Universidad de las Ciencias Informáticas

Faculta 8

Título: Aplicación de la minería de datos para la exploración y detección de patrones delictivos

Trabajo de Diploma para optar por el título de
Ingeniero Informático

Autores: ***Danier Marante Jacas***

Danner Marante Jacas

Tutor: ***Ing. Rafael Y Rodríguez Montero***

DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo al <nombre área > de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

"[Insertar nombre(s) de autor(es)]"

_____

"[Insertar nombre(s) de tutor(es)]"

_____

AGRADECIMIENTOS

Al compañero Javier Ignacio Chávez Barra, por ayudarnos con documentación y guiarnos en el difícil camino del minero.

A nuestro tutor Rafael y todos los compañeros de proyecto, por el apoyo brindado en todo este tiempo.

A nuestros amigos por el apoyo brindado en estos cinco largos años de carrera, especialmente a las jimaguas, a Lily y al resto del piquete.

A Natacha por su ayuda con el documento.

DEDICATORIA

A nuestro padre, por enseñarnos las reglas básicas de la vida.

A nuestro hermano, por dedicarnos toda su juventud.

A nuestra madre y sobrinas, por darnos su amor puro sin esperar nada a cambio.

A Sheila que se ha ganado un lugar especial en nuestros corazones.

Al resto de la familia por el apoyo brindado.

Danier y Danner.

RESUMEN

El uso de la Minería de Datos ha alcanzado una gran popularidad hoy en día. Las principales instituciones y empresas del mundo utilizan las ventajas que estas novedosas técnicas brindan para el análisis de la información. El siguiente trabajo resume la investigación realizada con el objetivo de montar un sistema minero, que analice la información almacenada en la base de datos policial del Cuerpo de Investigaciones Científicas, Penales y Criminalísticas (CICPC) de la República Bolivariana de Venezuela y sea capaz de detectar patrones delictivos. En el mismo se presenta un profundo estudio sobre las técnicas, metodologías y herramientas de Minería de Datos utilizadas en el mundo, con el objetivo de determinar cuáles brindan una mejor solución al problema en cuestión. Se expone además el montaje de dicho sistema y los resultados alcanzados mediante el análisis de un conjunto de datos de prueba facilitados por el cliente.

PALABRAS CLAVE

Data Mining, Minería de Datos, árbol de decisión, clúster, patrón, delito.

INDICE

INDICE DE FIGURAS

INTRODUCCIÓN

Al asumir la primera magistratura de la República Bolivariana de Venezuela el Teniente Coronel Hugo Rafael Chávez Frías, intenta cambiar el sistema de gobierno y la historia de Venezuela, creando una nueva constitución y realizando numerosos cambios en la infraestructura de la República. El 15 de mayo del 2003 se crea el Cuerpo de Investigación Científico Penales y Criminalística de Venezuela (CICPC), esta es una organización que absorbe al antiguo Cuerpo Técnico de Policía Judicial (CPTJ).

El CICPC es una Institución que garantiza la eficiencia en la investigación del delito, mediante su determinación científica; asegurando que el ejercicio de la acción penal conduzca a una administración de justicia. Con la visión de ser la institución indispensable, con la finalidad de alcanzar el más alto nivel de credibilidad nacional e internacional en la investigación del fenómeno delictivo y la criminalidad violenta; además de cumplir el decreto que instituye las competencias del CICPC como órgano principal de investigaciones penales al servicio del Estado; el Gobierno Bolivariano se empeña en superar las deficiencias que presenta en la actualidad.

Esta institución heredó una herramienta automatizada del Cuerpo Técnico de Policía Judicial (CPTJ) llamada SIIPOL (Sistema de Investigación e Información Policial). Esta es una herramienta desarrollada en el lenguaje de programación Natural y cuenta con una base de datos ADABAS (Adaptable Database System). En esta base de datos se encuentra almacenada información delictiva desde la década del 70 del pasado siglo.

Independientemente de las medidas tomadas por el gobierno de Hugo Rafael Chávez Frías el crimen en la República Bolivariana de Venezuela ha experimentado un aumento considerable en los últimos años, debidos fundamentalmente a la actividad de la oposición al movimiento revolucionario que ha estado experimentando Venezuela durante el mandato de Chávez. Al punto de que Caracas, la capital de Venezuela es una de las ciudades más violentas del mundo, en el año 2008, en la capital Caracas, murieron más personas por disparos de armas que en países en guerras como Irak.

Una de las medidas tomadas para contrarrestar esta ola de violencia fue la creación de un nuevo Sistema de Investigación e Información Policial (SIIPOL) el cual estará desarrollado en Java, y contará con una base de datos en ORACLE. Tanto este nuevo sistema como el viejo tratan la información utilizando

técnicas estadísticas, como el delito esta fuera del rango de la estadística lineal podemos decir que no se está aprovechando al máximo el conocimiento almacenado en dicha base de datos en la lucha contra el delito.

En función de utilizar el conocimiento almacenado en los datos de dicha base de datos se propone el siguiente **problema científico**: ¿Cómo aprovechar la información criminal almacenada en la base de datos del Cuerpo de Investigaciones Científico, Penales y Criminalística de Venezuela para la búsqueda de patrones delictivos?

El **objeto de estudio** de esta investigación lo constituyen las Técnicas de Minería de Datos, siendo el **campo de acción** las Técnicas de Minería de Datos aplicables a organismo policiales en la prevención del delito.

Por tal razón se plantea la **idea a defender**: Si se usa una metodología ideal con técnicas de Minería de Datos en la base de datos del Cuerpo de Investigación Científicas, Penales y Criminalísticas se aprovecharía el conocimiento almacenado en la misma para la detección y prevención del delito.

Se definió como **objetivo general** de la investigación realizar un análisis de los métodos y soluciones de Minería de Datos, con el objetivo de determinar cuáles brindan una mejor solución al montaje de una herramienta minera y analizar la información almacenada en la base de datos del Sistema de Investigación e Información Policial (SIIPOL) para detectar patrones delictivos.

Las **tareas** a desarrollar durante la investigación son:

- Hacer un análisis exhaustivo de la Minería de Datos y su aplicación en la detención de patrones delictivos.
- Diseñar un cronograma de tareas por el cual debe regirse la investigación.
- Analizar técnicas de Minería de Datos que ayudan en la búsqueda de patrones.
- Identificar una técnica de Minería de Datos que de solución al problema en cuestión, así como herramientas a utilizar.

- Analizar la información para determinar criterio de semejanza-desemejanza.
- Analizar métodos de agrupamiento de objetos (clúster).
- Elaborar una solución de Minería de Datos para la detección de patrones delictivos en la base de datos del Sistema de Investigación e Información Policial.
- Proponer herramientas de Minería de Datos.

Estrategia de Investigación.

Para la realización de este trabajo, se siguió una estrategia de investigación descriptiva, donde se le da menor importancia a las causas que originan el problema y el principal objetivo es la profundización teórica del planteamiento investigativo, describir el fenómeno, reflejar lo esencial y más significativo del mismo para llegar a los resultados científicos esperados.

Para llegar a un resultado concreto de la investigación, se hizo uso de los siguientes métodos investigativos:

Método Teórico-Histórico: Se hizo un estudio de las causas que originaron el problema, así como un análisis de las técnicas y algoritmos existentes en la actualidad para la Minería de Datos.

Método sistémico: A través este método se realizó el análisis de la solución propuesta mediante los patrones delictivos con el objetivo de facilitar la toma de decisiones de los expertos.

El siguiente trabajo está estructurado en tres capítulos principales.

Capítulo 1. **Fundamentación teórica.** En este capítulo se analizan elementos teóricos tales como: Sistemas similares existentes vinculados a las Técnicas de Minería de Datos aplicables a organismo policiales en la prevención del delito, selección de herramientas y algoritmos que se ajusten a lo requerido, entre otros.

Capítulo 2. **Fundamentación de la propuesta de solución.** En este capítulo se fundamenta la propuesta de solución explicando las técnicas, herramientas y algoritmos que se utilizarán en el proceso de minado.

Capítulo 3. **Resultados.** Este capítulo describe como se realiza el proceso de minado de un conjunto de datos utilizando la herramienta Weka. Se expondrán detalles de diferentes tipos y se analizarán los resultados obtenidos, demostrando que es fiable la utilización de la Minería de Datos para la detección de patrones delictivos

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA.

1.1 Introducción.

Aunque desde un punto de vista académico el término Minería de Datos es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos, en el entorno comercial, así como en este trabajo, ambos términos se usan de manera indistinta. Lo que en verdad hace la Minería de Datos es reunir las ventajas de varias áreas como la Estadística, Inteligencia Artificial, Computación Gráfica, Bases de Datos y el Procesamiento Masivo, usando por lo general como materia prima las bases de datos.

Uno de los principales problemas en la actualidad es la abrumadora cantidad información disponible, muchas empresas e instituciones en los últimos años presentan este problema, dado fundamentalmente por el gran poder de procesamiento de las máquinas y el bajo costo del almacenamiento de la información. Se estima que los datos almacenados en el mundo se duplican cada 20 meses. No es factible tener almacenada una gran cantidad de información ya que lo realmente valioso no es los datos en sí, sino el conocimiento almacenado en ellos. Este fenómeno unido a la enorme cantidad de bases de datos en todas las esferas de la humanidad ha provocado un incremento considerable de la popularidad de los procesos de Minería de Datos a nivel mundial.

Las aplicaciones de la Minería de Datos pueden identificar tendencias y comportamientos, no solo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son evidentes. Una solución de Minería de Datos se basa en la aplicación de diferentes técnicas, donde se destacan las de aprendizaje automático, con el objetivo de encontrar patrones y relaciones entre los datos. Se plantea que usando el SQL tradicional se puede extraer el 80% del conocimiento almacenado en una base de datos, pero el restante 20% que generalmente contiene la información más importante requiere de técnicas más avanzadas.

La Minería de Datos busca generar información similar a la que podría generar un experto humano, que además satisfaga el Principio de la Comprensibilidad. Es el proceso de descubrir conocimientos, como patrones, asociaciones, cambios, anomalías, estructuras significativas entre otros a partir de grandes volúmenes de datos almacenados en bases de datos, almacén de datos (Data Warehouse) o cualquier

otro medio de almacenamiento de la información. Siendo un ambiente pleno de desarrollo donde se aplican métodos de varias disciplinas como los presentes en sistemas de bases de datos, Aprendizaje Automático, estadísticas, visualización de la información entre otras. Además también se utilizan métodos de las áreas de redes neuronales, reconocimiento de patrones, programación lógica inductiva.

Minería de Datos es una tecnología compuesta por etapas que integra varias áreas por lo que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial. Actualmente existen aplicaciones o herramientas comerciales de Minería de Datos muy poderosas, que contienen un sinfín de utilidades que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

1.2 Definición de Minería de Datos.

Existen varias definiciones del término como se muestra a continuación:

- Minería de Datos no trivial es el proceso de identificación válida, novedosa, potencialmente útil y en última instancia comprensible en los patrones de tiempo: (1)

- Minería de Datos se utiliza para descubrir patrones y relaciones de datos, con énfasis en las grandes bases de datos de observación (2).

- Minería de Datos es el análisis de (grandes) conjuntos de datos observacionales para encontrar insospechadas relaciones y para resumir los datos en nuevas formas que son a la vez comprensibles y útiles para el propietario (3).

Desde el punto de vista empresarial se puede definir:

- La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión (3).

- El término Minería de Datos se refiere a un amplio espectro de técnicas de modelado matemático y herramientas de software utilizadas para encontrar patrones en los datos y construir modelos a partir de los mismos (4).

Con la base de las definiciones anteriores se puede definir la Minería de Datos como un proceso no trivial de análisis exploratorio de los datos en el cual se convierte la información en conocimiento. Utilizando entre otras técnicas la inteligencia artificial para detectar patrones y relaciones ocultas entre los datos, permitiendo la creación de modelos.

1.3 Historia y aplicaciones de la Minería de Datos.

La Minería de Datos no es un término nuevo. Desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology.

Con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido, a principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de Minería de Datos y Descubrimiento de conocimiento (KDD). Para los años ochenta solo existían un par de empresas que se dedicaban a la Minería de Datos, en el 2002 se incrementaron a más de 100 empresas que brindan más de 300 soluciones. Actualmente la Minería de Datos se encuentra en todas las esferas de la vida humana.

1.3.1 En la prevención del delito.

En muchos países se utiliza la Minería de Datos con el objetivo de prevenir el delito de una forma u otra. Dada la naturaleza de las aplicaciones de Minería de Datos no es posible crear una metodología aplicable a todas las aplicaciones de Minería de Datos que compartan el mismo objetivo.

A principios del mes de julio de 2002, el director del Federal Bureau of Investigación (FBI), John Aschcroft, anunció que el Departamento de Justicia comenzaría a introducirse en la vasta cantidad de datos comerciales referentes a los hábitos y preferencias de compra de los consumidores, con el fin de descubrir potenciales terroristas antes de que ejecutarán una acción.

A partir de la crisis del 2001 en argentina se desencadenó una ola de violencia e inseguridad social. Como respuesta a esto se creó el Sistema de Alerta Temprana (SAT) donde se realizó una aplicación de Minería de Datos con el objetivo de prevenir, crear metodologías y técnicas eficientes para la lucha contra el crimen.

1.3.2 En la empresa.

Detección de fraudes en las tarjetas de crédito.

En 2001, las instituciones financieras a escala mundial perdieron más de 2.000 millones de dólares estadounidenses en fraudes con tarjetas de crédito y débito. El Falcon Fraud Manager es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para detectar y mitigar fraudes.

Prediciendo el tamaño de las audiencias televisivas.

La British Broadcasting Corporation (BBC) del Reino Unido emplea un sistema para predecir el tamaño de las audiencias televisivas para un programa propuesto, así como el tiempo óptimo de exhibición.

1.3.3 En la universidad.

Se hizo un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II, en México, con el objetivo de observar si sus recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba conocer el perfil que caracterizaba a los ex alumnos durante su estancia en la universidad. El objetivo era conocer si con los planes de estudio de la universidad y el aprovechamiento del alumno se hacía una buena inserción laboral o si existían otras variables que participaban en el proceso.

1.3.4 En investigaciones espaciales.

Proyecto SKYCAT. Durante seis años, el Second Palomar Observatory Sky Survey (POSS-II) coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23.040 x 23.040 píxeles por

imagen. El objetivo era formar un catálogo de todos esos objetos. El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación (clúster) y árboles de decisión para la clasificación de los objetos en estrellas, planetas, sistemas, galaxias, etc. con una alta confiabilidad.

1.3.5 En los clubes deportivos.

El AC de Milán utiliza un sistema inteligente para prevenir lesiones. Este club usa redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta. Esto ayuda a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible lesión.

Los equipos de la NBA utilizan aplicaciones inteligentes para apoyar a su cuerpo de entrenadores. El Advanced Scout es un software que emplea técnicas de Minería de Datos, fue desarrollado por investigadores de IBM para detectar patrones estadísticos y eventos raros. Tiene una interfaz gráfica muy amigable orientada a un objetivo muy específico: analizar el juego de los equipos de la National Basketball Association (NBA). El software utiliza todos los registros guardados de cada evento en cada juego: pases, encestes, rebotes y doble marcaje (double team) a un jugador por el equipo contrario, entre otros. El objetivo es ayudar a los entrenadores a aislar eventos que no detectan cuando observan el juego en vivo o en película.

1.4 Fundamentos de la Minería de Datos.

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

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

- Algoritmos de Minería de Datos.

1.5 Características de la Minería de Datos.

Existen muchas características que definen la Minería de Datos, entre ellas se encuentran:

- Los datos a explorar se encuentran en las profundidades de las beses de datos y en ocasiones contienen información almacenada durante varios años y recolectadas de diferentes fuentes lo cual puede provocar que la información se encuentre con ruido.

- En ocasiones los datos no se encuentran almacenados en un solo lugar.

- El entorno de la Minería de Datos suele tener una arquitectura cliente servidor.

- Las herramientas de la Minería de Datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados

- El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias.

- Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.

- Las herramientas de la Minería de Datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.

- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.

- La Minería de Datos produce cinco tipos de información:

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

- Pronósticos.

La Minería de Datos es un proceso que invierte la dinámica del método científico en el siguiente sentido: En el método científico, primero se fórmula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis obteniendo de esta forma un nuevo conocimiento. En la Minería de Datos, se coleccionan los datos y se espera que de ellos emerjan hipótesis. Se busca que los datos describan o indiquen por qué son como son. Luego entonces, se valida esa hipótesis inspirada por los datos.

Los datos mismos, será numéricamente significativos, pero experimentalmente inválidos. De ahí que la Minería de Datos debe presentar un enfoque exploratorio, y no confirmador. Usar la Minería de Datos para confirmar las hipótesis formuladas puede ser peligroso, pues se realiza una inferencia poco válida.

1.6 Fases del proceso de Minería de Datos.

Los pasos a seguir para la realización de un proyecto de Minería de Datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

El proceso de Minería de Datos pasa por las siguientes fases:

- Filtrado de datos.
- Selección de Variables.
- Extracción de Conocimiento.
- Interpretación y Evaluación.

1.6.1 Filtrado de datos.

El formato de los datos contenidos en los contenedores de datos, ya sean bases de datos o Almacenes de datos nunca es el idóneo y en ocasiones ni siquiera es posible aplicarle algún algoritmo de Minería de Datos sobre estos datos en bruto. El proceso de filtrado o preparación de los datos es un proceso previo a

la Minería de Datos, vital para obtener buenos resultados, aunque suele ser costoso en ocasiones más que el proceso de análisis.

En el filtrado de datos suelen hacerse las siguientes tareas.

- Detección de valores anómalos.
- Eliminación de datos incorrectos.
- Manejo de los datos faltantes.
- Normalización.
- Clúster.

1.6.2 Selección de las variables.

Aún después de haber sido pre- procesados, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema.
- Los que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

1.6.3 Algoritmos de Extracción de Conocimiento.

Mediante una técnica de Minería de Datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de

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

1.6.4 Interpretación y evaluación.

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

1.7 Tratamiento de datos.

Los datos se almacenan en bases de datos en formato crudo, ya que la idea inicial es almacenarlo para consultarlos en el futuro, aunque estadísticamente se ha demostrado que en la mayoría de las empresas más del 60% de la información de sus bases de datos nunca es consultada.

En muchos proyectos de Minería de Datos esta etapa es aparentemente omitida. Lo que realmente sucede es que la fuente de los datos es un almacén de datos y en los procesos de construcción del almacén de datos se realiza un proceso similar de depuración de los datos.

Para que este proceso termine exitosamente se realizarán una serie de pasos.

- Tratamiento de valores omisos.
- Codificación numérica para valores categóricos (normalización).
- Eliminación de datos incorrectos.
- Pre-procesamiento de datos.

1.7.1 Tratamiento de valores omisos.

Es muy frecuente encontrar en las bases de datos actuales registros incompletos (valores nulos). La no existencia de valor es tan importante para la Minería de Datos como la existencia, esto es debido fundamentalmente por el comportamiento humano, un ejemplo claro de esto sería: si a un delincuente se le pregunta los ingresos anuales es muy probable que no conteste a esa pregunta.

En algunos casos no es necesario tomar esta información en cuenta, en estos casos podemos:

- No utilizar los registros que presentan valores omisos.
- No utilizar campos que presentan valores omisos.

En el segundo caso el peligro es relativamente obvio, se traduce en no utilizar variables importantes para la formulación del modelo explicativo. Podemos estar prescindiendo de variables que en realidad son importantes para modelar el fenómeno, siendo este el caso, el modelo probablemente no tenga la precisión que se lograría si se utilizara dicha variable.

La primera opción suele ser aún más peligrosa por ser más insidiosa. El hecho de que determinados registros no presenten valores puede traducirse en un aspecto importante. Al omitir estos registros se corre el riesgo de invertir la muestra, como consecuencia los modelos producirán siempre estimaciones poco precisas con señalamientos aleatorias para los individuos que presentan estas características. En estos casos hay que tomar en consideración que la falta de información en ocasiones es información.

Una posible solución podría ser revisar manualmente los casos que poseen valores omisos e introducirles valores manualmente de acuerdo al perfil de los registros. Esa solución no es práctica cuando los valores omisos son muchos. Además si no se introduce el valor opimo en los registros se corre el riesgo de introducir ruido en el conjunto de datos.

Otra alternativa para solucionar este problema es la predicción automática de los valores con una estimación de ellos. Hay varias técnicas para la de estimación de valores, una de las más obvias y simples consiste en adoptar una media a la tendencia central. Este método es simple, consiste en calcular el valor medio e introducir este valor en los campos que presentan campos con valores omisos.

Esta solución tiene un alto grado de riesgo ya que el valor medio no es siempre una solución óptima, introduciendo de esta forma ruido en el conjunto de datos.

Existen ocasiones en que la ausencia de datos puede ser un indicador importante en los procesos de extracción de conocimiento. En estos casos específicos los valores omisos serán codificados por otro posible valor para esta variable. Es decir, si se tiene un campo con posibles valores A, B, C y D en esta ocasión los valores omisos serán sustituidos por el valor E, esta nueva categoría será usada para categorizar valores omisos. Es importante resaltar que estos valores, la ausencia de datos es un indicador, son determinados por los especialistas en delito, el papel de los especialistas en esta etapa del desarrollo de un proyecto de Minería de Datos es de vital importancia.

Otra forma más sofisticada, también más trabajosa es crear un sistema predictivo que con una base de registros completos de las variables disponibles, lo constituye la estimación de los valores omisos. Supongamos que se tiene un modelo que presenta tres variables de entrada A, B y C para todas existen valores omisos. Una solución interesante consistiría en presentar un problema regresivo Figura 1. De esta forma se usan todos los registros completos de A y B para estimar los de C, luego los de A y C para estimar los de B y finalmente los de B y C para estimar los de A.

Figura 1. A través de los valores completos de A y B se estiman los valores omisos de C y así sucesivamente se repite el proceso para estimar los valores omisos de A y B.

Existen varios aspectos que hay que tener en cuenta para usar este método. Si el valor de la correlación es muy elevado, debe tomarse como que el valor presente en la variable dependiente se encuentra codificado ya en las variables independientes. Digamos que A y B son las variables independientes de C y la correlación entre estas variables es elevado un ejemplo de 0.95 entonces la variable C no debe ser tomada como variable de entrada para el modelado.

1.7.2 Codificación numérica para valores categóricos.

Una característica común de las herramientas de Minería de Datos, es que funcionan mejor con valores numéricos. Para lograr resultados óptimos es necesario pasar toda la información posible en forma númerica.

Los valores categóricos se puede pueden dividir en tres grupos:

- Variables ordinales.
- Variables binarias.
- Variables nominales.

Las variebles ordinales son aquellas que poseen valores con ordenamiento secuencial lógico. Un ejemplo seria una varíable que almacenes las categorías de tamaño(alto, medio, bajo).

Las variables binarias son aquellas que solo poseen dos valores.

Las variables nominales son aquellas que poseen un número finito de valores posibles. Un ejemplo puede ser una variable que almecena los colores(blanco , azul, verde).

10.7.2.1 Codificación (one of n).

Sea un número n de categorías relativamente pequeño, se puede codificar en n variables binarias. De esta forma cada nueva variable representa una única categoría asumiedo 1 como el valor original de esa categoría y 0 el caso contrario. Esta forma de codificación preserva la naturaleza independiente a cada valor que la variable asume, o sea no asume cualquier valor predeterminado. En la tabla a continuación se muestra el proceso de codificación.

Individuo	Ciudades	Var1	Var2

1	Caracas	0	1
2	Cumaná	1	0
3	San Carlos	1	0
4	Trujillo	0	1

Figura 2. Ejemplo de tabla de codificación.

Si el número de categorías n es demasiado grande se puede usar alguna técnica de agrupamiento para reducir el número de categorías y de ese modo poder usar el método one of n. Este método consiste en tener conocimiento del dominio al que pertenecen las categorías para proceder al agrupamiento manual de las categorías. Un ejemplo de esto sería codificar una variable que almacene todos los municipios del oriente de Cuba. Esto lo podemos hacer creando una variable por provincia y luego crear una serie de variables igual a la cantidad de municipios provincias que más municipos tenga y de esta forma se reduce considerablemente el número de categorías.

1.7.3 Eliminación de datos incorrectos y detección de datos anómalos.

La preparación de los datos debe incluir un análisis rápido de los datos para identificar y eliminar los registros incorrectos o resolver las incoherencias que pueden existir.

En el proceso de corrección de datos es importante tratar los datos “anómalos” (estos son los datos de la región fuera del interés de la zona de entrada). Es importante resaltar que estos casos aunque sean aislados pueden provocar resultados catastróficos a la hora del proceso de modelado.

Según esta definición los datos anómalos no necesariamente representan datos incorrectos. Los datos que son anómalos son datos distantes de los demás, son casos extremos en una o más dimensiones por lo que pueden tener un gran impacto en las configuraciones origen de los valores anómalos pueden ser

errores de las bases de datos aunque también pueden estar relacionados con la existencia de registros que son reales.

La dificultad de detección de estos registros es un proceso variable, si el registro se considera anómalo por tener un valor específico entonces es fácil su detección, sin embargo si el valor es una combinación atípica de varias variables, entonces la detección de estos registros puede ser complicada. Por lo que el aumento de la direccionalidad es proporcional al aumento de la dificultad de identificar un registro anómalo.

Unos de los métodos usados para la detección de valores anómalos son los métodos gráficos, como la "caja de parcela". También se pueden utilizar diferentes mecanismo como la técnica de la limitación, esta técnica es muy rápida y se basa en poner un máximo y un mínimo para acotar los valores, también suele llamarse "encuadramiento", si el valor entrante sobrepasa esos límites entonces es tratado como anómalo. Este método es extremadamente simple y útil. Básicamente lo que se hace es graficar un conjunto de datos, poniendo una especie de máximo y mínimo para detectar los valores que están fuera del rango lógico. Este método solo es aplicables cuando la direccionalidad es baja, cuando la gráfica generada por el conjunto de datos pasa de dos direcciones esta se complica y deja de ser expresiva, de esta forma no nos brinda casi o ninguna información.

Es importante resaltar que algunas de las herramientas de Minería de Datos ya traen implícitos métodos para el tratamiento de valores anómalos, un ejemplo de estas herramientas es SAS Enterprises Miner.

Cuando el nivel de complejidad aumente pueden que a las técnicas anteriores valores anómalos pasen desapercibidos, una técnica eficiente para solucionar este problema es el uso de los gráficos de dispersión.

1.7.4 Pre-procesamiento de datos.

El objetivo general de la fase de pre-procesamiento de datos es facilitar y simplificar el problema en cuestión, sin excluir o dañar información importante para el proceso de modelado. Esta etapa se puede dividir a gran escala en dos partes.

- Reducción de espacio de entrada.

- Trasformación de variables.

Es evidente que el pre-procesamiento de los datos es algo paradógico,ya que se quiere reducir la complejidad del problema, lo que se resume en reducir el área de entrada del conjunto de datos, pero a la vez se debe hacer sin dagnificar o excluir información importante.

Un término muy importante en el proceso de pre-procesamiento de datos es el espacio de input(cantidad de variables de entrada). Las dimensiones del espacio de entrada aumenta exponencialmente con el incremento de las dimensiones del problema(números de variables de entrada).

En esta etapa el minero necesita ser muy cuidadoso ya que errores como un gran números de variables de entrada combinado con poca información para minar, puede traer como consecuencia que se produzcan correlaciones erróneas entre variables de entrada y variables de salida, conduciendo el modelo a variables que supuestamente tienen gran importancia cuando en la realidad no la tienen. La disponibilidad de grandes cantidades de datos permite reducir la probabilidad de ocurrencias de estos problemas.

Los procesos de modelado que se utilizan en la Minería de Datos son internamente tareas de búsqueda, mientras más grande es el espacio de entrada mayor es la complejidad de busca del modelo óptimo. El proceso de pre-procesamiento de datos ayuda a reducir el espacio de entrada aumentando de esta forma el performance del modelado mediante la simplificación del proceso de búsqueda.

Con la disminución del área de entrada se pueden lograr mejoras significativas, sobre todo si el volumen de los regitros disponibles es limitado. La cantidad de registros disponibles condicionan el performance del resultado final de forma drástica. Tomando en cuenta que estos modelos aprenden con la experiencia que les pueda brindar el conjunto de datos, cuanto mayor sea la dimensionalidad del problema mayor será las necesidades de experiencia para brindar resultados óptimos.

1.7.5 Normalización de los datos.

El proceso de normalización es una de las tareas más importante en el proceso de tratamiento de datos. Esto se debe fundamentalmete a que la mayoría de los modelos que se utilizan en la Minería de Datos asumen implícitamente que la distancia en las diferentes direcciones del espacio de entrada tiene la

misma importancia. Un ejemplo de esto es en el proceso de análisis de clúster, el algoritmo k-means utiliza la distancia Euclidiana, esto depende del hecho de que las entradas se encuentren en escalas idénticas. Otros algoritmos como los de retro-propagación (RP) que utilizan redes neuronales multicapas funcionan de forma más eficientes cuando el conjunto de datos se encuentra correctamente normalizado.

Normalización Min-Max.

Este es un proceso de trasformación lineal para valores de una amplitud deteminada, en una escala menor (generalmente 0 y 1). Este proceso consiste en que si se toma la nueva escala 0 – 1 el entonces el min1 se trasforma en un min2 que sería en este caso 0 y el max1 se trasforma en el max2 que sería 1, y todos los demás números entre el min1 y el max1 son trasformados a las nueva escala que esta entre el min2 y el max2. Por ejemplo para el siguiente juego de datos 1000, 1100 , 1500 y 2000, estan en una escala predeterminada entre 1000 y 2000. Pasándolos a la nueva escala 0 – 1 quedarían de la siguiente forma: 0 , 0.1 , 1.5 y 1.

La fórmula matemática para el proceso de normalización max-min sería:

$$y' = \left(\frac{y - \min 1}{\max 1 - \min 1} \right) (\max 2 - \min 2) + \min 2$$

Donde “y” es el valor orginal, y´ el nuevo valor min1, max1 representan el mínimo y máximo de la escala inicial y min2, max2 representan mínimo y máximo de la nueva escala.

Una de las ventajas de este proceso es que preserva intacta las relaciones entre valores.

Normalización z-score

Es un proceso que trasforma los valores de las variables de entrada de forma tal que la media sea 0 y la varianza sea 1. El primer paso es calcular la media y la varianza (desviación estándar). Luego se procede a restar a cada entrada el valor de la media y dividir por la desviación estándar.

La fórmula es:

$$y' = \frac{y - média}{DE}$$

Donde “y” es el valor original y y’ el nuevo valor, de desviación estándar(varianza).

Este método es mucho más complejo que en Min-Max pero funciona particularmente bien cuando no se conocen los valores máximos y mínimos de la variables de entrada y en estos casos no es posible aplicar el método Min-Max.

La normalización sigmoidal.

Otra forma de normalización más compleja pero útil es la normalización sigmoidal. Este método realiza una trasformación no lineal de los datos de entrada en una escala de -1 a 1 utilizando una función sigmoidal.

La fórmula para este método es:

$$y' = \frac{1 - e^{-a}}{1 + e^{-a}}$$

Donde y’ es el nuevo valor y “a” es el resultado después de aplicar z-score.

Este método es generalmente bueno cuando existen valores extremos. Evita que los valores del rango de la media se compriman demaciado sin perder la capacidad de detectar como anómalos a los valores muy elevados.

La mayor diferencia de los tres métodos anteriormente expuestos es que el método signoidal trasforma en valores mucho más elevados que los dos métodos expuestos anteriormente. Pero es válido resaltar que el nivel de complejidad de este método es mucho más elevado que el de los anteriores. Es importante comprender que como en la Minería de Datos se trabaja con grandes conjuntos de datos, un milisegundo ganado en una operación, luego de analizar un diez millones de registros, es decir hacer esa misma operación diez millones de veces nos ahorramos 2.7 horas.

1.8 Tareas realizadas en el proceso de Minería de Datos.

A continuación se describirán las tareas a realizar en un proceso de Minería de Datos, es válido resaltar que para que un proceso de minería funcione correctamente no es necesario que cumpla con todas estas tareas, depende mucho de la naturaleza del problema en cuestión.

- Descripción de las clases: Es una clasificación resumida de un conjunto de datos y sus diferencias. El proceso de clasificación de los datos se conoce como caracterización y distinción entre ellos como comparación.

- Asociación: Es el descubrimiento de relaciones de asociación o correlaciones entre los datos. Las asociaciones se expresan como relaciones atributo-valor.

- Clasificación: Analiza un conjunto de datos cuya clasificación de clase se conoce y construye un modelo de objeto para cada clase. Dicho modelo suele representarse con un árbol de decisión o reglas de clasificación que muestran las características de los datos.

- Predicción: Es la función de la minería que predice los valores posibles de datos faltantes o la distribución de valores de ciertos atributos en un conjunto de objetos.

- Clustering: Identifica clúster en los datos, donde un clúster es una colección de datos “similares”. La similitud puede medirse mediante funciones de distancia, especificadas por los usuarios o por expertos.

- Análisis de series a través del tiempo: Analiza un gran conjunto de datos obtenidos con el correr del tiempo para encontrar en él regularidades y características interesantes, incluyendo la búsqueda de patrones secuenciales, periódicos, modas y desviaciones.

1.8 Métodos de Minería de Datos.

La Minería de Datos abarca un área muy extensa, no es solamente aplicar algoritmos existentes a un conjunto de datos. Las herramientas existentes actualmente incluyen mecanismos de depuración, preparación y normalización de los datos. Además brindan funcionalidades de visualización e

interpretación de los resultados. Muchas de estas herramientas funcionan bien en espacios de pocas dimensiones y con datos numéricos, pero sus limitaciones empiezan a aparecer en espacios con muchas dimensiones y con datos no numéricos. A continuación se presentan algunos métodos de Minería de Datos que resuelven distintos problemas inherentes a la misma.

- Aprendizaje activo/Diseño Experimental (Active Learning/Experimental design): el aprendizaje activo, por el lado de la Inteligencia Artificial, y el diseño experimental, por el lado de la Estadística, tratan de resolver el problema de la elección del método a aplicar durante el aprendizaje. Suponen que durante el proceso de aprendizaje, existe la oportunidad de influir sobre los datos, recordemos la diferencia entre la exploración pasiva y la experimentación activa. El aprendizaje activo afronta el problema de cómo explorar. En este caso el método descrito no es viable, ya que no existe ninguna necesidad de cambiar el proceso de aprendizaje en tiempo de ejecución.

- Aprendizaje acumulativo (Cumulative learning): Muchas bases de datos crecen continuamente. Si se toma por ejemplo, una base de datos sobre transacciones financieras en un banco. Aprender a partir de bases de datos de este tipo es difícil, ya que los datos deben ser analizados acumulativamente a medida que se incorporan a la base. Nos encontramos entonces ante el desafío de diseñar algoritmos que puedan incorporar nuevos datos y adaptarse a los cambios generados por la incorporación de los mismos. Este método es aplicable fundamentalmente en problemas bancarios donde la tasa de transacciones es extremadamente alta. En este caso el crecimiento de la información almacenada en la base de datos no es lo suficientemente elevado como para aplicar este método.

- Aprendizaje multitarea (Multitask learning): Muchos dominios se caracterizan por pertenecer a familias de problemas de aprendizaje relacionado o similar, si se toma por ejemplo el dominio médico. Mientras que cada enfermedad posee su aprendizaje individual con bases de datos dedicadas, muchas enfermedades tienen causas y/o síntomas en común, sería provechoso entonces favorecer el intercambio de información entre los distintos resultados de los algoritmos. Este método no es aplicable al problema en cuestión, ya que toda la información se encuentra almacenado en una base de datos central.

- Aprendizaje relacional (Relational Learning): En muchos problemas de aprendizaje las entidades no se describen a partir de un conjunto estático de atributos, sino a partir de las relaciones entre entidades. En las bases de datos inteligentes encontrar patrones o relaciones entre entidades es un problema primordial. Este es el método que se utilizará en la solución, ya que la nueva solución, el Sistema de Investigación e Información Policial (SIIPOL) , cuenta con la base de datos en el gestor Oracle10g que es un potente gestor de bases de dato relacional, la base de datos se encuentra diseñada en segunda forma normal. Esto provoca que mucha de la información importante para el proceso de aprendizaje se encuentre en las relaciones de las tablas.

1.9 Componentes de Minería de Datos.

La Minería de Datos cuenta con grandes componentes como clúster o clasificación, Reglas de Asociación y Análisis de Secuencias.

1.9.1 Análisis de clústers

Análisis de clúster es la nominación genérica atribuida a la gran variedad de metodologías utilizadas para la clasificación de entidades. Estas metodologías construyen grupos de entidades semejantes entre sí. Más específicamente el análisis de clúster es un conjunto de algoritmos, que a partir de un conjunto de información sobre un conjunto de objetos (entidades, individuos, ejemplos etc.) procura organizarlos en grupos homogéneos tanto como sea posible, determinado una estructura de semejanza-desemejanza.

El análisis de clúster es un método que permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori, pero que pueden ser útiles una vez que se han encontrado. Los resultados de un Análisis de Clúster pueden contribuir a la definición formal de esquemas de clasificación tal como una taxonomía para un conjunto de objetos, sugerir modelos estadísticos para describir poblaciones a asignar nuevos individuos a la clase para diagnóstico e identificación.

Es importante diferenciar el clúster de la clasificación ya que en la Minería de Datos estas palabras se utilizan para describir tareas diferentes. La clasificación se inicia con un conjunto de datos preclasificado, es decir se cuenta con un conjunto de datos que no solo se conoce las variables a utilizar en la clasificación, sino también se conoce las clases a las que pertenecen estas variables. El objetivo de la

clasificación es sobre la base de esta información, para crear un modelo capaz de predecir la clase de un nuevo registro. En el caso del análisis de clúster no se parte con un conjunto de datos preclasificados, se crean grupos de de objetos que tengan un comportamiento semejantes entre sí.

1.9.1.1 Características de los algoritmos de análisis de clúster.

1. Capacidad de tratar con grandes bases de datos, en el orden de los millones de registros, ya que no todos los algoritmos tienen una performance correcta cuando se trate de un número considerable de registros.
2. Capacidad para tratar con diferentes tipos de datos, ya que muchos de los algoritmos existentes se crearon para manejar cierto intervalo de datos, y no siempre cumple con las expectativas necesarias para realizar un correcto análisis de clúster.
3. Capacidad de tratar clúster de formas arbitraria, muchos de los algoritmos buscan agrupaciones utilizando como base de medida la Euclidiana o de Manhattan, lo que provoca que han tenido que encontrar agrupaciones esféricas de tamaño y densidad similares, sin embargo un grupo puede tomar cualquier forma, y por tanto la necesidad de desarrollar algoritmos capaces de detectar grupos de formas arbitrarias. En el caso del análisis de clúster no existen datos preclasificados, se crean grupos de objetos que sean semejantes entre sí.
4. Ser capaces de funcionar con un mínimo de conocimiento en el terreno, para poder determinar parámetros de entrada, ya que muchos algoritmos dependen de diferentes valores como el número de registros a formar entre otros; el problema es que estos parámetros de entrada los debe determinar el usuario y estos algoritmos pueden ser muy sensibles a los valores iniciales, lo que trae como consecuencia que se vea obstaculizada la labor del usuario y que sea difícil controlar la calidad del resultado.
5. Capacidad de liderar con ruido.
6. Capacidad de liderar con objetos de diferentes diseños. En las actuales Bases de Datos se manejan diferentes diseños de objetos con elevado número de atributos, y algunos algoritmos tienen un funcionamiento correcto en espacios de pequeñas dimensiones

1.9.1.2 Fases del análisis de clúster.

En muchos casos el proceso el análisis de clúster es reducido a la cuestión del análisis de los algoritmos de agrupación. Sin restarle importancia los algoritmos son solo una de las series de medidas para garantizar que los resultados sean útiles y fiables.

Las cuatro etapas fundamentales en el proceso de análisis de clúster.

1. Definición del conjunto de variables sobre las cuales se evaluará la similitud-disimilitud de las entidades.
2. Definir el criterio de Semejanza-Desemejanza entre entidades.
3. Aplicar el algoritmo de clúster.
4. Analizar y validar la solución.

La calidad del resultado final del proceso de conglomerado (clúster) se debe fundamentalmente a las variables seleccionadas para el proceso de análisis. Esto hecho conduce a una nota importante, que es la importancia del conocimiento del analista en el tema en cuestión. Esto es obviamente es un factor decisivo en el éxito de la aplicación de dicha metodología.

1.9.1.3 Criterios de Semejanza-Desemejanza entre entidades.

Este criterio es determinado por el software que se utiliza. Existen diversos métodos para medir el criterio de semejanza-desemejanza entre objetos. Las medidas basadas en el espacio euclidiano han dominado el análisis de las relaciones de similitud. Este método representa los objetos como puntos en el espacio multidimensional de manera tal que la desemejanza entre los objetos se representa con la distancia métrica entre los respectivos puntos. Es importante resaltar que estos índices de desemejanza respetan las propiedades métricas que son:

Simetría: dados dos objetos x e y con $X \neq y$, la distancia entre ambos objetos cumple la siguiente propiedad.

$$d(X,Y) = d(Y,X) > 0$$

Desigualdad triangular: dado tres objetos X, Y y Z la distancia entre ellos cumple con la siguiente propiedad.

$$d(X,Y) \leq d(X,Z) + d(Z,Y)$$

Diferencia entre los no idénticos: dado dos objetos X e Y.

$$d(X,Y) \neq 0 \Rightarrow X \neq Y$$

No diferencia entre los idénticos: sean dos objetos X e Y.

$$d(X,Y) = 0 \Rightarrow X = Y$$

1.9.1.4 Medidas geométricas.

Distancia Euclidiana.

Una de las medidas geométricas más utilizadas es la distancia Euclidiana, donde la distancia entre dos elementos (i,j) está dada por la raíz cuadrada de la sumatoria de los cuadrados de la diferencia entre los valores i , j de todas las variables (v 1,2,3, … p).

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

Si para cada variable se le asigna un peso de acuerdo a su importancia entonces la distancia euclidiana quedaría de la siguiente forma.

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

Distancia de Manhattan.

Esta es una de las medidas de distancia más utilizadas.

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

Distancia de Minkowski.

La distancia de Minkowski, definida a partir de la distancia absoluta, no es más que una generalización de la distancia Euclidiana y de Manhattan**.** Coincide con la distancia euclidiana cuando r = 2 y con la de Manhattan cuando r =1.

$$d_{ij} = \left(\sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right|^{r} \right)^{1/r}$$

Coeficiente de correlación de Paerson.

Este es un criterio se semejanza muy popular, especialmente cuando el problema se trata de ciencias sociales. El coeficiente de correlación de Paerson o coeficiente de correlación permite medir el grado de asociación lineal entre dos elementos.

$$r_{ij} = \frac{\sum_{v=1}^{p} \left(X_{iv} - \overline{X_i} \right)\left(X_{jv} - \overline{X_j} \right)}{\sqrt{\left(X_{iv} - \overline{X_i} \right)^2 \left(X_{jv} - \overline{X_j} \right)^2}}$$

Donde Xiv es el valor de la variable “v” para el individuo i v= (1, 2,3,…p); Xjv es el valor de la variable “v” para el individuo j, Xi es la media de todas las variables para el individuo i; *Xj* es la media de todas las variables para el individuo *j*; *p* es el número de variables.

1.9.1.5 Métodos de partición u optimización.

Estos son los métodos más utilizados en el área de la Minería de Datos. Las técnicas de optimización se basan en el criterio de agrupar n objetos en k clúster predeterminados. La optimización ya sea por maximización o minimización plantea que cada objeto pertenece a un único clúster.

Los métodos de partición u optimización plantean que:

Para una base de datos con n objetos, un método de partición construye k particiones donde cada partición representa a un clúster k < n satisfaciendo los siguientes criterios.

- Cada grupo contiene por lo menos un objeto.
- Cada objeto pertenece a un único clúster.

Dado k números de particiones a construir, los métodos de partición crean una partición inicial, luego utilizando una técnica interactiva de reubicación se procede a mejorar las particiones moviendo objetos de un grupo para otro. Los métodos de partición se basan en el criterio de que los objetos pertenecientes a un clúster tienen que ser semejantes, es decir tienen que encontrarse cercanos unos a otros. Mientras que los objetos que forman otros clúster se encuentran alejados de ellos.

Para obtener resultados óptimos con los algoritmos de partición se necesitaría hacer un análisis exhaustivo donde se tendrían que hacer una enumeración de todas las enumeraciones posibles, lo cual es inconcebible. Por eso la mayoría de los algoritmos adoptan una de las dos medidas heurísticas que son muy populares.

Algoritmo k-means donde cada clúster está representado por la media de los objetos que lo forman.

Algoritmo k-menoids donde cada clúster está representado por un objeto situado cerca del centro de la agrupación.

Estos métodos de análisis de clúster heurísticos funcionan bien cuando se forman clúster esféricos en grandes bases de datos. Para encontrar clúster de forma más complejas se necesitan hacerle algunos ajustes a estos algoritmos.

1.9.1.6 Algoritmos k-means

El algoritmos k-means recibe como parámetro de entrada “k” y procede a dividir en n objetos en “k” grupos, de forma tal que garantiza una elevada semejanza intra-clúster y desemejanza inter-clúster. La similitud entre los grupo se mide desde el punto medio de los grupos, que puede ser visto como el centro de gravedad de los clúster.

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

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

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

$\overline{X_j}$ Es el valor medo de la variable en el grupo j.

nj Es la dimensión del grupo j.

La media global del clúster se obtiene mediante la suma de los cuadrados de la diferencia entre la media de cada grupo j y la media global de las variables.

$$\sum_{j=1}^{m}\left(\overline{X}_j-\overline{X}\right)^2,$$

Donde: $\overline{X}_j$ es el valor medio de la variable X para el grupo j.

$\overline{X}$ Es el valor medio de la variable X para toda la población.

m: Es el número de grupos.

De hecho se puede demostrar que la suma de los cuadrados de la diferencia entre el individuo i del grupo j y la media global, puede ser dividida por la suma de los cuadrados de las diferencias de cada individuo y la media del grupo j más la suma de los cuadrados de la diferencia de la media de cada grupo j y la media global de los objetos.

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(X_{ij}-\overline{X}\right)^2=\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(X_{ij}-\overline{X}_j\right)^2+\sum_{j=1}^{m}n_j\left(\overline{X}_j-\overline{X}\right)^2$$

Obtener mediante un proceso iterativo una combinación de objetos en k grupos donde se minimice la variabilidad intra-grupo y maximice la variabilidad inter-grupo.

Cuando estamos en presencia de un espacio multidimensional, es decir, en los casos en que los objetos de la muestra se caracterizan por más de una variable p> 2, siendo p el número de variables, la lógica es la misma con la diferencia que se le añade la variabilidad a cada variable. La variabilidad total está dada por:

$$T=\sum_{i=1}^{n}\sum_{v=1}^{p}\left(X_{iv}-\overline{X}_v\right)^2$$

Donde: Xix es el valor de la variable v para el individuo i.

$\overline{X}_v$ Es el valor de la media para la variable v.

La variabilidad intra-grupo se obtiene:

$$W_j = \sum_{i=1}^{n_j} \sum_{v=1}^{p} \left(X_{imv} - \overline{X}_{mv} \right)^2$$

Donde: Ximv es el valor de la variable v del individuo i en el grupo m.

$\overline{X}_{mv}$ Es el valor medio de la variable v para el grupo m.

La variabilidad total intra-grupo está dada por:

$$W = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \sum_{v=1}^{p} \left(X_{ijv} - \overline{X}_{jv} \right)^2 ,$$

La variabilidad inter-grupo está dada por:

$$B = \sum_{j=1}^{m} \sum_{v=1}^{p} \left(\overline{X}_{jv} - \overline{X}_{v} \right)^2$$

En la práctica el algoritmo k-meas funciona de la siguiente forma:

- Inicialmente se seleccionan k objetos, donde cada uno sería inicialmente el centro de cada clúster.
- Cada uno de los restantes objetos se asignan a la agrupación con la que es más similar, es decir es asignada al grupo cuyo centroide es el más cercano.

- Una vez terminada el proceso de asignar los objetos a los grupos se procede a recalcular los centroides de los clúster.
- Este proceso continúa hasta que los grupos sean lo más compactos internamente y se encuentren lo más separado posible. Esto se puede determinar con la función objetivo.

Ventajas del algoritmo k-means.

- Velocidad, la cual puede ser considerable cuando se trata de grandes volúmenes de datos.
- Buenos resultados.
- Posibilidad de cambiar los puntos iniciales y obtener resultados diferentes.

Desventajas del k-mean.

- Dependencia del las posiciones de los puntos inicial.
- Dificultad para lidiar con clúster que no tiene forma convexa, ya que el algoritmo fue diseñado sobre la distancia óptima entre todos los puntos o centroide, por lo que el algoritmo promueve la construcción de grupos convexos aunque este no sea la forma más adecuada.

1.9.1.7 Métodos jerárquicos.

Los métodos jerárquicos como su nombre lo indica realizan una descomposición jerárquica al conjunto de objetos. Estos métodos crean una jerarquía de particiones P1, P2, P3… Ph, siendo un conjunto de n elementos y h grupos. La denominación de método jerárquico viene de que para cada partición Pi y Pi+1, cada grupo de la partición Pi+1 se encuentra incluido en la partición Pi.

El punto de partida de estos métodos es la creación de una matriz de semejanza-desemejanza entre los objetos. Sobre la base de las variables seleccionadas cada elemento de la matriz describe el grado de similitud o diferencias entre dos objetos. Esta información es combinada para dar lugar a los nuevos objetos o clúster.

Los métodos jerárquicos se dividen en dos grupos.

- Método jerárquico aglomerativo: Se comienza con los objetos o individuos de modo individual. De este modo, se tienen tantos clúster iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único clúster.
- Métodos jerárquicos divididos: Se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén.

Cuando se inicia el análisis cada objeto representa un clúster diferente, en ese momento es fácil medir la distancia entre ellos, a partir del momento que se crean los primeros objetos se hace necesario crear una regla para determinar la distancia entre los objetos recién formados. Existen varias reglas de agrupamiento todas presentan ventajas y desventajas.

Ventajas de los métodos jerárquicos.

- Alta velocidad.
- Facilidad de visualizar los resultados en tiempo de ejecución.

Desventajas de los métodos jerárquicos.

- Imposibilidad de retroceder en el proceso para corregir errores.

- Resultados poco flexibles.

1.9.1.8 Métodos para determinar la distancia intra-clúster en los algoritmos jerarquicos.

Single linkage (vecino más cercano) Figura 3. La distancia entre dos clúster es determinada por la distancia entre los objetos más próximos de los diferentes clúster. Esta regla tiende a crear clúster de forma alongada.

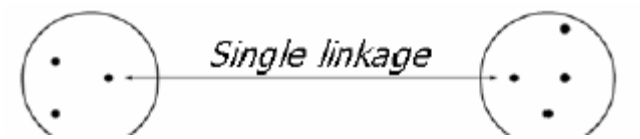

Figura 3. Representación del concepto de Single linkage. En este caso la distancia entre los dos clúster es la distancia entre los elementos de ambos clúster que se encuentre más próximo.

Complete linkage (vecino más alejado) Figura 4. La distancia entre dos clúster es determinada por la distancia entre los dos objetos más alejados de ambos clúster. Esta regla tiene buenos resultados cuando los clúster se encuentran aislados y bien definidos. Si los clúster se encuentras alongados, el resultado será pobre.

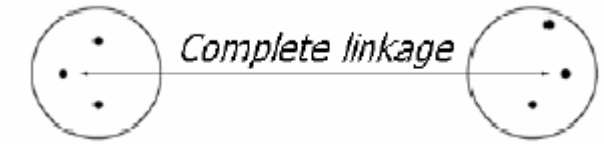

Figura 4. Representación del concepto Complete linkage: En este caso la distancia entre dos clúster está dada por la distancia de los dos objetos más lejanos de cada uno de los grupos.

Unweighted pair-group average: La distancia entre los dos clúster está definida como la media entre todas las distancias entre todos los objetos de ambos clúster. Este método tiende a tener buenos resultados cuando los clúster se encuentran en diferentes grupos o cuando forman una cadena alongada.

Unweighted pair-group centroid: La distancia entre los dos clúster es definida como la distancia entre los centroides de los dos clúster. El centroide es el punto medio del especio multidimensional definido por las variables. El centroide puede ser visto como el centro de gravedad del clúster.

Ward's method: A diferencia de los demás métodos este utiliza el cálculo de la varianza para determinar la distancia entre clúster. El método Ward en general tiene es considerado bastante eficaz, aunque tiende a crear grupos muy pequeños.

Una vez definida la regla de agrupamiento o algoritmo jerárquico se inicia con una matriz de datos (Figura 5), donde las líneas representan los objetos y las columnas las variables. A partir de esta primera matriz es construida una segunda (Figura 6) matriz de semejanza /desemejanza en la que las filas y las columnas representan a los objetos y las intercepciones representan la distancia que existe entre ellos.

	X_1	X_2	...	X_p
I_1				
I_2				
...				
I_n				

Figura 5. Matriz donde las filas representan a los objetos y las columnas a las variables.

	I_1	I_2	...	I_n
I_1	0			
I_2	$d(I_2,I_1)$	0		
...	$d(I_{...},I_1)$	$d(I_{...},I_2)$	...	
I_n	$d(I_n,I_1)$	$d(I_n,I_2)$	$d(I_n,I_{...})$	0

Figura 6. Matriz donde las filas y columnas representan a los objetos y las intercepciones a la distancia que hay entre ellos.

Una vez creada la matriz de semejanza-desemejanza el proceso es muy simple basta con buscar el menor elemento de la matriz y construir un nuevo clúster. El paso siguiente es reconstruir la matriz ahora incluyendo el nuevo clúster recién formado.

1.9.2 Algoritmos de Clasificación (Classification Algorithms).

En la Clasificación de Datos se desarrolla una descripción o modelo para cada una de las clases presentes en la base de datos. Existen muchos métodos de clasificación como aquellos basados en los

árboles de decisión TDIDT como el ID3 y el C4.5, los métodos estadísticos, las redes neuronales, y los conjuntos difusos, entre otros.

A continuación se explicarán algunos de los algoritmos de clasificación que se han usado satisfactoriamente en los procesos de Minería de Datos.

- Algoritmos estadísticos: Muchos algoritmos estadísticos han sido utilizados por los analistas para detectar patrones inusuales en los datos y explicar dichos patrones mediante la utilización de modelos estadísticos, como, por ejemplo, los modelos lineales. Estos métodos se han ganado su lugar y seguirán siendo utilizados en los años venideros.

- Redes Neuronales: las redes neuronales imitan la capacidad de la mente humana para encontrar patrones. Han sido aplicadas con éxito en aplicaciones que trabajan sobre la clasificación de los datos.

- Algoritmos genéticos: técnicas de optimización que utilizan procesos como el entrecruzamiento genético, la mutación y la selección natural en un diseño basado en los conceptos de la evolución natural.

- Método del vecino más cercano: es una técnica que clasifica cada registro de un conjunto de datos en base a la combinación de las clases de los k registros más similares. Generalmente se utiliza en bases de datos históricas.

- Reglas de inducción: la extracción de reglas si-entonces a partir de datos de importancia estadística.

- Visualización de los datos: la interpretación visual de las relaciones entre datos multidimensionales Clasificadores basados en instancias o ejemplos: Una manera de clasificar un caso es a partir de un caso similar cuya clase es conocida, y predecir que el caso pertenecerá a esa misma clase. Esta filosofía es la base para los sistemas basados en instancias, que clasifican nuevos casos refiriéndose a casos similares recordados. Un clasificador basado en instancias necesita teorías simbólicas. Los problemas centrales de este tipo de sistemas se pueden resumir

en tres preguntas: ¿cuáles casos de entrenamiento deben ser recordados?, ¿cómo puede medirse la similitud entre los casos?, y ¿cómo debe relacionarse el nuevo caso a los casos recordados?

- Los métodos de aprendizaje basados en reglas de clasificación buscan obtener reglas o árboles de decisión que particionen un grupo de datos en clases predefinidas. Para cualquier dominio real, el espacio de datos es demasiado grande como para realizar una búsqueda exhaustiva en el mismo.

- En cuanto a los métodos inductivos, la elección del atributo para cada uno de los nodos se basa en la ganancia de entropía generada por cada uno de los atributos. Una vez que se ha recopilado la información acerca de la distribución de todas las clases, la ganancia en la entropía se calcula utilizando la teoría de la información o bien el índice de Gini [Joshi, 1997].

1.9.3 Algoritmos de reglas de asociación.

Una regla de asociación es una regla que implica ciertas relaciones de asociación entre distintos objetos de una base de datos, como puede ser: “ocurren juntos” o “uno implica lo otro”. Dado un conjunto de transacciones, donde cada transacción es un conjunto de ítems, una regla de asociación es una expresión de la forma XY, donde X e Y son conjuntos de ítems. Un ejemplo de regla de asociación sería: “30% de las transacciones que contienen niños, también contienen pañales; 2% de las transacciones contienen ambas cosas”. En este caso el 30% es el nivel de confianza de la regla y 2% es la cantidad de casos que respaldan la regla. La cuestión está en encontrar todas las reglas de asociación que satisfagan los requerimientos de confianza mínima y máxima impuestos por el usuario.

1.9.4. Análisis de Secuencias.

En este caso se trabaja sobre datos que tienen una cierta secuencia entre sí. Cada dato es una lista ordenada de transacciones (o ítems). Generalmente, existe un tiempo de transacción asociado con cada dato. El problema consiste en encontrar patrones secuenciales de acuerdo a un límite mínimo impuesto por el usuario, dicho límite se mide en función al porcentaje de datos que contienen el patrón. Por ejemplo, un patrón secuencial puede estar dado por los usuarios de un video club que alquilan “Arma Mortal”, luego

“Arma Mortal 2”, “Arma Mortal 3” y finalmente “Arma Mortal 4”, lo cual no implica que todos lo hagan en ese orden.

1.10 Árboles de decisión.

Podemos clasificar los árboles de desafición como herramientas de clasificación y predicción. Una de las ventajas de los árboles de decisión es que representan las relaciones mediante reglas que son fáciles de interpretar. Permitiendo representar los resultados de los árboles de decisión en lenguajes naturales. Los conocimientos obtenidos por un árbol de decisión pueden ser representados en cualquier leguaje de base de datos como SQL.

Uno de los conceptos clásicos de árboles de decisión es: Un árbol de decisión es un algoritmo de toma de decisiones, que representa la información en forma de conocimiento.

1.10.1 Principios básicos de un árbol de decisión.

A pesar del interés de los árboles de decisión como herramientas de representación de conocimiento, el principal interés desde el punto de vista de la Minería de Datos es la capacidad de inducir estos árboles automáticamente de grandes volúmenes de datos. La idea básica de asociar la creación de estos árboles de decisión es bastante simple partiendo del cuadro habitual de un conjunto de datos de entrada y una de salida. Utilizando varias variables de entrada se crean reglas que permiten aislar subconjuntos de observaciones que posen valores idénticos para la variable de salida.

Para determinar en cada momento que partición tomar se utilizan criterios para determinar cual seleccionar independientemente de las demás particiones posibles. Siempre se intentará tomar la partición que se encuentre más homogénea en torno a la variable de salida. Para evaluar la calidad de las distintas particiones se utilizan diferentes medidas las más comunes son:

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

1.10.2 Medición de la taza de error.

Una ver creado el árbol es necesario evaluar la calidad de las particiones resultantes. La forma de medir la calidad global del árbol de decisión es muy simple, se aplica un conjunto de datos con el que el árbol nunca haya tenido contacto, se mide el porcentaje de registros que son correctamente clasificados, este proceso se conoce como validación cruzada.

En cualquier caso puede ser relevante analizar el error asociado a cada uno de los nodos. De cada uno de estos se puede medir:

- Número de registros que entrenan o no.
- La forma como los registros serían clasificados en caso de que este fuera un nodo final.
- El porcentaje de registros clasificados correctamente.

Al final del proceso de construcción del árbol, todos los registros del conjunto de entrenamiento son atribuidos a un nodo que forman el árbol final. A cada uno de estos nodos se le puede asociar una clase y una taza de error.

El error asociado a cada nodo final es la probabilidad de que salga ese resultado. La taza de error aparente del árbol de decisión es la suma ponderada de las tasas de error de todos los nodos hojas. Es decir la tasa de error del nodo hoja multiplicada por la probabilidad de un elemento de ser clasificado como ese resultado.

1.10.3 Criterio de parada.

Existen diversos criterios que pueden ser seguidos para parar el algoritmo de inducción. La mayoría del software permite seleccionar el criterio de parada a utilizar, Los criterios de parada de inducción tienen como principal objetivo prevenir el sobre-aprendizaje. Al establecerse un criterio de parada se reduce el tiempo de procesamiento en situaciones donde las particiones adicionales no producen conocimiento generalizable.

Independientemente del criterio de parada se definen otros criterios para evitar el sobre-aprendizaje del árbol. Uno de ellos es definir un límite en la medida de la diversidad. De esta forma cuando el árbol alcanza un valor por debajo de ese límite deja de crecer.

1.10.4 Desbaste del árbol.

El árbol de decisión crece mediante nuevas particiones, que mejora la capacidad discriminatoria del árbol, particiones capaces de separar el conjunto de datos de entrenamiento del árbol. Si se utiliza un conjunto de datos en entrenamiento como si se fuese a calcular la Tasa de error asociada (TEA), para validad la cualidad de cualquier árbol podado solo se encontrará un aumento en la taza de error.

La mayoría de los algoritmos para inducir árboles de decisión utilizan la mejor partición no raíz, donde existe una gran población de registros. Cuando crece el proceso cada partición tiene una sub-población menos representativa para trabajar, a medida que se acerca al nodo hoja hay variables que empiezan a tomar más pesos que otras y determinan el proceso.

La mejor forma de desbastar el árbol consiste en guardar un conjunto de datos para identificar el mejor de los subárboles. Este proceso es conocido como validación cruzada. Una vez terminado el proceso de construcción del árbol se utiliza el conjunto de validación para determinar cuáles de los subárboles tiene una menor estimación de error. De esta forma se obtendrá un gráfico como el de la Figura 7 donde se podrá observar la evolución del error tanto del conjunto de entrenamiento como del conjunto de validación mientras el árbol se va construyendo.

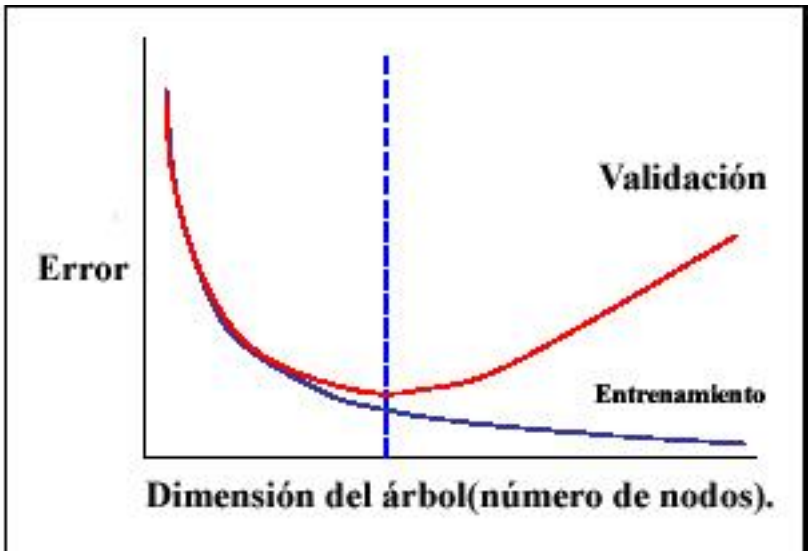

Figura 7. Evolución típica del error en un conjunto de entrenamiento y en un conjunto de validación. A partir de una determinada dimensión el conocimiento extraído por el árbol deja de ser útil en la medida que no es generalizable.

En el eje “y” podemos ver el valor del error y en el eje “x” el número de particiones realizadas. Como se puede observar el valor del error disminuye en el conjunto de formación durante todo el proceso, no así para el conjunto de validación. A partir de un determinado punto el error deja de disminuir y empieza a aumentar. Normalmente, este es el árbol que es elegido para hacer predicciones sobre el comportamiento de la población. Como se puede ver en la Figura 8 en árbol que se utiliza para las estimaciones es el que minimiza el error en el conjunto de validación que en general implica un adelgazamiento del árbol general.

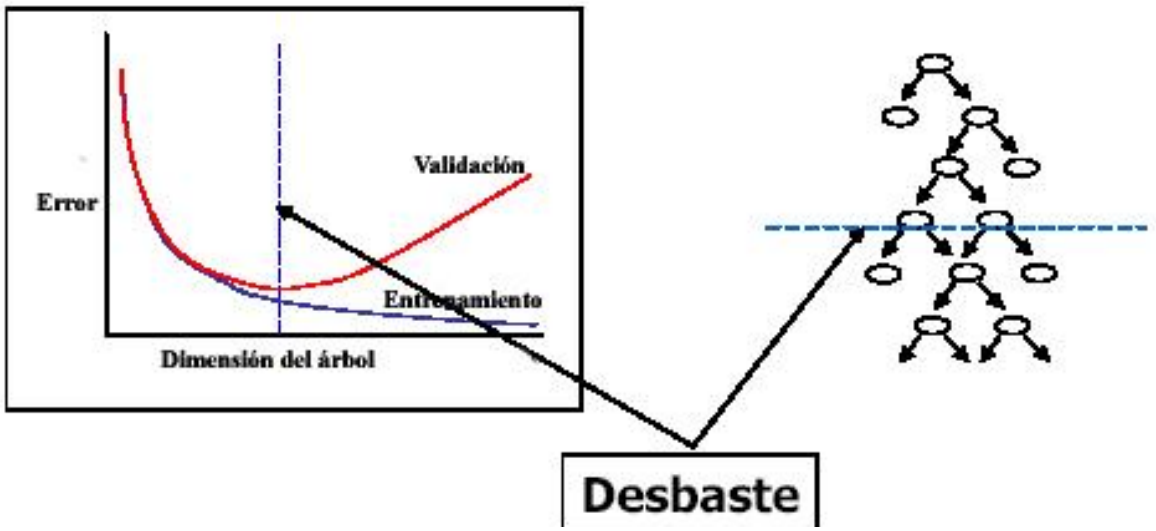

Figura 8. Punto de detención determinado por el conto de datos de validación implica en desbaste del árbol creado, los nodos por debajo de la línea no son utilizados.

1.10.5 Ventajas de los árboles de decisión.

Los árboles de decisión son herramientas capaces de construir buenos modelos predictivos siendo particularmente relevantes cuando en la clasificación la variable de salida es discreta y asume un número relativamente reducido de valores. Tiene la capacidad de utilizar datos binarios y no categóricos sin necesidad de transfórmalos, esta es una característica muy importante en la medida que la mayoría de las restantes aplicaciones tienen problemas para tratar con este tipo de variables. Además el árbol de decisión es insensible a variables de escala, estas variables pueden ser usadas sin ningún tipo de normalización.

Una de las mayores ventajas de los árboles de decisión es la facilidad de interpretación de los resultados. De hecho los resultados son fáciles de interpretar y de implementar en cualquier lenguaje de base de datos.

En grandes rasgos se pueden mencionar las siguientes ventajas:

- Facilita la interpretación de la decisión adoptada.
- Proporciona un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.
- Las reglas de asignación son simples y legibles, por tanto la interpretación de resultados es directa e intuitiva.
- Es robusto frente a datos atípicos u observaciones mal etiquetadas.
- Es válida sea cual fuera la naturaleza de las variables explicativas: continuas, binarias nominales u ordinales.
- Es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.
- Es computacionalmente rápido

1.10.6 Desventajas de los árboles de decisión.

- Las reglas de asignación son bastantes sensibles a pequeñas perturbaciones en los datos (inestabilidad).

- Dificultad para elegir el árbol óptimo.

- Ausencia de una función global de las variables y como consecuencia pérdida de la representación geométrica.

- Los árboles de clasificación requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

1.10.7 Algoritmo CART (Classification And Regression Trees).

El algoritmo CART **(**Classification And Regression Trees) es uno de los métodos más populares para la construcción de árboles. Este algoritmo desarrollado por Briemen en 1984 fue todo un hito en el área de aprendizaje automático. Este algoritmo construye un árbol binario. El CART separa los registros de cada uno de los nodos de acuerdo a una función con un único parámetro de entrada. La primera tareas es decidir cuáles de las variables de entrada produce la mejor partición.

Para seleccionar la mejor partición se consideran las variables de entrada de forma independiente. Luego se evalúan cada partición posible de acuerdo al valor de la diversidad en este caso sobre la base del coeficiente de Gini.

Para determinar el coeficiente de pureza de este árbol se calcula de la siguiente forma:

$$I_G(i) = \sum_{c=1}^{m} P_c(1 - P_c)$$

Donde Pc corresponde al porcentaje del individuo c para en el nodo i, m la dimensión del árbol.

Para decidir si se debe aceptar una partición o si solo se necesita comparar la diversidad de los nodos antes de la partición con la diversidad de los nodos después de la partición. En otras palabras se busca maximizar.

$$\Delta_j^m = I_G(i) - p_d I_G(i_d) - p_e I_G(i_e)$$

Donde Ig representa la diversidad inicial del nodo, Ig (id) la diversidad del nodo descendiente directo, Ig (ie) la diversidad del nodo descendiente de la Izquierda y Pd, Pe la proporción de registros clasificados en cada nodo.

La primera partición produce dos nodos, cada uno se trata de dividir de la misma manera que se divide el nodo raíz. Luego se examinan todas las variables de entrada para encontrar particiones. Si el campo es solo un valor se elimina ya que no hay forma de obtener particiones de él. Si no se pueden hacer más particiones se convierte el nodo en un nodo final.

1.10.7 Algoritmos C4.5.

El algoritmo C4.5 es muy simple y suele ser muy rápido, su estrategia básica es la siguiente.

El árbol comienza con un solo nodo, en representación de todos los ejemplos en formación.

Si todos los ejemplos son de la misma clase el nodo se convierte en un nodo final y todos los ejemplos se clasifican en esa categoría; en caso contrario el algoritmo utiliza una medida de entropía como forma de seleccionar la variable de entrada, la que mejor separa los ejemplos de clases individuales. La entropía varía entre un 0.5 máximo y 0 ente los elementos de una misma clase.

Cada valor de la variable de entrada es utilizado simultáneamente en cada partición.

La mejor partición es comparada con el nodo que le da origen para determinar si se ha ganado en conocimiento.

Si la partición mejora los resultados de la variable seleccionada En esta versión del algoritmos todas las variables de entrada son categorizadas, las variables expresadas en valores continuos deben ser discretizadas.

Se crea un nodo por cada valor de la variable de entrada.

El algoritmo continúa con todo el proceso hasta que el árbol este completo.

El algoritmo se detiene cuando una de las siguientes condiciones en verdadera.

Todos los ejemplos de un nodo pertenecen a una misma clase.

No existen más variables de entrada a las que se le pueden realizar particiones adicionales. En este caso el nodo asume la clase de los elementos dominantes en el.

La medida de ganancia de información es utilizada para determinar las variables de entrada que servida de base para la partición de cada nodo del árbol. La variable de entrada con mayor ganancia de información (mayor reducción de entropía) es seleccionada como base de prueba para el nodo. Esta variable minimiza la información necesaria para clasificar los ejemplos de las particiones resultantes,

minimiza las impurezas de las particiones. Este enfoque minimiza en número de pruebas previstas para clasificar un objeto y asegura un árbol simple, no necesariamente el más simple.

Siendo S un conjunto constituido por “s” ejemplos, supongamos que la variable de output asume m valores distintos. Definiendo m clases distintas Ci (par =1… i=m). Siendo “si” el número de ejemplo S para la clase Ci. La información necesaria para clasificar un ejemplo está dada por:

$$Ent(S) = \sum_{i=1}^{m} - p_i \log_2(p_i)$$

Donde pi es la probabilidad de que un ejemplo arbitrario pertenezca a la clase Ci está dada por si/s.

Debido al problema del sobre aprendizaje es necesario realizar un desbaste del árbol construido. La metodología de desbaste inicialmente propuesta por Quilan (1) para este algoritmo difiere mucho del método analizado anteriormente. En lugar de guardar un conjunto de datos para determinar el tamaño ideal del árbol el algoritmo C4.5 procede a desbastar el árbol sin algún conjunto de datos adicional.

El algoritmo C4.5 procede al desbaste mediante un análisis de la taza de error de cada nodo asumiendo que el verdadero error sustancial es peor que el observado. Si n registros pertenecen a un nodo y e son clasificados erradamente la taza de error está dada por e/n. Este algoritmo utiliza una analogía con la teoría de muestreo estadístico con el fin de construir la estimación más alta de error que puede tener un nodo. Esta analogía funciona pensando que los datos observados en un nodo son el resultado de un conjunto de pruebas en la que cada uno puede tomar dos resultados. Si los ejemplos son todos de la misma clase el nodo se convierte automáticamente en un nodo final.

1.11 Redes neuronales.

1.11.1 Introducción.

Una red neuronal artificial puede considerarse como la representación computacional del funcionamiento del cerebro humano. El principal desafío consiste en simular a través de representaciones

computacionales el proceso de aprendizaje con el objetivo de crear aplicaciones inteligentes, las cuales no son más que la combinación de elementos simples del proceso (neuronas, y su funcionamiento) interconectados que operando de forma paralela consiguen resolver problemas relacionados con reconocimiento de formas, patrones, predicciones etc.

Aunque en la actualidad existen computadoras que tienen un alto grado de procesamiento, y pueden realizar operaciones miles de veces más rápidas que el cerebro humano, estas siguen siendo una herramienta inferior, ya que el funcionamiento del cerebro se basa en paradigmas diferentes al del ordenador. Además de un uso masivo de diez millones de neuronas en procesamiento paralelo, esto le brinda la capacidad de poder adaptarse para el reconocimiento de patrones, asociaciones de conceptos con información incompleta, capacidad de desarrollar estrategias para hacer frente a situaciones complejas, y una gran tolerancia a información con ruido. Estos entre otros han sido los principales motivos que han inspirado al desarrollo de las redes neuronales artificiales

Como las redes neuronales artificiales están inspiradas en las estructura del sistema nervioso contienen las mismas características básicas, lo que significa que la unidad central es la neurona y tiene que cumplir con la capacidad de comunicarse. Para establecer esta similitud sináptica y analógica se puede definir que las señales que llegan a la sinapsis son las entradas a las neuronas; estas son ponderadas (atenuadas o simplificadas) a través de un parámetro denominado peso asociado a la sinapsis correspondiente, esta señal puede excitar a la neurona (sinapsis con peso positivo) o inhibirla (peso negativo). El efecto es la sumatoria de las entradas ponderadas, si la sumatoria es mayor que el umbral, la neurona se activa (da salida), de lo contrario se cierra. La sinapsis es susceptible a diversos eventos como la fatiga, deficiencia de oxígeno, la falta de uso etc., esta habilidad de ajustar la señal es un mecanismo de aprendizaje.

1.11.2 Tipos de redes neuronales.

Aunque hoy en día existen miles de tipos de redes neuronales se puede decir que existen dos tipos básicos de redes, basados en los algoritmos de aprendizaje que estos usan:

- Redes neuronales supervisadas: Las redes neuronales supervisadas utilizan para sus algoritmos de aprendizaje, para la fase de entrenamiento del modelo, un conjunto de datos en donde la salida deseada se conoce de antemano. La formación se basa en los valores de entrada, dando lugar a

una cierta salida, y a continuación una comparación de esta con los valores reales de producción fijos y variables. Durante el modelo de aprendizaje, uno de los patrones que se presentan de entrada a la red se extiende a través de ella (sin importar la estructura específica de la red) a la capa de salida. Esta capa genera un valor de resultados que se compara con el valor real (también llamado objetivo (meta)). La diferencia entra la salida y la meta es el error en la previsión. El error puede dar una indicación a fin de permitir ajustar la red para el aprendizaje, cuanto mayor sea el error mayor es el ajuste que se toma en los valores de los pesos.

- Redes neuronales no supervisadas: En la formación de redes neuronales no supervisadas no existen variables de salida, es decir , no existe orientación al proceso de aprendizaje de la red, el objetivo de este tipo de red es organizar los datos por semejanza/ desemejanza (organización de clúster) un ejemplo de este tipo de red es el Self-Organizing Map

1.11.3 Elementos de una red neuronal artificial.

Debido a la utilidad de las redes neuronales, existen en la actualidad decenas de miles de estas, pero todas mantienen los elementos básicos, ya que estos son lo que le permiten reproducir un comportamiento parecido al cerebro humano. Estos elementos realizan una simplificación y averiguación de los elementos relevantes del sistema, dado a que por lo general la cantidad de información que se maneja es excesiva o redundante. Por lo que cada red neuronal contiene los siguientes elementos:

- Unidad de proceso (neurona artificial): Por lo general existen tres tipos de neuronas en los sistemas, las de entrada, salida y las ocultas. Las neuronas de entrada reciben las señales del entorno, las de salida envían las señales fuera de la red, y las ocultas son las neuronas cuyas entradas y salidas se encuentran dentro del sistema.
- Estado de activación: Son estados del sistema que en un tiempo t representan un vector A (t). Estos ventores pueden ser continuos o discretos, limitados o ilimitados
- Función de Salida o de Trasferencia: Asociada con cada unidad hay una función de salida, que trasforma el estado actual de activación en una señal de salida. Existen cuatro tipos de funciones de trasferencia, los que determinan determinados tipos de neuronas.
- Función Escalón.
- Función Lineal Mixta .

- Sigmoidal.

- Función Gaussiana.
- Conexión entre neuronas: Las conexiones que unen las neuronas que forman la red neuronal artificial (RNA) tiene asociado un peso, que es lo que hace que la red adquiera conocimiento. Se considera que el efecto de cada señal es aditivo, de tal forma que la entrada neta que recibe una neurona es la suma del producto de cada señal individual por el valor de la sinapsis que conecta ambas neuronas, este fenómeno se conoce como red de propagación.
- Función o Regla de Activación: Se requiere de una regla que combine las entradas con el estado actual de la neurona para producir un nuevo estado de activación. Esta función F produce un nuevo estado de activación en una neurona a partir del estado que existía y la combinación de las estradas con los pesos de las conexiones Esta función se denomina función de activación, y la salida que se obtienen en una neurona para las diferentes formas de la función serán:
    - Función de Activación Escalón.
    - Función de Activación Identidad.
    - Función de Activación Lineal-Mixta.
    - Función de Activación Sigmoidal.
- Regla de aprendizaje: El aprendizaje puede ser comprendido como la modificación del comportamiento inducido por la interacción con el entorno, y como resultado de experiencias conduce al establecimiento de nuevos modelos de respuesta a estímulos externos. En el cerebro humano el conocimiento se encuentra en la sinapsis, en caso de las redes neuronales artificiales se encuentra en los pesos de la conexión entre las neuronas. Todo proceso de aprendizaje implica cierto número de cambios en estas conexiones. En realidad, puede decirse que se aprende modificando los valores de los pesos de la red.

1.11.4 Características de las Redes Neuronales.

Existen cuatro aspectos que caracterizan una red neuronal: su topología, el mecanismo de aprendizaje, tipo de asociación realizada entre la información de entrada y salida, y la forma de representación de esta información.

Topología de las Redes Neuronales: La arquitectura de las redes neuronales consiste en la organización y disposición de las neuronas formando capas más o menos alejadas de la entrada y salida de la red. En este sentido, los parámetros fundamentales de la red son: el número de capas, el número de neuronas por capa, el grado de conectividad y el tipo de conexiones entre neuronas.

Redes Mono capa: Se establecen conexiones laterales, cruzadas o auto recurrentes entre las neuronas que pertenecen a la única capa que constituye la red. Se utilizan en tareas relacionadas, lo que se conoce como auto asociación; por ejemplo, para generar información de entrada que se presentan a las redes incompletas o distorsionadas.

Redes Multicapa: Son aquellas que disponen de conjuntos de neuronas agrupadas en varios niveles o capas. Una forma de distinguir la capa a la que pertenece la neurona, consiste en fijarse en el origen de las señales que recibe a la entrada y el destino de la señal de salida. Según el tipo de conexión, como se vio previamente, se distinguen las redes feedforward, y las redes feedforward/feedback.

Mecanismo de Aprendizaje: El aprendizaje es el proceso por el cual una red neuronal modifica sus pesos en respuesta a una información de entrada. Los cambios que se producen durante el proceso de aprendizaje se reducen a la destrucción, modificación y creación de conexiones entre las neuronas, la creación de una nueva conexión implica que el peso de la misma pasa a tener un valor distinto de cero, una conexión se destruye cuando su peso pasa a ser cero. Se puede afirmar que el proceso de aprendizaje ha finalizado (la red ha aprendido) cuando los valores de los pesos permanecen estables ($dw_{ij} / dt = 0$).

Un criterio para diferenciar las reglas de aprendizaje se basa en considerar si la red puede aprender durante su funcionamiento habitual, o si el aprendizaje supone la desconexión de la red.

Otro criterio suele considerar dos tipos de reglas de aprendizaje, las de aprendizaje supervisado y las correspondientes a un aprendizaje no supervisado, estas reglas dan pie a una de las clasificaciones que se realizan de las RNA. Redes neuronales con aprendizaje supervisado y redes neuronales con aprendizaje no supervisado. La diferencia fundamental entre ambos tipos estriba en la existencia o no de un agente externo (supervisor) que controle el aprendizaje de la red.

Redes con Aprendizaje Supervisado: El proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor comprueba la salida de la red y en el caso de que ésta no coincida con la deseada, se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida se aproxime a la deseada.

Se consideran tres formas de llevar a cabo este tipo de aprendizaje:

- Aprendizaje por corrección de error: Consiste en ajustar los pesos en fusión de la diferencia entre los valores deseados y los obtenidos en la salida de la red; es decir, en función del error.
- Aprendizaje por refuerzo: Se basa en la idea de no indicar durante el entrenamiento exactamente la salida que se desea que proporcione la red ante una determinada entrada. La función del supervisor se reduce a indicar mediante una señal de refuerzo si la salida obtenida en la red se ajusta a la deseada (éxito=+1 o fracaso=-1), y en función de ello se ajustan los pesos basándose en un mecanismo de probabilidades.
- Aprendizaje estocástico: Este tipo de aprendizaje consiste básicamente en realizar cambios aleatorios en los valores de los pesos de las conexiones de la red y evaluar su efecto a partir del objetivo deseado y de distribuciones de probabilidades.

Redes con Aprendizaje no Supervisado: Estas redes no requieren influencia externa para ajustar los pesos de las conexiones entre neuronas. La red no recibe ninguna información por parte del entorno que le indique si la salida generada es o no correcta, así que existen varias posibilidades en cuanto a la interpretación de la salida de estas redes.

En algunos casos, la salida representa el grado de familiaridad o similitud entre la información que se le está presentando en la entrada y las informaciones con las que a interactuado en el pasado. En otro caso podría realizar una codificación de los datos de entrada, generando en la salida una versión codificada de la entrada, con menos bits, pero manteniendo la información relevante de los datos; en algunas redes con aprendizaje no supervisado lo que realizan es un mapeo de características, obteniéndose en las neuronas de salida una disposición geométrica que representa un mapa topográfico de las características de los datos de entrada, de tal forma que si se presentan a la red información similar, siempre sean afectadas las mismas neuronas de salidas próximas entre sí, en la misma zona del mapa.

En general en este tipo de aprendizaje se suelen considerar dos tipos:

- Aprendizaje Hebbiano: Consiste básicamente en el ajuste de los pesos de las conexiones de acuerdo con la correlación, así si las dos unidades son activas (positivas), se produce un reforzamiento de la conexión. Por el contrario cuando un es activa y la otra pasiva (negativa), se produce un debilitamiento de la conexión.
- Aprendizaje competitivo y cooperativo: Las neuronas compiten (y cooperan) unas con otras, con el fin de llevar a cabo una tarea dada. Con este tipo de aprendizaje se pretende que cuando se presente a la red cierta información de entrada, solo una de las neuronas de salida se active (alcance su valor de respuesta máximo). Por tanto las neuronas compiten por activarse, quedando finalmente una, o una por grupo, como neurona vencedora.

1.11.5 Ventajas de las Redes Neuronales.

Debido a su constitución y a sus fundamentos, las RNA presentan un gran número de características semejantes a las del cerebro. Por ejemplo, son capaces de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas. Estas ventajas incluyen:

1. Aprendizaje Adaptativo: Es una de las características más atractivas de las redes neuronales, es la capacidad de aprender a realizar tareas basadas en un entrenamiento o una experiencia inicial. En el proceso de aprendizaje, los enlaces ponderados de las neuronas se ajustan de manera que se obtengan unos resultados específicos. Una RNA no necesita un algoritmo para resolver un problema, ya que ella puede generar su propia distribución de los pesos de los enlaces mediante el aprendizaje. También existen redes que continúan aprendiendo a lo largo de su vida, después de completado e período inicial de entrenamiento. La función del diseñador es únicamente la obtención de la arquitectura apropiada. No es problema del diseñador el cómo la red aprenderá a discriminar; sin embargo, si es necesario que desarrolle un buen algoritmo de aprendizaje que proporcione la capacidad de discriminar de la red mediante un entrenamiento con patrones.
2. Auto organización: Las redes neuronales usan su capacidad de aprendizaje adaptativo para organizar la información que reciben durante el aprendizaje y/o la operación. Una RNA puede crear

su propia organización o representación de la información que recibe mediante una etapa de aprendizaje. Este auto organización provoca la facultad de las redes neuronales de responder apropiadamente cuando se les presentan datos o situaciones a los que no habían sido expuestas anteriormente.

3. Tolerancia a Fallos: Comparados con los sistemas computacionales tradicionales, los cuales pierden su funcionalidad en cuanto sufren un pequeño error de memoria , en las redes neuronales, si se produce un fallo en un pequeño número de neuronas, aunque el comportamiento del sistema se ve influenciado, no sufre una caída repentina.
4. Operación en Tiempo Real: Los computadores neuronales pueden ser realizados en paralelo, y se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.
5. Fácil inserción dentro de la tecnología existente. Debido a que una red puede ser rápidamente entrenada, comprobada, verificada y trasladada a una implementación hardware de bajo costo , es fácil insertar RNA para aplicaciones específicas dentro de sistemas existentes (chips, por ejemplo). De esta manera, las redes neuronales se pueden utilizar para mejorar sistemas de forma incremental, y cada paso puede ser evaluado antes de acometer un desarrollo más amplio.

1.11.6 Aplicaciones de las Redes Neuronales.

Las redes neuronales son una tecnología computacional emergente que puede utilizarse en un gran número y variedad de aplicaciones, tanto como comerciales como militares.

Hay muchos tipos diferentes de redes neuronales, cada uno de los cuales tiene una aplicación particular más apropiada. Separándolas según las distintas disciplinas algunos ejemplos de sus aplicaciones son:

Biología.

- Aprender más acerca del cerebro y otros sistemas.
- Obtención de modelos de la retina.

Empresa.

- Reconocimiento de caracteres escritos.

- Identificación de candidatos para posiciones específicas.
- Optimización de plazas y horarios en líneas de vuelo.
- Explotación de bases de datos.
- Evaluación de probabilidad de formaciones geológicas y petrolíferas.
- Síntesis de voz desde texto.

Medio Ambiente.

- Analizar tendencias y patrones.
- Previsión del tiempo.

Finanzas.

- Previsión de la evolución de los precios.
- Valoración del riesgo de los créditos.
- Identificación de falsificaciones.
- Interpretación de firmas.

Manufacturación.

- Robots automatizados y sistema de control (visión artificial y sensores de presión, temperatura, gas, etc.)
- Control de producción en líneas de proceso.
- Inspección de calidad.

- Filtrado de señales.

Medicina.

- Analizadores del habla para la ayuda de audición de sordos profundos.
- Diagnóstico y tratamiento a partir de síntomas y/o de datos analíticos (encefalograma, etc.).
- Monitorización en cirugía.
- Predicción de reacciones adversas a los medicamentos.
- Lectores de Rayos X.
- Entendimiento de causa de ataques epilépticos.

Militares.

- Clasificación de las señales de radar.
- Creación de armas inteligentes.
- Optimización del uso de recursos escasos.

1.12 Herramientas usadas en el proceso de Minería de Datos.

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

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

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

Generalmente las aplicaciones actuales que se utilizan para los procesos de Minería de Datos tienen implementadas funcionalidades que las ubican en ambos grupos.

A continuación se expondrán algunas de las herramientas usadas en la Minería de Datos.

Weka es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software ha sido desarrollado en la universidad de Waikato (Nueva Zelanda) bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años.

Weka 3.6 es la última versión estable que ha salido al mercado de esta poderosa aplicación, tiene incluido un conjunto de funcionalidades que facilitan el tratamiento y la normalización de variables. Además pose implementados sistemas capases de generar árboles de decisión, redes neuronales y reglas de asociación de problemas específicos. Tiene implementado algoritmos que permiten realizar diferentes actividades. Esta herramienta permite manejar la información a tratar directamente de la base de datos mediante sentencias SQL o mediante un fichero .arff. Esto es una gran ventaja ya que permite realizar las operaciones de minado fuera del servidor de base de datos evitando que se cuelguen los procesos de la base de datos.

MicroStrategy Data Mining Services: Es un componente completamente integrado a la Plataforma de Inteligencia de negocio de MicroStrategy con capacidades para brindar modelos predictivos de Minería de Datos a todos los usuarios. Permite a los usuarios realizar funciones de Minería de Datos por medio de métricas construidas a partir de funciones predictivas instantáneas o importadas de modelos de Minería de Datos de herramientas de terceros. Esta es una aplicación desarrollado bajo la licencia de software propietario. Este componente es utilizado fundamentalmente en tareas de Business Intelligence.

Oracle Data Mining: La base de datos Oracle incluye funcionalidad para la Minería de Datos en la edición Enterprise. Esta funcionalidad está totalmente integrada y bajo el mismo motor que la parte relacional de

la misma. Se puede acceder a toda la funcionalidad Minería de Datos a través de la API Java que incluye la base de datos, de manera que las aplicaciones puedan sacar el máximo partido de las funciones disponibles.

Al estar integrado en la base de datos, Oracle Data Mining simplifica el proceso de extracción de conclusiones basadas en grandes cantidades de datos, ya que se elimina la necesidad de movimientos de datos para el proceso de análisis. Todas las operaciones de preparación, creación de modelos y análisis permanecen en la base de datos lo que resulta en una mejora de la productividad, automatización e integración. Oracle Data Mining acepta tablas transaccionales y no transaccionales (resúmenes, registros únicos). Oracle Data Mining hace todas las transformaciones necesarias automáticamente de forma interna, liberando así de este trabajo a los usuarios o desarrolladores.

Es bueno resaltar que la base de datos Oracle brinda funcionalidades limitadas para la Minería de Datos, además el hecho de que las herramientas estén integradas al motor produce más problemas de los que soluciona. Otra desventaja es la licencia de Oracle ya que este se distribuye bajo los términos de software propietario y su licencia de uso es una de las más caras en el mercado actual.

1.13 Propuesta de solución.

Debido a las necesidades actuales que tiene el gobierno de Venezuela de combatir el crimen y la inseguridad social proponemos la creación de una metodología para el montaje de un sistema de Minería de Datos en el Centro de Investigaciones Científicos, Penales y Criminalísticas (CICPC) de Venezuela. Utilizando como fuente de información la base de datos del Cuerpo de Investigación e Información Policial (SIIPOL). Para el montaje de este sistema minero proponemos realizar un profundo análisis de la información, con el objetivo de eliminar las irregularidades que pueden introducir errores en el proceso de aprendizaje, y utilizar la herramienta Weka para el proceso de aprendizaje. Por la naturaleza del problema será tratado de dos formas diferentes, como un problema de agrupación utilizando el algoritmo k-means para el agrupamiento de objetos (clúster) y como un problema de clasificación, utilizando el algoritmo C4.5 para la generación de un árbol de decisión.

CAPÍTULO 2: FUNDAMENTACIÓN DE LA PROPUESTA DE SOLUCIÓN.

2.1 Introducción.

Para obtener un resultado final óptimo es necesario que el proceso de minado pase por las siguientes fases.

- Filtrado de datos.
- Selección de Variables.
- Extracción de Conocimiento.
- Interpretación y Evaluación.

Estas son las fases básicas por las que debe pasar un proyecto de Minería de Datos, para que los resultados finales sean relevantes. Es importante recordar que el proceso de minería es meramente exploratorio de los datos. Este incumple con el principio tradicional del conocimiento, ya que se analizan los datos en búsqueda de patrones, y no con el objetivo de refutar o probar la validez de un patrón determinado.

Dado a que la información se encuentra almacenada en una base de datos relacional el método de Minería de Datos que se utilizará para el desarrollo de la aplicación será el Aprendizaje Relacional (Relational Learning). El hecho de que la base de datos sea relacional implica que gran parte de la información se encuentre en las relaciones entre las tablas. Estas relaciones serán extraídas como información mediante código PL-SQL.

2.2 Filtrado de datos.

La fuente de información que se utilizará en el proceso de minado, estará montada sobre el gestor Oracle 10g. Esta es un sistema de bases de datos objeto-relacional que utilizará como lenguaje de programación el Lenguaje Procedural (PLSQL).

El proceso de filtrado inicialmente se seleccionarán mediante un proceso de análisis las variables que son significativas para el proceso de aprendizaje o para el proceso de modelaje.

Seguidamente se extraerán todos estos datos de la base de datos utilizando código PL-SQL ejecutados de un conjunto de procedimientos almacenados.

2.2.1 Tratamiento de los datos.

Para garantizar el éxito del proceso de filtrado de datos los datos serán procesados de forma tal que se realicen las siguientes tareas sobre ellos:

- Tratamiento de valores omisos.
- Normalización.
- Eliminación de datos incorrectos.
- Pre-procesamiento de datos.

Dado que la mayoría de los datos almacenados en esta base de datos fueron recolectados con una antigua aplicación desarrollada sobre una base de datos Adaptable Database System (ADABAS) y como lenguaje de programación Natural, existen datos que no se encuentran almacenados con un formato correcto, lo cual es necesario corregir para el proceso de minado.

2.2.1.1 Tratamiento de valores omisos.

Es muy común encontrar en bases de datos actuales ausencia de registros, la base de datos del SIIPOL no es una excepción. En ocasiones la ausencia de valores en el proceso de aprendizaje es un hecho significativo.

Para solucionar este problema se tienen las siguientes opciones:

- No utilizar los registros que presentan valores omisos.

- No utilizar campos que presentan valores omisos.
- Corregirlos manualmente.
- Predecir automáticamente los valores faltantes con una estimación.
- Ignorar la presencia de valores omisos.
- Codificar los datos omisos a un valor predeterminado.

Estas son algunas de las técnicas utilizadas para solucionar este problema. Para este caso específico se utilizarán varias técnicas, ya que la solución varía independientemente del nivel de importancia del campo que se esté analizando. Es decir que no importa que el segundo nombre de la persona sea un valor nulo, ya que este dato no brinda información relevante para el aprendizaje, no ocurre lo mismo para otros valores.

La solución que se brindará será:

Los atributos que no son significativos en el proceso de aprendizaje serán ignorados si contienen o no valores nulos.

Para los atributos que son significativos la solución será codificar los datos faltantes a un valor predeterminados, para que la herramienta los analice correctamente.

Un ejemplo de esto es el siguiente fragmento de código PLSQL.

```
CASE (SELECT DD.CEDULA FROM DIEX DD
WHERE DD.ID = DIEX.ID)
WHEN '' THEN '000000000'
END CASE
```

Este fragmento de código lo que hace es que cuando la cedula de la persona no se encuentra definida retorna como valor predeterminado ‘000000000’.

2.2.1.2 Codificación numérica sobre datos categóricos.

Como el algoritmo k-means solo es aplicable sobre datos numéricos, es necesario trasformar los atributos significativos para el proceso de modelado no numéricos en valores numéricos. Este proceso se realizará utilizando código PLSQL.

El siguiente fragmento de código codifica el sexo, que en la base de datos del SIIPOL es un atributo categórico, lo cual implica que no puede ser usado en su formato estándar en el proceso de modelado. Para poder utilizarlo hay que aplicar la técnica de codificación (m.n). Esto se puede ver desarrollado en el siguiente fragmento de código:

```
CASE (SELECT D.SEXPER FROM DIEX D
WHERE D.ID = DIEX.ID)
WHEN 'M'
THEN '0'
WHEN 'F'
THEN '1'
ELSE '2'
END CASE
```

En este fragmento de código, se le asigna un valor numérico para cada valor categórico del sexo;

2.2.1.3 Normalización de los datos.

Esta es una de las tareas que más peso tiene en el proceso de tratamiento de datos. Está dado a que el algoritmo que se utilizará para clusterizar será el k-mean, este algoritmo determina la semejanza desemejanza entre objetos mediante términos de distancia. Esto significa que los datos deben de encontrarse todos en la misma escala, entre otras cosas. Este proceso se realizará implementando la técnica (one of n) en el lenguaje PLSQL. Otro problema es el en los datos nominales, los valores de cadenas no pueden presentar espacios, ya que el algoritmo J48 que será utilizado para analizar el comportamiento de los datos no lo soporta.

Para garantizar que los datos estén en la escala correcta, inicialmente, se determinarán manualmente en la etapa de selección de variables los atributos que pueden presentar este tipo de problemas. Luego en el proceso de extracción de datos mediante código PLSQL serán corregidos.

Un ejemplo de esto sería el siguiente fragmento de obtención:

select replace(personacaso.delito,' ','') from personacaso

Este código elimina los espacios en blanco en la cadena de caracteres.

2.2.1.4 Eliminación de datos incorrectos.

Para la detección y eliminación de los datos incorrectos existen diversas técnicas, muchas de ellas son mediante gráficos que ayudan a la comprensión que los valores que se encuentran fuera de rango lógico.

Para dar una solución óptima para el problema en cuestión inicialmente se determinarán cuáles son las variables significativas que son las que se la aplicarán técnicas para garantizar que los datos sean correctos.

Si el atributo en análisis tiene un mínimo y un máximo definido el proceso es simple, solo hay que incluir una cláusula en el código SQL que filtre los datos de ese atributo por un rango determinado.

Si el atributo no tiene límites definidos se procederá a seleccionarlo de forma independiente, luego se realizará una gráfica de dispersión para ver de forma gráfica cuáles son las posibles tendencias de los valores incorrectos. Una vez determinados estas tendencias los datos serán filtrados mediante expresiones regulares en el código SQL, para evitar que se introduzcan ruido en el conjunto de datos de prueba.

Esto se puede ver en el siguiente fragmento de código, donde se garantiza que las personas seleccionadas no tengan más de 100 años de edad:

AND (DIEX.ANNONACE > (**SELECT** TO_CHAR (**SYSDATE**,'RRRR') - 100 **FROM** DUAL))

2.2.1.5 Pre-procesamiento de datos.

El objetivo general de la fase de pre-procesamiento de datos es facilitar y simplificar el problema en cuestión, sin excluir o dañar información importante para el proceso de modelado.

La reducción de espacio de entrada es un proceso de suma importancia desde el punto de vista del performance de la aplicación.

Las variables serán seleccionadas mediante un análisis de los datos, para evitar las variables que no brindan conocimiento. Además los datos seleccionados serán procesados con la herramienta Weka en la pestaña de trabajo *SelectAtributes* del entorno de trabajo Explorer, aquí se evaluarán los datos con el filtro *cfsSubsetEval* para determinar los atributos relevantes en el proceso de aprendizaje.

En caso de seleccionar una variable que no aporte conocimiento como la cedula de las persona, pero puede ser útil en el modelado de los datos, la herramienta Weka permite filtrarlos datos para eliminar un columna y en el caso específico del análisis de clúster permite seleccionar los atributos que se desean ignorar en el proceso.

2.3 Análisis de clúster, problema de agrupamiento.

El análisis de clúster es la nominación genérica atribuida a la gran variedad de metodologías utilizadas para la clasificación de entidades. Estas metodologías construyen grupos de entidades semejantes entre sí. Más específicamente el análisis de clúster es un conjunto de algoritmos, que a partir de un conjunto de información sobre un conjunto de objetos (entidades, individuos, ejemplos etc.) procura organizarlos en grupos homogéneos tanto como sea posible, determinado una estructura de semejanza-desemejanza.

Dada a la naturaleza de este problema se puede tatar como un problema de agrupamiento, para darle solución se utilizara la herramienta Weka. Esta herramienta tiene implementado los siguientes algoritmos de análisis de clúster.

- CLOPE.

- Cobweb.
- DBScan.
- EM.
- FasthesFirst.
- FilteredCluster.
- MakedensityBasedCluster.
- OPTICS.
- Sib.
- SimpleKMeans.
- XMeans.

Esta tarea se realizará utilizando el algoritmo *SimpleKMeans,* que pertenece al grupo de algoritmos de partición-optimización. Este algoritmo fue seleccionado por que cumple con las siguientes características:

1. Capacidad de tratar con grandes bases de datos, en el orden de los millones de registros, ya que no todos los algoritmos tienen una performance correcta cuando se trate de un número considerable de registros.
2. Capacidad para tratar con diferentes tipos de datos, ya que muchos de los algoritmos existentes se crearon para manejar cierto intervalo de datos, y no siempre cumple con las expectativas necesarias para realizar un correcto análisis de clúster.
3. Capacidad de tratar clúster de formas arbitraria, muchos de los algoritmos buscan agrupaciones utilizando como base de medida la Euclidiana o de Manhattan, lo que provoca que han tenido que encontrar agrupaciones esféricas de tamaño y densidad similares, sin embargo un grupo puede tomar cualquier forma, y por tanto la necesidad de desarrollar algoritmos capaces de detectar

grupos de formas arbitrarias. En el caso del análisis de clúster no existen datos preclasificados, se crean grupos de objetos que sean semejantes entre sí.

4. Ser capaces de funcionar con un mínimo de conocimiento en el terreno, para poder determinar parámetros de entrada, ya que muchos algoritmos dependen de diferentes valores como el número de registros a formar, entre otros. El problema es que estos parámetros de entrada los debe determinar el usuario y estos algoritmos pueden ser muy sensibles a estos valores iniciales, lo que trae como consecuencia que se vea obstaculizada la labor del usuario y que sea difícil controlar la calidad del resultado.
5. Capacidad de liderar con ruido.
6. Capacidad de liderar con objetos de diferentes diseños. En las actuales Bases de Datos se manejan diferentes diseños de objetos con elevado número de atributos, y algunos algoritmos tienen un funcionamiento correcto en espacios de pequeñas dimensiones.

El algoritmo k-means o k-medias es el método más utilizado en las soluciones científicas y empresariales. Este algoritmo solo puede ser aplicado sobre datos numéricos.

2.4 Extracción de conocimiento, problema de predicción.

Para extraer conocimiento de los datos almacenados en el fichero .arff se utilizará la herramienta Weka en el entorno de trabajo *Explorer* visto anteriormente en el proceso de análisis de clúster. Para ello se utilizará la pestaña *Clasify* Para el proceso se podrá seleccionar algoritmos, que se encuentran agrupados a grandes rasgos en las siguientes familias:

- Bayes: Métodos basados en el paradigma del aprendizaje de Bayes.
- Funciones, Métodos “matemáticos”, Redes neuronales, regresiones, SVM.
- Lazy: Métodos que utilizan el paradigma de aprendizaje perezoso, es decir no construyen un modelo.
- Meta: Métodos que permiten combinar diferentes métodos de aprendizaje.
- Trees: Métodos que aprenden mediante la generación de árboles de decisión.

- Rules: Métodos que aprenden modelos que se pueden expresar como reglas.

Para realizar este proceso se seleccionó un algoritmo clásico de la familia de árboles de decisión *C4.5, j48* es el nombre que de la Weka, este es uno de los algoritmos de árboles de decisión más populares en el mundo y es seleccionado por las siguientes razones:

- El algoritmo *J48* no es afectado por la introducción de datos que no son altamente significativos en el proceso de aprendizaje.
- Posibilidad de modelar el resultado del árbol de decisión en lenguaje SQL.
- Velocidad computacional.
- Fiabilidad de los resultados.

2.5 Proceso de minado.

Selección de los datos: Se seleccionaron mediante un proceso de análisis los atributos que se consideró que pueden brindar conocimiento al sistema.

Extracción de los datos: Mediante código PLSQL ejecutado en procedimientos almacenados seextraera la información de la base de datos. Estos procedimientos serán ejecutados por un programa realizado en el lenguaje C# y almacenará esa información en un fichero .arff con una estructura determinada.

Este fichero es cargado con la herramienta Weka.

Validar que la información cargada es relevante para el proceso de aprendizaje utilizando el filtro cfsSubsetEval.

Realizar el proceso de análisis de clúster.

Analizar el resultado del análisis de grupo.

Aplicar el filtro no supervisado Remove para eliminar los campos que no se pueden utilizar en el proceso de construcción del árbol de decisión.

Construir el árbol de decisión.

Analizar las reglas o patrones obtenidos en el árbol.

Realizar de forma iterativa el proceso para un subconjunto de datos determinado para aumentar la interpretación de una patrón determinado.

2.6 Conclusiones.

En este capítulo se define el mecanismo que se utilizará para el desarrollo del proceso de minado. Como resultado de la investigación se lograron definir los métodos, algoritmos y herramientas que mejor satisfacen las necesidades del problema.

CAPÍTULO 3: RESULTADOS.

3.1 Introducción.

Para aprovechar al máximo en conocimiento almacenado en la base de datos del CICPC se montará un sistema de Minería de Datos utilizando la herramienta Weka para la exploración y detección de patrones delictivos. Este es un proceso costoso que requiere de tiempo y mucha capacidad de procesamiento computacional. Pero brindará ventajas significativas en la lucha contra el delito en todo el territorio venezolano. Este sistema preverá a los expertos de información que desconocen o que es extremadamente difícil de deducir, ya que se encuentra almacenada en el corazón de los datos. Este es un problema que se interpretará de diferentes formas por su naturaleza.

Para probar el sistema, se analizará un conjunto da datos de prueba para demostrar que es posible utilizar las técnicas de Minería de Datos para la detección de patrones delictivos. Para esto se cuenta con una réplica de la base de datos del Sistema de Información e Investigación Policial, en la que se tiene un fragmento de la base de datos que contiene información relevante para el proceso de aprendizaje.

3.2 Selección de variables.

En el desarrollo de este sistema minero inicialmente se seleccionarán las variables que aporten conocimiento en el proceso de aprendizaje. Para esto se escogen las principales variables del SIIPOL que estén relacionados con los delitos.

Esta selección será realizada mediante código PLSQL, ejecutado en un procedimiento almacenado con el objetivo de ganar tiempo y rendimiento en esta etapa.

El código PL-SQL que seleccionará los datos es el siguiente:

```
PROCEDURE SP_DATAMINIG(CURS OUT CUR)IS
CUR1 TES.CUR;
BEGIN
OPEN CUR1 FOR
```

```
SELECT DPERSONA.CEDULA,
DPERSONA.ESTADO,
DPERSONA.SEXO,
DPERSONA.ESTCIVIL,
PERSONACASO.DELITO,
PERSONACASO.NATURALEZADELITO,
PERSONACASO.EDAD,
DCASO.SEXOVISTIMA,
DCASO.NOCASO,
DDIRECCIONES.IDMUN,
DDIRECCIONES.IDES
FROM DPERSONA
INNER JOIN DDIRECCIONES
ON DDIRECCIONES.PERSONA=DPERSONA.ID
INNER JOIN PERSONACASO
 INNER JOIN DCASO
 ON DCASO.ID=PERSONACASO.CASO
ON DPERSONA.ID=PERSONACASO.PERSONA
WHERE (SELECT TO_CHAR(SYSDATE,'RRRR') FROM DUAL)-
TO_NUMBER(TO_CHAR(DPERSONA.FENACIMIENTO,'RRRR')) < 112;
CURS:=CUR1;
END SP_DATAMINIG;
```

En la etapa de selección de datos es necesaria una preparación de los mismos para que sean compatibles con los algoritmos de Minería de Datos.

Una gran parte de este pre procesamiento de los datos será ejecutada en el mismo código SQL. En este código se validarán mediante límites definidos las variables categóricas, con el objetivo de disminuir el ruido en el conjunto de datos. Un ejemplo de esto es la siguiente restricción aplicada a uno de los procedimientos almacenados que valida que la edad de las personas no sobrepasen los 112 años.

```
WHERE (SELECT TO_CHAR (SYSDATE,'RRRR') FROM DUAL)
TO_NUMBER(TO_CHAR(P.FENACIMIENTO,'RRRR')) < 112
```

Las herramientas de Minería de Datos no funcionan de forma correcta para un conjunto de variables como son campos extensos de texto o variables numéricas excesivas. Una buena práctica para solucionar este problema es mediante la discretización de las variables. Campos como la edad de las personas brindan incluso más conocimiento expresado en valores como (niños, jóvenes, adultos, viejos) que con un simple número que indique la edad. Se puede solucionar este problema de dos formas, una sería aplicando el filtro no supervisado Discretize de la herramienta Weka, con este filtro crearíamos un campo nuevo con datos de forma discreta del campo en cuestión. La mayor desventaja de este filtro es que no se controlaría el rango específico para cada valor y que las nomenclaturas utilizadas no serían descriptivas. La otra solución por la que nos inclinamos sería mediante clausulas PL_SQL, de esta forma se establecen los rangos específicos y la nomenclatura utilizada para nombrar a las personas que se encuentren en un rango de edad determinado.

La solución a este problema mediante código PL_SQL se encuentra en el siguiente fragmento.

```
DECLARE
CURSOR CUR IS
SELECT DPERSONA.ID,DCASO.ID,TO_CHAR(DCASO.FEDELITO,'RRRR')-
TO_CHAR(DPERSONA.FENACIMIENTO,'RRRR')
FROM DPERSONA
INNER JOIN PERSONACASO
 INNER JOIN DCASO
 ON DCASO.ID=PERSONACASO.CASO
ON DPERSONA.ID=PERSONACASO.PERSONA;
IDP INT;
IDC INT;
EDAD VARCHAR2(4);
BEGIN
OPEN CUR;
LOOP
```

```
FETCH CUR INTO IDP,IDC,EDAD;
EXIT WHEN CUR%NOTFOUND;
IF EDAD > 13 AND EDAD <18 THEN
UPDATE PERSONACASO
SET PERSONACASO.EDAD='ADOLECENTE'
WHERE PERSONACASO.PERSONA =IDP
AND PERSONACASO.CASO=IDC;
ELSIF EDAD >= 18 AND EDAD <= 30 THEN
UPDATE PERSONACASO
SET PERSONACASO.EDAD='JOVEN'
WHERE PERSONACASO.PERSONA =IDP
AND PERSONACASO.CASO=IDC;
ELSIF EDAD > 30 AND EDAD < 60 THEN
UPDATE PERSONACASO
SET PERSONACASO.EDAD='ADULTOMAYOR'
WHERE PERSONACASO.PERSONA =IDP
AND PERSONACASO.CASO=IDC;
ELSIF EDAD > 60 THEN
UPDATE PERSONACASO
SET PERSONACASO.EDAD='TERCERAEDAD'
WHERE PERSONACASO.PERSONA =IDP
AND PERSONACASO.CASO=IDC;
END IF;
END LOOP;
END;
```

Este fragmento de código anónimo PL-SQL compara las fechas de nacimiento de la persona y la fecha en la que se cometió el delito, codificando la edad de forma nominal en rangos de 13 a 18 adolecente, 18 a 30 joven, 30 a 60 adulto, mayor que 60 tercera edad.

3.3 Codificación numérica sobre datos categóricos.

Dado a la naturaleza del problema será tratado con diferentes técnicas para aprovechar al máximo el conocimiento de los datos.

En una de estas etapas el problema será analizado como un problema de agrupación, para determinar los patrones de comportamiento de los individuos. Este análisis se realizará utilizando el algoritmo k-means, este algoritmo brinda muy buenos resultados y es eficiente desde el punto de vista computacional. Pero tiene muchas limitantes ya solo funciona con datos numéricos. Esto es debido que para determinar la semejanza-desemejanza entre los objetos utiliza la Distancia Euclidiana. Para solucionar este problema es necesario codificar los datos categóricos a datos numéricos.

Este proceso será realizado mediante código PL_SQL, de forma que los datos categóricos que brindan información en el proceso de aprendizaje sean codificados de forma numérica, de forma tal que puedan interpretarse los resultados sin mucha complejidad.

Este proceso será manejado de la siguiente forma:

DDIRECCIONES.IDMUN

Aquí se selecciona el Id que es un valor numérico que identifica a cada municipio, de esta manera con una simple consulta SQL se puede conocer el nombre del municipio.

3.4 Tratamiento de los datos omisos.

Los datos omisos serán trabajados con dos de las técnicas expuestas anteriormente.

Serán ignorados los datos que contienen información relevantes mediante clausulas PLSQL. Un ejemplo de esto es AND (PTJ_PERSONA.ESTADO IS NOT NULL).

Codificando con un valor predeterminado en caso que los datos sean nulos. Ej:

```
CASE DD.EDOCIV
WHEN '1' THEN 'SOL'
WHEN '2' THEN 'CAS'
WHEN '3' THEN 'DIV'
WHEN '4' THEN 'VIU'
WHEN '5' THEN 'SOL'
WHEN '6' THEN 'CAS'
WHEN '7' THEN 'DIV'
WHEN '8' THEN 'VIU'
WHEN NULL THEN 'ND'
```

Este código coloca el estado civil de valor que ponga el estado civil como no definido (NF).

3.4 Extracción de conocimiento.

Para el proceso de extracción de las personas, en la información almacenada en la actual base de datos de trabajo del SIIPOL tiene el estado civil como un dato numérico con valores de 1 a 8, los primeros 4 pertenecen a personas de sexo masculino y las restantes a las personas de sexo femenino, en caso de no existir ningún conocimiento se utilizará el entorno de trabajo Explorer de la herramienta Weka, este entorno permite utilizar casi todas las funcionalidades de la herramienta.

El primer paso de la tarea consiste en cargar los datos guardados en el archivo .arff. Para esto dentro de la ventana *Preprocess* se selecciona la opción *Open File* Figura 9.

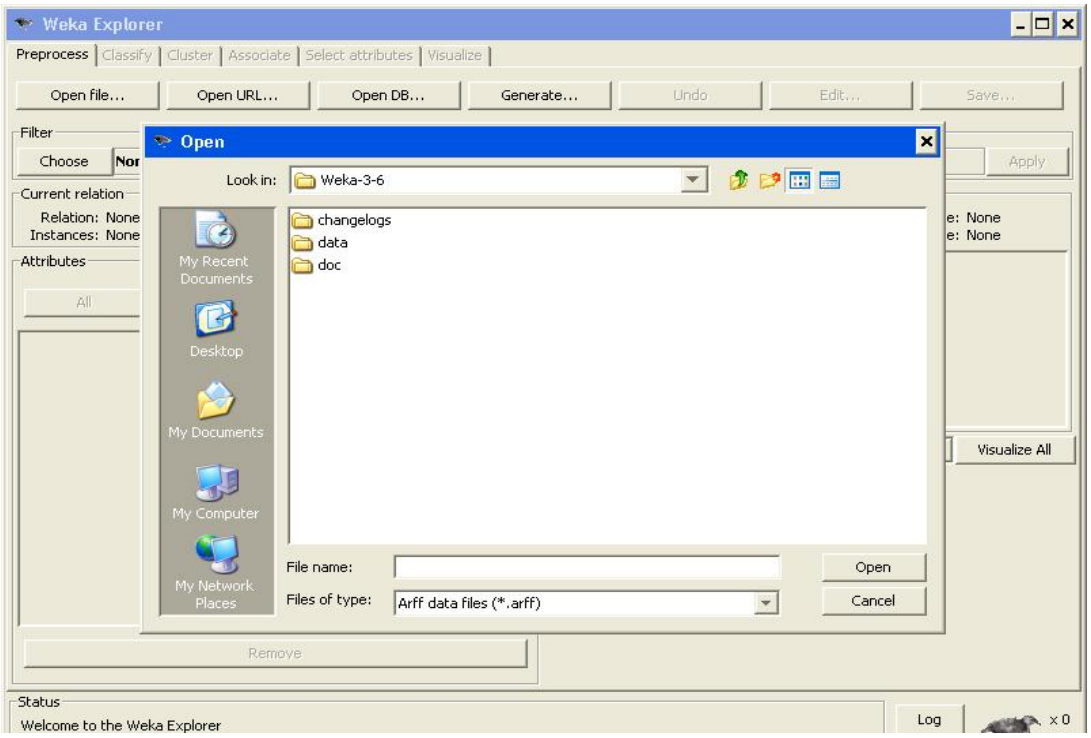

Figura 9. Entorno de trabajo Explorer, utilizado para cargar los datos a analizar.

3.5 Análisis de los datos.

Ya cargados los datos en la herramienta Weka, se procede a analizar de forma gráfica la información almacenada para tratar de detectar alguna relación entre ellos.

Para esto se va a la pestaña *Visualize* del entorno de trabajo Explorer de la herramienta Weka Figura 10 Aquí se puede observar de forma gráfica los datos que serán sometidos al proceso de aprendizaje. Con el modo gráfico es posible descubrir posibles relaciones entre variables que pueden aumentar la efectividad de los procesos posteriores.

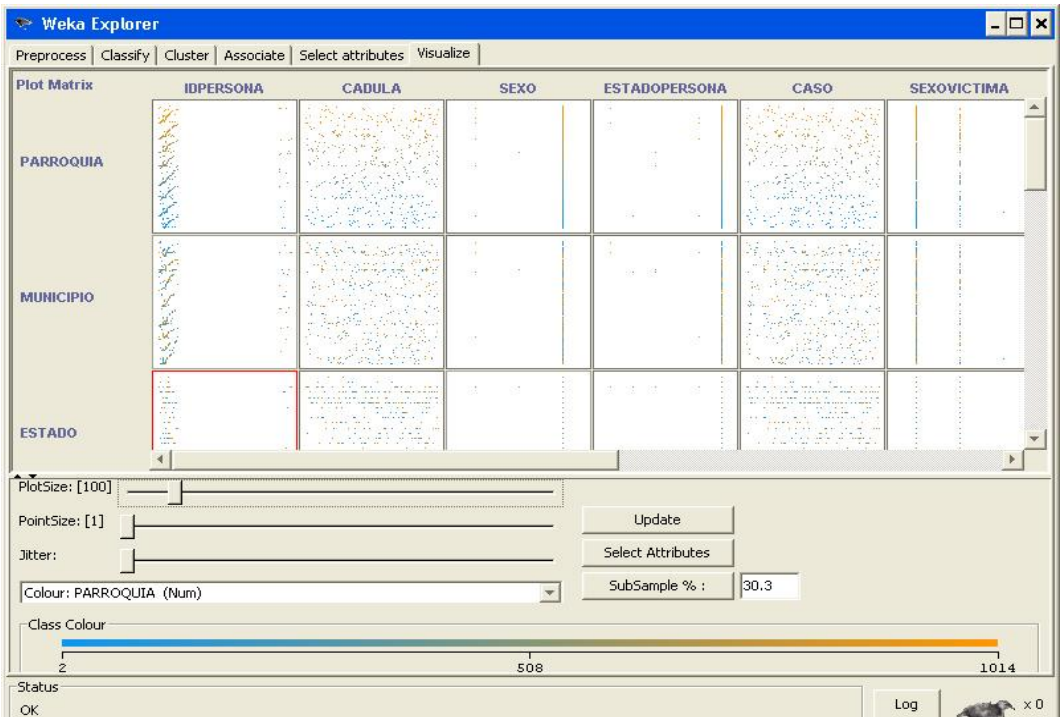

Figura 10. Ventana de visualización de los datos.

Si se encuentra una relación de dependencia entre dos variables específicas como en este caso se tiene una relación evidente. En estos casos se puede utilizar este cociente para aumentar la fiabilidad del modelo generado. Para esto se procede a crear un nuevo atributo derivado (llamado pick&mix, mediante el uso de filtros del sub-entrono de trabajo Preprocess. Se aplica el filtro no supervisado *AddExpression* seleccionando los atributos correlacionados. Figura 11.

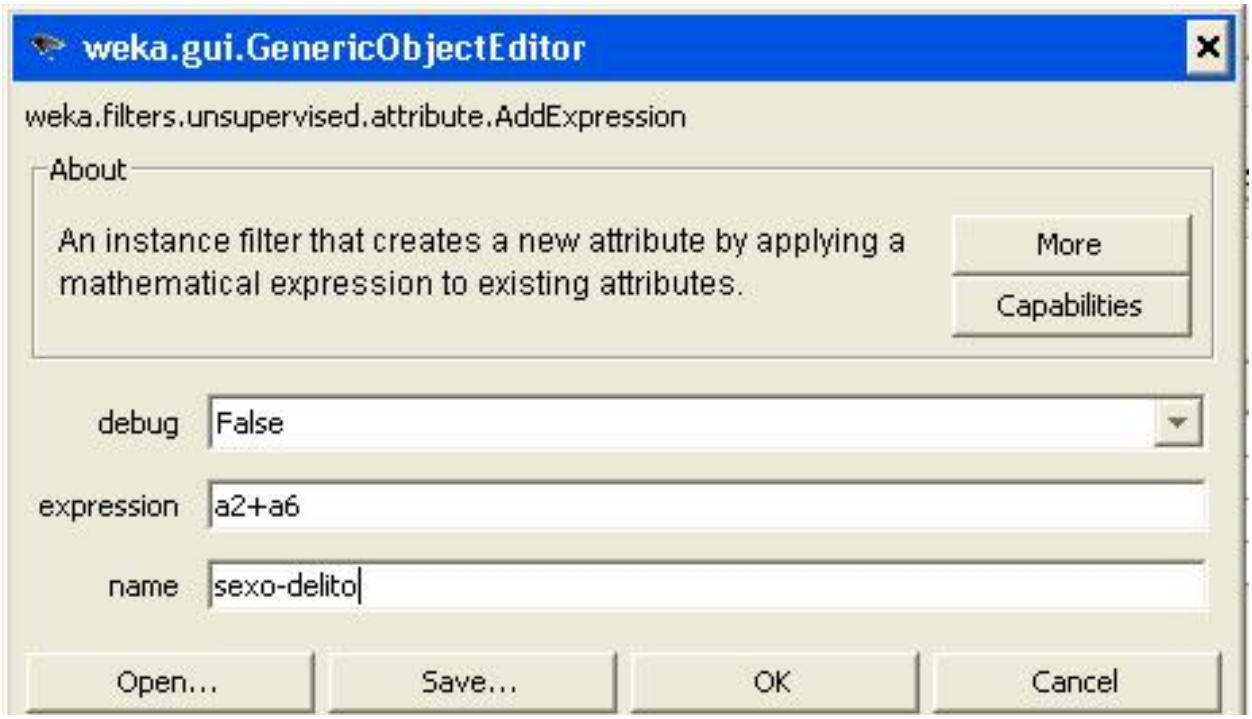

Figura 11. Filtro no supervisado AddExpression utilizado para crear un nuevo atributo que represente la correlación de un subconjunto de atributos.

- Debug: Si se usa el modo depuración el atributo generado será nombrado por el tipo de relación que se aplique, el valor de ese parámetro será false.
- Expresión:_ Expresión por la cual serán evaluados los atributos.
- Name: Nombre del nuevo atributo.

Esa técnica es eficiente, pero hay que tener mucho cuidado al aplicarla, ya que un error puede cambiar la filosofía de aprendizaje completamente.

3.6 Problema de agrupamiento (análisis de clúster)

Inicialmente se tratará el problema como un problema de agrupamiento para analizar las relaciones entre los individuos. Para realizar este proceso primeramente se abrirá la pestaña del entorno de trabajo *Explorer Cluster*. Seguido se seleccionará el algoritmo a utilizar Figura 12, el algoritmo que se seleccionará será el k-meas.

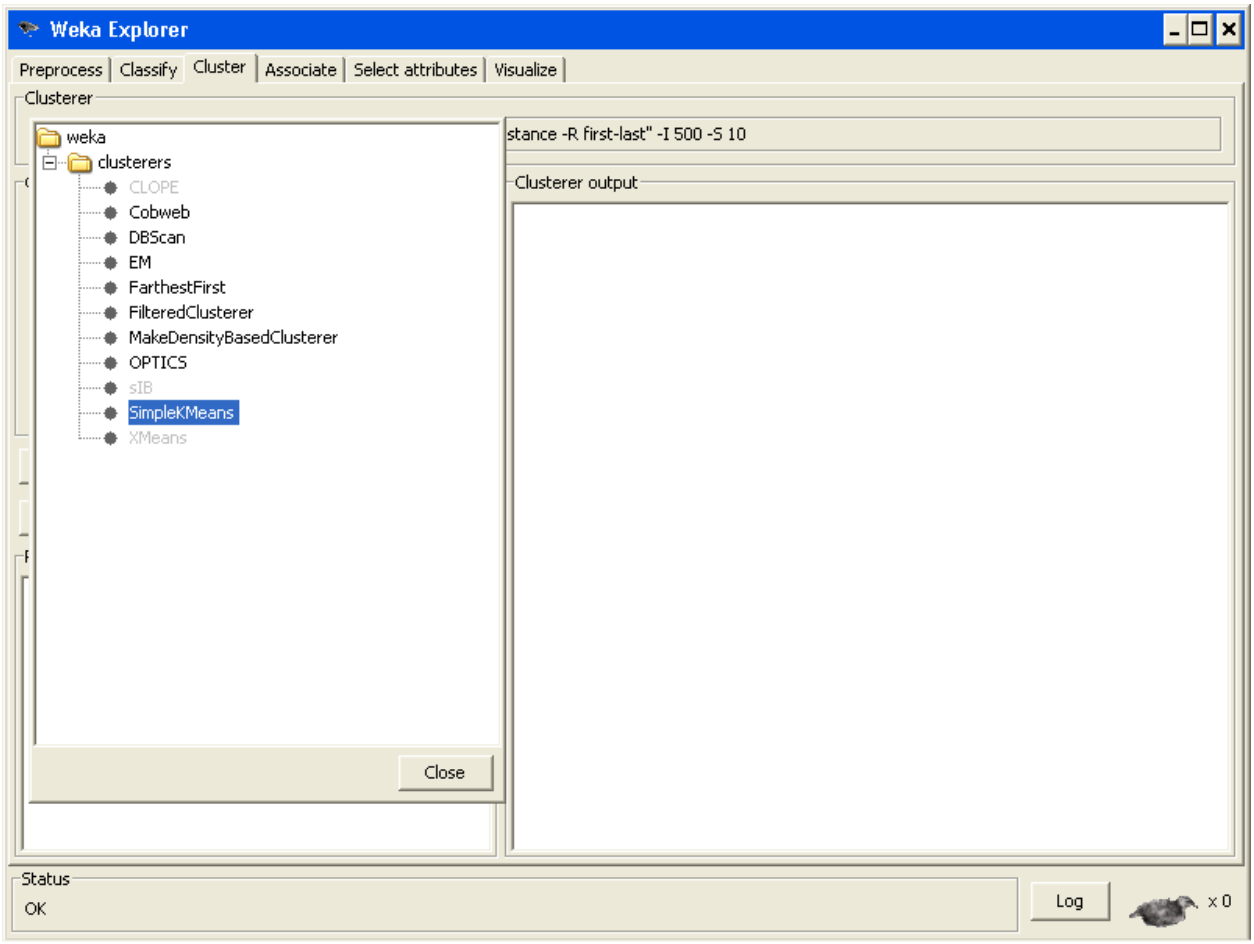

Figura 12. Ventana donde se selecciona los algoritmos de clúster a utilizar.

3.6.1 Configuración del k-meas

Luego se procede a configurar los parámetros de entrada de dicho algoritmo Figura 13, el algoritmo SimpleKmeas tiene siete parámetros configurables:

- displayStdDevs: Muestra las desviaciones de los atributos numéricos y cuenta los atributos nominales. Se debe activar, ya que por defecto aparece desactivada.
- distanceFunction: Este atributo selecciona la función de distancia, por lo que se regirá la semejanza/desemejanza de los objetos, mantener el valor que se carga por defecto.
- AttributeIndices: Se seleccionarán los índices de los atributos que se utilizarán en el proceso de clusterización.
- dontNormalize: Opción para normalizar los atributos, como los atributos están normalizados se coloca en true.
- invertSelection: Establece el modo de selección de los atributos, el valor que se le pondrá es falso, para que solo se tomen en cuenta los atributos seleccionados para el cálculo de distancia.
- dontRemplaceMissingValues: Sustituir los valores faltantes por la media. Este atributo se coloca en false, ya que los datos omisos son tratados en la etapa de pre-procesamiento de datos.
- numClusters: Cantidad de clúster a generar, se generarán 3 clúster.
- maxIterations: Número máximo de iteraciones, se pondrá en 1000 para aumentar la fiabilidad del método.
- preserveIntanceOrder: Preservar el orden de las instancias, mantener el valor que trae por defecto.
- Sed: La semilla del número aleatorio que se utiliza, mantener en diez, que es el valor que trae por defecto.

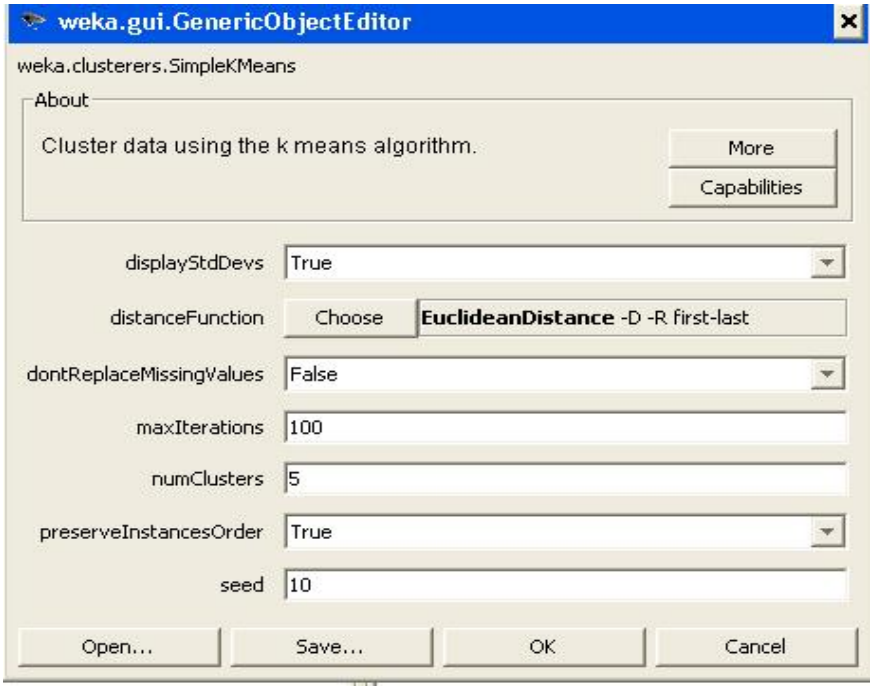

Figura 13. Ventana de configuración del algoritmo k-means.

3.6.2 Validación del entrenamiento.

Ya con configurado el algoritmo a utilizar tenemos que selecciona el método de validar los resultados. Figura 14.

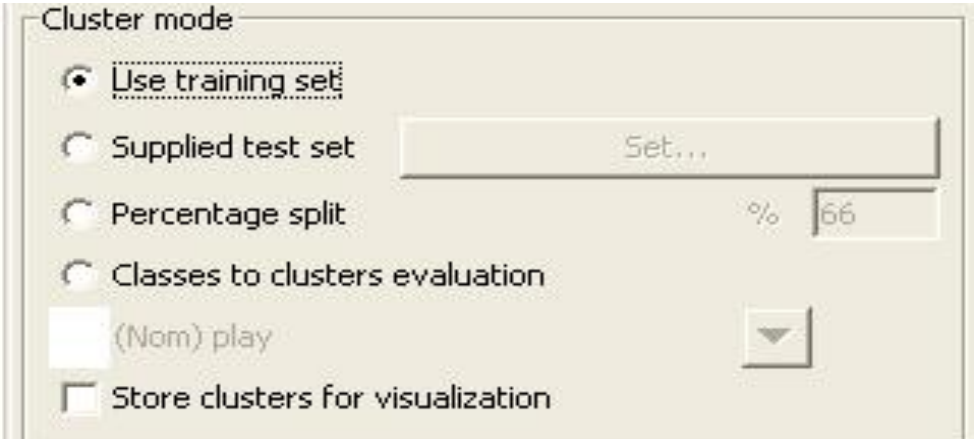

Figura 14. Modos de evaluación de los resultados.

Use training set: El clasificador se evalúa según las predicciones de las clases del set de entrenamiento.

Supplied test set: Utilizar un nuevo set de datos con los que el algoritmo no haya tenido contacto para validar el análisis realizado (validación cruzada).

Porcetaje Split: El clasificador se evalúa según las predicciones que realice de un porciento de los datos de prueba.

Classes to cluster evaluation: Valida clúster a clúster utilizando los niveles de homogeneidad entre ellos.

En este caso se validará como se puede apreciar en la figura utilizando la opción Use tainig set. Se recomienda esta opción por la velocidad computacional entre otras cosas. Utilizando este método los clúster son evaluados correctamente y no hay necesidad de otro conjunto de dato.

3.6.3 Ignorar atributos.

Debido a que el algoritmo utilizado en este proceso es el k-means y que la forma de calcular la distancia que se utilizará será la Distancia Euclidiana solo se puede aplicar el análisis de clúster cuando todos los atributos sean numéricos. En este caso tenemos atributos como la cédula, que son cadenas de caracteres. Estos atributos son incluidos en el conjunto de datos para darle mayor capacidad de entendimiento a los resultados, pero excluidos a la hora de realizar el proceso.

Para solucionar este problema se ignorarán estos atributos en el proceso de clusterización, ya que no brindan ningún tipo de conocimiento. Para realizar esta tarea se da clic en el botón Ignore attributes y se seleccionarán en la ventana los atributos que serán ignorados en el proceso que serían todos los atributos de cadenas de caracteres.

Resultados del análisis de clúster.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 1000 -S 10

Relation: DataMinig-weka.filters.unsupervised.attribute.Remove-R1-2,4-5

Instances: 1718

Attributes: 8

SEXO

SEXOVICTIMA

FEDELITO

DELITO

ESTADO

MUNICIPIO

PARROQUIA

EDAD

Test mode: evaluate on training data

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 22

Within cluster sum of squared errors: 1838.8072452672018

Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#			
Attribute	Full Data	0	1	2
	(1718)	(800)	(725)	(193)
SEXO	M	M	M	F
SEXOVICTIMA	M	M	M	F
FEDELITO	15838017.9884	16048900.7138	14488856.3531	20031986.9275
DELITO	HURTO	HURTO	HURTO	HURTO
ESTADO	13.734	11.9225	15.9834	12.7927
MUNICIPIO	143.2288	144.8063	143.12	137.0984

PARROQUIA	503.3155	522.3825	470.4317	547.8083
EDAD	JOVEN	JOVEN	ADULTOMAYOR	ADULTOMAYOR

Clustered Instances

0 800 (47%)

1 725 (42%)

2 193 (11%)

De este resultado se pueden deducir a grandes rasgos los siguientes patrones:

- Las personas que cometen delitos son generalmente de edad joven.
- El delito predominante es el hurto.
- El 89% de las personas que comenten delitos son de sexo masculino.
- El 47% de los delitos son cometidos por personas de sexo masculino a una edad joven.
- El 42% de los delitos son cometidos por personas de sexo masculino a una edad adulta.
- El 11% de las personas que cometen delitos son mujeres y generalmente cuando pueden ser clasificadas en la edad de adulto mayor.

3.7 Problema de clasificación.

Una vez terminado el proceso de análisis de clúster, el problema se tratará como un problema de clasificación. Para esto se utilizará un algoritmo predictivos de la herramienta Weka. Para generar un árbol de decisión utilizando el conjunto de datos anterior.

3.7.1 Creando un árbol de decisión.

Para detectar patrones de comportamiento en los datos se generará un árbol de decisión utilizando como fuente de aprendizaje de los de prueba citado anteriormente. Para generar este árbol se utilizará el algoritmo J48, este es el clásico algoritmo C4.5 mejorado. Para ello será necesario realizar un conjunto de pasos.

3.7.2 Preparación de los datos para aplicar el J48.

El algoritmos J48 tiene algunas restricciones en el conjunto de datos a utilizar, una de ellas es que necesita una variable nominal como valor de salida, y no puede tener valores de cadena al igual que el k-meas.

Para poder aplicar dicho algoritmo de forma exitosa es necesario eliminar dichos valores de nuestro conjunto de datos. Además es válido resaltar que el atributo cedula es seleccionado para darle una mayor claridad al conjunto de datos pero no aporta ningún conocimiento descriptivo.

Para eliminar estos campos del conjunto de datos es necesario aplicar un filtro no supervisado. El filtro que se aplicará será el Remove. Para ello se irá a la ventana Classify se dará clic en el botón Chosse y se seleccionará el filtro a utilizar Figura 15.

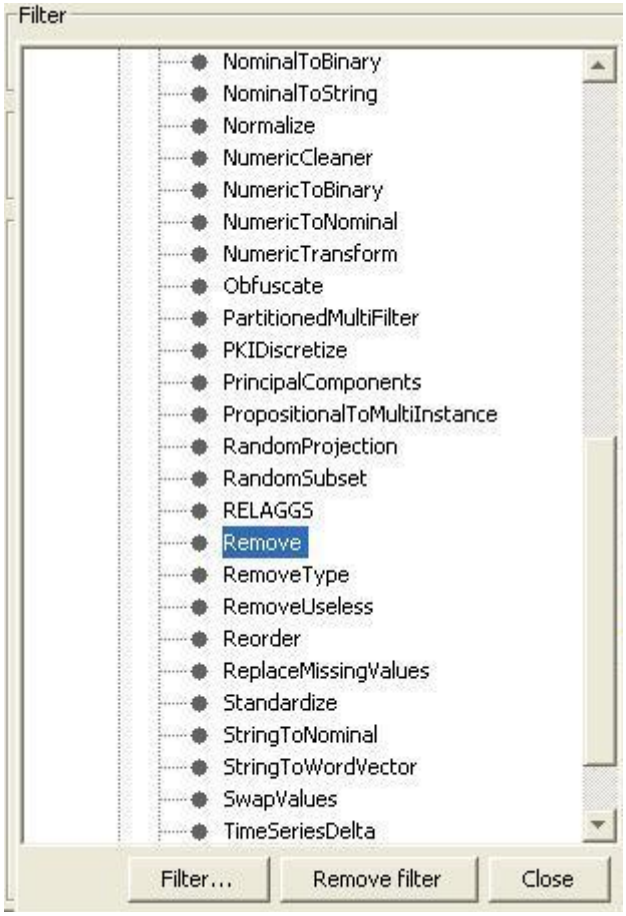

Figura 15. Filtro no supervisado Remove.

Luego de seleccionar el filtro a utilizar se procede a configurarlo, lo cual es extremadamente simple. Se da clic sobre el nombre del filtro y se despliega una ventana de configuración, se ponen los índices de los atributos que se desean eliminar que en este caso serán el atributo 2 Figura 16 y por último se aplica dicho filtro.

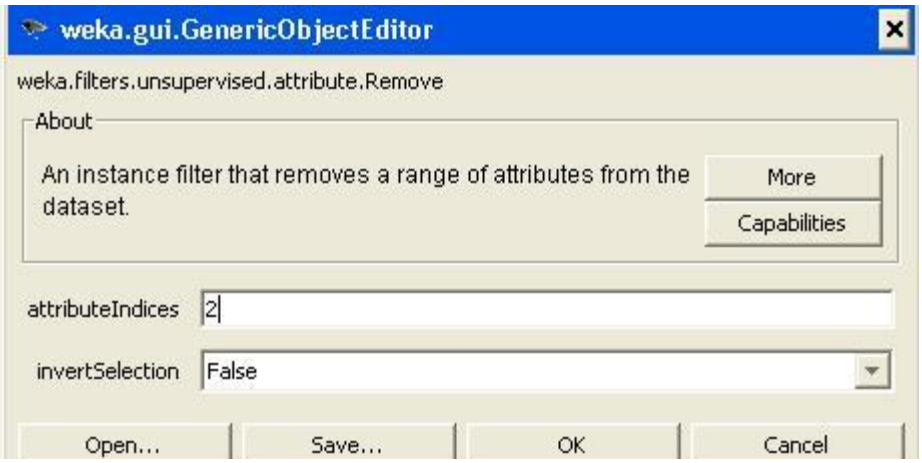

Figura 16. Ventana de configuración del filtro no supervisado Remove, utilizado para eliminar uno o varios elementos de un conjunto de datos.

Una vez preparados los datos para la creación del árbol de decisión, se irá a la pestaña Classify del entorno de trabajo Explorer de la herramienta Weka Figura 17.

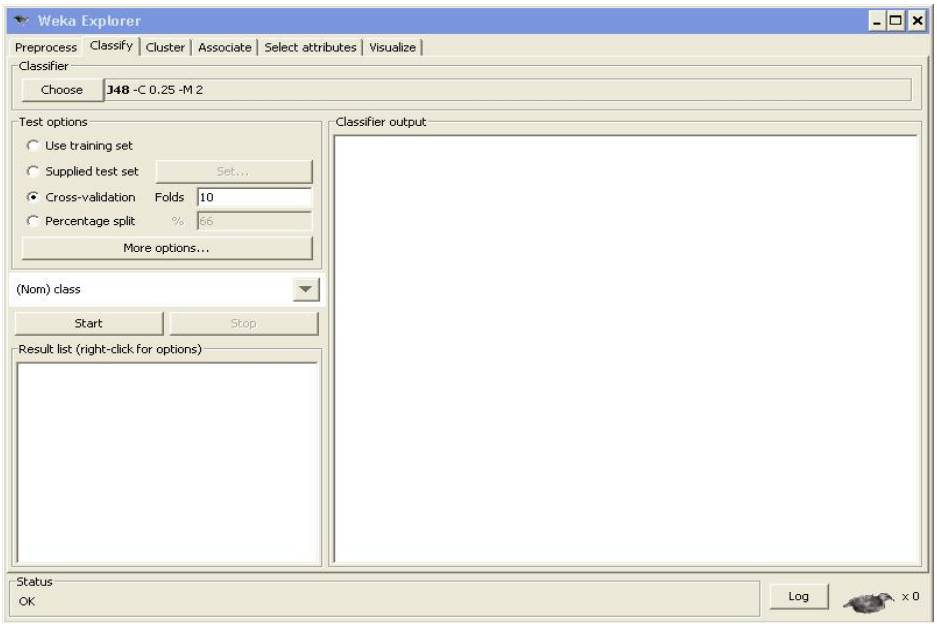

Figura 17. Ventana de trabajo Classify del entorno Explorer.

Ya en la pestaña Classify es necesario seleccionar el método que se utilizará para la extracción de conocimiento. Para esto se seleccionará la opción classifier Figura 18, el algoritmo J48 que pertenece a la familia árbol (Tree).

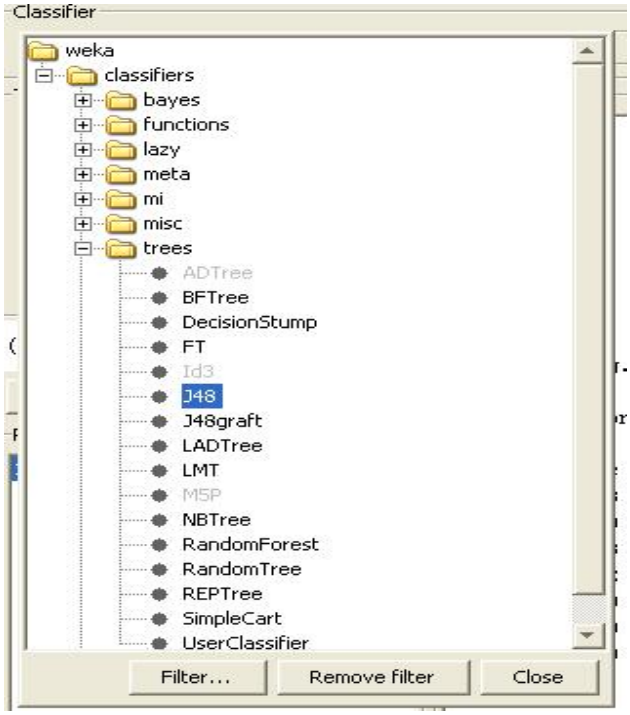

Figura 18. Ventana de Classifier donde se selecciona el algoritmo a utilizar para seleccionar el método de tratar los datos**.**

3.7.3 Configuración del algoritmo J.48

Seguidamente se procederá a configurar el algoritmo *J48*. Este proceso es muy parecido al realizado anteriormente en la fase de análisis de clúster. Para proceder a la configuración se da clic en el cuadro donde está el nombre del algoritmo y se despliega la siguiente ventana Figura 19.

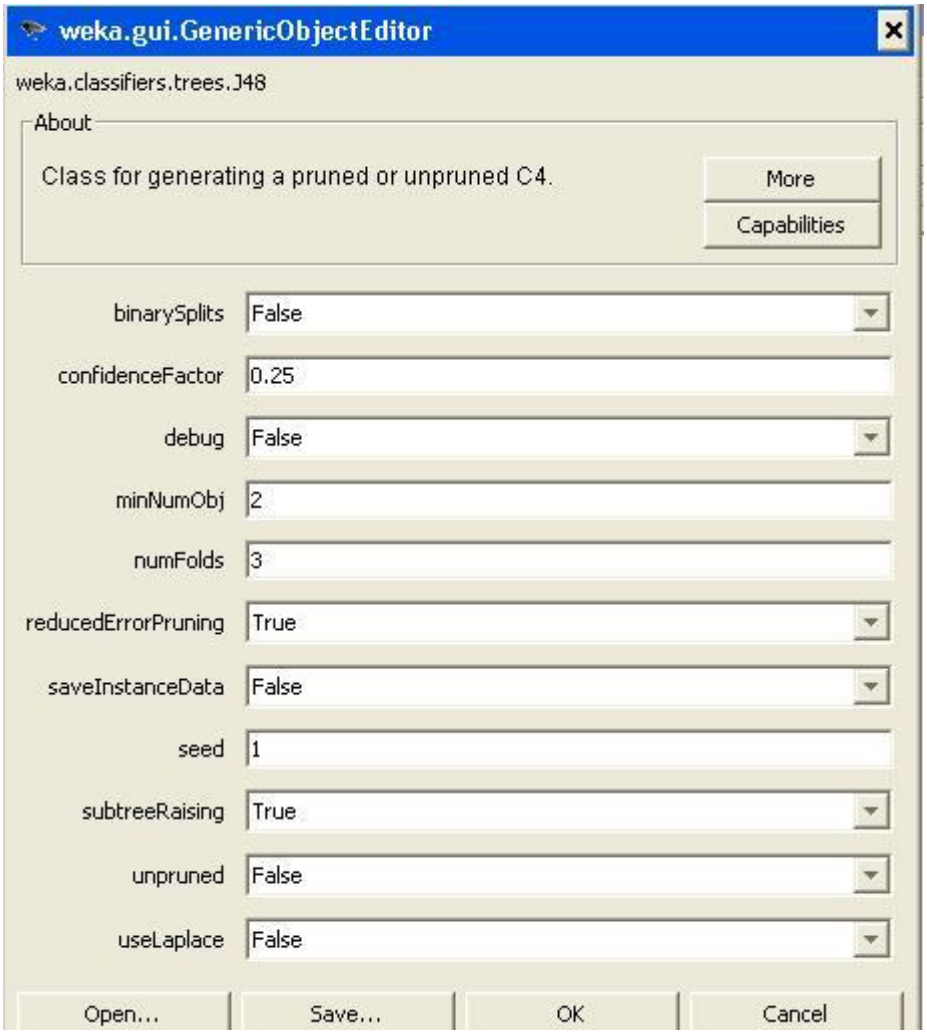

Figura 19. Ventana de configuración del J48.

- *binarySplits*: Usa los valores de los atributos nominales como variables binarias en la construcción del árbol. Colocar en true.
- *confidenceFactor*: Factor de confianza para la poda del árbol mientras el valor sea más pequeño mayor será la poda. Dejar 0.35 para que el árbol sea legible.
- *Debug*: Mostrar información extra del proceso de construcción, dejar en faso.
- *minNumObj*: Mínimo de instancias por hojas. Dejar el valor 2 que trae por defecto. Esto es debido a la poca cantidad de datos con la que se cuenta.

- *numFolds*: Determinar la cantidad de datos que se utilizará para la poda. Colocar el valor en 5.
- *reducedErrorPruning*: Reducción del error mediante la poda del árbol colocar el true. *saveIntanceData*: Guarda las instancias de datos para visualizarlas, colocar en false.
- *Seed*: Semillas de los datos seleccionados al azar par para reducir el error en el proceso de poda, dejar valor por defecto 1.
- *subbtreeRaising*: Aumenta la recolección de subárboles en el proceso de poda, colocar valor true.
- *Unprune:* Si se realiza la poda, dejar en false.
- *useLapace*: Determina si se suaviza las hojas sobre la base de Lapace, dejar en falso.

Terminado el proceso de configuración del algoritmo J.48 se procede a seleccionar el atributo por el correr el algoritmo Weka dando clic en el botón Start.

Una vez generado el árbol de decisión se muestra la información en la pestaña Classifier output la información del árbol generado.

3.7.4 Análisis del resultado.

El siguiente ejemplo que se verá a continuación es la información de un árbol generado con un conjunto de datos de prueba sobre la variable delito que contiene almacenada la información de los tipos de delitos cometidos.

=== Run information ===

Scheme: weka.classifiers.trees.J48 -R -N 3 -Q 1 -M 2

Relation: DataMinig-weka.filters.unsupervised.attribute.Remove-R1-2-
weka.filters.unsupervised.attribute.Remove-R3

Instances: 1718

Attributes: 9

SEXO

ESTADOPERSONA

SEXOVICTIMA

FEDELITO

DELITO

ESTADO

MUNICIPIO

PARROQUIA

EDAD

Test mode: 10-fold cross-validation

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

J48 pruned tree

SEXOVICTIMA = M

| ESTADO <= 23

| | ESTADO <= 5

| | | PARROQUIA <= 106: HURTO (21.0/12.0)

| | | PARROQUIA > 106

| | | | EDAD = TERCERAEDAD

| | | | | ESTADOPERSONA = DET

| | | | | | SEXO = M

| | | | | | | PARROQUIA <= 749: HURTO (12.0/1.0)

| | | | | | | PARROQUIA > 749: ESTAFA (6.0/3.0)

| | | | | | SEXO = F: HURTO (3.0)

| | | | | ESTADOPERSONA = DDT: HURTO (0.0)

| | | | | ESTADOPERSONA = DAG: HURTO (14.0/4.0)

| | | | | ESTADOPERSONA = DEN: HURTO (2.0)

| | | | | ESTADOPERSONA = SOL

| | | | | | SEXO = M

| | | | | | | MUNICIPIO <= 255: HURTO (22.0)

| | | | | | | MUNICIPIO > 255: ESTAFA (2.0/1.0)

| | | | | | SEXO = F: HURTO (2.0)

| | | | EDAD = ADOLECENTE: HURTO (0.0)

| | | | EDAD = JOVEN

| | | | | ESTADOPERSONA = DET: HURTO (12.0/3.0)

| | | | | ESTADOPERSONA = DDT: HURTO (0.0)

| | | | | ESTADOPERSONA = DAG: HURTO (19.0/3.0)

| | | | | ESTADOPERSONA = DEN: HURTO (2.0)

| | | | | ESTADOPERSONA = SOL

| | | | | | SEXO = M

| | | | | | | ESTADO <= 3

| | | | | | | | ESTADO <= 2

| | | | | | | | | MUNICIPIO <= 65: HURTO (5.0/1.0)

| | | | | | | | | MUNICIPIO > 65

| | | | | | | | | | FEDELITO <= 17121991: HOMICIDIOINTENCIONAL (2.0/1.0)

| | | | | | | | | | FEDELITO > 17121991: HURTO (2.0/1.0)

| | | | | | | | ESTADO > 2: HURTO (3.0)

| | | | | | | ESTADO > 3: HURTO (10.0/6.0)

| | | | | SEXO = F: DELITOMENOR (3.0/2.0)

| | | | EDAD = ADULTOMAYOR: HURTO (72.0/23.0)

| | ESTADO > 5

| | | ESTADOPERSONA = DET: HURTO (179.0/36.0)

| | | ESTADOPERSONA = DDT: HURTO (8.0/2.0)

| | | ESTADOPERSONA = DAG: HURTO (212.0/38.0)

| | | ESTADOPERSONA = DEN: HURTO (8.0/2.0)

| | | ESTADOPERSONA = SOL

| | | | MUNICIPIO <= 24: HURTO (30.0/10.0)

| | | | MUNICIPIO > 24

| | | | | EDAD = TERCERAEDAD: HURTO (74.0/13.0)

| | | | | EDAD = ADOLECENTE: HURTO (0.0)

| | | | | EDAD = JOVEN

| | | | | | SEXO = M

| | | | | | | ESTADO <= 22: HURTO (91.0/16.0)

| | | | | | | ESTADO > 22

| | | | | | | | FEDELITO <= 14041992: HURTO (7.0/2.0)

| | | | | | | | FEDELITO > 14041992

| | | | | | | | | FEDELITO <= 16121998: ESTAFA (2.0/1.0)

| | | | | | | | | FEDELITO > 16121998: HURTO (3.0/1.0)

| | | | | | SEXO = F: HURTO (11.0/3.0)

| | | | | EDAD = ADULTOMAYOR: HURTO (91.0/11.0)

| ESTADO > 23: HURTO (64.0/19.0)

SEXOVICTIMA = F

| ESTADO <= 22: HURTO (120.0/45.0)

| ESTADO > 22

| | EDAD = TERCERAEDAD: HURTO (2.0/1.0)

| | EDAD = ADOLECENTE: HURTO (0.0)

| | EDAD = JOVEN

| | | SEXO = M: RAPTOCONSENSUAL (9.0/6.0)

| | | SEXO = F: HURTO (4.0/2.0)

| | EDAD = ADULTOMAYOR: HURTO (16.0/4.0)

SEXOVICTIMA = A: HURTO (1.0)

Number of Leaves : 43

Size of the tree : 69

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	1295	75.3783 %
Incorrectly Classified Instances	423	24.6217 %
Kappa statistic	0.0221	
Mean absolute error	0.0541	

Root mean squared error 0.1673

Relative absolute error 97.4479 %

Root relative squared error 101.0273 %

Total Number of Instances 1718

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0.001	0	0	0	0.526	PERSONAEXTRAVIADAODESAPAREC
0.031	0.001	0.333	0.031	0.057	0.659	HOMICIDIOINTENCIONAL
0	0	0	0	0	0.336	ABUSODEFIRMAENBLANCO
0	0	0	0	0	0.402	VIOLENCIAFISICAMUJER-FAMILIA
0	0	0	0	0	0.323	FRAUDE
0.5	0.004	0.333	0.5	0.4	0.982	RAPTOCONSENSUAL
0	0	0	0	0	0.307	ADULTERACIONDESUSTANCIASALI
0.009	0.007	0.077	0.009	0.017	0.547	ESTAFA
0	0.001	0	0	0	0.547	DELITOMENOR
0	0	0	0	0	0.628	SECUESTRO
0	0	0	0	0	0.267	FALSIFICACION
0	0	0	0	0	0.298	COMERDETENTSUSTESTUPEFPSIC

0.989 0.971 0.763 0.989 0.861 0.569 HURTO

0 0 0 0 0 0.513 LESIONES

0 0 0 0 0 0.772 VIOLACION

Weighted Avg. 0.754 0.738 0.592 0.754 0.658 0.566

=== Confusion Matrix ===

a b c d e f g h i j k l m n o <-- classified as

0 0 0 0 0 0 0 0 0 0 0 0 45 0 0 | a = PERSONAEXTRAVIADAODESAPAREC

0 1 0 0 0 0 0 0 0 0 0 0 31 0 0 | b = HOMICIDIOINTENCIONAL

0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 | c = ABUSODEFIRMAENBLANCO

0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 | d = VIOLENCIAFISICAMUJER-FAMILIA

0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 | e = FRAUDE

0 0 0 0 0 3 0 0 0 0 0 0 3 0 0 | f = RAPTOCONSENSUAL

0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 | g = ADULTERACIONDESUSTANCIASALI

0 1 0 0 0 2 0 1 0 0 0 0 104 0 0 | h = ESTAFA

0 0 0 0 0 0 0 0 0 0 0 0 65 0 0 | i = DELITOMENOR

0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 | j = SECUESTRO

0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 | k = FALSIFICACION

0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 | l = COMERDETENTSUSTESTUPEFPSIC

1 1 0 0 0 2 0 10 1 0 0 0 1290 0 0 | m = HURTO

0 0 0 0 0 0 0 1 0 0 0 0 119 0 0 | n = LESIONES

0 0 0 0 0 2 0 1 0 0 0 0 20 0 0 | o = VIOLACION

Con un simple análisis de este árbol se pueden detectar un conjunto de patrones delictivos, los que pueden ser traducidos al lenguaje SQL si es necesario seleccionar más información.

Terminado el proceso de construcción del árbol se analizarán las reglas obtenidas, se valora los resultados y si es necesario se puede traducir dicha regla al lenguaje SQL. Luego se analizará si es necesario realizar el mismo proceso de forma iterativa sobre el subconjunto de datos seleccionados para lograr una mejor comprensión del resultado.

A continuación se explicarán algunos de los patrones delictivos obtenidos del análisis del conjunto de datos:

- Cuando el sexo de la víctima es femenino y el estado es Zulia, Mérida o Vargas, si la persona que cometió el delito es de la tercera edad, es un 0.5 de probabilidad que sea un hurto, si es un adolecente la probabilidad de que sea hurto es de 1, si es un joven y el sexo es masculino la probabilidad es de 0.66 de que sea un rapto consensual, si el sexo es femenino la probabilidad es de 0.5 de que sea un hurto.
- En los estados Mérida o Vargas, la mayoría de los delitos cometidos donde la víctima es una persona de sexo masculino son hurtos.
- En los estados (Miranda o Aragua o Distrito Capital o Carabobo o Lara o Barinas o Táchira o Cojedes o Sucre o Monagas o Amazonas o Bolívar o Nueva Esparta o Portuguesa o Apure o Delta Amacuro o Trujillo o Guárico o Zulia), el sexo de la víctima es masculino y el estado de la persona es solicitado, y el municipio es (Alto Orinoco o Atabapo o Atures o Maroa o Manapiare o Río Negro o Achaguas o Rómulo Gallegos o Camatagua o Girardot o José Ángel Lamas o José Félix Ribas o San Sebastián o Santos Michelena o Urdaneta o Zamora o Barinas o Obispos o Rojas o Sosa o El Callao o Piar o Rocío o Padre Pedro Chien o Bejuma o Juan José Mora o Miranda o Montalbán o Puerto Cabello o San Diego o Valencia o Ricaurte o Rómulo Gallegos o

San Carlos o Tinaco o Anzoátegui o Girardot o Lima Blanco o Pao de San Juan Bautista o Casa coima o Leonardo Infante o Julián Mellado o Francisco de Miranda o José Tadeo Monagas o José Félix Ribas o Juan Germán Rocío o San José de Guaribe o Crespo o Morán o Simón Planas o Urdaneta o El Hatillo o Guaicaipuro o Los Salías o Páez o Paz Castillo o Plaza o Urdaneta o Zamora o Acevedo o Andrés Bello o Buros o Chacao o Piar o Punceres o Santa Bárbara o Díaz o García o Gómez o Maneiro o Nariño o Península de Macanao o Villalba o Esteller o Guanare o Guanarito o Papelón o Santa Rosalía o Turén o Valdez o Bermúdez o Cajigal o Mariño o Mejía o Francisco de Miranda o García de Hevia o Guasimos o Jáuregui o Lobatera o Michelena o Panamericano o Pedro María Ureña o Rafael Urdaneta o Samuel Darío Maldonado o San Cristóbal o San Judas Tadeo o Andrés Bello o Antonio Rómulo Costa o Cárdenas o Córdoba o Fernández Feo o Andrés Bello o Boconó o Miranda o Motatán o Pampón o Pampa nito o Rafael Rangel o San Rafael de Carvajal o Urdaneta o Valera o Valmore Rodríguez o Almirante Padilla o Baralt o Catatumbo o Francisco Javier Pulgar o Jesús Enrique Lossada o Jesús María Semprún o Maracaibo o Miranda o Rosario de Perijá o San Francisco). Si la persona que comete el delito es un joven y si su sexo es femenino el delito probablemente es un hurto, si el sexo es masculino el delito es hurto o estafa.

Para más información ver anexo 1.

3.8 Conclusiones.

La Minería de Datos es un proceso costoso, ya que requiere de esfuerzos en la recolección de los datos, preparación de la información, la integración del software, la formulación del problema y la construcción del modelo de análisis. Todo este proceso es bastante caro, sin tener en cuenta los profesionales necesarios. A pesar de esto es capaz de brindar resultados extraordinarios que pueden ayudar a la toma de decisiones. Mediante estos procesos salen a la luz relaciones en los datos que son desconocidas y se pueden ver relaciones a gran escala que son extremadamente difíciles de descifrar. En modo de resumen se puede decir que es un método que brinda ventajas a las personas encargadas de la toma de decisiones, a gran escala en la cuestión de la seguridad social venezolana.

CONCLUSIONES

La investigación recorre todo el proceso de desarrollo del sistema de Minería de Datos para la detección de patrones delictivos en la base de datos del Sistema de Investigación e Información Policial, realizando un análisis profundo en las tecnologías actuales, de las cuales se obtuvo la propuesta de solución. Dando paso al estudio de cada uno de los subprocesos llevados a cabo mediante el análisis de los datos para la obtención de conocimiento, teniendo en cuenta cada uno de los algoritmos y herramientas existentes en el mercado actual.

Basados en las características del conocimiento, a lo largo de todo el trabajo se utilizaron un conjunto de datos seleccionados y los algoritmos que brinda la herramienta Weka. Se logró obtener un sistema de Minería de Datos capaz de detectar patrones delictivos ocultos en el interior de los datos.

RECOMENDACIONES

1. Aplicar el sistema minero con la base de datos real del Sistema de Investigación e Información Policial.

2. Realizar un estudio para determinar cuáles de los restantes procesos que se manejan en el Sistema de investigación e Información Policial (SIIPOL) se les puede aplicar técnicas de inteligencia de negocio.

3. Realizar un sistema similar para la Policía Nacional Revolucionaria (PNR).

BIBLIOGRAFÍA

1. **Ussama M, Fayyad.** *Advances in Knowledge Discovery and Data Mining.* 1996.

2. **Friedman, Nir.** *Beyesian Netwok Classifiers.* Boston : kluwe Academic Publishers, 1997. ,,1-37().

3. **David J. Hand, Heikki Mannila, Padhraic Smyth.** *Principles of data mining.* s.l. : illustrated, 2001. ISBN 026208290X.

4. *Torturando a los datos hasta que confiesen.* **Félix, Luis Carlos Molina.** s.l. : UOC, 2001.

5. **Hernádez Orallo, José.** *Introducción a la Minería de Datos.* s.l. : PEARSON, 2004.

6. DDE,PTH and Eggshell Thinning In Phaesant ad Ring Dove. Ohio : Columbus, 1966. 43210.

7. **Quinlan, J.Ross.** *Machine Learning.* s.l. : Springer Netherlands, 1994. 0885-6125 (Print) 1573-0565.

8. **Bação, Fernando Lucas.** *Introducción Minería de Datos.* s.l. : Nova.

9. pentaho. [Online] 3 10, 5. http://www.pentaho.com/.

10. **Frank, Eibe and Witten, Ian.** *Data Mining: Practical Machine Learning Tools and Techniques.* s.l. : Amazon, 2005.

11. **Berry, Michael J and Linoff, Gordon.** *Data Mining Techniques: For Marketing, Sales, and Customer Support .* 1997. 10158-0012.

12. **Han, Jiawei and Kamber, Micheline.** *Data Mining: Concepts and Techniques.* California : s.n., 2001.

ANEXOS

Anexo1. Código PLSQL utilizado.

```
create or replace procedure TESIS3 is – Procedimiento para poblar las persona.

CURSOR CUR IS

SELECT PE.PR_NOMBRE,

PE.SG_NOMBRE,

PE.PR_APELLIDO,

PE.SG_APELLIDO,

PE.CEDULA,

PE.FE_NACIMIENTO,

PE.ESTADO,

CASE DD.EDOCIV – Se codifica el estado civil.

WHEN '1' THEN 'M'

WHEN '2' THEN 'M'

WHEN '3' THEN 'M'

WHEN '4' THEN 'M'

WHEN '5' THEN 'F'

WHEN '6' THEN 'F'
```

```
WHEN '7' THEN 'F'

WHEN '8' THEN 'F'

WHEN NULL THEN 'ND'

END CASE,

CASE DD.EDOCIV – Se codifica el estado de la persona.

WHEN '1' THEN 'SOL'

WHEN '2' THEN 'CAS'

WHEN '3' THEN 'DIV'

WHEN '4' THEN 'VIU'

WHEN '5' THEN 'SOL'

WHEN '6' THEN 'CAS'

WHEN '7' THEN 'DIV'

WHEN '8' THEN 'VIU'

WHEN NULL THEN 'ND'

END CASE

FROM PTJ_CASOS CA

INNER JOIN PTJ_CASOS_ENLACE EN

 INNER JOIN PTJ_PERSONAS PE
```

```
 INNER JOIN DIEX DD

 ON DD.CEDULA=PE.CEDULA

 ON PE.ID_PERSONA = EN.ID_REFERENCIA

ON EN.NUMERO_CASO = CA.NUMERO_CASO

WHERE EN.CODI_REG='P'

AND PE.ESTADO IS NOT NULL;

PER TESIS.DPERSONA%ROWTYPE;

– Se seleccionan los datos de las tablas PTJ_CASOS, PTJ_PERSONAS y PTJ_CASOS_ENLACE que contienen la información de los casos, las personas y la relación de ambas.

begin

OPEN CUR;

LOOP

FETCH CUR INTO –Se itera el cursor que contiene los datos seleccionados anteriormente.

PER.PNOMBRE,

PER.SNOMBRE,

PER.PAPELLIDO,

PER.SAPELLIDO,

PER.CEDULA,
```

```
PER.FENACIMIENTO,
PER.ESTADO,
PER.SEXO,
PER.ESTCIVIL;
EXIT WHEN CUR%NOTFOUND;
INSERT INTO TESIS.DPERSONA(TESIS.DPERSONA.ID,
TESIS.DPERSONA.PNOMBRE,
TESIS.DPERSONA.SNOMBRE,
TESIS.DPERSONA.PAPELLIDO,
TESIS.DPERSONA.SAPELLIDO,
TESIS.DPERSONA.CEDULA,
TESIS.DPERSONA.FENACIMIENTO,
TESIS.DPERSONA.ESTADO,
TESIS.DPERSONA.SEXO,
TESIS.DPERSONA.ESTCIVIL)
VALUES(TESIS.SQ_DPERSONA.NEXTVAL,
PER.PNOMBRE,
PER.SNOMBRE,
```

```
PER.PAPELLIDO,

PER.SAPELLIDO,

PER.CEDULA,

PER.FENACIMIENTO,

PER.ESTADO,

PER.SEXO,

PER.ESTCIVIL);-- Se insertan los datos en la tabla persona.

END LOOP;

COMMIT;

end TESIS3;
```

Procedimiento utilizado en el poblado de la tabla persona utilizando los datos actuales de pruebas del actual SIIPOL.

```
create or replace procedure TESIS2 is

CURSOR CUR IS – Se seleccionan los datos a utilizar.

 SELECT distinct EN.NUMERO_CASO,P.ID_PERSONA,CC.DELITO

 FROM PTJ_PERSONAS P

 INNER JOIN PTJ_CASOS_ENLACE EN

  INNER JOIN PTJ_CASOS CC
```

```
 ON CC.NUMERO_CASO = EN.NUMERO_CASO

ON EN.ID_REFERENCIA = P.ID_PERSONA

WHERE EN.NUMERO_CASO IN (SELECT TESIS.DCASO.NOCASO FROM TESIS.DCASO)

and p.id_persona IN(SELECT PER.ID FROM TESIS.DPERSONA PER)

ORDER BY P.ID_PERSONA;

CAD VARCHAR2(100);

CAD1 number(15);

IDPERSONA NUMBER(15);

DELITO VARCHAR(4);

AA VARCHAR (60);

DEL NUMBER (13);

BEGIN

OPEN CUR;

LOOP – Se itera el cursor con los datos seleccionados.

FETCH CUR INTO CAD,IDPERSONA,DELITO;

EXIT WHEN CUR%NOTFOUND;

SELECT min(TESIS.DCASO.ID) INTO CAD1 FROM TESIS.DCASO

WHERE TESIS.DCASO.NOCASO = CAD;
```

```
SELECT COUNT(CO.TX_DESCRIPCION)INTO DEL FROM PTJ_CODIGOS CO

WHERE TRIM(CO.CL_TABLAS) = '0021'||DELITO;

IF DEL = 1 THEN

SELECT CO.TX_DESCRIPCION INTO AA FROM PTJ_CODIGOS CO

WHERE TRIM(CO.CL_TABLAS) = '0021'||DELITO;

END IF;

 INSERT INTO TESIS.PERSONACASO(PERSONA,CASO,DELITO)

 VALUES(

 IDPERSONA,

 CAD1,

 AA

 );

COMMIT;

END LOOP;

end TESIS2;
```

Procedimiento utilizado en el poblado de la tabla personacaso utilizando los datos actuales de pruebas del actual SIIPOL.

```
 create or replace procedure TESIS4 is
```

```
CURSOR CUR IS

SELECT PTJ_CASOS.NUMERO_CASO,PTJ_CASOS.FE_DELITO,PTJ_CASOS.SEXO_VICTIMA

FROM PTJ_CASOS

INNER JOIN PTJ_CASOS_ENLACE

 INNER JOIN PTJ_PERSONAS

 ON PTJ_PERSONAS.ID_PERSONA = PTJ_CASOS_ENLACE.ID_REFERENCIA

ON PTJ_CASOS_ENLACE.NUMERO_CASO = PTJ_CASOS.NUMERO_CASO

WHERE PTJ_CASOS_ENLACE.CODI_REG = 'P' AND PTJ_PERSONAS.ID_PERSONA IN

(SELECT TESIS.DPERSONA.ID FROM TESIS.DPERSONA);

-- Selecciona los datos de las tablas PTJ_CASOS que es la que contiene los casos almacenados y PTJ_CASOS_ENLACE que contiene la relación entre los casos.

NOCASO VARCHAR(13);

FECHA DATE;

SEXO VARCHAR(1);

BEGIN

OPEN CUR;

LOOP –Itera el cursor que contiene los datos seleccionados anteriormente.

 FETCH CUR INTO NOCASO,FECHA,SEXO;
```

```
 EXIT WHEN CUR%NOTFOUND;

 INSERT INTO TESIS.DCASO(ID,NOCASO,FEDELITO,SEXOVISTIMA)

 VALUES(

 TESIS.CASO.NEXTVAL,

 NOCASO,

 FECHA,

 SEXO);--Inserta los datos en la tabla caso.

END LOOP;

COMMIT;

END TESIS4;
```

Procedimiento utilizando para el poblado de la tabla caso, para esto se usan los datos de prueba de los venezolanos del actual Sistema de Investigación e Información Policial (SIIPOL).

Anexo2. Análisis de clúster.

A continuación se expondrán algunas de las gráficas generadas en el proceso de análisis de clúster, las cuales contiene información valiosa para le extracción de patrones delictivos.

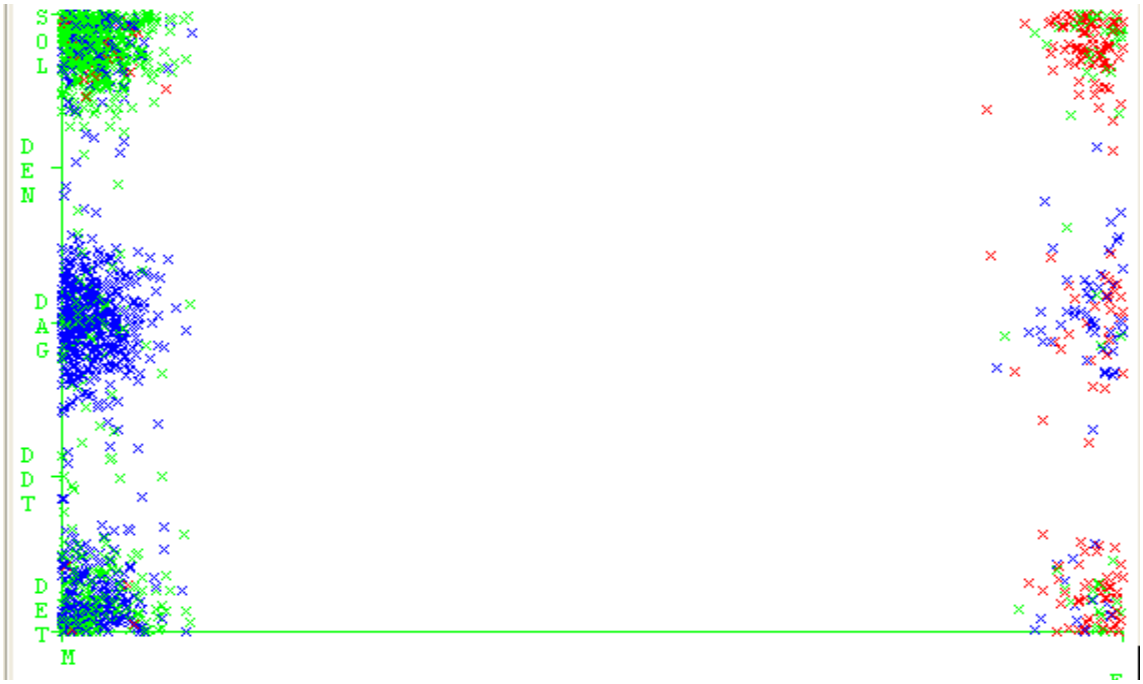

Figura 20. Gráfica que representa la relación entre el estado de la persona y el sexo.

En esta gráfica se puede apreciar la relación entre el sexo de la persona y el estado de la misma. Se puede determinar el predominio de las personas de sexo masculino, lo que indica que la mayoría de las personas que cometen delitos en Venezuela son hombres.

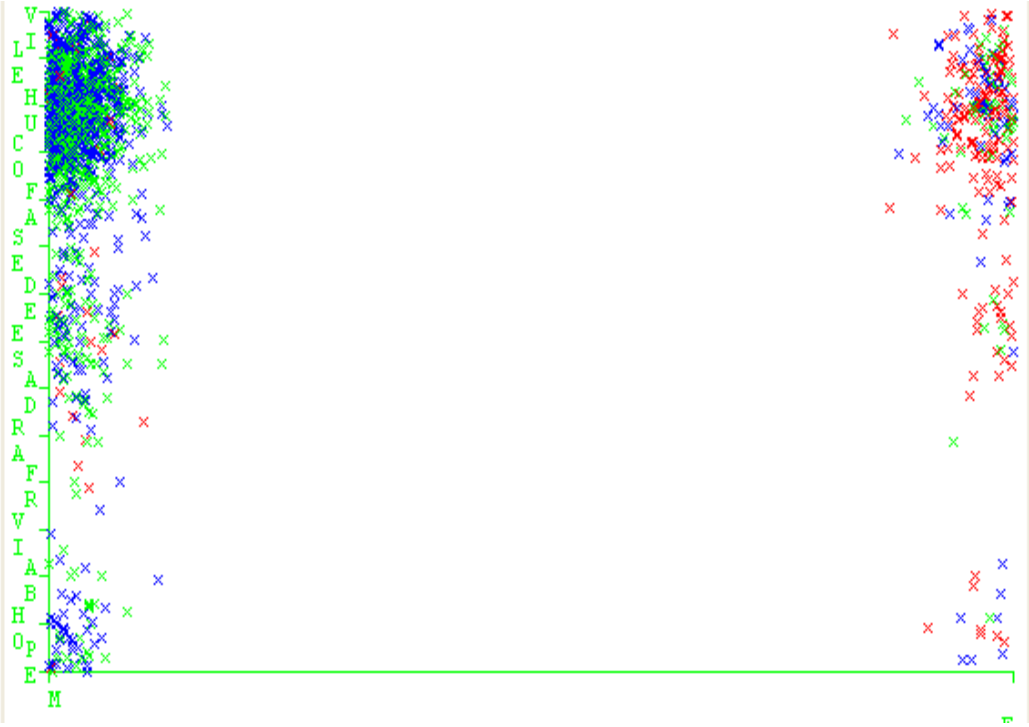

Figura 21. Gráfica que representa la relación entre el delito y el sexo de las personas.

En esta gráfica se puede apreciar que las personas que más delitos cometen son de sexo masculino, y que los delitos más comunes son violación, hurto y falsificación. Estos delitos son los más comunes tanto en personas con sexo femenino como masculino.

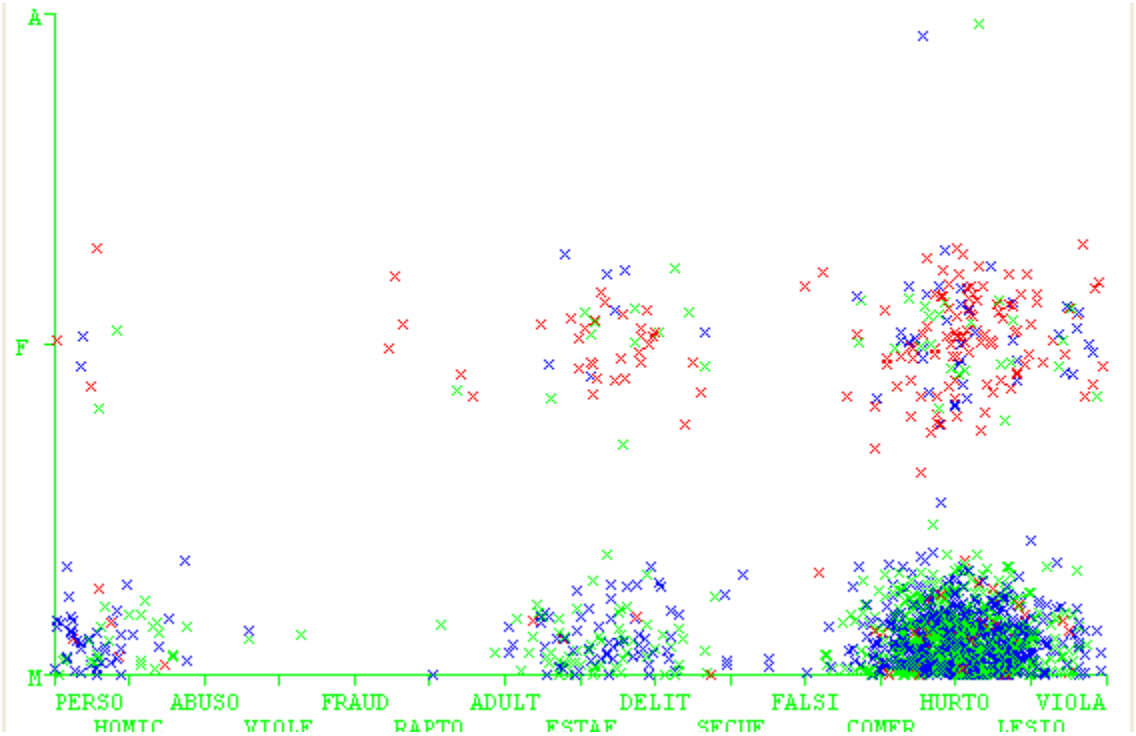

Figura 22. Relación entre el delito y el sexo de la víctima.

Aquí se puede determinar que las víctimas son en su mayoría de sexo masculino. Que la mayoría de los delitos cometidos son Hurto, Lesiones, Violación.

Anexo3. Árbol de decisión.

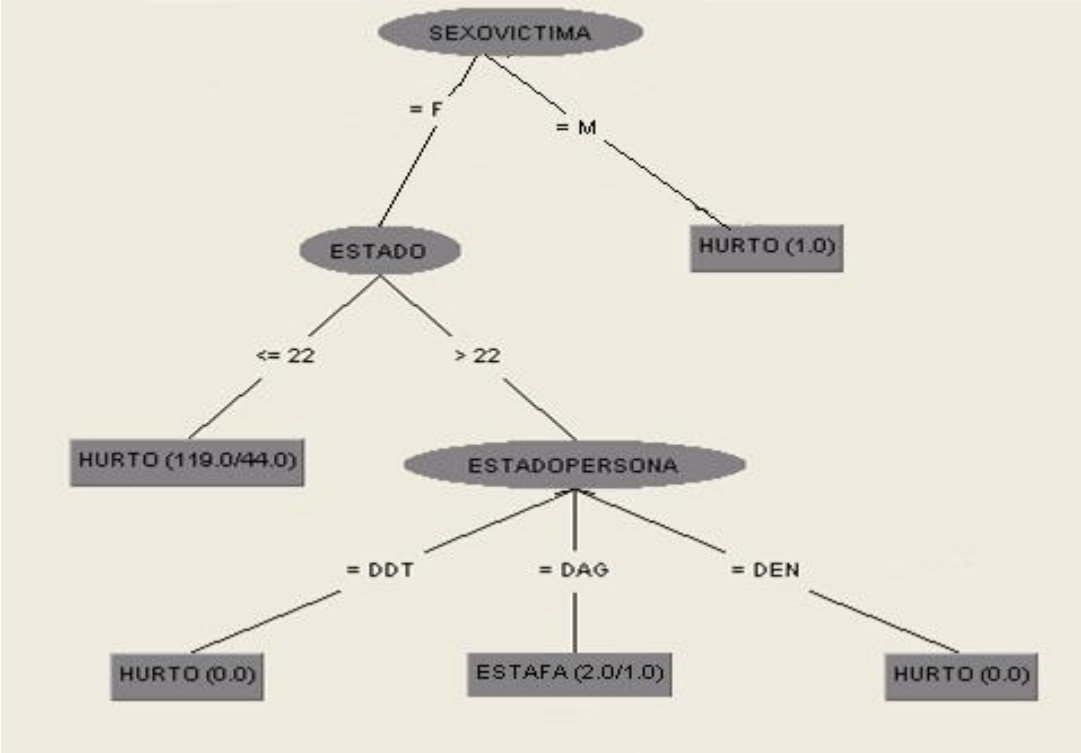

Figura 23. Fragmento del árbol de decisión generado utilizando el algoritmos C4.5 implementado en la herramienta Weka sobre el conjunto de datos de prueba del Sistema de Investigación e Información Policial.

En este fragmento del árbol se puede apreciar que las víctimas de sexo femenino que pertenecen a los estados de Acosta, Falcón, Escuque, Juan Vicente Campo Elías, La Ceiba, Monte Carmelo, Mario Briceño Iragorry, Autana, El Socorro, San José de Guanipa, Carlos Arvelo, La Cañada de Urdaneta, Iribarren, Torres, Alberto Adriani, Territorio Nacional 3, Guaraque, Julio César Salas, Tubores, Agua Blanca, Santiago Mariño. El delito cometido sobre ellos es generalmente hurto. Si los estados son (Urdaneta, Zamora, Atures, Unión, Mariño, Mejia, Manuel Ezequiel Bruzual, Piar, Juan José Mora, Palmasola, Piritu, Jesús María Semprún, José Félix Ribas, Juan Germán Roscio, Morán, Santa Bárbara, Bruzual, Aricagua, Caracciolo Parra Olmedo, Maneiro, Villalba, Esteller, Guanare, Guanarito, Nirgua, Cárdenas, Rómulo Gallegos, Girardot, José Angel Lamas, Cajigal, Valencia, Pampán, El Callao, Montalbán, Lobatera, San Cristóbal, Silva, Buroz, El Hatillo, Buchivacoa, Francisco de Miranda, Colina, Atabapo, Urumaco, Julián

Mellado, Pedro María Freites, Jacura, Los Taques, Bejuma, Petit, Punceres, Díaz, Gómez, Andres Bello, Antonio Romulo Costa, Territorio Nacional 1, San Sebastián, Andrés Bello, Maroa, Manapiare, Río Negro, Anaco, Baralt, Barinas, Catatumbo, Francisco Javier Pulgar, Casacoima, Guasimos, Turén, Sosa, Pedro María Ureña, San Judas Tadeo, Lima Blanco, Pueblo Llano, Santos Marquina, Tulio Febres Cordero, Rosario de Perijá, Santa Rosalía, Montes, Ribero, Dabajuro, Carache, Simón Bolívar, José María Vargas, Junín, Libertad, Diego Bautista Urbaneja, Mauroa, Heres, Raúl Leoni, Diego Ibarra, Lagunillas, Mara, Ortiz, Pedro Zaraza, Sotillo, Arismendi, Territorio Nacional 2, San Casimiro, Antonio Pinto Salinas, Marcano, Ospino, Manuel Monge, Urachiche, Veroes, Caripe, Tovar, Antonio José de Sucre, Cruz Paredes, San Fernando, Aguasay, Antonio Díaz, Benítez, Naguanagua, Uribante, Rangel, Cristobal Rojas, Simon Bolívar, Biruaca, Páez, Pedernales, Tucupita, Bolívar, Sucre, Jose Felipe Marquez Cañizales, Camaguán, San Gerónimo de Guayabal, Las Mercedes, Santa Ana, Simón Rodríguez, Sifontes, Los Guayos, Santa María de Ipire, Ayacucho, Araure, Monseñor José Vicente de Unda, Cedeño, Ezequiel Zamora, Maturín, Francisco Linares Alcántara, Pedraza, José Rafael Revenga, Colón, Cruz Salmerón Acosta, San Joaquín, Caroní, Torbes, Tocópero, Obispo Ramos de Lora, Zea, Brion, Muñoz, Independencia, Juan Antonio Sotillo, Juan Manuel Cajigal, José Gregorio Monagas, Candelaria, Pedro Gual, Píritu, San Juan de Capistrano, Guacara, Libertador, Machiques de Perijá, Andrés Eloy Blanco, Jiménez, Palavecino, Uracoa, Antolín del Campo, Trujillo, José Antonio Páez, Vargas, Arzobispo Chacón, Cardenal Quintero, Justo Briceño, Francisco del Carmen Carvajal, Peña, Pedro Camejo, Cabimas, Andrés Mata, Gran Sabana, Seboruco, Padre Noguera, Baruta, Carrizal, Lander, Santa Rita, San Genaro de Boconoito, San Rafael de Onoto, Chaguaramas, Acosta, Falcón, Escuque, Juan Vicente Campo Elías, La Ceiba, Monte Carmelo, Mario Briceño Iragorry, Autana, El Socorro, San José de Guanipa, Carlos Arvelo, La Cañada de Urdaneta, Iribarren, Torres, Alberto Adriani, Territorio Nacional 3, Guaraque, Julio César Salas, Tubores, Agua Blanca, Santiago Mariño, Ocumare de La Costa de Oro, Alberto Arvelo Torrealba, Obispos, Rojas, Garcia de Hevia, Rafael Rangel, Samuel Darío Maldonado, Tinaco, Pao de San Juan Bautista, Cacique Manaure, Boconó, Paz Castillo, Plaza, Leonardo Infante, Sir Arthur Mc Gregor, Achaguas, Maracaibo, Federación, San José de Guaribe, Arístides Bastidas, La Trinidad, Alto Orinoco, Campo Elías, Fernando de Peñalver, Península de Macanao, San Felipe, Almirante Padilla, Bermúdez, San Diego, Motatán, Pampanito, San Rafael de Carvajal, Anzoátegui, Ricaurte, San Carlos, Acevedo, Los Salias, Papelón, Valdez, Guanta, Miranda, Jose Tadeo Monagas, Jáuregui, Roscio, Padre Pedro Chien, Monseñor Iturriza, Crespo, Simón Planas, García, Valera, Cocorote, Valmore Rodríguez, Aragua, Santos Michelena, Camatagua, Jesús Enrique Lossada, Córdoba, Fernández Feo, Puerto Cabello, Michelena,

Panamericano, Rafael Urdaneta, San Francisco, Rivas Davila, Chacao, Guaicaipuro, Carirubana, Democracia, Paez), si el estado de la persona es detenido (DDT) entonces el delito cometido es un hurto, si el estado es diagnosticado (DAG) entonces el delito cometido es una estafa.

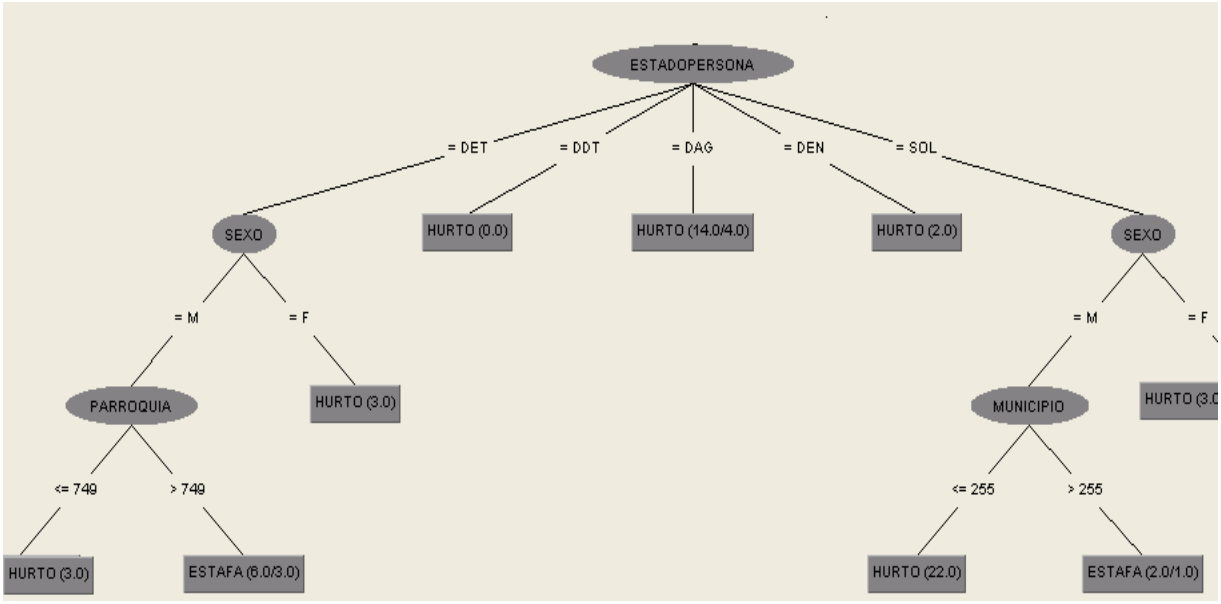

Figura 24. Fragmento del árbol de decisión generado utilizando el algoritmo C4.5 y el conjunto de datos de prueba del Sistema de Investigación e Información Policial.

En este subárbol de se puede determinar el siguiente patrón: Las personas que se encuentran en estado detenido si su sexo es femenino, el delito que cometieron es hurto, su es sexo es masculino y si zona de residencia pertenece a la parroquia con id inferior a 470 el delito cometido es hurto, si el id es superior a 740 el delito cometido es estafa.

GLOSARIO DE TÉRMINOS.

- CICPC: Cuerpo de Investigaciones Científicas Penales y Criminalísticas.
- SIIPOL: Sistema de Investigación e Información Policial.
- Diagrama: Representación gráfica de una colección de elementos de modelado.
- Data Mining: Minería de Datos, proceso en el que se convierten los datos en conocimiento.
- PLSQL: Lenguaje de programación, estructurado de consulta utilizado para realizar la consultas en Oracle.
- Conocimiento: Es el resultado del proceso de minado.
- Data Warehouse: Almacén de datos, es una colección de datos orientada a un ámbito, son un gran conjunto de datos, que por lo general se divide en pequeños subconjuntos.
- Patrones: Un objeto o sustancia que se emplea como muestra para medir alguna magnitud o para replicarla.
- Clases: Abstracción que representa un atributo y los posibles valores que este puede tomar.
- Valores categóricos: Variables no numéricas, que tienen un conjunto de valores finito,
- Clúster: Conjunto o racimo de objetos, que tienen características comunes.
- Normalización: Estandarización de los datos, con el objetivo de lograr un mejor acoplamiento de estos.
- Sesgo: Propiedad de una muestra estadística que hace que los resultados sea representativo para toda la población.
- Ingente: Enorme, inmenso.