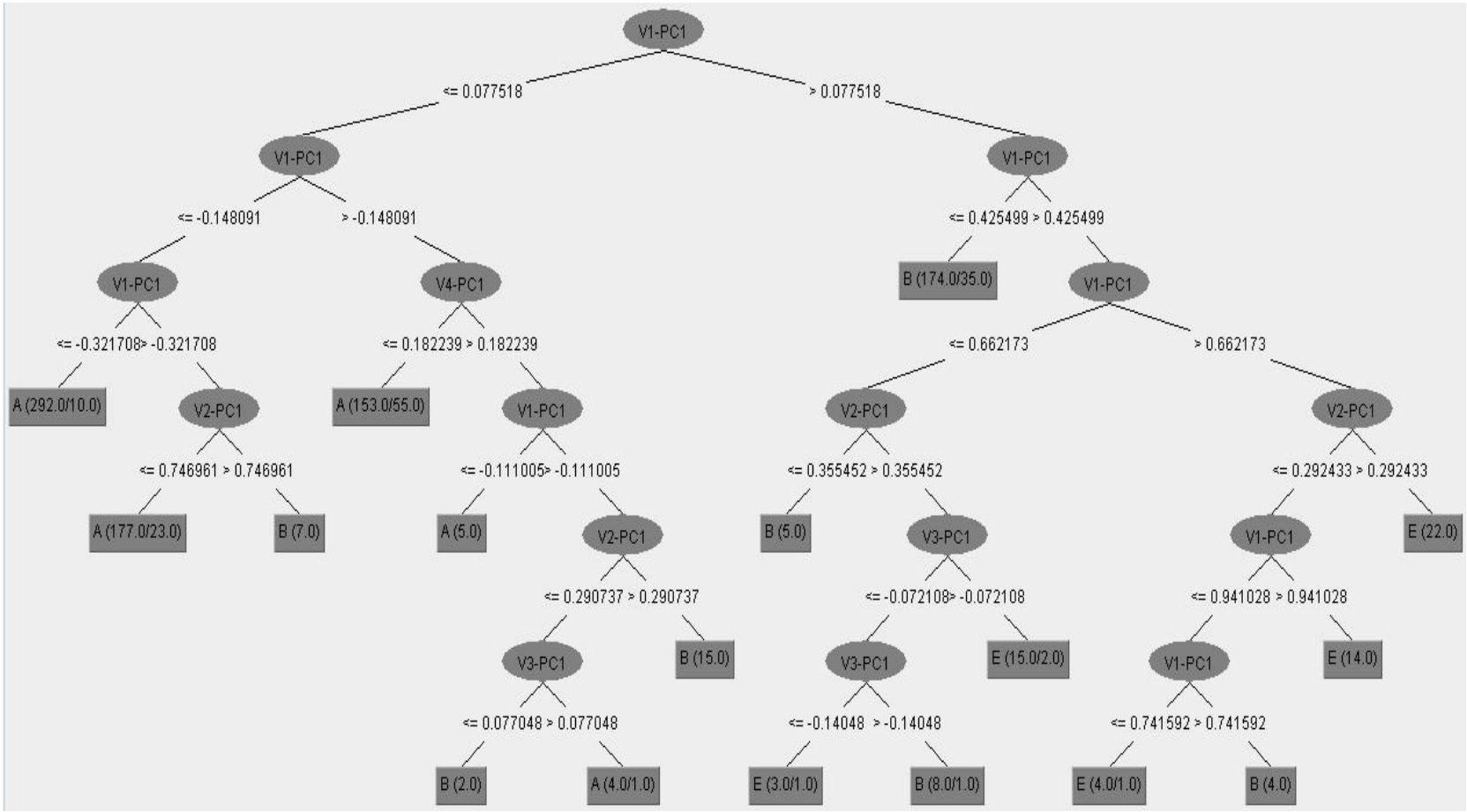
SI V1-PC1 > 0.077518, V1-PC1 > 0.425499, V1-PC1 <= 0.662173, V2-PC1 > 0.355452 **Y** V3-PC1 > -0.072108 **ENTONCES** POR (porosidad) = E (Excelente).

SI V1-PC1 > 0.077518, V1-PC1 > 0.425499, V1-PC1 > 0.662173, V2-PC1 <= 0.292433, V1-PC1 <= 0.941028 **Y** V1-PC1 <= 0.741592 **ENTONCES** POR (porosidad) = E (Excelente).

SI V1-PC1 > 0.077518, V1-PC1 > 0.425499, V1-PC1 > 0.662173, V2-PC1 <= 0.292433 Y V1-PC1 > 0.941028 **ENTONCES** POR (porosidad) = E (Excelente).

SI V1-PC1 > 0.077518, V1-PC1 > 0.425499, V1-PC1 > 0.662173 **Y** V2-PC1 > 0.292433 **ENTONCES** POR (porosidad) = E (Excelente

Figura 31.Árbol de decisión Grafico aplicado a Componentes Principales



Fuente: SATMD

2.6.8 ***Resultado Aplicando Árboles de decisión mediante el Método Percentage Split.*** Aplicamos la técnica, esta vez usando la opción de "Percentage Split", pero con los mismos resultados, así que revisamos la teoría y probamos con la versión oficial del WEKA utilizando datos entrenados aplicando filtros, pero con los mismos resultados.

Se define un porcentaje con el que se asimila el modelo y la evaluación se realiza con los datos restantes, se ha definido un porcentaje de 70%, 50% y 30% y la evaluación se hará con los porcentajes restante respectivamente. (Ver Figuras 32, 33 y 34)

Figura 32. Percentage Split 70 %

Correctly Classified Instances	**234**	**86.0294 %**
Incorrectly Classified Instances	**38**	**13.9706 %**
Kappa statistic	**0.6999 (Sustancial)**	
Mean absolute error	**0.1331**	
Root mean squared error	**0.2981**	
Relative absolute error	**39.4494 %**	
Root relative squared error	**74.3287 %**	
Total Number of Instances	**272**	

Fuente: WeKa 3.4.12

Figura 33. Percentage Split 30 %

Correctly Classified Instances	**473**	**74.7235 %**
Incorrectly Classified Instances	**160**	**25.2765 %**
Kappa statistic	**0.505 (Moderado)**	
Mean absolute error	**0.1738**	
Root mean squared error	**0.3993**	
Relative absolute error	**49.6756 %**	
Root relative squared error	**97.2133 %**	
Total Number of Instances	**633**	

Fuente: WeKa 3.4.12

Figura 34. Percentage Split 50 %

Correctly Classified Instances	**379**	**83.8496 %**
Incorrectly Classified Instances	**73**	**16.1504 %**
Kappa statistic	**0.6678 (Sustancial)**	
Mean absolute error	**0.1465**	
Root mean squared error	**0.3013**	
Relative absolute error	**42.4948 %**	
Root relative squared error	**74.1234 %**	
Total Number of Instances	**452**	

Fuente: WeKa 3.4.12

Los resultados obtenidos en las anteriores pruebas mediante el método de "Percentage Split" se enseñan en la siguiente tabla haciendo una comparación de cada uno de los resultados obtenidos en los diferentes porcentajes de prueba. (Ver tabla 19)

Tabla 19. Comparación de Percentage Split

Percentage Split	TP Rate	FP Rate	Precisión	Class
70%	0.94	0.213	0.901	A
	0.662	0.06	0.797	B
	0.833	0.028	0.682	E
30%	0.811	0.261	0.848	A
	0.611	0.194	0.556	B
	0.717	0.022	0.717	E
50%	0.897	0.184	0.906	A
	0.723	0.117	0.688	B
	0.727	0.014	0.8	E

Fuente: Autores del Proyecto

TP Rate: Significa "True Positive Rate". Indica el porcentaje de instancias de una clase clasificadas positivamente dentro de esa clase.

FP Rate: Significa "False Positive Rate". Indica el porcentaje de instancias de una clase que estando en esa clase son clasificadas incorrectamente.

En la tabla 20 se interpreta los resultados obtenidos anteriormente:

Tabla 20. Comparación de Percentage Split

Percentage Split	Correctly Classified Instances	Kappa statistic
70%	86,0294%	0.6999
30%	74,7235%	0.505

50%	83,8496%	0.6678

Fuente: Autores del Proyecto

Percentage Split nos quiere decir que el archivo original de datos será dividido en dos archivos: Uno para entrenamiento del árbol generado (Con el 70% de los ejemplos escogidos aleatoriamente) y otro para probar dicho árbol y poder generar unas medidas de rendimiento (con el restante 30% de los ejemplos).

Se puede concluir que las instancia correctamente clasificadas son alta y que el índice kappa que es una medida de concordancia entre clasificaciones cuyo valor debería ser de uno en el caso optimo; Según el grado de acuerdo el índice de kappa en las prueba es moderado, sustancial, moderado respectivamente lo que indica un grado significante de concordancia entre las clasificaciones.

2.7 SOLUCIÓN

En esta etapa el usuario podrá concluir y tomar decisiones o realizar nuevos análisis, (no necesariamente ligados a la serie de pasos) y comparar entre ellos el mejor. Por ejemplo, el usuario puede dividir la base de datos en la mitad o en lo que considere necesario y hacer un análisis de estos nuevos datawarehouse; o incluir nuevas categorías como (No-Aceptable, Aceptable, Bueno, Muy Bueno, Excelente), y así tendrá más bases para determinar las variables independientes que hacen más óptima la Porosidad.

Para el modelo A puede escoger cualquiera de los conjuntos dados ya que como se pudo observar a través de las pruebas dan buenos resultados. Para el modelo B sólo puede observar cómo se agrupan las diferentes variables y determinar cuál conglomerado da una mejor aproximación en el análisis de regresión.

3. CONCLUSIONES

La primera técnica en ser aplicada es Análisis de Componentes Principales, lo cual permite eliminar variables no correlacionadas con las otras variables y agruparlas en factores que facilitará a través de las otras técnicas establecer cuáles son las decisiones correctas a tomar.

La técnica de Clústers solo permite determinar cómo están agrupadas las variables y en qué medida afectan o contribuyen en el rendimiento de la Porosidad (POR).

Las variables POR (Porosidad) y GR (Gamma Ray), son las variables más representativas al momento de generar nuevo conocimiento, en razón a su relación con las demás variables, esto se resalta al momento de hacer uso de las técnicas Árboles de decisión.

El proceso de Descubrimiento de información en bases de datos (KDD) no se limita solo a la minería de datos, va desde el tratamiento y preparación de los datos hasta la toma de decisiones.

El prototipo final no genera decisiones que deban ser seguidas por el usuario, al contrario, es el usuario quien debe tomar decisiones en base a los resultados arrojados por el prototipo.

La técnica de árbol de decisión permite tomar claras y correctas decisiones, mediante las reglas de decisión obtenidas del árbol Grafico.

El almacén de datos (Datawarehouse) debe contener información suficiente y sus campos deben ser los adecuados para poder realizar un correcto análisis y, por ende, tomar una correcta decisión debido a que sobre esta información recae toda la investigación realizada.

Al modificar la variable POR (Porosidad), se obtiene un árbol de decisión mucho más fácil de leer e interpretar en razón a que la variable es cambiada de numérica a nominal, por cuanto, se realiza determinando intervalos evaluados con los expertos en petróleos.

Al aplicar Análisis de Regresión a las Variables del data set se puede percibir que la variable RHOB tiene una alta correlación con la variable Porosidad (POR). Igualmente, en las variables generadas por Componentes Principales, el componente que tiene una alta correlación con la variable Porosidad (POR) es al que pertenece la variable RHOB que en este caso en el Componente 1 (V1-PC1).

Los resultados de la técnica de Componentes Principales fueron positivos, por cuanto, se pudo extraer un buen número de componentes basados en la Varianza Total Explicada, tomando además como referencia el eigenvalue superior a la unidad.

Se logró eficacia al realizar los análisis en forma oportuna, pues con la ayuda del prototipo los datos son extraídos desde sus orígenes, lo que facilita obviar el PRE procesado de los mismos, es decir, filtrando los datos innecesarios.

Igualmente, al aplicar las diferentes Técnicas de minería de datos se comprobó que la técnica de mayor utilidad es la de componentes principales, por cuanto, permitió reducir el número de instancias de 14 a 10 componentes, perdiendo la menor cantidad de información.

BIBLIOGRAFÍA

About EMC: Leadership and Innovation: The Digital Universe. [Online]. [Visitada Marzo 16 de 2008]. Disponible en Internet: <http://www.emc.com/digital_universe

Análisis Multivariados, Dr. Roberto Mercado Hernandez. [Online]. [Visitada en Octubre de 2007] Disponible en Internet <www.feoc.ugto.mx/super/obtenga/D0003.ppt >.

Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. [Online]. [Visitada en Octubre de 2007] .Disponible en Internet:< www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf >.

Arboles de clasificación. [Online]. [Visitada en Octubre de 2007].Disponible en Internet: < http://math.uprm.edu/~edgar/treeDaza.html >.

Bienvenidos: ECOPETROL: Disponible en internet. [Online]. [Visitada Agosto 21 de 2007]. Disponible en Internet: <http://www.ecopetrol.com.co/Informe_Anual/....htm >.

DAEDALUS - Proceso de minería de datos. Disponible en internet. ? [Online]. [Visitada Septiembre de 2007]. Disponible en Internet: URL: < http://www.daedalus.es.../datos >.
Data Mining – Enfoque [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < http://datamining.iespana.es/dataminig_enfoque.htm >.

Data Mining & Knowledge Discovery in Databases (KDD). [Online]. [Visitada Octubre 3 de 2007]. Disponible en Internet: <http://elvex.ugr.es/etexts/spanish/kdd/KDD.html >

Descubrimiento de Conocimiento en Bases de Datos [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/ MonografiaMD.PDF

El análisis de Regresión – Joaquín Aldás Manzano.

El Análisis Factorial. [Online]. [Visitada en Diciembre de 2007] Disponible en Internet: URL: < www.ciberconta.unizar.es/.../.pdf >.

eGobierno e Innovación Gubernamental. [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet:< www.tidap.gob.mx/enc_dataw/JulioVargasdwh.ppt >.

Formación y prevención de riesgos laborales: Bases para la adquisición de una cultura preventiva en los centros educativos. [Online]. [Visitada en Diciembre de 2007] Disponible en Internet:< http://hera.ugr.es/tesisugr/17116788.pdf >.

HERNÁNDEZ ORALLO. Introducción a la Minería de Datos. 2005. Editorial Pearson,

Informática y Salud.BIOCHIPS Y BIOINFORMÁTICA [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: <http://www.conganat.org/SEIS/i_s/i_s25/i_s25_1.htm >.

Mati – Minería de datos. ¿Para qué nos sirve en una publicación digital? [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet:< http://www.mati.unam.mx/...51>

Métodos Estadísticos de Clasificación. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet:< http://www.seh-lelha.org/clasifica.htm >.

Minería de datos inteligente [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: < http://www.itba.edu.ar...datos.htm >

Minería de datos para control de robots. [Online]. [Visitada Octubre 27 de 2007]. Disponible en Internet:< http://www.daedalus.es/...Robotica.pdf >

OHRN, A. 1999. “Discernibility and Rough Sets in Medicine: Tools and Applications”, Department of Computer and Information Science, Norwegian University of Science and Technology, N-7941 Trondheim, Norway

Principal – Tipos de técnicas Multivariantes. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet: < http://www.virtual.unal.edu.co/.../tiposmulti.htm

Quality Control for the Manufacturing of Industrial Materials. [Online]. [Visitada Octubre 27 de 2007] Disponible en Internet: < http://www.daedalus.es/...l.php >

Red Española de Minería de Datos y Aprendizaje. [Online]. [Visitada Septiembre 11 de 2007]. Disponible en Internet: http://www.lsi.us.es/redmidas/

Reglas de asociación aplicadas a la detección de fraudes con tarjetas de crédito. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet:<www.ucu.edu.uy/.../.pdf >.

The Expanding Digital Universe. [Online]. [Visitado en Marzo 16 de 2008]. Disponible en Internet: < http://www.emc.com/.../Expanding...022507.pdf >

Universidad de Salamanca. Departamento de Informática y Automática. [Online]. [Visitada Octubre 27 de 2007]. Disponible en Internet:< www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf >

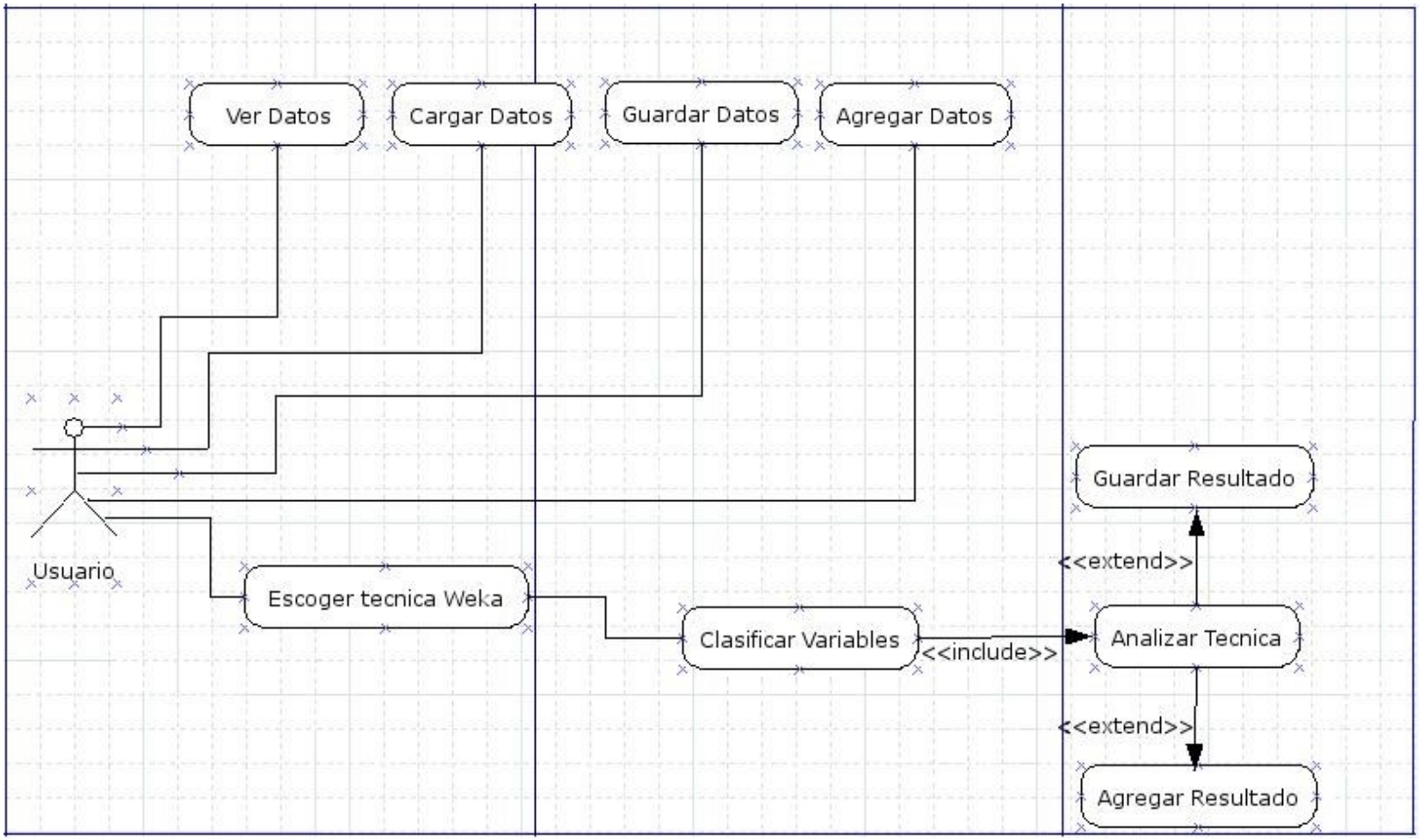
Universidad de Antioquia – Facultad de Ingenieria electronica[Online]. [Visitada Octubre 27 de 2007]. Disponible en Internet: < http://electronica.udea.edu.co/...pdf

Uso del método de análisis de componentes principales para la caracterización de fincas agropecuarias. [Online]. [Visitada en Octubre de 2007]. Disponible en Internet: < www.ceniap.gov.ve/.../htm >.

WOLFF Carmen Gloria. La Tecnología Datawarehousing. 1999 [online, Articulo]. [Citado el 27 de agosto 2006]. Disponible en Internet: <http://www.inf.udec.cl/revista/ediciones/edicion3/cwolff.PDF> p. 2.

ANEXOS.

Anexo A. DIAGRAMA DE CASOS DE USO



Nombre de Caso de Uso
1. Ver Datos
Descripción:
Este caso de uso es ejecutado por el usuario cuando desea ver los datos que está utilizando, es decir, los datos que están listos para aplicarle las técnicas de minería de datos. Finaliza una vez el usuario cierra la ventana, este caso de uso es de solo lectura.
Flujo Básico:
Inicia desde la opción ver Datos representados por un botón en la ventana principal.
Flujo(s) Alterno(s):
No tiene
Pre-condiciones:
– El usuario debe C*argar Datos* previamente desde un archivo especifico ó con una consulta a la base de datos (Similar a la consulta desde WEKA a la base de datos)
Post-condiciones:

Nombre de Caso de Uso
2. Cargar Datos
Descripción:
Este caso de uso es inicializado por el usuario cuando desea cargar los datos que va a utilizar, es decir, son los datos a los que se le aplicarán las técnicas seleccionadas. Finaliza una vez el usuario termina de usar la aplicación ya que estos siempre estarán disponibles.
Flujo Básico:
Inicia desde la opción Cargar Datos del menú Archivo en la ventana principal.

El usuario selecciona el submenú Archivo de datos y esta abre una ventana para que el usuario seleccione el archivo con extensión .arff (Extencion también compatible con WEKA) el cual contiene los datos a usar.
Flujo(s) Alterno(s):
– Si el archivo no es correcto no se activa ninguna opción anterior
Pre-condiciones:
– Tener el archivo .arff creado y con los datos correspondientes.
Post-condiciones:
– El usuario puede seleccionar la técnica que desea utilizar

Nombre de Caso de Uso
3. Guardar Datos
Descripción:
Este caso de uso es ejecutado por el usuario cuando desea Guardar los datos actualmente cargados en el Prototipo. Finaliza una vez que se termine este proceso.
Flujo Básico:
Inicia desde la opción Guardar Datos del menú Archivo en la ventana principal. Esta opción abre una nueva ventana para que el usuario seleccione la ubicación donde desea guardar el archivo con extensión .arff.
Flujo(s) Alterno(s):
Pre-condiciones:
– Tener cargados los datos previamente.
Post-condiciones:
Nombre de Caso de Uso
4. Agregar Datos

Descripción:
Este caso de uso es ejecutado por el usuario cuando desea Agregar al dataset más datos desde un archivo. Finaliza una vez que se termine este proceso.
Flujo Básico:
Inicia desde la opción Archivo > Archivo de Datos > Agregar datos de Datos en la ventana principal. Esta opción abre una nueva ventana para que el usuario seleccione la ubicación donde está el archivo con extensión .arff con los datos que desea agregar.
Flujo(s) Alterno(s):
– Si el archivo no es correcto no se activa ninguna opción anterior
Pre-condiciones:
– Haber Cargado datos previamente desde un archivo o desde una base de datos
Post-condiciones:
– Los datos nuevos se agregan después de los datos previamente cargados.

Nombre de Caso de Uso
5. Escoger Técnica Weka
Descripción:
Este caso de uso es inicializado por el usuario cuando escoge entre las 5 técnicas de weka disponibles (Análisis de componentes principales, Análisis de clúster, Análisis de regresión, Árbol de decisión, Reglas de asociación). Finaliza una vez el usuario termina de usar la aplicación.
Flujo Básico:
Inicia desde la opción seleccionada del menú Técnicas en la ventana principal.
Flujo(s) Alterno(s):
– Si el usuario no ha seleccionado datos no se activara la opción analizar
Pre-condiciones:
– Tener datos cargados.
Post-condiciones:
– Se inicia el caso de uso clasificar variables

Nombre de Caso de Uso
6. Clasificar Variables
Descripción:
Este caso de uso es inicializado por los casos de uso Escoger Técnica Weka. Finaliza una vez el usuario cierra la ventana.
Flujo Básico:
Una vez se seleccione la técnica que se quiere trabajar el usuario da clic en el botón analizar se abrirá una ventana donde se le solicita al usuario que indique cuales son las variables independientes y cual la dependiente, así como las opciones especificas de cada técnica seleccionada.
Flujo(s) Alterno(s):
– Se activa el botón Analizar
Pre-condiciones:
– Haber escogido alguna técnica previamente
Post-condiciones:
– Si se selecciona el botón analizar se da inicio al caso de uso analizar técnica

Nombre de Caso de Uso
7. Analizar Técnicas
Descripción:
Este caso de uso es inicializado por el caso de uso clasificar variables una vez se seleccione el botón analizar. Se muestran los resultados en una nueva ventana y finaliza una vez el usuario cierra la ventana.
Flujo Básico:
Una vez se seleccione el botón analizar de la ventana clasificar variables se activa el proceso de analizar la técnica correspondiente.
Flujo(s) Alterno(s):
– Se muestra la ventana con los resultados de la técnica seleccionada – Se activa el botón guardar resultados – Se activa el botón agregar resultados
Pre-condiciones:
– Haber activado el botón analizar
Post-condiciones:
Ninguna

Anexo B. AMBIENTE DE DESARROLLO

El Lenguaje de programación seleccionado para el diseño de prototipo computacional es JAVA y para el motor de la base de datos es MySQL. Esta elección se realizó debido a que las dos herramientas están bien posicionadas en cada una de las áreas en las cuales se desempeñan; además, están catalogados como software gratuito, por lo cual el costo del proyecto se disminuye considerablemente.

JAVA

Java es un lenguaje de programación orientado a objetos con el que se puede realizar cualquier tipo de programa. Actualmente es un lenguaje que cada vez cobra más importancia tanto en el ámbito de Internet como en la informática en general. Este lenguaje de programación es desarrollado por Sun Microsystems y siempre enfocado a cubrir las necesidades tecnológicas más recientes.

Una de las principales características por las que Java se ha hecho muy famoso es que es un lenguaje independiente de la plataforma. Eso quiere decir que si hacemos un programa en Java podrá funcionar en cualquier ordenador del mercado. Es una ventaja significativa para los desarrolladores de software, pues antes tenían que hacer un programa para cada sistema operativo, por ejemplo Windows, Linux, Apple, etc. Esto lo consigue porque se ha creado una Máquina de Java para cada sistema que hace de puente entre el sistema operativo y el programa de Java y posibilita que este último se entienda perfectamente.
“La independencia de plataforma es una de las razones por las que Java es interesante para Internet, ya que muchas personas deben tener acceso con

ordenadores distintos” [1]. Pero no se queda ahí, Java está desarrollándose incluso para distintos tipos de dispositivos además del ordenador como móviles, agendas y en general para cualquier cosa que se le ocurra a la industria.

MySQL

MySQL es un motor de bases de datos con diferentes características que hacen de esta una excelente alternativa a otras bases de datos comerciales como por ejemplo Oracle y SQL Server.

MySQL usa el estándar para bases de datos, puede manipulas las bases de datos insertando, borrando, leyendo y actualizando información en las bases de datos previamente creadas.

Dentro de las características de MySQL se encuentra que soporta múltiples usuarios, facilidad de conexión por medio de diferentes lenguajes de programación como C++, C, JAVA, PHP, además cuenta con la opción de protección de media contraseña con lo cual las bases de datos están bastante seguras. Existen muchos tipos de bases de datos, desde un simple archivo hasta sistemas relacionales orientados a objetos. MySQL, como base de datos relacional, utiliza múltiples tablas para almacenar y organizar la información.

MySQL fue escrito en C y C++ y destaca por su gran adaptación a diferentes entornos de desarrollo, permitiendo su ínter actuación con los lenguajes de programación más utilizados como PHP, Perl y Java y su integración en distintos sistemas operativos.

También es muy destacable, la condición de open source de MySQL, que hace que su utilización sea gratuita e incluso se pueda modificar con total libertad,

pudiendo descargar su código fuente. Esto ha favorecido muy positivamente en su desarrollo y continuas actualizaciones, para hacer de MySQL una de las herramientas más utilizadas[33].

Las principales características de este gestor de bases de datos son las siguientes:

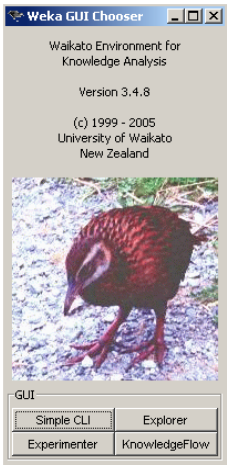
1. Aprovecha la potencia de sistemas multiprocesador, gracias a su implementación multihilo.
2. Soporta gran cantidad de tipos de datos para las columnas.
3. Dispone de API's en gran cantidad de lenguajes (C, C++, Java, PHP, etc).
4. Gran portabilidad entre sistemas.
5. Soporta hasta 32 índices por tabla.
6. Gestión de usuarios y passwords, manteniendo un muy buen nivel de seguridad en los datos

[33]Que es Java – Desarrollo Web [Disponible en Internet]
URL: < http://www.desarrolloweb.com/articulos/497.php >

Anexo C. WEKA

WEKA (Waikato Environment for Knowledge Analysis) es una suite de algoritmos de aprendizaje automático escritos en JAVA, desarrollada por la universidad de Waikato. WEKA es software libre disponible bajo la licencia GNU [34] [35].

Ventana inicial WEKA



[34] GNU, General Public License

[35] WEKA (Machine learning) – Wikipedia, the free encyclopedia. [Online]. [Visitada Marzo 16 de 2008]. Disponible en Internet: < http://en.wikipedia.org/wiki/Weka_(machine_learning) >