UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 7

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

Técnicas de Minería de Datos aplicadas al estudio de la Hipertensión Arterial

Autores:

Kirenia Helen León Rodríguez

Frank Davila Hernández

Tutor:

Ing. Yovannys Sánchez Corales

Asesor:

Dr. Edilberto Fernández Cumbá

La Habana, junio 2011

"Año 53 del Triunfo de la Revolución"

Datos de Contacto

Nombre: Ing. Yovannys Sánchez Corales: Profesor asistente graduado en el año 2005 de Ingeniero en Informática en la CUJAE. Pertenece al Centro de Desarrollo de Software para la Salud. Ha impartido las asignaturas de Inteligencia Artificial, Programación III y Práctica Profesional. Forma parte del proyecto de Atención Primaria de la Salud.

Correo electrónico: yscorales@uci.cu

Nombre: Dr. Edilberto Fernández Cumbá: Profesor asistente. Graduado en 1994 como Médico General. Especialista de I Grado en Medicina General Integral, año 2000. Vice-Director Docente y de Investigaciones Policlínico Docente Ernesto Che Guevara. Ha impartido en el pre-grado las asignaturas de Medicina General Integral I, II, III y IV, Ciencias de la Salud e Informática Medica II (Bioestadísticas). En el posgrado los cursos con los siguientes títulos Hacia una Universidad Promotora de Salud e Informática y Salud, en la actualidad coordinador y profesor principal del Diplomado Informática y Salud. Master en Tecnología de los procesos educativos.

Correo electrónico: efdezc@uci.cu, edilberto.fernandez@infomed.sld.cu

Resumen

La minería de datos se ha convertido en una herramienta muy potente en el mundo actual por la necesidad de encontrar métodos que descubran la información oculta dentro de grandes volúmenes de datos. Las ventajas de estas técnicas son muy usadas por las grandes empresas para el análisis de información, siendo estas de gran aplicación en el campo de la medicina.

El siguiente trabajo de diploma está basado en la investigación realizada para encontrar dos modelos que contribuyan al estudio y diagnóstico de la Hipertensión Arterial usando técnicas de minería de datos. Para ello se plantea la extracción del conocimiento a partir del almacén de datos perteneciente al Sistema Integral para la Atención Primaria de la Salud (alas SIAPS), departamento que se encuentra en el Centro de Informática Médica (CESIM), encargado del desarrollo de aplicaciones para el sector de la salud, este a su vez se ubica en la Universidad de las Ciencias Informáticas (UCI). El desarrollo de la investigación se rige por la metodología más utilizada actualmente en los procesos de Knowledge Discovery in Databases (KDD): CRISP-DM 1.0 y se apoya en la herramienta de libre distribución WEKA 3.6.2 de gran prestigio entre las utilizadas para el modelado de minería de datos. Se espera como resultado obtener dos modelos usando técnicas de minería de datos que contribuyan a la detección y diagnóstico de la Hipertensión Arterial; apoyando a los especialistas en el proceso de toma de decisiones y mejorando de esta forma el servicio sanitario ofrecido.

Palabras claves: minería de datos, CRISP-DM, Hipertensión Arterial, KDD, Weka.

Tabla de Contenido

TABLA DE CONTENIDO

Introducción

Con el triunfo de la Revolución Cubana el primero de enero de 1959, el Comandante en Jefe Fidel Castro Ruz, ha expresado la necesidad de un Sistema de Salud Pública multidisciplinario, donde toda la sociedad se involucre en aras de elevar la calidad de vida de la población cubana. Las ansias por cumplir este deseo han servido de aliento para promover, desde entonces, un conjunto de medidas y acciones prácticas en aras de lograr ese objetivo.

Actualmente la salud en Cuba está dando gigantescos pasos en los procesos de informatización de sus servicios; la Universidad de las Ciencias Informáticas (UCI), conjuntamente con el Ministerio de Salud Pública (MINSAP) y otras empresas de software y hardware se encuentran a la vanguardia en este sentido.

Aunque el país no cuenta con la infraestructura tecnológica suficiente para responder a las demandas actuales en este sector, la informatización permite lograr la utilización futura de estos servicios y la formación de recursos humanos capaces de utilizar la informática de forma eficiente en función del desarrollo socio-económico del país. A partir de esto, no cabe duda alguna que lograr una eficiente gestión de la información en los servicios de salud en Cuba se impone, así como el establecimiento de un mecanismo inteligente que sirva de apoyo a los especialistas en el proceso de toma de decisiones.

Disminuir el error médico, mejorar los procesos de salud y garantizar el cuidado de los pacientes ha sido foco de preocupación constante de todos los miembros del equipo de salud. En este contexto surgen los Sistemas Clínicos de Soporte para la Toma de Decisiones (*Clinical Decisión Support System – CDSS*) los cuales son un componente fundamental del proceso que conlleva la informatización de la capa clínica (1).

Con la evolución de las tecnologías gran cantidad de datos han podido ser estudiados y clasificados a partir de la minería de datos, creando para ello un gran número de estrategias. Una de las principales ventajas en la utilización de la minería de datos en el desarrollo de los Sistemas Clínicos de Soporte para la Toma de Decisiones ha sido su capacidad de generar nuevo conocimiento.

La Universidad de las Ciencias Informáticas (UCI) cuenta con varios centros de desarrollo de software. El Centro de Informática Médica (CESIM) es uno de ellos, encargado del desarrollo de aplicaciones para el sector de la salud; entre estas se encuentra el Sistema Integral para la Atención Primaria de la Salud (alas SIAPS), el cual posee un componente de tipo CDSS (Sistema Clínico de Soporte para la Toma de

Decisiones), para que facilite el procesamiento analítico en línea y la minería de datos y que servirá además al resto de los ambientes bajo un escenario tecnológicamente sólido (2). Actualmente en el Centro de Toma de Decisiones se está manejando la información con técnicas estadísticas, sin embargo con estas técnicas no se está aprovechando al máximo el conocimiento almacenado.

Las Historias Clínicas Electrónicas (HCE) pertenecientes al Sistema Integral para la Atención Primaria de la Salud, se encuentran almacenadas en un gran repositorio y su información se envía periódicamente a un Almacén de Datos. Dado el gran volumen de datos acumulado en él, y la incapacidad de los especialistas de identificar patrones de comportamiento y extraer conocimiento oculto en los datos almacenados para apoyar sus decisiones, surge la necesidad de aplicar la minería de datos a dicho proyecto.

En la actualidad la Hipertensión Arterial se ha convertido en una de las primeras causas de muertes en el mundo, según las bibliografías esta enfermedad no es más que el aumento de la presión arterial de forma crónica. Algunos autores como (3), coinciden que anualmente existen 15 millones de muertos por enfermedades circulatorias, 7.2 millones de muertes por enfermedades del corazón y 4.6 millones de muertes por Accidentes Vasculares Encefálicos (AVE).Todas las patologías mencionadas son producidas por la HTA, cuando no se logra un control adecuado. (4)

Las dificultades más frecuentes asociadas al diagnóstico de la Hipertensión Arterial se describen a continuación:

- ✓ Las decisiones importantes que se toman para el diagnóstico de la Hipertensión Arterial se hacen en base a la experiencia y la intuición de los especialistas, más que aprovechando la rica información almacenada.
- ✓ Se dificulta la toma de decisiones de los especialistas para realizar un análisis rápido y efectivo y de esta manera encontrar información útil y valiosa oculta en los datos.
- ✓ La no predicción del comportamiento futuro de la Hipertensión Arterial con un alto porcentaje de certeza, basado en el entendimiento del pasado de personas que han sufrido esta enfermedad.
- ✓ No existe un componente inteligente que clasifique a los pacientes en cuanto a las similitudes de los factores de riesgos y así ayudar a los especialistas en el diagnóstico de la Hipertensión Arterial.

En función de utilizar el conocimiento almacenado en el repositorio se propone el siguiente **problema a resolver**: ¿Cómo aprovechar la información almacenada en el repositorio de Historias Clínicas

Electrónicas perteneciente al Sistema Integral para la Atención Primaria de la Salud en materia de toma de decisiones?

Para dar respuesta a esta interrogante, este trabajo incluye como **objeto de estudio** el aprovechamiento de la información almacenada en el repositorio del Sistema Integral para la Atención Primaria de la Salud. Se concibió como **objetivo general:** Definir un modelo empleando técnicas de minería de datos que tribute al diagnóstico y detección de pacientes con riesgos a sufrir Hipertensión Arterial, siendo el **campo de acción** las técnicas de minería de datos aplicables a la rama de la salud.

Por tal razón se plantea la **idea a defender**: Utilizando una metodología ideal con técnicas de minería de datos en el repositorio del Sistema Integral para la Atención Primaria de la Salud se aprovechará el conocimiento almacenado en la misma para la detección y diagnóstico de la Hipertensión Arterial.

Para lograr el objetivo propuesto se le dará cumplimiento a las siguientes tareas:

- ✓ Consultar bibliografías sobre las técnicas, herramientas y metodologías a emplear en el proceso de minería de datos.
- ✓ Aplicar las técnicas de minería de datos a los conjuntos de datos seleccionados para el estudio.
- ✓ Evaluar los patrones obtenidos a partir de las técnicas de minería de datos aplicadas.
- ✓ Realizar las pruebas a los algoritmos utilizados
- ✓ Comprobar la veracidad del modelo propuesto y el (los) algoritmos escogidos.
- ✓ Validar la propuesta de solución mediante comparaciones.

Estrategia de Investigación.

Para la realización de esta investigación, se siguió una estrategia descriptiva, donde se le da menor importancia a las causas que originan el problema en el cual, el principal objetivo es la profundización teórica del planteamiento investigativo, describir el fenómeno y, reflejar lo esencial y más significativo del mismo para llegar a los resultados esperados. Para obtener una solución concreta de la investigación, se hace uso de los siguientes métodos investigativos:

Métodos Teóricos.

- ✓ **Análisis Histórico-Lógico:** Se pone de manifiesto en la realización de los estudios de las causas que originaron el problema, así como en el análisis de las técnicas y algoritmos existentes.

- ✓ **Analítico-Sintético:** A través este método se realiza el análisis de la solución propuesta mediante los patrones médicos con el objetivo de facilitar la toma de decisiones de los expertos.

Métodos Empíricos

- ✓ **Entrevista:** Se basa en las entrevistas para validar cuáles de los datos que se tenían en el almacén de datos eran necesarios controlar y cuáles se podían descartar, además para la obtención de un conocimiento manejando términos, diagnósticos y tratamiento a la enfermedad que se analiza.

Esta investigación tendrá un aporte práctico basado en que el Sistema Integral para la Atención Primaria de la Salud contará con un soporte de toma de decisiones que lo convertirá en un sistema más robusto. El mismo permitirá acelerar el proceso de análisis de la información de los especialistas en la toma de decisiones médicas.Una vez demostrada su aplicación y garantía de funcionamiento, este sistema puede ser extendido a todas las esferas de la medicina.

El documento está estructurado por tres capítulos:

CAPÍTULO 1. Fundamentación Teórica: Contiene los aspectos esenciales para entender el entorno del problema a resolver. Se describen los conceptos fundamentales asociados al dominio del problema, sistemas similares existentes vinculados a las técnicas de minería de datos, así como las tendencias y las tecnologías más usadas.

CAPÍTULO 2. **Fundamentación de la propuesta de solución.** Se realiza una descripción de las 3 primeras fases propuestas por la metodología CRISP-DM: *Análisis del problema, Análisis de los datos y Preparación de los datos*; así como las actividades implícitas en las mismas, para realizar el proceso de minería.

CAPÍTULO 3. **Resultados y Discusión.** En este capítulo se analizan los resultados obtenidos explicando las técnicas, herramientas y algoritmos que se utilizarán en el proceso de minado. Se realiza una descripción de las 3 últimas fases propuestas por CRISP-DM: *Modelado, Evaluación y Explotación*; así como las actividades que se ejecutan dentro de estas. Se demuestran las ventajas de la utilización de la minería de datos.

Capítulo 1. Fundamentación Teórica

En el presente capítulo se profundiza en el fundamento teórico de la minería de datos, tareas y algoritmos empleados para obtener modelos o patrones a partir de los datos. También se incluye un estudio sobre las metodologías, tecnologías, herramientas, lenguajes y notaciones empleadas para obtener el conocimiento de un conjunto de datos.

1.1. Tipos de Conocimiento

Antes de adentrarse en el tema del descubrimiento del conocimiento y de minería de datos se deben comprender e identificar los tipos de conocimientos que se poden extraer de una Base de Datos.

El conocimiento se puede clasificar según las siguientes categorías:

- ✓ ***Evidente:*** esta información se puede obtener de las Bases de Datos a través de consultas SQL.
- ✓ ***Multidimensional:*** modela una tabla con *n* atributos como un espacio de *n* dimensiones, lo que permite detectar varias situaciones difíciles de observar. Este tipo de análisis se logra utilizando herramientas OLAP (Online Analytical Processing) o Procesamiento Analítico en Línea.
- ✓ ***Oculto:*** es la información no evidente, desconocida hasta el momento, pero potencialmente útil, que puede obtenerse a través de técnicas de minería de datos. Esta información tiene un gran valor, ya que hasta el momento no se conocía, y descubrirla permite tener una nueva visión del problema y de su solución. (5)

1.2. Descubrimiento de Conocimiento en Bases de Datos (KDD).

Las siglas KDD provienen de Knowledge Discovery in Databases, que significa Descubrimiento de Conocimiento en Bases de Datos, la misma se define como “la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos”. (6)

El proceso de KDD toma los resultados tal como vienen de los datos, los transforma en información útil y entendible. KDD puede usarse como un medio de recuperación de información, de la misma manera que los agentes inteligentes realizan la recuperación de información en la Web. En la imagen 1 se presentan los procesos que involucra el descubrimiento de conocimiento en las bases de datos, los mismos serán explicados en mayor o menor medida en el transcurso del presente documento.

Imagen 1. Procesos que involucra el Descubrimiento de Conocimiento.

1.2.1. Fases de KDD.

KDD define un conjunto de fases para guiar el desarrollo de un proyecto como se muestra en la Imagen 2. En cada una de ellas se generan tareas; estas, aunque son flexibles, deben tenerse en cuenta en su totalidad pues de su estricto cumplimiento depende la calidad del conocimiento que se obtenga una vez finalizado el proceso.

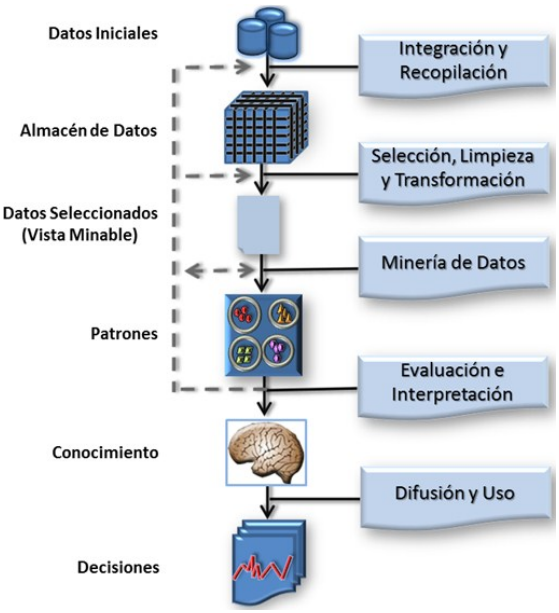

Imagen 2. Fases de KDD.

1.3. Minería de datos

La minería de datos es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Aunque se basa en varias disciplinas, algunas de ellas más tradicionales, se distingue de ellas en la orientación más hacia el fin que hacia el medio, hecho que permite nutrirse de todas ellas sin prejuicios. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido a la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones. (7)

Pese a la popularidad del término, la minería de datos es sólo una etapa, si bien la más importante, de lo que se ha venido llamando el proceso de extracción de conocimiento a partir de datos. Este proceso consta de varias fases e incorpora diferentes técnicas de los campos del aprendizaje automático, la estadística, las bases de datos, los sistemas de toma de decisión, la inteligencia artificial y otras áreas de la informática y de la gestión de información. (8)

La minería de datos no aparece por el desarrollo de tecnologías esencialmente diferentes a las anteriores, sino que se crea, en realidad, por la aparición de nuevas necesidades y, especialmente, por el reconocimiento de un nuevo potencial: el valor, hasta ahora generalmente infrautilizado, de la gran cantidad de datos almacenados informáticamente en los sistemas de información de instituciones, empresas, gobiernos y particulares. Los datos pasan de ser un "producto" (el resultado histórico de los sistemas de información) a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento; un conocimiento que ha de ser especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos. (9)

Entre las múltiples definiciones que identifican a la minería de datos se encuentran:

“…el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en Bases de Datos, Data-Warehouse, o cualquier otro medio de almacenamiento de información”. (10)

“…término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos”. (11)

Los retos de la minería de datos son: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información con los problemas que ello conlleva (ruido; datos ausentes, intratabilidad, volatilidad de los datos...), y por el otro: usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil.

1.3.1. Ciclo de vida en espiral de un proyecto de minería de datos.

El primer programa de minería de datos no ha de contemplar todos los aspectos mejorables en la gestión de la organización. En primer lugar, porque sería imposible contemplarlos todos y, en segundo lugar, porque haría inviable el proyecto. La alternativa consiste en elegir aquellos aspectos o necesidades más claras y relevantes y, posteriormente, una vez realizados esos objetivos, plantearse otros. (12)

Por tanto, la primera implantación debe marcarse unos problemas concretos y que tengan unos beneficios manifiestos. Esta primera "ronda" puede constituir lo que vulgarmente se denomina proyecto "piloto". Esta aproximación cíclica es similar al ciclo de vida en espiral de la ingeniería del software. (13)

Como se muestra en la Imagen 3 el primer ciclo completo engloba las fases de identificación de problemas del negocio, la planificación y organización, identificando los problemas de minería de datos, las fases de extracción de conocimiento, de difusión, despliegue y explotación de modelos y, finalmente, la evaluación de resultados, midiendo los costes y beneficios.

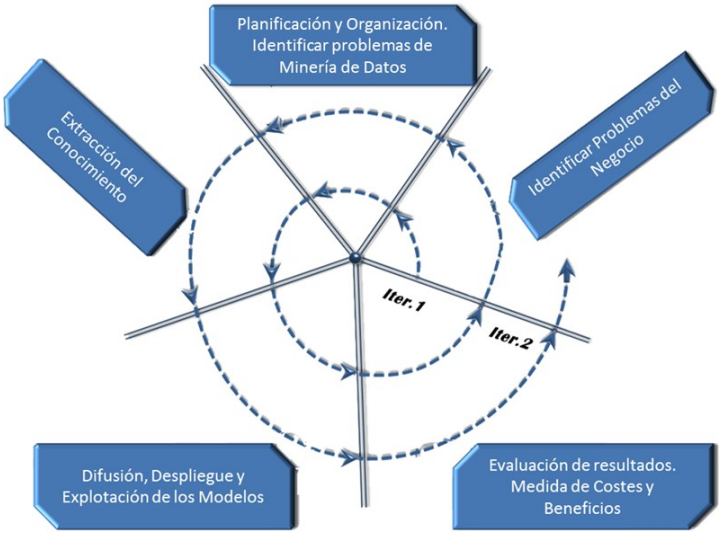

Imagen 3. Progreso en espiral de las acciones de Minería de datos en una organización.

Este ciclo corresponde con un mini proyecto piloto de minería de datos, donde se recomienda que las tareas de la minería de datos sean principalmente dirigidas. Del primer proyecto y, especialmente, de lo aprendido (de la evaluación de costes, beneficios obtenidos, etc.), se pueden plantear objetivos más ambiciosos y, en sucesivos ciclos, se puede plantear una minería de datos no dirigida con este ciclo en espiral, la pregunta de cuánto tiempo se necesita para implantar un programa de minería de datos es baldía. (14)

A medida que va avanzando el tiempo se van planteando objetivos más generosos. Aun así, subsiste la pregunta de ¿cuánto debe durar el primer ciclo? la respuesta es difícil de responder, pero debería durar entre unas semanas a unos meses. En general, no se debe dejar esperar más de seis meses para tener los primeros resultados, aunque fueran simplemente los más sencillos. Ser capaz de dar una vuelta completa al ciclo anterior es quizás la garantía de que se va a poder continuar con las siguientes vueltas. También el primer ciclo piloto representa, en cierto modo, un experimento controlado (con beneficios y costes pequeños) a partir del cual se puede tomar la decisión de ampliar el programa de una manera más ambiciosa. (15)

1.3.2. Fases de un proceso típico 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.

- **Selección y pre-procesado de datos**

El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse...) nunca es el idóneo y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto". Mediante el pre-procesado se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos, según las necesidades y el algoritmo que va a usarse), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reduce el número de valores posibles (mediante redondeo, clustering...). (16)

- **Selección de 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:

1. Aquellos basados en la elección de los mejores atributos del problema.
2. Aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos. (17)

- **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. (18)

- **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. (19)

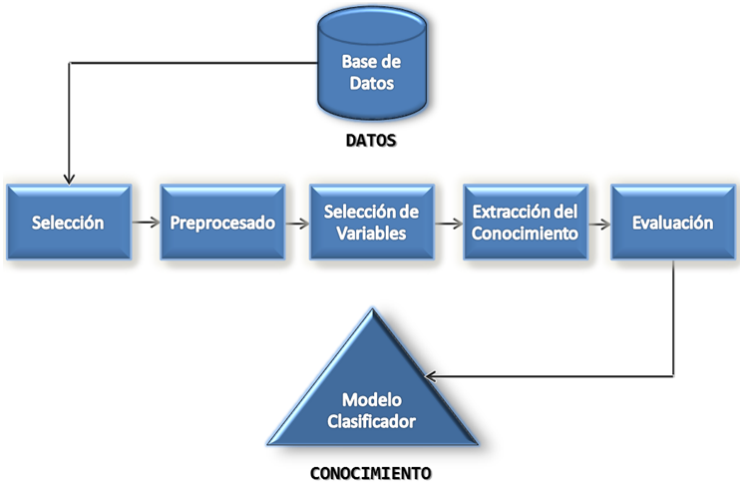

Imagen 4. Fases de la Minería de datos.

1.3.3. Técnicas de minería de datos

Las técnicas de la minería de datos provienen de la Inteligencia Artificial y de la Estadística. Dichas técnicas no son más que algoritmos, más o menos sofisticados, que se aplican sobre un conjunto de datos para obtener unos resultados. Las técnicas más representativas son:

- **Redes neuronales:** procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:
    1. El perceptrón.
    2. El perceptrón multicapa.
    3. Los mapas auto-organizados, también conocidos como redes de Kohonen. (20)

- ✓ **Árboles de decisión:** un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Como por ejemplo:
    1. Algoritmo ID3
    2. Algoritmo C4.5. (21)

- ✓ **Modelos estadísticos:** es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta. (22)

- ✓ **Agrupamiento o Clustering:** es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia. Se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tenga características comunes. Ejemplos:
    1. Algoritmo K-medias
    2. Algoritmo K-medoides. (23)

- ✓ **Técnicas bayesianas:** Son fáciles de usar, muy eficientes, pueden tratar muchos atributos (cientos o miles), son muy robustos al ruido, la expresividad es limitada y depende de la discretización, son estables a la muestra. Al igual que las técnicas anteriores no construyen modelos, sólo estiman una serie de probabilidades (a excepción de los modelos gráficos probabilísticos, donde la red bayesiana creada puede ser muy informativa). (24)

- ✓ **Técnicas relacionales y declarativas:** Son técnicas muy expresivas que permiten tratar datos con estructuras y capturas patrones relaciones y recursivos, así como expresar el conocimiento previo en forma de reglas. Los mayores inconvenientes de estas técnicas son la dificultad de manejo (hay que saber expresar los ejemplos convenientes) y la poca eficiencia, en general, de los métodos

existentes. Sobre el resto de rasgos (robustez al ruido, estabilidad ante la muestra, precisión, etc.), la variedad de métodos difieren en muchos de ellos. Quizá la ventaja más importante, además de la gran expresividad, sea que los modelos son comprensibles. (25)

- ✓ **Técnicas Supervisadas.**

Expresado en una forma breve, el objetivo del aprendizaje supervisado es: a partir de un conjunto de ejemplos, denominados de entrenamiento, de un cierto dominio D de ellos, construir criterios para determinar el valor del atributo clase en un ejemplo cualquiera del dominio. Esos criterios están basados en los valores de uno o varios de los otros pares (atributo; valor) que intervienen en la definición de los ejemplos. Es sencillo transmitir esa idea al caso en el que el atributo que juega el papel de la clase sea uno cualquiera o con más de dos valores. Dentro de este tipo de aprendizaje se pueden distinguir dos grandes grupos de técnicas: la predicción y la clasificación. (26)

- ✓ **Técnicas No Supervisadas.**

El aprendizaje inductivo no supervisado estudia el aprendizaje sin la ayuda del maestro; es decir, se aborda el aprendizaje sin supervisión, que trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la guía del atributo especial clase. Este es el proceder de los sistemas que realizan agrupamiento conceptual y de los que se dice también que adquieren nuevos conceptos. Otra posibilidad contemplada para estos sistemas es la de sintetizar conocimiento cualitativo o cuantitativo, objetivo de los sistemas que llevan a cabo tareas de descubrimiento. (27)

1.3.4. Algoritmos de Extracción de Conocimiento.

1.3.4.1. Descripción de los Algoritmos a utilizar.

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. (28)

Cada técnica de minería trae asociado uno o más algoritmos destinados a realizar, unos de forma más eficientes que otros, el proceso que conlleva al descubrimiento del conocimiento a partir de los datos que se estudian. A partir de esto se seleccionan dentro de la técnica supervisada Árbol de decisión el

algoritmo J48 y dento de la técnica no supervisada Agrupamiento el algoritmo Simple K-Means, estos algoritmos se describen a continuación:

- **J48:**

El algoritmo J48 implementado en Weka es una versión del clásico algoritmo de árboles de decisión C4.5 propuesto por Quilan 6. Los árboles de decisión entran dentro de los métodos de clasificación supervisada, es decir, se tiene una variable dependiente o clase, y el objetivo del clasificador es determinar el valor de dicha clase para casos nuevos. (29)

El proceso de construcción del árbol comienza por el nodo raíz, el que tiene asociados todos los ejemplos o casos de entrenamiento. Lo primero es seleccionar la variable o atributo a partir de la cual se va a dividir la muestra de entrenamiento original (nodo raíz), buscando que en los subconjuntos generados haya una mínima variabilidad respecto a la clase. Este proceso es recursivo, es decir, una vez que se haya determinado la variable con la que se obtiene la mayor homogeneidad respecto a la clase en los nodos hijos, se vuelve a realizar el análisis para cada uno de los nodos hijos. Aunque en el límite este proceso se detendría cuando todos los nodos hojas contuvieran casos de una misma clase, no siempre se desea llegar a este extremo, para lo cual se implementan métodos de pre-poda y post-poda de los árboles. (30)

El algoritmo J48 amplía las funcionalidades del C4.5, tales como permitir la realización del proceso de post-poda del árbol mediante un método basado en la reducción del error (reduced Error Pruning) o que las divisiones sobre las variables discretas sean siempre binarias (binary Splits) 4.5. Algunas propiedades concretas de la implementación son las siguientes:

1. El algoritmo J48 no es afectado por la introducción de datos que no son altamente significativos en el proceso de aprendizaje.
2. Posibilidad de modelar el resultado del árbol de decisión en lenguaje SQL.
3. Velocidad computacional.
4. Fiabilidad de los resultados. (31)

- **Simple K-Means:**

Para obtener un modelo no supervisado usando agrupamientos, se realizará utilizando el algoritmo Simple K-Means, que pertenece al grupo de algoritmos de partición-optimización. El algoritmo 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.

Este algoritmo fue seleccionado por las ventajas que presenta:

- Velocidad, la cual puede ser considerable cuando se trata de grandes volúmenes de datos.
- Buenos resultados.
- Posibilidad de cambiar los puntos iníciales y obtener resultados diferentes. (32)

1.3.5. Aplicaciones y Tendencias actuales de la minería de datos.

Actualmente el panorama es alentador con respecto al desarrollo de aplicaciones que utilizan la minería de datos. A medida en que los negocios se mueven, las organizaciones aprovechan al máximo las herramientas para mejorar sus servicios. La información es la clave para prosperar en un mercado competitivo. Existen un conjunto de técnicas y herramientas capaces de ayudar a la toma de decisiones de los expertos. A continuación se describen algunas de las técnicas, proyectos y software realizados y que se encuentran operativos en el mercado de la información.

1.3.5.1. Ámbito Internacional.

- **Aplicaciones en el sector de la Salud:**

- ✓ *Aplicación de la minería de datos al estudio de las alteraciones respiratorias durante el sueño. (Santiago de Compostela, España)*. Se desarrolló un almacén de datos teniendo en cuenta los procesos y guías clínicas implicadas en la atención de los pacientes con síndrome de apnea del sueño. Con la información de este estudio se puede decidir el alta, iniciar un tratamiento específico con las consiguientes revisiones dentro del Servicio o ser remitido a otro Servicio. (33)

- ✓ *Aplicación de técnicas de minería de datos para el diagnóstico prematuro de Cáncer. (Madrid, España)*.En el experimento, se consideraron 322 imágenes de una base de datos para ambos sistemas de clasificación. Se utilizaron subconjuntos de la colección de datos y se calcularon los resultados para todos ellos con el objetivo de obtener un resultado sobre la calidad del sistema más exacta. La mamografía es uno de los mejores métodos empleados en la detección de cáncer de mama, pero en algunos casos, los especialistas en radiología no son capaces de detectar

tumores a pesar de su experiencia. Por ello los métodos presentados en este documento pueden ser de gran ayuda a la hora de realizar un diagnóstico por parte del personal médico. (34)

1.3.5.2. Ámbito Nacional.

- **Aplicaciones en el sector de la Salud:**

✓ *Aplicación de la minería de datos para el análisis de información clínica. Estudio Experimental en cardiopatías isquémicas. (La Habana, Cuba)*. Es un trabajo elaborado por especialistas cubanos donde se expone un ejemplo de aplicación de la minería de datos para el apoyo a la toma de decisiones en esta especialidad de la medicina, a partir del estudio de las coronariografías realizadas a pacientes con cardiopatía isquémica. (35)

✓ Una Colaboración entre México y Cuba sobre *Redes Neuronales para la minería de datos y Textos: Aplicación al Análisis Exploratorio y Descubrimiento de Conocimiento en Grandes Bases de Datos de Información Biomédica*, este trabajo señala que tuvo el objetivo de proponerse, agenciarse, diseñar y construir herramientas teóricas y computacionales, basadas en métodos multivariados y adaptativos (ej. redes neuronales), entrenarse en su uso y aplicarlas para hacer análisis exploratorio de datos y textos para el descubrimiento de conocimiento en bancos digitales de información biomédica (en la perspectiva de la ciencia infométrica).

✓ *Empleo de minería de datos en la predicción de diabetes. Pre-procesado de datos*. Donde se expone el desarrollo de las fases de comprensión y pre-procesado de los datos, dentro de la metodología para desarrollar procesos de minería de datos, CRISP-DM 1.0. Como caso práctico, se refleja el trabajo con una serie de datos producto de encuestas realizadas en la localidad de Jaruco (Provincia Mayabeque, Cuba); con el objetivo de determinar factores influyentes en el padecimiento de diabetes.

1.3.5.3. Universidad de las Ciencias Informáticas (UCI).

- **Aplicaciones en el sector de la Salud.**

✓ *"Diagnóstico de Enfermedades de Transmisión Sexual mediante técnicas de Inteligencia Artificial"* (Junio-2009) de la Facultad 5. En este se creó una Base de Hechos para la Blenorragia y otra para la Clamidia, las cuales se procesaron en el software WEKA mediante la técnica Árboles de Decisión (AD), usando para esto el algoritmo ID3. Una vez obtenido el conocimiento se utiliza el

mismo para la creación de una aplicación basada en reglas, que permite diagnosticar si una persona está infectada por las Enfermedades de Transmisión Sexual Blenorragia o Clamidia. (36)

- ✓ *"Proceso de análisis y gestión del conocimiento a partir de los datos obtenidos en la conducción de los de Ensayos Clínicos del Centro de Inmunología Molecular, aplicando técnicas de minería de datos"* (Junio-2009) de la Facultad 6. Aquí se emplean un grupo de técnicas de minería de datos como la clasificación; con el objetivo de predecir el tiempo de supervivencia de un paciente con cáncer de pulmón, para posteriormente encontrar patrones ocultos y reglas que los caractericen; a partir de la variable surrogada "evaluación de la respuesta", basado en las relaciones que se establecen entre las variables de control (sexo, edad, color de piel, estadio clínico, clasificación histológica, peso). (37)

1.4. Metodologías para el Diseño y la Implementación

Cuando se va a realizar un proyecto de minería siempre es necesario contar con una metodología que guíe todo el proceso. De esta manera diversas empresas han especificado y propuesto procesos de modelado con el objetivo de guiar al desarrollador a través de una serie de pasos dirigidos a obtener buenos resultados.

El instituto de Sistemas de Análisis estadísticos (SAS) fue el desarrollador de la metodología SEMMA (*Sample, Explore, Modify, Model, Assess*) para la realización de proyectos de Minería. Por otra parte, en 1999 varias empresas europeas como la NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para desarrollar la metodología CRISP-DM (*Cross-Industry Standard Processfor Data Mining*). Estas metodologías son las más utilizadas en la actualidad para realizar proyectos de minería de datos. (38)

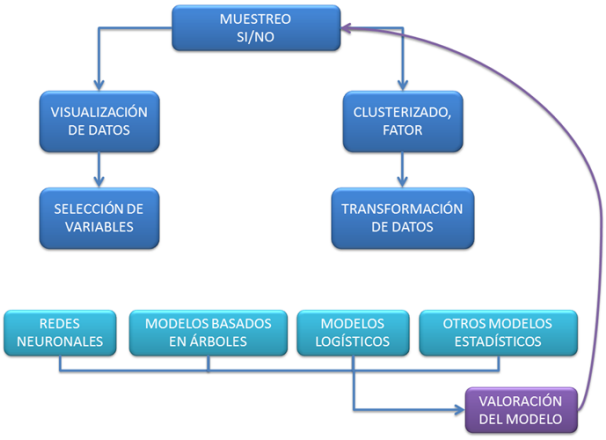

Imagen 5. Metodología SEMMA.

1.4.1. Metodología CRISP-DM

CRISP-DM (*Cross Industry Standard Processfor Data Mining*) es una metodología de libre distribución que puede trabajar con cualquier herramienta para desarrollar cualquier proyecto .Esta metodología estructura el ciclo de vida de un proyecto de minería de datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. (39)

La metodología CRISP-DM es una herramienta de trabajo que surge con la necesidad de aprender nuevas técnicas para aplicar y comprender de mejor manera la minería de datos y sus resultados basándose en un proceso jerárquico ya que está compuesta por diferentes niveles o tareas. Es válido resaltar que esta metodología no es ni la "más actual" ni "la mejor", pero es muy útil para comprender estas técnicas o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características, por lo que es considerada en diferentes textos que tratan sobre inteligencia de negocio como la metodología de minería de datos para inexpertos. (40)

En la imagen 6 se muestra el ciclo de vida que propone la metodología CRISP-DM para un proyecto de minería de datos, cada fase que la misma propone engloba tareas que permiten adentrarse en las próximas, de esta manera se explican de forma general que debe hacerse y cuales son los pasos a seguir surante todo el proceso.

Imagen 6. Fases de la metodología CRISP-DM**.**

Comprensión del negocio: (Objetivos y requerimientos desde una perspectiva no técnica): Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

- ✓ Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito).
- ✓ Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio,…).
- ✓ Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito).
- ✓ Generación del plan del proyecto (plan, herramientas, equipo y técnicas). (41)

Comprensión de los datos: (Familiarizarse con los datos teniendo presente los objetivos del negocio): La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

- ✓ Recopilación inicial de datos.
- ✓ Descripción de los datos.
- ✓ Exploración de los datos.
- ✓ Verificación de calidad de datos. (42)

Preparación de los datos: (Obtener la vista minable o dataset): La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final los datos que serán provistos en las herramientas de modelado de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

- ✓ Selección de los datos.
- ✓ Limpieza de datos.
- ✓ Construcción de datos.
- ✓ Integración de datos.
- ✓ Formateo de datos. (43)

Modelado: (Aplicar las técnicas de minería de datos a los dataset): En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

- ✓ Selección de la técnica de modelado.
- ✓ Diseño de la evaluación.
- ✓ Construcción del modelo.
- ✓ Evaluación del modelo. (44)

Evaluación: (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio): En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos. Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida.

- ✓ Evaluación de resultados.
- ✓ Revisar el proceso.
- ✓ Establecimiento de los siguientes pasos o acciones. (45)

Despliegue: (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización): Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo.

- ✓ Planificación de despliegue.
- ✓ Planificación de la monitorización y del mantenimiento.
- ✓ Generación de informe final.
- ✓ Revisión del proyecto. (46)

Se selecciona CRISP-DM como metodología de desarrollo a utilizar en el proceso de minería de datos. La selección está sustentada principalmente por las siguientes ventajas:

- ✓ Concibe el proyecto de minería de datos de forma global y estrechamente relacionado al negocio en cuestión.
- ✓ Fue diseñada de forma neutra a la herramienta que se utilice para el desarrollo del proyecto.
- ✓ Es de distribución libre y se encuentra en constante perfeccionamiento por parte de la comunidad internacional.

- ✓ Presenta una precisa y sólida distribución de tareas de carácter general con sus resultados, así como una guía para su desarrollo.
- ✓ Muchas de las metodologías que podemos encontrar en la actualidad se basan en este estándar.
- ✓ Es la que cuenta con mayor aceptación por parte de los desarrolladores de procesos de extracción de conocimientos a partir de datos.
- ✓ CRISP-DM es producto de la experiencia de varias empresas que se dedican a la minería de datos (SPSS, Daimler-Chrysler y NCR; entre otras) y no de un simple estudio teórico.

1.5. Herramientas utilizadas.

Para realizar el pre-procesado, los que deseen extraer conocimientos a partir de datos deben, además de contar con una metodología adecuada: apoyarse en herramientas software que les faciliten la tarea. Para ello se puede emplear todo un arsenal de herramientas.

Se pueden encontrar tanto en ámbitos comerciales como académicos una serie de entornos software diseñado para dar soporte al ejercicio de minería de datos. A continuación se muestran algunos de los entornos de minería de datos más populares que actualmente se encuentran disponibles para el usuario.

- ✓ Yale: RapidMiner
- ✓ Clementine
- ✓ Enterprise Miner
- ✓ IntelligentMiner de IBM. Armonk, NY, USA.
- ✓ Decision Series, de NeoVista Software. Cupertino CA, USA.
- ✓ Darwin de Thinking Machines. Bedford MA, USA.
- ✓ MineSet, de Silicon Graphics. Mountain View, CA, USA.
- ✓ SAS Solution for Data Mining de SAS Institute. Cary, NC, USA.
- ✓ Weka, Universidad de Waikato, Nueva Zelanda.

1.5.1. Weka V3.6.2.

WEKA (*Waikato Environment for Knowledge Analysis*) es una herramienta visual de libre distribución (licencia GNU) desarrollada por un equipo de investigadores de la Universidad de Waikato (Nueva Zelanda). Como entorno de minería de datos conviene destacar:

Acceso a datos: Los datos son cargados desde un archivo en formato ARFF (archivo plano organizado en filas y columnas). El usuario puede observar en los diferentes componentes gráficos, información de

interés sobre el conjunto de muestras (talla del conjunto, número de atributos, tipo de datos, medidas y varianzas de los atributos numéricos, distribución de frecuencias en los atributos nominales, etc.). (47)

Pre-procesado de datos:

- ✓ Selección de atributos.
- ✓ Discretización.
- ✓ Tratamiento de valores desconocidos.
- ✓ Transformación de atributos numéricos. (48)

Modelos de aprendizaje:

- ✓ Arboles de decisión (J4.8, versión propia del método C4.5).
- ✓ Tablas de decisión.
- ✓ Vecinos más próximos.
- ✓ Máquinas de vectores soporte (método *sequential minimal optimization*).
- ✓ Reglas de asociación (método *Apriori*).
- ✓ Métodos de agrupamiento (*K medias, EM y Cobweb*).
- ✓ Modelos combinados (*bagging, boosting, stacking, etc.*). (49)

Visualización: La interfaz gráfica se compone de 4 entornos: Explorer, Consola (CLI), Experimenter y Knowledge Flow.

Después de un estudio comparativo entre las herramientas, se seleccionó WEKA como herramienta a utilizar en el proceso de minería de datos. Está implementada en Java, luego no presenta problemas de portabilidad, siempre y cuando el sistema disponga de la máquina virtual apropiada. Es interesante remarcar, que dado que se trata de una herramienta bajo licencia GNU, es posible actualizar su código fuente para incorporar nuevas utilidades o modificar las ya existentes, de ahí que se puedan encontrar toda una serie de proyectos asociados a WEKA (*Spectral clustering, Kea, WEKA Metal, etc.*) que permiten garantizar la continua evolución y adaptación de dicha herramienta. (50)

Resumiendo, se puede llegar a la conclusión de que la herramienta presenta como ventajas relevantes las que a continuación se listan:

- ✓ De libre distribución.
- ✓ Multiplataforma.
- ✓ Tiene muchos algoritmos de regresión/clasificación.
- ✓ Incluye meta-algoritmos de aprendizaje.

- ✓ Tiene pre-procesado de datos.
- ✓ Incorpora herramientas para la visualización de los datos y resultados.
- ✓ Se distribuye también su código fuente JAVA.
- ✓ Se pueden añadir nuevas clases de clasificadores y filtros.
- ✓ Tiene versiones de consola y con interfaz gráfico. (51)

1.5.2. NetBeans V6.9.

NetBeans es un entorno de desarrollo o IDE (Integrated Development Environment) para todo tipo de tecnologías Java e incluso permite la codificación de programas en C, C++ y otros (aunque está pensado para Java). (52)

Sus funciones son:

- ✓ Editor de código sensible al contenido. Con soporte para autocompletar el código, coloreado de etiquetas, auto tabulación y uso de abreviaturas para varios lenguajes de programación.
- ✓ Soporte para Java, C, C++, XML y lenguajes HTML.
- ✓ Soporte para JSP, XML, RMI, CORBA, JINI, JDBC y tecnologías Servlet.
- ✓ Incluye CVS (control de versiones) y Ant (compilación avanzada).
- ✓ Posibilidad de utilizar otras versiones de compiladores, depuradores, etc.
- ✓ Creación visual de componentes gráficos.
- ✓ Herramientas con asistentes para facilitar la escritura de código. (53)

Con la ayuda de este IDE se pudo realizar un software intermedio de forma tal que se pudieran realizar las transformaciones necesarias en la base de datos para facilitar el trabajo al minero y ahorra tiempo en el proceso de minado.

A partir de todo lo analizado anteriormente puede afirmarse que la minería de datos es una herramienta eficaz para dar respuestas a preguntas complejas de Inteligencia de Negocios. Es una buena manera de convertir datos en información, y esta a su vez en conocimiento, para la correcta toma de decisiones. Además las herramientas disponibles permiten automatizar gran parte de la tarea de encontrar los patrones de comportamiento ocultos en los datos. Se seleccionó como herramienta a usar Weka 3.6.2. De esta misma forma las metodologías exponen un modelo de referencia y una guía para el usuario con orientaciones y consejos más detallados para el desarrollo de cada fase y tarea. Se seleccionó como metodología a utilizar CRISP-DM 1.0 para dar solución al problema planteado.

Capítulo 2. Fundamentación de la propuesta de solución.

Una vez conocida la herramienta y metodología a utilizar y, siguiendo los pasos establecidos por la misma para la construcción del sistema, se hace necesaria la identificación de aquellos indicadores que aporten información relevante acerca del paciente y su estado de salud. En este capítulo se realiza una descripción de las 3 primeras fases propuestas por la metodología CRISP-DM 1.0: *Análisis del problema, Análisis de los datos y Preparación de los datos*; así como las actividades implícitas en las mismas, para realizar el proceso de minería, asimismo se elabora una descripción de los resultados de cada una de estas fases. Los datos seleccionados para realizar el proyecto de minería de datos corresponden a la información del Almacén de Datos de Historias Clínicas Electrónicas perteneciente al Proyecto (alas SIAPS).

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.

2.1. Análisis del problema.

2.1.1. Comprensión del negocio.

- **Resultados del Proyecto.**

Los resultados de esta investigación se deben presentar a los especialistas y serán los siguientes:

1. Descripción y resultados obtenidos en cada una de las fases de CRISP-DM 1.0.
2. Presentar un informe con los resultados obtenidos de la investigación donde se muestren los resultados de los modelos de minería de datos obtenidos.

2.1.2. Criterios de éxito del Negocio.

Los criterios para lograr el éxito de la investigación desde el punto de vista del objetivo del negocio son:

2. Obtener dos modelos de conocimiento y comprobar su validez**.**

3. Desarrollar el proyecto usando la herramienta Weka para la minería de datos.

4. Realizar un proyecto de minería de datos guiado por la metodología CRISP-DM.

2.1.3. Evaluación de la situación.

- **Requerimientos y Restricciones.**

Confiabilidad: Proteger la información en el almacén de datos y restringir el acceso no autorizado a la misma.

Documentación: Se documentará cada una de las fases de la metodología CRISP-DM 1.0.

Fecha de entrega: Junio 2011.

Modalidad de resultados: Entrega de la documentación de forma digital y presentación de los resultados.

Terminología: La terminología se explicara en el glosario de término.

2.1.4. Objetivos de la minería.

Criterios de éxito de la minería: Obtener las predicciones con un valor de certeza igual o superior al 80%.

2.2. Comprensión de los Datos.

2.2.1. Recopilación inicial de los datos.

Los datos provienen de un Almacén de Datos que está montado sobre el gestor PostgreSQL 8.3. Es un sistema de bases de datos objeto-relacional que utilizará como lenguaje de programación el Lenguaje Procedural (PLSQL).Para recolectar los mismos se realizó un análisis de las variables más significativas para el proceso de aprendizaje o para el proceso de modelaje.

Para recolectar los datos necesarios en la investigación se hizo un análisis profundo de la Hipertensión Arterial, para lo cual se le realizaron encuestas a los especialistas en este tema .Las preguntas aplicadas fueron:

- ✓ ¿Cuáles son los principales factores de riesgos de la Hipertensión Arterial?
- ✓ ¿Qué relación existe entre estos factores?

- ✓ ¿Cuáles son los síntomas más frecuentes de la Hipertensión Arterial?
- ✓ ¿Principales Causas de la Hipertensión Arterial?

➢ Criterio de Selección de las variables.

Las variables que serán necesarias para la obtención de la vista minable y su posterior modelado se obtienen a partir de entrevistas realizadas a personal calificado y especialistas en Hipertensión Arterial.

➢ Resultados de las entrevistas.

Una vez concluidas y analizadas las encuestas se evidencia que las respuestas de los especialistas tienen un alto porcentaje de coincidencia. A continuación se exponen los resultados obtenidos:

Tabla 1. Listado de los factores de riesgos mayores.

Antecedentes Personales de Enfermedad Renal.	Antecedentes Personales de Síndrome de apnea de sueño
Antecedentes Personales de Enfermedad Endocrinológica. • Cushing • Hiperaldosteronismo • Feocromocitoma • Acromegalia • Diabetes • Obesidad	Antecedentes Familiares de Hipertensión Arterial. Antecedentes Familiares de Muerte súbita. Antecedentes Familiares de Diabetes Antecedentes Familiares de Gota Ingesta habitual de fármacos. Antecedentes Familiares de Dislipemia Antecedentes Familiares de Enfermedad renal.
Antecedentes Personales de Enfermedad cardiovascular.	Antecedentes Familiares de Enfermedad cardiovascular.
Antecedentes Personales de Enfermedad del sistema nervioso.	

Tabla 2. Listado de los factores de riesgos menores.

Grasas	Café	Sal	Sexo
Ejercicio físico.	Droga	Edad	
Consumo de Tabaco	Raza	Alcohol	

En el Anexo 1 se muestran las tablas del almacén de datos que fueron seleccionadas por contener los atributos expuestos anteriormente.

Según el criterio de los especialistas todos estos factores están fuertemente relacionados entre sí, y mientras más combinados se encuentren en un paciente, mayor será la probabilidad de que sea diagnosticado hipertenso. Es necesario aclarar que algunas personas que padecen la enfermedad son asintomáticas, lo que quiere decir que la ausencia de ellos o alguno de ellos, no exime la posibilidad de padecer hipertensión. Afirman además que está comprobado mediante estudios e investigaciones que los antecedentes patológicos familiares de hipertensión familiar tienen un peso importante en el desarrollo y padecimiento de la enfermedad por el individuo.

Sobre las causas de la enfermedad los especialistas coincidieron en que en aproximadamente de un 90% a un 95% son idiopáticas (de origen desconocido), y solo un 5% se conocen, estas últimas aparecen en edades tempranas como la infancia o adolescencia y son tratables y hasta curables. Entre ellas, por solo mencionar algunas se destacan: Tumores Cerebrales, Estenosis de la Arteria Renal, Tumores de Glándulas Suprarrenales, etc.

Por otro lado las investigaciones recientes demuestran que si un padre sufre de HTA, un hijo entonces tendrá el 28% de padecer la enfermedad, si dos padres son hipertensos, entonces el 50% de sus hijos tendrán genes relacionados con esta patología. (54)

2.2.2. Descripción de los datos.

- **Informe de la descripción de los datos.**

En esta tarea de la fase se desarrollará un Informe de Descripción de los Datos con un doble propósito: primero, que el usuario tenga la posibilidad de tener de antemano la información que se manejará y segundo, permitir que el usuario tenga una mayor claridad sobre el tipo de información que se utilizará en el desarrollo de los modelos de minería de datos.

Para el desarrollo del Informe de Descripción de los Datos, se seleccionarán de las tablas de mayor interés y los atributos más significativos que intervendrán en el desarrollo de los modelos. De ellos se realizará un análisis donde se deben exponer explícitamente el nombre del atributo, una breve descripción de su utilización y el tipo de dato que almacena (Numérico, Nominal, Booleano).

A continuación se describen los atributos más relevantes para el proceso de minería depositados en el Almacén de Datos:

Tabla d_antecedentes.

Tabla 3. Descripción de los atributos seleccionados de la tabla *d_antecedentes.*

Atributo	*Descripción*	*Tipo de Datos*
nombre	Almacena el nombre de el antecedente	Nominal
tipo_antecedente	Almacena el tipo de antecedente que puede ser *personal* o *familiar*.	Nominal

Tabla d_habitos_personales

Tabla 4. Descripción de los atributos de la Tabla *d_habitos_personales.*

Atributo	*Descripción*	*Tipo de Datos*
nombre_habito	Nombre del hábito personal del paciente que pueden ser: *Grasas, Ejercicio físico, Consumo de Tabaco, Alcohol, Café, Drogas, Sal*.	Nominal

Tabla d_problema_salud.

Tabla 5. Descripción de los atributos de la Tabla *d_problema_salud.*

Atributo	*Descripción*	*Tipo de Datos*
problema_salud	Almacena el nombre del problema de salud	Nominal

Tabla d_caracteristicas_consulta.

Tabla 6. Descripción de los atributos de la Tabla *d_caracteristicas_consulta.*

Atributo	*Descripción*	*Tipo de Datos*
edad_paciente	Almacena la edad del paciente	Numérico

Tabla d_datos_personales.

Tabla 7. Descripción de los atributos de la Tabla *d_datos_personales.*

Atributo	*Descripción*	*Tipo de Datos*
genero_paciente	Almacena el género del paciente: *M* o *F*	Nominal
etnia_ paciente	Almacena la raza del paciente: *Blanca, Negra, Mestiza*	Nominal

Para cada campo de la tabla se expone:

- Identificador del campo.
- Breve descripción: Conocer semánticamente qué refleja dicho campo para poder interpretar el resultado mostrado por una herramienta de análisis.

- Tipo de dato que almacena el campo (Numérico, Nominal, Booleano): Conocer esto es importante pues todas las herramientas de análisis no tratan de igual forma cada tipo de datos. Incluso existen algoritmos que implementan técnicas de minería de datos que no pueden tratar con determinados tipos.

2.2.3. Exploración de los datos.

La exploración exhaustiva de los datos es algo previo a cualquier análisis que tiene por finalidad el hacernos una idea de las características de los datos. Esta exploración abarca diferentes aspectos como son la tabulación de datos en frecuencias absolutas y relativas, diferentes tipos de gráficos, índices que caracterizan una distribución de frecuencias, etc.

En esta tarea se debe realizar un Reporte de la Exploración de los Datos, en el mismo se utilizan combinaciones de algunas técnicas de visualización, análisis de correlación y técnicas estadísticas. Esta etapa de exploración se podría dividir en varias fases, dependiendo de los tipos de análisis y de herramientas a utilizar. En este apartado no se podrá realizar un análisis de composición de variables para cada uno de los objetivos fijados en la etapa de definición del problema y por tanto, no se generará el reporte a consecuencia de la información confusa y escasa que se encuentra en el almacén de datos.

2.2.4. Verificar la calidad de los datos.

Esta tarea tiene un peso importante dentro de la fase de Comprensión pues tiene como objetivo corroborar si la información recolectada es lo suficientemente sólida o no para satisfacer las necesidades del minero. Para decidir esta cuestión es necesario analizar si los datos contienen errores, y si el contenido del campo describe realmente lo que este almacena.

Los datos contenidos en el almacén fueron sometidos a un riguroso análisis basado fundamentalmente en los siguientes aspectos:

- ✓ Representación de la realidad.
- ✓ Consistencia.
- ✓ Campos innecesarios.
- ✓ Campos vacíos.
- ✓ Datos de naturaleza híbrida, poco genuina.

Para esto se realizaron consultas a la Base de Datos con código PL/PGSQL como la que se muestra a continuación:

```
SELECT tipo_antecedente FROM dwh.d_antecedentes WHERE
dwh.d_antecedentes.tipo_antecedente = 'Tipo desconocido'

SELECT tipo_antecedente FROM dwh.d_antecedentes WHERE
dwh.d_antecedentes.tipo_antecedente = NULL
```

2.3. Preparación de los datos.

Una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente. (55)

2.3.1. Selección de datos.

El objetivo de esta fase es listar los atributos que serán incluidos o excluidos del proceso de minería de datos. La selección de datos no sólo se realizó sobre los atributos (columnas), sino también sobre las tuplas (filas) de la base de datos. Los atributos seleccionados para el proyecto de Minería coinciden con los descritos en la fase de Comprensión de los Datos, por lo que no se realizará una descripción detallada de los campos en esta tarea.

Tabla 8. Listado de los atributos (columnas).

genero_paciente	edad_paciente
etnia_paciente	

Tabla 9. Listado de las tuplas (filas).

Antecedentes Familiares de Enfermedades Cardiovasculares	Antecedentes Personales de Enfermedades Renales
Antecedentes Familiares de Enfermedades Renales	Antecedentes Personales de Enfermedades Endocrinológicas
Antecedentes Familiares de Diabetes Mellitus	Antecedentes Personales de Enfermedades Cardiovasculares
Cefalea	Antecedentes Familiares de Hipertensión Arterial
Disnea	Náuseas
Palpitaciones	Síntomas de Dolor Abdominal
Edemas	

Todos los datos que no se exponen anteriormente fueron excluidos del proceso de minería por no ser de interés para este.

2.3.2. Construir los datos.

A partir de este momento se comienzan a actualizar valores de columnas, crear nuevas columnas, introducir nuevos registros que se componen de valores agregados u ordenados en caso que la minería lo exija.

Se adiciona el atributo **hta**, el cual almacenará elementos con tipo de datos nominal de en el caso de árboles de decisión y de tipo numérico para el agrupamiento. Como atributo nominal **hta** toma los valores **(SI/NO): SI** para el caso de las personas que han sido diagnosticadas de Hipertensión Arterial y **NO** en caso contrario. Posteriormente se integran en ***nombre_antecedente*** las columnas ***nombre*** y ***tipo_antecedente*** de la tabla ***d_antecedentes***. Esta tarea se realiza con el objetivo de facilitar el posterior uso de estos atributos y simplificar el entendimiento.

Para el caso de la construcción del modelo de Agrupamiento el atributo **hta** tomará los valores **(0/1): 0** para las personas que no han sido diagnosticada de Hipertensión Arterial y **1** en caso contrario; también se realizó la integración de la tabla **d_antecedentes** descrita anteriormente**.** Así mismo se seleccionaron las tuplas que intervendrán en el proceso de minería y fueron transformadas para proporcionar una fácil manipulación de las mismas.

- **Atributos derivados**

Para un correcto cumplimiento de los objetivos del proceso de modelado se crea la columna ***rango_edad***, la misma está intrínsecamente relacionada con la columna ***edad_paciente***, esto cumple con el principio de agrupar las edades de los pacientes por rangos estimados para un mejor entendimiento en las futuras fases. A continuación se representan los valores que puede tomar la columna ***rango_edad*:**

- ✓ ***Rango1:*** Pertenece a los pacientes cuya edad está entre 15 y 40 años.
- ✓ ***Rango2:*** Pertenece a los pacientes cuya edad está entre 41 y 65 años.
- ✓ ***Rango3:*** Pertenece a los pacientes cuya edad es mayor que 65 años.

En el Anexo 2 se expone una tabla donde se especifican los significados de cada valor numérico empleado en la tabla vista_minable_skm.

- **Transformaciones en los Nombres de las Tuplas.**

Tabla 10. Transformaciones en los Nombres de las Tuplas.

Nombre Original de las Tuplas(NOT)	**Contracción del NOT**	**Nombre Original de las Tuplas(NOT)**	**Contracción del NOT**
Antecedentes Patológicos Personales de Enfermedades Renales	**APP_ER**	Consumo de Alimentos con Abundantes Grasas	**GR**
Antecedentes Patológicos Personales de Enfermedades Endocrinológicas	**APP_EE**	Realización de Ejercicio Físico.	**EF**
Antecedentes Patológicos Personales de Enfermedades Cardiovasculares	**APP_EC**	Consumo de Tabaco	**CT**
Antecedentes Patológicos Familiares de Hipertensión Arterial	**APF_HTA**	Consumo de Alcohol	**AL**
Antecedentes Patológicos Familiares de Enfermedades Cardiovasculares	**APF_EC**	Consumo de Café	**CA**
Antecedentes Patológicos	**APF_ER**	Consumo de Drogas	**DR**

Familiares de Enfermedades Renales			
Antecedentes Patológicos Familiares de Diabetes Mellitus	**APF_DM**	Consumo de Alimentos con Sal en demasía	**S**
Cefalea	**CEF**	Mareos	**MAR**
Disnea	**DIS**	Vértigo	**VER**
Palpitaciones	**PAL**	Dolor Torácico	**DT**
Edemas	**EDE**	Alteraciones de la Visión	**AV**
Síntomas de Dolor Abdominal	**DA**	Padecimiento de Diabetes Mellitus	**DIA**
Náuseas	**NAU**	Insuficiencia Cardiaca	**IC**

2.3.3. Limpieza de datos.

El formato de los datos contenidos en el Almacén de Datos no es el idóneo. Haciendo simples consultas al mismo, se obtienen campos con valores desconocidos. Este proceso se realizó en las tablas seleccionadas permitiendo determinar las tuplas inconsistentes y que por ende no aportan ninguna información provechosa, una vez localizadas las tuplas con estas características se procede a eliminar dichos valores. Un ejemplo de las mismas se muestra a continuación:

```
SELECT nombre, tipo_antecedente FROM dwh.d_antecedentes
WHERE tipo_antecedente = 'Tipo desconocido' OR   nombre = 'Nombre desconocido';
```

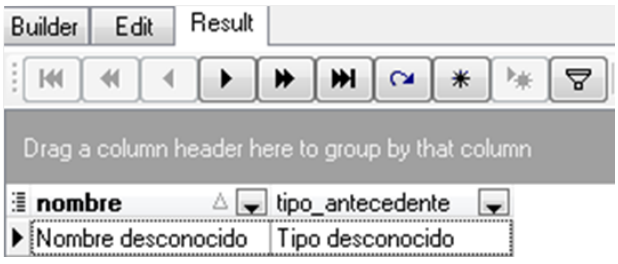

Imagen 7. Visualización de la Consulta a la tabla *dwh.d_antecedentes.*

Además, durante este proceso se detectaron valores no válidos, como es el caso en la tabla Hábitos Personales, donde existían tuplas contenidas en el campo ***nombre_habito*** con la especificación indebida. A continuación se muestra un ejemplo:

```
SELECT nombre_habito FROM dwh.d_habitos_personales
WHERE nombre_habito='Nombre desconocido';
```

Imagen 8. Visualización de la Consulta a la tabla *dwh.d_datos_personales.*

2.3.4. Integración de los datos.

Se analizan los datos que son necesarios para el proyecto y se combinan con el objetivo de obtener la información que proviene de diferentes dimensiones del Almacén de Datos integradas en una sola tabla: ***pre_vista_minable*** (Anexo 3, Imagen 11). A esta tabla se le aplicaron un conjunto de transformaciones para las cuales se hizo necesaria la creación de un software desarrollado en el IDE NetBeans, el cual funciona como intermediario entre la tabla ***md.pre_vista_minable*** y ***md.vista_minable_j48*** **–** tabla que almacena los datos que serán utilizados para la creación del modelo mediante árboles de decisión-**.** Este software hace uso de 2 procedimientos almacenados (Anexo 3, Imagen 12): el primero, llamado ***agrupando_por_id***, tiene la función de agrupar a los pacientes por su identificador; el segundo, denominado ***insertando_en_tb_vista_minable*** se encarga de insertar los valores en la tabla final. Al mismo tiempo el programa se encarga de convertir las tuplas o filas significativas para el proceso de minería en atributos o columnas.

Se generó además de la tabla ***md.vista_minable_j48*** (Anexo 3, Imagen 13), una llamada ***md.vista_minable_skm*** (Anexo 3, Imagen 14) **–** tabla que almacena los datos que serán utilizados para la creación del modelo mediante agrupamiento- la misma es un duplicado de ***md.vista_minable_j48***, la diferencia radica en que sus tuplas serán numéricas, esto permite una mejor asignación de las variables a la hora de calcular los centroides de los grupos. Esta transformación se logra aplicando simples consultas SQL, un ejemplo de estas se presentan en el Anexo 3, Imagen 15.

Capítulo 3. Resultados y discusión.

En este capítulo se realiza una descripción de las 3 últimas fases propuestas por CRISP-DM 1.0: *Modelado, Evaluación y Explotación*; así como las actividades que se ejecutan dentro de estas, de igual forma se elabora una descripción de los resultados obtenidos en cada una de estas fases.

4.1. Modelado.

En esta fase se crean los modelos necesarios para el proyecto para lo cual se ejecutará la herramienta Weka sobre el conjunto de datos preparados para crear los dos modelos previstos.

4.1.1. Selección de la técnica de modelado.

Esta tarea consiste en la selección de la técnica de minería de datos más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de minería de datos existentes. Las técnicas de minería de datos a aplicar en el presente estudio son: clasificación (algoritmo J48) y agrupamiento (Simple K-Means).

Técnicas Supervisadas.

- **Árbol de Decisión construido con el Algoritmo J48 que implementa WEKA:**
  - ✓ Admite atributos simbólicos y numéricos, aunque la clase debe ser simbólica.
  - ✓ Se permiten ejemplos con valores desconocidos.
  - ✓ Se pueden tratar registros que tienen valores de atributos desconocidos, evaluando la ganancia o la relación de ganancia de un atributo considerando sólo los registros que tienen definidos ese atributo.
  - ✓ Se pueden clasificar registros que tienen no definido el valor de algún atributo estimando la probabilidad de los posibles resultados.
  - ✓ Se puede trabajar con atributos que tienen valores continuos. (56)

Técnicas No Supervisadas.

- **Grupos construidos con el Algoritmo Simple K-Means que implementa WEKA:**
  - ✓ Admite atributos simbólicos y numéricos.

- ✓ Para obtener los centroides iniciales se emplea un número aleatorio obtenido a partir de la semilla empleada. Los *k* ejemplos correspondientes a los *k* números enteros siguientes al número aleatorio obtenido serán los que conformen dichos centroides.
- ✓ No se estandarizan los argumentos, sino que se normalizan. (57)

4.1.2. Construcción del modelo.

En esta fase se ejecutará la herramienta Weka sobre el conjunto de datos en la fase de Modelado para crear los dos modelos previstos.

4.1.2.1. Construcción del modelo aplicando J48.

A continuación se muestra el modelo obtenido después de haber aplicado el algoritmo J48 sobre los datos de entrenamiento almacenados en la tabla *vista_minable*. Las reglas se construyen de arriba a abajo y de izquierda a derecha de manera escalonada desde el nodo raíz hasta los nodos hojas. El nodo ubicado más a la izquierda en la representación constituye la raíz del árbol. Los nodos hojas por su parte son aquellos a los que le sigue el valor alcanzado por el tiempo de vida (variable a predecir).

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -R -N 3 -Q 1 -M 2
Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1
Instances:    676
Attributes:   30
Test mode:    evaluate on training data

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

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

VER = No
|   DA = No
|   |   S = No
|   |   |   DIS = No
|   |   |   |   PAL = No
|   |   |   |   |   DT = No
|   |   |   |   |   |   APP_EC = No
```

```
|   |   |   |   |   |   |   GR = No
|   |   |   |   |   |   |   |   APP_ER = No
|   |   |   |   |   |   |   |   |   EDE = No
|   |   |   |   |   |   |   |   |   |   IC = No
|   |   |   |   |   |   |   |   |   |   |   EF = No
|   |   |   |   |   |   |   |   |   |   |   |   DIA = No
|   |   |   |   |   |   |   |   |   |   |   |   |   CT = No
|   |   |   |   |   |   |   |   |   |   |   |   |   |   APF_EC = Si: No (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   APF_EC = No
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   rango_edad = Rango2
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   APF_ER = No: No (15.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   APF_ER = Si: Si (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   rango_edad = Rango1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   CA = No
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   APF_DM = No: No (13.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   APF_DM = Si: Si (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   CA = Si: No (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   rango_edad = Rango3: Si (13.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   CT = Si: No (42.0/6.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   DIA = Si: Si (24.0/6.0)
|   |   |   |   |   |   |   |   |   |   |   |   EF = Si: Si (13.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   IC = Si: Si (12.0)
|   |   |   |   |   |   |   |   |   |   EDE = Si: Si (27.0)
|   |   |   |   |   |   |   |   |   APP_ER = Si: No (12.0/1.0)
|   |   |   |   |   |   |   |   GR = Si: Si (15.0/2.0)
|   |   |   |   |   |   |   APP_EC = Si
|   |   |   |   |   |   |   |   APP_ER = No
|   |   |   |   |   |   |   |   |   CA = No: Si (24.0)
|   |   |   |   |   |   |   |   |   CA = Si
|   |   |   |   |   |   |   |   |   |   APF_HTA = No: No (2.0)
|   |   |   |   |   |   |   |   |   |   APF_HTA = Si: Si (7.0)
|   |   |   |   |   |   |   |   APP_ER = Si: No (3.0/1.0)
|   |   |   |   |   |   DT = Si: Si (28.0)
|   |   |   |   |   PAL = Si: Si (73.0/2.0)
|   |   |   |   DIS = Si: Si (80.0)
|   |   |   S = Si: No (7.0)
|   |   DA = Si: No (10.0)
VER = Si: No (22.0)
```

```
Number of Leaves:  24
Size of the tree:  46
Time taken to build model: 0.06 seconds
```

4.1.2.2. Construcción del modelo aplicando Simple K-Means.

A continuación se muestra el modelo obtenido después de haber aplicado el algoritmo Simple K-Means sobre los datos de entrenamiento almacenados en la tabla vista_minable. Se procedió a agrupar el data set en 3 grupos. Para la ejecución de este algoritmo es necesario seleccionar un número, denominado semilla, para realizar una distribución aleatoria inicial a partir de la cual el algoritmo comience las sucesivas iteraciones. Para la selección de este número se realizaron 20 corridas consecutivas probando distintas semillas y se seleccionó aquella que minimizaba la suma del error cuadrático. Si bien este método heurístico no garantiza la semilla óptima, asegura una relativamente buena asignación (58). En el Anexo 5 se presentan los resultados obtenidos para las 20 corridas.

Como se puede ver la menor suma de error cuadrático se obtuvo con una semilla igual 8. El resultado obtenido con *Weka* tras la ejecución de Simple K-Means con 3 clúster y una semilla de 8 se sintetiza a continuación:

```
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -D -R first-last" -I
1500 -O -S 8
Relation: QueryResult-weka.filters.unsupervised.attribute.Remove-R1
Instances: 676
Attributes:   30
Test mode: evaluate on training data

=== Model and evaluation on training set ===
KMeans
======
Number of iterations: 5
Within cluster sum of squared errors: 3074.0
Missing values globally replaced with mean/mode
Cluster centroids:
Cluster#
```

Attribute	Full Data (676)	0 (270)	1 (175)	2 (231)
rango_edad	1	2	3	1
genero_paciente	1	1	2	1
etnia_paciente	2	3	1	2
APP_ER	0	0	0	0
APP_EE	0	0	0	0
APP_EC	0	0	0	0
APF_HTA	0	0	0	1
APF_EC	0	0	0	0
APF_ER	0	0	0	0
APF_DM	0	0	0	0
CEF	0	0	0	0
DIS	0	0	0	0
PAL	0	0	0	0
EDE	0	0	0	0
DA	0	0	0	0
NAU	0	0	0	0
GR	0	0	0	0
EF	0	0	0	0
CT	0	0	0	1
AL	0	0	0	0
CA	0	0	0	0
DR	0	0	0	0
S	0	0	0	0
MAR	0	0	0	0
VER	0	0	0	0
DT	0	0	0	0
AV	0	0	0	0
DIA	0	0	0	0
IC	0	0	0	0
HTA	1	1	1	1

4.1.3. Evaluación del modelo.

En esta tarea, los ingenieros de minería de datos interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema

juzgan los modelos dentro del contexto del dominio y expertos en minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc...). (59)

4.1.3.1. Evaluación del modelo generado por J48.

La herramienta Weka contiene método de validar los resultados como los que se muestran a continuación:

Use training set *(Usar el conjunto de entrenamiento):* El clasificador se evalúa según las predicciones de las clases del conjunto de entrenamiento.
Supplied test set *(Suministrado del conjunto de prueba):* Utilizar un nuevo conjunto de datos con los que el algoritmo no haya tenido contacto para validar el análisis realizado (validación cruzada).
Percentage Split *(Porcentaje de división):* El clasificador se evalúa según las predicciones que realice de un porciento de los datos de prueba.
Classes to cluster evaluation *(Clases para la evaluación de grupo):* Valida grupo a grupo utilizando los niveles de homogeneidad entre ellos.

En este caso se validará con la opción «Use tainig set ». Se recomienda esta opción por la velocidad computacional entre otras cosas. El conjunto de patrones que se presentan en el modelo fueron obtenidos con una precisión de 0.933432, lo que equivale decir que se clasificaron correctamente el 93,3432% del total de casos. La herramienta arrojó los siguientes resultados.

```
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         631               93.3432 %
Incorrectly Classified Instances        45                6.6568 %
Kappa statistic                          0.8371
Mean absolute error                      0.1069
Root mean squared error                  0.2307
Relative absolute error                 26.5075 %
Root relative squared error             51.4146 %
Total Number of Instances              676
```

Para una mejor comprensión del modelo, en el Anexo 4 se muestra el árbol obtenido a partir de aplicar el algoritmo J48. Los nodos representan atributos, las ramas representan valores de dichos atributos y los nodos finales representan los valores de la clase. Cada camino del árbol representa una regla, las reglas o patrones más relevantes obtenidos en el anterior modelo son las siguientes:

1. Si un paciente tiene *vértigo* entonces no tiene Hipertensión Arterial.
2. Si un paciente no tiene *vértigo* y tiene dolor *abdominal* entonces no tiene Hipertensión Arterial.
3. Si un paciente no tiene *vértigo*, *dolor abdominal* y consume abundante *sal* entonces no tiene Hipertensión Arterial.
4. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal* y tiene *disnea* entonces tiene Hipertensión Arterial.
5. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea* y presenta *palpitaciones* entonces tiene Hipertensión Arterial.
6. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, y presenta *dolor torácico* entonces tiene Hipertensión Arterial.
7. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico* y tiene *antecedentes patológicos personales de enfermedades cardiovasculares* y *enfermedades renales* entonces tiene Hipertensión Arterial.
8. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*", *dolor torácico* y tiene *antecedentes patológicos personales de enfermedades cardiovasculares* y no de *enfermedades renales*, además no tienes hábitos personales de *consumo de café* entonces tiene Hipertensión Arterial.
9. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico* y tiene *antecedentes patológicos personales de enfermedades cardiovasculares* y no de *enfermedades renales* además tiene hábitos personales de *consumo de café* y posee *antecedentes patológicos familiares de Hipertensión Arterial* entonces tiene Hipertensión Arterial.
10. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico* y tiene *antecedentes patológicos personales de enfermedades cardiovasculares* y no de *enfermedades renales* además tiene hábitos personales de *consumo de café* y no posee *antecedentes patológicos familiares de Hipertensión Arterial* entonces no tiene Hipertensión Arterial.
11. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares* y además tiene hábitos personales de altos *consumo de grasas* entonces tiene Hipertensión Arterial.
12. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas* y presenta

antecedentes patológicos personales de enfermedades renales entonces no tiene Hipertensión Arterial.

13. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas* ni presenta *antecedentes patológicos personales de enfermedades renales*, y tiene *edemas* entonces tiene Hipertensión Arterial.

14. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume con abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, y tiene *insuficiencia cardiaca* entonces tiene Hipertensión Arterial.

15. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca* y realiza *ejercicios físicos* entonces tiene Hipertensión Arterial.

16. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas* no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca,* no realiza *ejercicios físicos* y tiene *diabetes* entonces tiene Hipertensión Arterial.

17. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca*, no realiza *ejercicios físicos*, no tiene *diabetes* y *consume tabaco* entonces no tiene Hipertensión Arterial.

18. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca,* no realiza *ejercicios físicos*, no tiene *diabetes*, ni *consume tabaco*, y no presenta *antecedentes patológicos familiares de enfermedad cardiovascular* entonces no tiene Hipertensión Arterial.

19. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca,* no realiza *ejercicios físicos*, no tiene *diabetes*, ni *consume tabaco*, presenta *antecedentes patológicos familiares de enfermedad cardiovascular*, está en una *edad en el rango de 45 a 65 años*, y presenta *antecedentes patológicos familiares de enfermedades renales* entonces no tiene Hipertensión Arterial.

20. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca*, no realiza *ejercicios físicos* ,no tiene *diabetes* ni *consume tabaco*, presenta *antecedentes patológicos familiares de enfermedad cardiovascular*, está en una *edad en el rango de 45 a 65 años*, y no presenta *antecedentes patológicos familiares de enfermedades renales* entonces no tiene Hipertensión Arterial.

21. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca*, no realiza *ejercicios físicos* ,no tiene *diabetes* ,ni *consume tabaco*, presenta *antecedentes patológicos familiares de enfermedad cardiovascular,* es *menor de 45 años de edad*, y toma *café* entonces no tiene Hipertensión Arterial.

22. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas*, no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia cardiaca* no realiza *ejercicios físicos*, no tiene *diabetes*, ni *consume tabaco*, presenta *antecedentes patológicos familiares de enfermedad cardiovascular*, es *menor de 45 años de edad*, no toma *café* y no tiene *antecedentes patológicos familiares de diabetes mellitus* entonces no tiene Hipertensión Arterial.

23. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas,* no presenta *antecedentes patológicos personales de enfermedades renales*, no tiene *edemas*, *insuficiencia*

cardiaca, no realiza *ejercicios físicos* ,no tiene *diabetes*, ni *consume tabaco*, presenta *antecedentes patológicos familiares de enfermedad cardiovascular* ,es *menor de 45 años de edad*, no toma *café* y tiene *antecedentes patológicos familiares de diabetes mellitus* entonces tiene Hipertensión Arterial.

24. Si un paciente no tiene *vértigo*, *dolor abdominal*, no consume abundante *sal*, no tiene *disnea*, *palpitaciones*, *dolor torácico*, no tiene *antecedentes patológicos personales de enfermedades cardiovasculares*, no tiene hábitos personales de altos *consumo de grasas* y no presenta *antecedentes patológicos personales de enfermedades renales*, y no tiene *edemas* , *insuficiencia cardiaca* no realiza *ejercicios físicos*, no tiene *diabetes*, ni *consume tabaco*, y presenta *antecedentes patológicos familiares de enfermedad cardiovascular*, es *mayor de 65 años de edad*, entonces tiene Hipertensión Arterial.

4.1.3.2. Evaluación del modelo generado por Simple K-Means.

Los modelos descriptivos en general, son complicados de evaluar debido a la ausencia de una clase determinada donde medir el grado de acierto del modelo. La mejor evaluación de este tipo de modelos es saber si el modelo resultado de la fase de aprendizaje tiene un comportamiento útil cuando se utilice en su área de aplicación. (60)

Al igual que para el modelo generado por el árbol de decisión J48, para este se validará con la opción «Use training set ». El conjunto de patrones que se presentan en el modelo fueron obtenidos a partir de los resultados que se muestran a continuación, la herramienta arrojó:

```
==Clustered Instances==
    0       270 (40%)
    1       175 (26%)
    2       231 (34%)
```

Antes de realizar un análisis a profundidad sobre este modelo, primero es necesario observar las características de cada grupo obtenido una vez aplicado el algoritmo. En el Anexo 5.1 se exponen las gráficas de dispersión de alguna de los atributos más importantes a la consideración de los especialistas.

A partir de la interpretación conjunta de las imágenes anteriores podemos descubrir en el conjunto de datos lo siguiente:

- ✓ **Grupo 0 (40%):** se destacan las personas que se encuentran entre 45 y 65 años *de edad,* predomina el sexo masculino y la mayoría de ellos son de raza mestiza. La distribución de los pacientes que tienes antecedentes patológicos familiares de Hipertensión Arterial es bastante uniforme, sin embargo se puede apreciar una ligera mayoría de personas que no tienen este tipo de antecedente.
- ✓ **Grupo 1 (26%):** muy concentrado por personas de más de 65 años de edad, generalmente del sexo femenino y hay mayor concentración de personas de raza blanca. En este grupo, aunque al igual que en el grupo 0 existe una distribución relativamente uniforme de casos de antecedentes familiares de Hipertensión Arterial, es más notable que en la generalidad de los casos tampoco presentan antecedentes de esta índole.
- ✓ **Grupo 2 (34%):** representa en su mayoría a las personas que son menores de 45 años de edad, generalmente masculinos de raza negra. Se puede apreciar una notable concentración de personas que si presentan antecedentes de Hipertensión Arterial en su familia.

Se puede apreciar que en los 3 grupos la generalidad de los pacientes que se encuentran agrupados son personas que tienen Hipertensión Arterial.

Una vez analizado el contenido de cada grupo se deducen a grandes rasgos los siguientes patrones:

- ✓ En el 40% de los casos los pacientes que padecen Hipertensión Arterial están entre 45 y 65 años de edad, son de sexo masculino y de raza mestiza.
- ✓ El 34% de las personas que padecen Hipertensión Arterial tienen antecedentes patológicos familiares de la enfermedad y consumen tabaco.
- ✓ El 66 % de los casos con Hipertensión Arterial fueron asintomáticos.

4.2. Evaluación

Después de que uno o varios modelos han sido construidos y estos cuentan con una alta calidad desde la perspectiva del análisis de los datos, el modelo es evaluado con respecto a los objetivos del negocio. También se realiza una revisión de los pasos ejecutados durante la construcción del modelo. El objetivo principal es determinar si alguna problemática del negocio no ha sido suficientemente considerada. Al final de esta etapa se decide si los resultados obtenidos cubren satisfactoriamente los objetivos planteados. Las actividades principales de esta etapa son: evaluación de los resultados, revisión del proceso y el determinar si se continúa con la siguiente etapa. (61)

4.2.1. Evaluación de resultados.

Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional. (62)

- **Evaluación de los resultados J48.**

```
=== Confusion Matrix ===
  a     b   <-- classified as
171    18    |   a = No
 27   460    |   b = Si
```

Matriz de confusión: Para cada clase real registra el número de casos en los cuales el clasificador ha predicho una clase. La suma de la diagonal principal (Traza) corresponde al número total de aciertos. De la anterior matriz se puede afirmar que hubo 631 aciertos en la clasificación y quedaron mal clasificados 45 de los 676 casos.

Se analizó el modelo en detalle con el objetivo de obtener otros resultados de interés. Otro de los resultados fue el siguiente:
Kappa Statistic = 0.8371

WEKA calcula el «Kappa Statistic (Coeficiente de Kappa)» para mostrar la concordancia entre los datos de prueba y la clasificación hecha por el modelo. Cuando todas las instancias son clasificadas correctamente se obtiene la máxima concordancia, es decir, Kappa Statistic = 1.
Para el caso del árbol de decisión obtenido, el «Coeficiente de Kappa» resultó ser igual a 0. 8371, lo cual indica un alto nivel de concordancia entre los datos de prueba y la clasificación hecha por el modelo.

- ✓ **Descripción de la precisión por clases.**

```
=== Detailed Accuracy by Class ===
               TP Rate   FP Rate   Precision   Recall   F-Measure     ROC   Area Class
                 0.905     0.055      0.864     0.905     0.884      0.972     No
                 0.945     0.095      0.962     0.945     0.953      0.972     Si
Weighted Avg.    0.933     0.084      0.935     0.933     0.934      0.972
```

En la tabla anterior se observa que la clase “No” presenta una Tasa de Verdaderos Positivos (TP Rate) del 90.5% y la clase “Si” del 94.5%, por lo cual se puede interpretar que: dada una instancia el árbol de decisión la clasifica como “No” el 90.5% de las veces y “Si” el 94.5% de las veces. En el mismo sentido, se tiene que la Tasa de Falsos Positivos (FP Rate) para las instancias clasificadas por el modelo es del 5.5% para la clase “No” y del 9.5% para la clase “Si”.

Con las medidas anteriores, se obtiene la precisión del modelo para las clases “No” y “Si”, la cual es del 86.4% y 96.2 % respectivamente. Este porcentaje indica la proporción de aciertos del modelo obtenido.

«F-Measure» representa la media armónica entre la precisión y el Recall. Entre más cercana sea a 1, mayor será la confiabilidad del modelo en la clase. Para la clase “No” se tiene una confiabilidad (F-Measure) del 88.4% y la clase “Si” del 95.3%. De esta forma, se tiene que el árbol de decisión obtenido clasifica de manera aceptable a las instancias pertenecientes a las clases “No” y “Si”. En otras palabras, se cuenta con un conjunto de reglas de decisión que clasifican con un buen grado de confiabilidad a las instancias.

En la siguiente tabla se muestran las reglas obtenidas a partir de los modelos de Árboles de Decisión generados y el valor de precisión de cada una de ellas.

Tabla 11. Reglas generadas del Árbol de Decisión con la precisión obtenida.

Reglas	***Precisión***	***Reglas***	***Precisión***
Regla 1	0.864	**Regla 13**	0.962
Regla 2	0.864	**Regla 14**	0.962
Regla 3	0.864	**Regla 15**	0.962
Regla 4	0.962	**Regla 16**	0.962
Regla 5	0.962	**Regla 17**	0.864
Regla 6	0.962	**Regla 18**	0.864
Regla 7	0.962	**Regla 19**	0.864
Regla 8	0.962	**Regla 20**	0.864
Regla 9	0.962	**Regla 21**	0.864
Regla 10	0.864	**Regla 22**	0.864
Regla 11	0.962	**Regla 23**	0.962
Regla 12	0.864	**Regla 24**	0.962

- **Propuesta de Mejora J48.**

Modelo: Arboles de Decisión: J48. Estudio de la Hipertensión Arterial.
Objetivo del experimento: Definir dos modelos empleando técnicas de minería de datos que tributen al diagnóstico y detección de pacientes con riesgos de sufrir Hipertensión Arterial.
Criterios de éxito del modelo: Obtener las predicciones con un valor de certeza igual o superior al 80%.

Propuesta de Mejoras:
Realizar los modelos de minería de datos utilizando el algoritmo de árboles de decisión J48 con una mayor cantidad de datos de las Historias Clínicas con datos reales de un área de salud específica.
En el proyecto no se propone repetir ningún paso, ya que después de un nuevo análisis no se han encontrado fallas, ni se ha omitido ninguna variable que pudiera limitar el éxito de los resultados. Las propuestas de mejoras en los modelos se dejan para próximas iteraciones o proyectos.

- **Propuesta de Mejora Simple KMeans.**

Modelo: Agrupamiento**:** Simple KMeans. Estudio de la Hipertensión Arterial.
Objetivo del experimento: Definir dos modelos empleando técnicas de minería de datos que tribute al diagnóstico y detección de pacientes con riesgos a sufrir Hipertensión Arterial.
Criterios de éxito del modelo: Obtener grupos con un alto grado de similitud, basándose en las semejanzas y diferencias existentes entre los componentes de la muestra.

Propuesta de Mejoras:
Realizar los modelos de minería de datos utilizando el algoritmo Simple KMeans con una mayor cantidad de datos de las Historias Clínicas con datos reales de un área de salud específica.

- **Resumen de la evaluación de los resultados.**

A continuación se describen los criterios de éxitos del negocio así como la valoración de su cumplimiento.

Tabla 12. Cumplimiento de los criterios de éxito.

Criterios de Éxitos	Porciento de Cumplimiento
Realizar un proyecto de minería de datos guiado por la metodología CRISP-DM.	100%

Desarrollar el proyecto usando la herramienta Weka para la minería de datos.	100%
Obtener dos modelos modelo de conocimiento y comprobar su validez.	100%

- **Aprobar modelos**

Después de la evaluación los modelos con respecto a los criterios de éxito del negocio, los modelos generados que satisfacen los criterios seleccionados se convierten en modelos aprobados. Para esto se debe entender el resultado de la minería de datos para poder interpretarlos y comprobar que los mismos cumplen las metas iniciales del negocio, comprobar que estos resultados son novedosos y útiles.

De acuerdo a los resultados obtenidos en el proyecto para dar cumplimiento al objetivo del negocio; el modelo realizado es aprobado, dado que el mismo permitió encontrar patrones significativos en los datos.

4.2.2. Revisar el proceso.

El proceso de revisión, se refiere a calificar al proceso entero de minería de datos, a objeto de identificar elementos que pudieran ser mejorados. Después de la revisión del proceso el modelo es satisfactorio ya que cumple con las necesidades, en consecuencia con esto no se repetirá ningún paso por no haberse encontrado fallas durante la ejecución del mismo, es válido mencionar además que no se omitió ninguna de las variables utilizadas y ninguna tarea fue pasada por alto. Las propuestas de mejoras en los modelos se dejan para próximas iteraciones o proyectos.

4.2.3. Establecimiento de los siguientes pasos o acciones.

Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de minería de datos. (63)

Después de la revisión del proceso de Minería se decide finalizar el proyecto y pasar a la fase de despliegue.

Para las posteriores iteraciones o desarrollos de procesos de minería de datos se proponen las siguientes acciones a seguir:

- ✓ Realizar los modelos de minería de datos utilizando el algoritmo de agrupamiento K-Means con mayor cantidad de datos para poder realizar de forma satisfactoria una evaluación del modelo.

- ✓ Realizar los modelos j48 y K-Means con datos reales de un área de salud específica.

4.3. Despliegue

El despliegue del proyecto no significa que el propósito haya terminado, ya que se deben hacer pruebas de su implementación así como un seguimiento periódico. Los resultados que presenta deben ser claros de tal manera que se puedan comprender fácilmente por personas que no tengan mucha intervención en el área de la Informática, sino también de profesionales o no que pertenezcan a otras áreas.

La fase de despliegue será tan compleja como los requisitos que se hayan propuesto, así que pueden ir desde la simple presentación de un informe hasta la demostración del sistema ya implementado. En este caso se presentará un informe con el resumen del trabajo realizado.

4.3.1. Generación de Informe Final.

El informe final del proyecto se basa en un resumen de los pasos que se han implementado a lo largo del proyecto, así como los inconvenientes encontrados. El presente documento se considera como el Informe Final.

Conclusiones

A partir de los resultados del estudio realizado se llegaron a las siguientes conclusiones fundamentales:

- ✓ La investigación recoge todo el proceso de obtención de los modelos usando técnicas de minería de datos para el estudio y diagnóstico de la Hipertensión Arterial para lo cual se realizó una investigación profunda a los expertos en el tema de dicha enfermedad donde se encontraron patrones significativos para la investigación como, factores de riesgos, causas y síntomas .

- ✓ Se obtuvieron los modelos requeridos con los algoritmos Árboles de Decisión (J48) y Agrupamiento (Simple K-Means) que brinda la herramienta Weka a partir de los datos contenidos en el almacén de datos del (alas SIAPS)

- ✓ Se lograron obtener modelos de minería de datos con un alto grado de validez encontrando información oculta en el interior de los datos sobre Hipertensión Arterial.

Recomendaciones

Las recomendaciones de la investigación están dirigidas a sugerir acciones para complementar el producto obtenido. Por lo que para el buen desempeño y puesta en marcha de la investigación se hacen las siguientes recomendaciones:

- ✓ Realizar los modelos de minería de datos utilizando el algoritmo de agrupamiento Simple K-Means con mayor cantidad de datos para poder realizar una mejor evaluación del modelo obtenido.

- ✓ Realizar los modelos J48 y Simple K-Means con datos reales de un área de salud específica para el estudio de la Hipertensión Arterial.

Referencias Bibliográficas

1. **Autores, Colectivo de.** *Curso Universitario. Sistema de Información en los Sistemas de Salud.* Italia: Instituto Universitario del Hospital Italiano.

2. **Sánchez Corales, Yovannys.** *Documento de Arquitectura de Software v1.0.* La Habana: s.n., 2010.

3. **Fernández Cumbá, E.** *Propuesta didáctica para la promoción de salud en el caso de la Hipertensión Arterial en los pacientes de la Universidad de las Ciencias Informáticas.* La Habana: Instituto Superior Politécnico José A. Echeverría, 2008.

4. *Ídem a la referencia 3.*

5. **Pautsch, J.G.A.** *Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación, Facultad de Ciencias Exactas, Químicas y Naturales.* Argentina: Universidad Nacional de Misiones, 2009.

6. **Bressán, G.E.** *Almacenes de datos y minería de datos, T.y.s. distribuidos.* Buenos Aires, Argentina: Universidad Nacional del Nordeste. Facultad de Ciencias Exactas y Naturales y Agrimensura, 2003.

7. **Hernández Orallo, Jose, Ramírez Quintana, Maria José y Ferri Ramírez, César.** *Introducción a la minería de datos.* Valencia, España: Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia, 2004. 84-105-409 1-0.

8. *Ídem a la referencia 7.*

9. *Ídem a la referencia 7.*

10. **M., Servente.** Algoritmos TDIDT aplicados a la minería de datos Inteligente. *Tesis de Grado en Ingeniería Informática, 2002.* [En línea] Universidad de Buenos Aires, Facultad de Ingeniería. [Citado el: 17 de Noviembre de 2010.] http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf.

11. **Molina López, J. M. y García Herrero, J.** *Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA.* Madrid, España: Universidad Carlos III, 2006.

12. *Ídem a la referencia 7.*

13. *Ídem a la referencia 7.*

14. *Ídem a la referencia 7.*

15. *Ídem a la referencia 7.*

16. DAEDALUS. *Data, Decisions and Language, S. A.* [En línea] [Citado el: 15 de Noviembre de 2010.] http://www.daedalus.es/mineria-de-datos/proceso-de-mineria-de-datos/.

17. *Ídem a la referencia 16.*

18. *Ídem a la referencia 16.*

19. *Ídem a la referencia 16.*

20. **Soto, Lauro.** Técnicas Herramientas De minería de datos. [En línea] [Citado el: 9 de Mayo de 2011.] http://www.mitecnologico.com/Main/TecnicasHerramientasDeMineriaDeDatos.

21. *Ídem a la referencia 20.*

22. *Ídem a la referencia 20.*

23. *Ídem a la referencia 20.*

24. **Witten, I.H. y Frank, E.** *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Diego, EE.UU: Morgan Kaufmann Publishers, 2000.

25. *Ídem a la referencia 24.*

26. *Ídem a la referencia 24.*

27. **Suils, Pilar Sopena.** *Extracción y Recuperación de información no supervisada.* s.l. : Valladolid, 2008.

28. **González Gómez, Jose Ignacio.** *Generalidades de la minería de datos.* España: Universidad de La Laguna.

29. **J., Quinlan.** *C4.5. Programs for machine learning,* s.l. : Morgan Kaufmann Pub., 1993. 1558602380.

30. *Ídem a la referencia 29.*

31. **Marante Jacas, Danier y Marante Jacas, Danner.** *Aplicación de la minería de datos para la exploración y detección de patrones delictivos.* La Habana: Universidad de las Ciencias Informáticas, Facultad 8, 2008.

32. *Ídem a la referencia 31.*

33. **Zamarrón Sanz, Carlos, et al.** *Aplicación de la minería de datos al estudio de las alteraciones respiratorias durante el sueño.* Santiago de Compostela. España: Hospital Clínico Universitario de Santiago de Compostela, 2006.

34. **Rodríguez Jara, Félix y Vallejo Delgado, Nieves.** *Aplicación de técnicas de minería de datos para el diagnóstico prematuro de Cáncer.* Madrid, España: Universidad Carlos III.

35. **Rosete Suárez, Alejandro, Rodríguez Díaz, Alfredo y Acosta Sánchez, Rolando.** Revista Cubana de Informática Médica. [En línea] [Citado el: 9 de Noviembre de 2010.] http://www.rcim.sld.cu/revista_18/articulos_htm/prediccionpaciente.htm#t. 1684-1859.

36. **Bañobre Corpas, Yanet y Brossard González, Yulie.** *Diagnóstico de Enfermedades de Transmisión Sexual mediante técnicas de Inteligencia Artificial.* La Habana: Universidad de las Ciencias Informáticas, Facultad 5, 2009.

37. **Baró, D.P. and E.R. Alonso.** *Proceso de análisis y gestión del conocimiento a partir de los datos obtenidos en la conducción de los de Ensayos Clínicos del Centro de Inmunología Molecular,*

aplicando técnicas de minería de datos. La Habana: Universidad de las Ciencias Informáticas, Facultad 6, 2009.

38. **Corría Ramírez, Isidro Manuel y Shelton Nadal, Ronald.** *Estrategia de trabajo para el desarrollo del módulo de minería de datos de un Call Center, aplicando la metodología CRISP-DM.* La Habana: Universidad de la Habana, Facultad de Matemática y Computación, 2004.

39. **Chapman y Pete (NCR), et al.** *CRISP-DM 1.0. Guía paso a paso de minería de datos.* Estados Unidos: s.n., 2000.

40. *Ídem a la referencia 39.*

41. *Ídem a la referencia 39.*

42. *Ídem a la referencia 39.*

43. *Ídem a la referencia 39.*

44. *Ídem a la referencia 39.*

45. *Ídem a la referencia 39.*

46. *Ídem a la referencia 39.*

47. Group, Machine Learning. Weka. *Universidad de Waikato.* [En línea] [Citado el: 9 de Mayo de 2011.] http://www.cs.waikato.ac.nz/ml/weka/.

48. *Ídem a la referencia 47.*

49. *Ídem a la referencia 47.*

50. *Ídem a la referencia 47.*

51. *Ídem a la referencia 47.*

52. **Sánchez, Jorge (www.jorgesanchez.net).** Scribd. *NetBeans-Guía Rápida.* [En línea] 2004. [Citado el: 9 de Mayo de 2011.] http://www.scribd.com/doc/6615294/NetBeans.

53. *Ídem a la referencia 52.*

54. *Ídem a la referencia 3.*

55. **Gallardo, Jose Alberto Arencibia.** *Metodología para el Desarrollo de Proyectos en minería de datos CRISP-DM. Sistemas del Conocimiento.* EPB603.

56. **Escribano Barreno, Julio.** *Minería de datos, Análisis de Datos mediante WEKA.*

57. *Ídem a la referencia 38.*

58. **Perversi, Ignacio.** Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina. [En línea] Instituto Tecnológico de Buenos Aires, Departamento de Ingeniería y Computación, 2007. [Citado el: 7 de Diciembre de 2010.] http://laboratorios.fi.uba.ar/lsi/rgm/tesistas/PERVERSI-tesisdegradoeningenieria.pdf.

59. *Ídem a la referencia 55.*

60. **Gómez Gómez, Omar Salvador.** Propuesta de mejora sobre la primera etapa del modelo de proceso KDDM, CRISP-DM. [En línea] [Citado el: 6 de Abril de 2011.] http://osgg.net.

61. *Ídem a la referencia 60.*

62. *Ídem a la referencia 60.*

63. *Ídem a la referencia 58.*

Bibliografía

1. **[SPS02]**.*"SPSS Home Page".* Disponible en: http://www.spss.com (2002). [Citado el: 16 de Noviembre de 2010.]

2. **A. Botey Puig, A. Coca Payeras; I. J. Ferreira Montero**. *Medicina Interna [Sección en CD ROM]*, Capítulo 55. Decimocuarta Edición. Editorial Harcourt. España. 2000.

3. *American Heart Association*. *Home Monitoring of High Blood Pressure*. Disponible en: http://www.americanheart.org/presenter.jhtml?identifier=576. [Citado el: 11 de Febrero de 2011.]

4. *Aplicación de minería de datos para el diagnóstico de accidentes cerebrovasculares agudos (ACVAs)*. Daedalus. Sector Medicina. Disponible en: http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/DAEDALUS-MD19-Accidentes_Cardiovasculares.pdf. [Citado el: 14 de Noviembre de 2010.]

5. **Autores, Colectivo de. Curso Universitario**. *Sistema de Información en los Sistemas de Salud*. Italia: Instituto Universitario del Hospital Italiano. pág. 38.

6. **Bañobre Corpas, Yanet; Brossard González, Yulie.** *Diagnóstico de Enfermedades de Transmisión Sexual mediante técnicas de Inteligencia Artificial.* La Habana : Universidad de las Ciencias Informáticas, Facultad 5, 2009.

7. **Baró, D.P.; E.R. Alonso**, *Proceso de análisis y gestión del conocimiento a partir de los datos obtenidos en la conducción de los de Ensayos Clínicos del Centro de Inmunología Molecular, aplicando técnicas de minería de datos*, Facultad 6. 2009, Universidad de las Ciencias Informáticas: Ciudad de la Habana. p. 77.

8. **Bressán, G.E.**, *Almacenes de datos y minería de datos, T.y.s. distribuidos*, Editor. 2003, Universidad Nacional del Nordeste. Facultad de Ciencias Exactas y Naturales y Agrimensura: Buenos Aires, Argentina.

9. **Cabrera Cruz**, **Niviola**; *"Retos y posibilidades de los Ensayos Clínicos Controlados para los pacientes"*, Ministerio de Salud Pública de Cuba, junio 2008 Costa Rica. Disponible en:

http://www.eventos.bvsalud.org/agendas/BVS-COR/public/documents/Niviola_RETOS%20_POSIBILIDADES_EC-150413.pdf. [Citado el: 13 de Noviembre de 2010.]

10. **Calderón, Montero, Alberto**. *Información sobre HTA para pacientes y familiares*. Disponible en: http://www.medynet.com/hta/3.htm#1-2. [Citado el: 11 de Febrero de 2011.]

11. *Centro de Estudios de Reconocimiento de Patrones y minería de datos (CERPAMID)*,- Disponible en: http://www.cerpamid.co.cu, [Citado el: 7 de Diciembre de 2010.]

12. **Chapman; Pete (NCR), et al**. *CRISP-DM 1.0. Guía paso a paso de minería de datos*. Estados Unidos: s.n., 2000.

13. **Chobanian, Aram, Brakis, George, Black, Henry, Cushman, William, Green, Lee, Izzo, Joseph et all**. *Joint -7 Complete Version Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of the High Blood Pressure*. Disponible en: http://www.sld.cu/servicios/hta/doc/CopiadeVIIJNC.pdf . [Citado el: 10 de Febrero de 2011.]

14. **Corría Ramírez, Isidro Manuel; Shelton Nadal, Ronald**, *Estrategia de trabajo para el desarrollo del módulo de minería de datos de un Call Center*, aplicando la metodología CRISP-DM, Facultad de Matemática y Computación. 2004, Universidad de la Habana: Ciudad Habana. p. 93.

15. **DAEDALUS** - *Data, Decisions and Language, S. A*. Disponible en: http://www.daedalus.es/mineria-de-datos/proceso-de-mineria-de-datos/ [Citado el: 15 de Noviembre de 2010.]

16. **Dopico Mateo, Dra.C. Ileana; Placencia Salgueiro**. *Diplomado "minería de datos para las organizaciones, la industria y el medio ambiente."* Pertinencia y concepción curricular. La Habana, Cuba: Ministerio de Educación Superior de Cuba e Instituto de Cibernética, Matemática y Física. CITMA, 2009.

17. *Environmental Systems Research Institute, Inc*. Disponible en: http://www.esri.com/software/arcview. [Citado en: Noviembre de 2010.]

18. **Escribano Barreno, Julio.** *minería de datos,Análisis de Datos mediante WEKA.*

19. **Fayyad, D. M.; Piatetsky-Shapiro, G.; Smyth, P**. *From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996. Disponible en: http://elvex.ugr.es/etexts/spanish/kdd/KDD.html. [Citado en: Noviembre de 2010.]

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

21. **Febles Rodríguez, Juan Pedro y González Pérez, Abel**; ACIMED 02 2002; *"Aplicación de la minería de datos en la bioinformática"*. Disponible en: http://bvs.sld.cu/revistas/aci/vol10_2_02/aci03202.htm, [Citado el: 12 de Noviembre de 2010.]

22. **Fonseca Reyes, Salvador, Parra Carrillo, José Z**. *Las guías de Tratamiento en Hipertensión Arterial*. Capítulo 9.3. Universidad de Guadalajara. Centro Universitario de Ciencias de la Salud. Disponible en: http://virtual.cucs.udg.mx/recursos/capitulo9.3.pdf. [Citado el: 1 de Febrero de 2011.]

23. **Gómez Gómez, Omar Salvador .** *Propuesta de mejora sobre la primera etapa del modelo de proceso KDDM, CRISP-DM.*

24. **González Gómez, Jose Ignacio**. *Generalidades de la minería de datos*. s.l.: Departamento de Economía Financiera y Contabilidad, Universidad de La Laguna.

25. **Graham J. Williams; Simeon J. Simoff .** *Data Mining Theory, Methodology, Techniques,and Applications.* Springer-Verlag Berlin Heidelberg : s.n., 2006.

26. *Guía Colombiana de Atención de la Hipertensión Arterial*. Disponible en: http://www.sld.cu/galerias/pdf/servicios/hta/guia_colombiana_de_hta_para_medicos.pdf. [Citado el: 4 de Febrero de 2011.]

27. **Hernández Orallo, Jose, Ramírez Quintana, Maria José; Ferri Ramírez, César**. *Introducción a la minería de datos*. Valencia, España: Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia, 2004. ISBN: 84-105-409 1-0.

28. *Hipertensión Arterial del MINSAP. Programa Nacional de Prevención, Diagnóstico, Evaluación y Control de la Hipertensión Arterial. Guía para la atención médica*. Octubre 2004. Biblioteca Médica Nacional. La Habana.

29. *Hipertensión Arterial*. Disponible en URL http://www.sld.cu/servicios/hta/. [Citado el: 28 de Enero de 2011.]

30. *I2 Analyst´s Workstation*. Disponible en: http://www.i2.co.uk/Products/Analysts_Workstation/default.asp. [Citado en: Noviembre de 2010.]

31. Information Builders. Disponible en: http://www.informationbuilders.com. [Citado en: Noviembre de 2010.]

32. IT Innovation Centre. *"CRITIKAL. European Project for Large Scale Data Mining"*. Disponible en: http://www.attar.com/pages/critikal.htm (1999). [Citado el: 13 de Noviembre de 2010.]

33. Joint National Committee. Report of the Joint National Committee on Detection, *Evaluation and Treatment of High Blood Pressure*. A Cooperative Study. JAMA 1977; 237:255-261.

34. Larose, Daniel T. *DATA MINING METHODS AND MODELS.* s.l. : Department of Mathematical Sciences Central Connecticut State University, 2006. ISBN-13 978-0-471-66656-1.

35. Last Ben-Gurion, Mark; Abraham Kandel, Israel . *DATA MINING IN TIME SERIES DATABASES.* Florida : Israel University of South Florida. ISBN 981-238-290-9.

36. Macías, Castro, Ignacio, Cordies, Jakson, Liliam, Landrove, Rodriguez, Orlando, Vazquez, Vigoa, Alfredo, Alfonso, Guerra, Jorge et all. *Programa Nacional de Prevención, Diagnóstico, Evaluación y Control de la Hipertensión Arterial*. Biblioteca Médica Nacional. La Habana.

37. *MapInfoCorporation*. Disponible en: http://www.mapinfo.com/location/integration. [Citado en: Noviembre de 2010.]

38. Marante Jacas, Danier; Marante Jacas, Danner. *Aplicación de la minería de datos para la exploración y detección de patrones delictivos*. La Habana: Universidad de las Ciencias Informáticas, Facultad 8, 2008.

39. Martín Rodríguez, Diana; Socorro Llanes, Raisa; Wilford Rivera, Ingrid. *Herramienta de minería de datos para usuarios no expertos basada en bibliotecas de Weka*. UCIencia 2008. Instituto Superior Politécnico José Antonio Echeverría (CUJAE).

40. **Molina López, J. M.; García Herrero, J**. *Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA*. Madrid, Universidad Carlos III, 2006.

41. **Molina López, José Manuel; García Herrero, Jesús.** *Técnicas de Análisis de Datos: Aplicaciones prácticas utilizando Microsoft Excel y WEKA.* Madrid: s.n., 2004.

42. **Molina, Rafael, Martí, Juan Carlos**. *JNC 7mo*. Disponible en: http://www.sld.cu/servicios/hta/doc/JNC-7esp.pdf. [Citado el: 11 de Febrero de 2011.]

43. **Oatley, G.C., B.W. Ewart, J. Zeleznikow**, 2004. *Decision Support Systems for Police: Lessons from the application of Data Mining Techniques to "Soft" forensic Evidence*. Disponible en:http://www.aic.gov.au/conferences/occasional/2005-04.zeleznikow.html. [Citado en: Noviembre de 2010.]

44. **Pautsch, J.G.A**., *minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación*., Facultad de Ciencias Exactas, Químicas y Naturales. 2009, Universidad Nacional de Misiones: Argentina. p. 171.

45. **Pérez, Roberto**. *Hipertensión Arterial*. Disponible en: http://www.monografias.com/trabajos10/confind/confind.shtml. [Citado el: 10 de Febrero de 2011.]

46. **Perversi, Ignacio**., *Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina:* Departamento de Ingeniería y Computación. 2007, Instituto Tecnológico de Buenos Aires: Argentina, Buenos Aires. p. 117. Disponible en: http://laboratorios.fi.uba.ar/lsi/rgm/tesistas/PERVERSI-tesisdegradoeningenieria.pdf [Citado el: 7 de Diciembre de 2010.]

47. **Pinal Borges, M**. *Curso sobre Hipertensión arterial (En 11 grupos de presentaciones en power point).Primera parte*. Disponible en: http://www.sld.cu/servicios/hta/. [Citado el: 19 de Enero de 2011.]

48. **Rodríguez Jara, Félix; Vallejo Delgado, Nieves**, *Aplicación de técnicas de minería de datos para el diagnóstico prematuro de Cáncer*, Universidad Carlos III. Madrid: Madrid, España. p. 10.

49. **Román, Oscar, Alvo, Miriam, Prat, Hernán, Fasce, Oscar**. *Guías Clínicas para el tratamiento del Adulto Mayor con Hipertensión en el nivel primario de Atención*, 1999. Disponible en: http://www.sld.cu/galerias/pdf/servicios/hta/guias_clinicas_para_el_adulto_mayor_1999_chile_1.pdf [Citado el: 10 de Febrero de 2011.]

50. **Rosete Suárez, Alejandro, Rodríguez Díaz, Alfredo; Acosta Sánchez, Rolando**. *Revista Cubana de Informática Médica*. Acceso Octubre de 2010. [Citado el: 9 de Noviembre de 2010.]. Disponible en: http://www.rcim.sld.cu/revista_18/articulos_htm/prediccionpaciente.htm#t. ISSN: 1684-1859.

51. **Rueda-Clausen Gómez, Christian Federico, Villa-Roel Gutiérrez, Cristina, Rueda-Clausen Pinzón, Christian Eduardo**; *"Indicadores bibliométricos: origen, aplicación, contradicción y nuevas propuestas"*; Disponible en: http://74.125.155.132/search?q=cache:K3clqaB2qoMJ:caribdis.unab.edu.co/pls/portal/url/ITEM/20BED9CE888965B8E0440003BA3D5405+indicadores+bibliometricos&cd=3&hl=es&ct=clnk&gl=cu, [Citado el: 12 de Noviembre de 2010.]

52. **Sánchez Corales, Yosvannys**. *Documento de Arquitectura de Software v1.0*. La Habana, 2010.

53. **Scott, C J and Al-Attar, A and Schneider, W and Nisbet, D; Barth, T and Schwarz, H.***"CRITIKAL Final Report"*. Department of ECS. University of Southampton, (1999).

54. **Servente M.** *"Algoritmos TDIDT aplicados a la minería de datos Inteligente"*, Prof. Dr. Ramón García Martínez (Dir.). Universidad de Buenos Aires, Facultad de Ingeniería. Tesis de Grado en Ingeniería Informática, 2002. Disponible en: http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf. [Citado el: 17 de Noviembre de 2010.]

55. **Silveira Martineaux, Karina; Fernández Pérez, Reidelendy**., *Comparación de algoritmos de clasificación y agrupamiento aplicando técnicas de minería de datos*, Facultad 5. 2008, Universidad de las Ciencias Informáticas: Ciudad de la Habana. p. 75.

56. **Suils, Pilar Sopena.** *Extracción y Recuperación de información no supervisada*. Valladolid: s.n., 2008.

57. **Thomas P, Soumen Chakrabarti; Richa, Nadeau Earl Cox.** *Data Mining Know It All.* ISBN 978-0-12-3746.

58. **Toledo Curbelo, G, J**. *Fundamentos de Salud Pública 1*. Editorial de Ciencias Médicas. La Habana. 2005. Biblioteca Médica Nacional.

59. **Vilches González, Erika; Escobar Broitman, Iván A**. *minería de datos*. Septiembre 2007. Disponible en: http://www.erikavilches.com/km/mineria_datos.pdf [Citado el: 14 de Noviembre de 2010.]

60. **Virseda Benito, Fernando; Román Carrillo, Javier**, *minería de datos y aplicaciones*, Universidad Carlos III: Madrid, España. p. 8.

61. **Witten, I.H.; Frank, E**. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. EE.UU, San Diego: Morgan Kaufmann Publishers, 2000.

62. **Witten, Ian H.; Eibe , Frank .** *Data Mining Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems).* s.l. : Jim Gray, Microsoft Research, 2005.

63. **Yunyue Zhu .** *High Performance Data Mining in Time Series: Techniques and Case Studies.* January 2004.

64. **Zamarrón Sanz, Carlos, et al**. *Aplicación de la minería de datos al estudio de las alteraciones respiratorias durante el sueño*, Santiago de Compostela. España: Hospital Clínico Universitario de Santiago de Compostela, 2006.p. 11.

Glosario de Términos

- ✓ **Aplicación o Sistema Informático:** Programas con los cuales el usuario final interactúa a través de una interfaz y que realizan tareas útiles para éste.

- ✓ **Aprendizaje Automático:** Es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan crear programas capaces de generalizar comportamientos a partir de una información no estructurada.

- ✓ **Árboles de Decisión:** Representan reglas donde atributos independientes determinan los valores finales. En estos árboles cada nodo representa una propiedad que puede tomar diversos valores, cada uno de los cuales genera una rama.

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

- ✓ **Bases de Datos:** Se define como una serie de datos organizados y relacionados entre sí, los cuales son recolectados y explotados por los sistemas de información de una empresa o negocio en particular.

- ✓ **CENATAV:** Sus siglas significan Centro de Aplicaciones de Tecnologías de Avanzada y es un centro orientado a las investigaciones teóricas y aplicadas en el área del Reconocimiento de Patrones y la minería de datos.

- ✓ **CERPAMID:** Sus siglas significan Centro de Estudios de Reconocimiento de Patrones y Minería de Datos y está orientado a la investigación básica y aplicada en el área del Reconocimiento de Patrones y su aplicación a la minería de datos y Textos.

- ✓ **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.

- ✓ **Clúster:** Conjunto o racimo de objetos, que tienen características comunes.

GLOSARIO DE TÉRMINOS

- ✓ **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.

- ✓ **Componente:** Parte física y reemplazable de un sistema que se ajusta a, y proporciona la realización de, un conjunto de interfaces.

- ✓ **Data Warehouse:** Es una colección de datos en la cual se encuentra integrada la información de la Institución y que se usa como soporte para el proceso de toma de decisiones gerenciales.

- ✓ **Deficiencia:** Es toda pérdida o anormalidad de una estructura o función psicológica, fisiológica o anatómica.

- ✓ **Dependencia:** Relación semántica entre dos elementos, en la cual un cambio en uno puede afectar al otro.

- ✓ **Dominio:** Área de conocimiento o actividad caracterizada por un conjunto de conceptos y terminología comprendidos por los practicantes de ese dominio.

- ✓ **Equipo Básicos de Salud:** Binomio conformado por el médico y enfermera de la familia, que atiende una población geográficamente determinada, que puede estar ubicado en la comunidad, centros laborales o educacionales.

- ✓ **Informática:** Disciplina que estudia el tratamiento automático de la información utilizando dispositivos electrónicos y sistemas computacionales.

- ✓ **Informatizar:** Proceso de aplicar sistemas o equipos informáticos al tratamiento de la información.

- ✓ **Inteligencia Artificial:** Es la ciencia que enfoca su estudio a lograr la comprensión de entidades inteligentes.

- ✓ **Internet:** Método de interconexión de redes de computadoras implementado en un conjunto de protocolos denominado TCP/IP y garantiza que redes físicas heterogéneas funcionen como una red (lógica) única.

GLOSARIO DE TÉRMINOS

- ✓ **KDD [Knowledge Discovery in Databases]:** Su término en español significa Extracción de Conocimientos en Bases de Datos.

- ✓ **MD (minería de datos):** Extracción de conocimientos ocultos en grandes bases de datos.

- ✓ **PL/PGSQL:** Lenguaje de programación, estructurado de consulta utilizado para realizar la consultas en PostgreSQL.

- ✓ **Policlínico:** Unidad de salud donde se brindan servicios médicos a una población geográficamente determinada perteneciente al nivel asistencial de Atención Primaria de Salud.

- ✓ **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.

- ✓ **Redes Neuronales:** son utilizadas para la predicción, la minería de datos, el reconocimiento de patrones y los sistemas de control adaptativo. Constituyen una parte muy importante en el estudio y desarrollo de la inteligencia artificial y el de la vida artificial.

- ✓ **Servicio:** Unidad de software que encapsula alguna funcionalidad de negocio y proporciona estas a otros servicios a través de interfaces públicas bien definidas.

- ✓ **Software:** Conjunto de programas y procedimientos necesarios para hacer posible la realización de una tarea específica, en contraposición a los componentes físicos del sistema.

- ✓ **Unidad de Salud:** Centro de trabajo que pertenece al Ministerio de Salud Pública (MINSAP).

- ✓ **Weka:** Herramienta de MD.

Anexos

Anexo 1. Tablas de mayor interés.

A continuación se muestran las tablas de mayor interés para el proceso de minería, las mismas pertenecen al Almacén de Datos del (alasSIAPS):

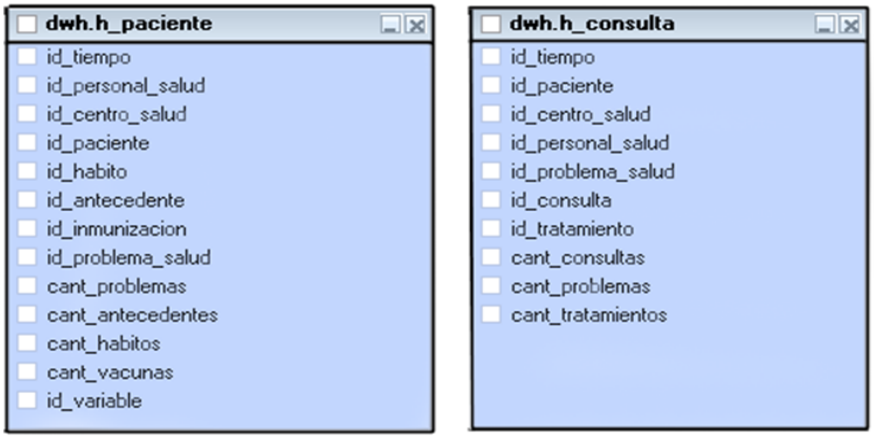

Imagen 9. Hechos necesarios para la integración de las tablas de interés

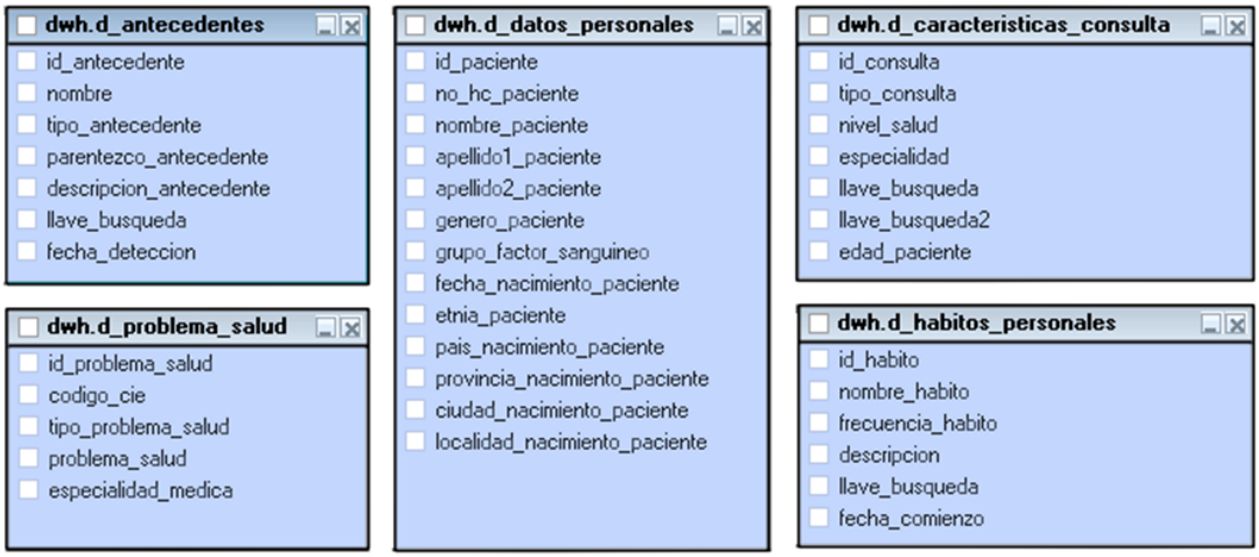

Imagen 10. Tablas Seleccionadas del Almacén de Datos del (alasSIAPS).

Anexo 2. Significado de los valores numéricos empleados.

Tabla 13. Significado de los valores numéricos empleados en la tabla *vista_minable_skm*.

Atributos	Valor Numérico	Significado
rango_edad	1	Rango1: Menores de 45 años.
	2	Rango2: Entre 45 y 65 años.
	3	Rango3: Mayores de 65 años.
genero_paciente	1	M: Sexo masculino.
	2	F: Sexo femenino.
etnia_paciente	1	Blanca: Raza blanca.
	2	Mestiza: Raza mestiza.
	3	Negra: Raza negra.
Resto de las variables	0	No: Casos negativos.
	1	Si: Casos positivos.

Anexo 3. Datos Combinados.

id_paciente	rango_edad	genero_paciente	etnia_paciente	nombre_antecedent	nombre_habito	problema_salud
57	Rango2	F	Mestiza	APF Dislipemia	Ejercicio Fisico	Mareos
33	Rango1	M	Negra	APF Asma	Ejercicio Fisico	No Problema Salud
19	Rango2	F	Blanca	APF Diabetes Mellitus	Sedentaria	Mareos
114	Rango2	M	Mestiza	APP Asma	No Habitos	Mareos
28	Rango1	M	Blanca	APP HTA	No Habitos	Mareos
28	Rango1	M	Blanca	APP Asma	No Habitos	Mareos
49	Rango2	M	Negra	APF Diabetes Mellitus	Drogas	Vertigos
10	Rango2	M	Blanca	APF Cefalea	Sedentaria	Calambres
35	Rango2	M	Mestiza	No Antecedentes	Sedentaria	Mareos
29	Rango2	M	Blanca	APP Salud Mental	No Habitos	Afectaciones Sistema Nervioso
43	Rango2	M	Negra	No Antecedentes	Sedentaria	Mareos
29	Rango2	M	Blanca	APP HTA	No Habitos	Afectaciones Sistema Nervioso

Imagen 11. Fragmento de la tabla *pre_vista_minable*.

Función agrupando_por_id:

```
DECLARE
  res md.pre_vista_minable%rowtype;
BEGIN
  for res in
      SELECT md.pre_vista_minable.*
      FROM md.pre_vista_minable
      WHERE  md.pre_vista_minable.id_paciente = $1
      loop
            return next res;
    end loop;
END;
```

Función insertando_en_tb_vista_minable:

```
insert into md.vista_minable
values ($1, $2, $3,$4,
        $5, $6, $7, $8,
        $9, $10, $11, $12,
        $13, $14, $15, $16,
        $17, $18, $19, $20,
        $21, $22, $23, $24,
        $25, $26, $27, $28,
        $29, $30,
        $31);
```

Imagen 12. Procedimientos almacenados.

id_paciente	rango_edad	ge	etnia_pa	AF	AF	AF	AF	AF	AF	AF	CE	DIS	PA	ED	DA
2	Rango2	M	Mestiza	No	No	No	No	No	Si	No	Si	No	No	No	No
3	Rango2	F	Mestiza	No	No	No	No	No	No	No	Si	No	No	No	No
4	Rango1	M	Negra	No	No	No	Si	No	No	No	Si	No	No	No	No
5	Rango2	M	Negra	No	No	No	Si	No	No	No	Si	No	No	No	No
6	Rango2	M	Blanca	No	No	No	No	No	No	No	Si	No	No	No	No
7	Rango1	F	Blanca	No	No	No	Si	No	No	No	Si	No	No	No	No
8	Rango1	F	Mestiza	No	No	No	No	No	No	Si	Si	No	No	No	No
9	Rango2	M	Blanca	No	No	No	No	No	No	No	No	No	No	No	No
10	Rango2	M	Blanca	No	No	No	No	No	No	No	Si	No	No	No	No
11	Rango3	F	Negra	No	No	No	Si	No	No	No	No	No	No	No	No
12	Rango3	M	Negra	No	No	No	No	No	No	Si	Si	No	No	No	No

Imagen 13. Fragmento de la tabla *vista_minable_j48*.

id_pacier	rango_edi	ge	etnia_pa	AF	AF	AF	AF	AF	AF	AF	CE	DIS	PA	ED	DA
50	2	1	1	0	0	0	0	1	0	0	0	0	0	0	0
32	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
689	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0
254	3	1	1	1	0	0	0	1	0	0	0	0	1	1	0
575	3	1	1	0	1	0	0	0	0	0	0	0	1	0	0
355	3	1	1	0	1	0	0	0	0	0	1	0	0	0	0
384	3	1	1	0	1	0	0	0	0	0	1	0	0	0	0
395	3	1	1	0	1	0	0	0	0	0	1	0	0	0	0
696	3	2	1	0	0	0	1	1	0	1	1	0	0	0	0
574	3	2	1	1	1	1	1	0	0	0	0	1	0	1	0
658	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0
31	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0

Imagen 14. Fragmento de la tabla *vista_minable_skm*.

```
UPDATE md.vista_minable
SET genero_paciente = 1
WHERE md.vista_minable.genero_paciente ='M';

UPDATE md.vista_minable
SET "HTA" = 1
WHERE md.vista_minable."HTA" ='Si';
```

Imagen 15. Dos de las Consultas SQL utilizadas para la transformación de la tabla *md.vista_minable_j48* a la tabla *md.vista_minable_skm.*

Anexo 4. Árbol de Decisión J48.

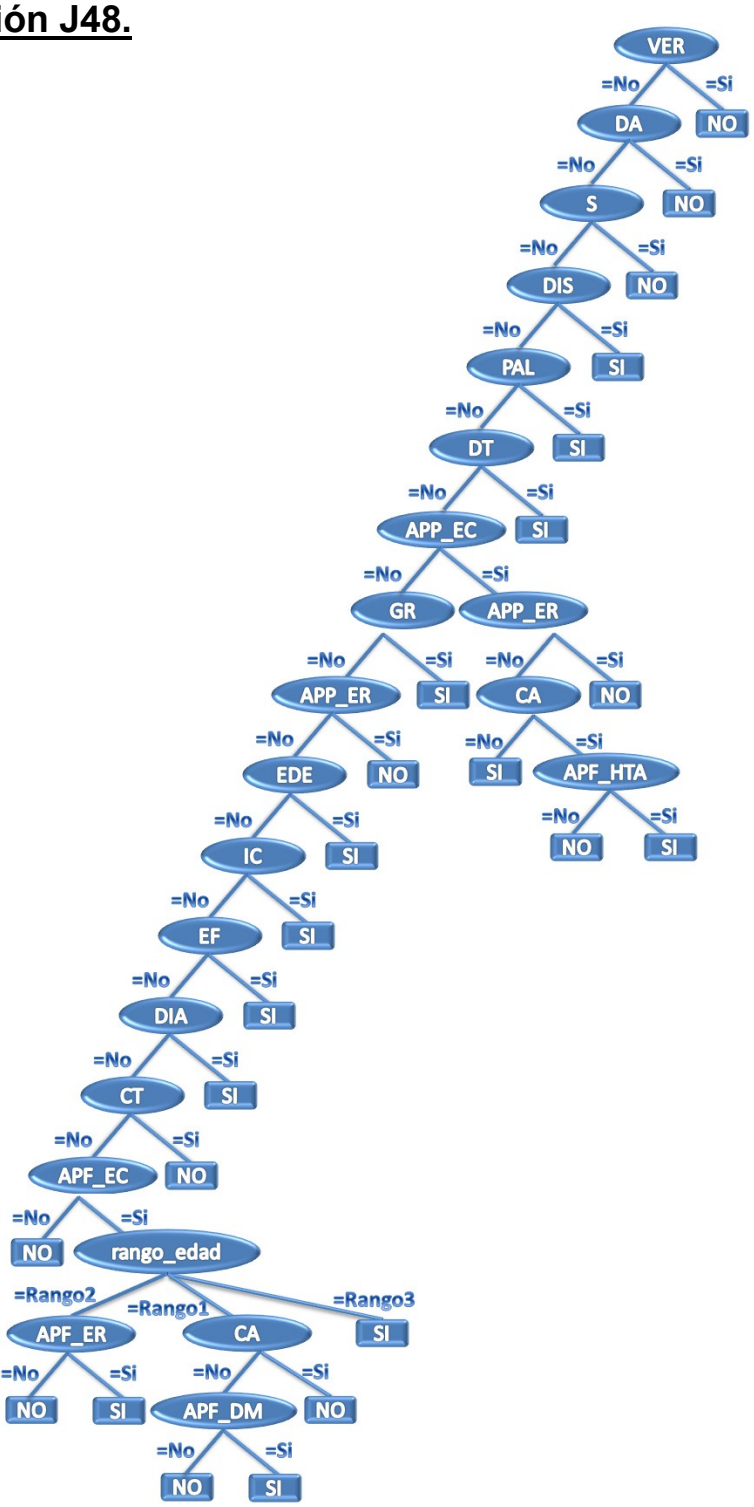

Imagen 16. Árbol de decisión generado por Weka aplicando el algoritmo J48.

Anexo 5. Agrupamientos.

Tabla 14. Resultado de K-Means para 3 clúster con varias semillas.

Semilla	*Error Cuadrático*	*Núm. Iteraciones*	*Semilla*	*Error Cuadrático*	*Núm. Iteraciones*
1	3100	4	**11**	3215	3
2	3182	4	**12**	3108	6
3	3091	5	**13**	3093	7
4	3102	3	**14**	3312	3
5	3350	4	**15**	3180	4
6	3140	7	**16**	3080	3
7	3133	3	**17**	3245	5
8	3074	5	**18**	3226	3
9	3121	3	**19**	3325	3
10	3335	3	**20**	3179	4

Anexo 5.1. Gráficos de Distribución.

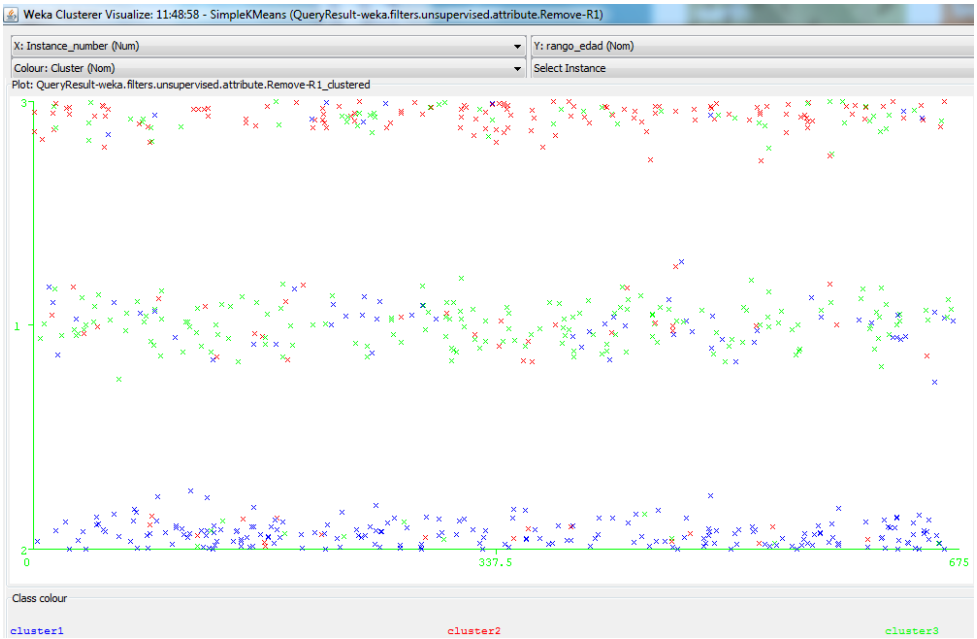

Imagen 17. Distribución de grupos según atributo *rango_edad.*

La imagen 17 puede describirse de la siguiente manera: en el grupo azul se destacan las personas que se encuentran entre 45 y 65 años de edad, el grupo rojo está muy concentrado de personas de más de 65 años de edad y el verde representa en su mayoría a las personas que son menores de 45 años.

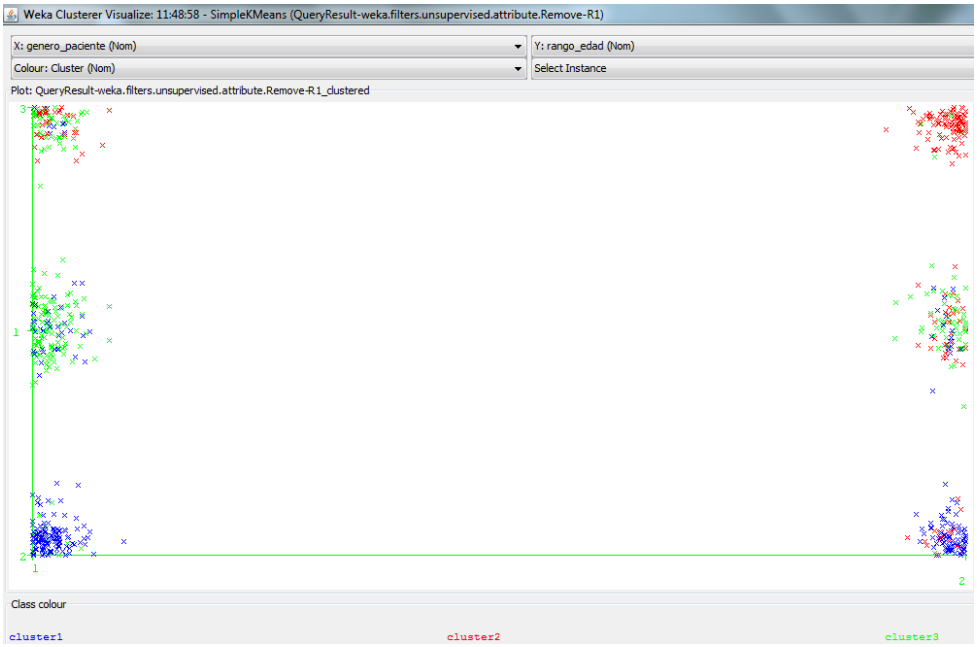

Imagen 18. Distribución de grupos según atributo *genero_paciente y rango_edad.*

En la imagen 18, se analiza de forma fija la variable *genero_paciente* y se ajusta la variable *rango_edad* a los grupos obtenidos, por ejemplo:

- ✓ Para sexo masculino; en el grupo verde se destacan las personas que son menores de 45, de igual forma este grupo está representado, generalmente para aquellas que son mayores de 65 años; en el grupo azul se encuentran los pacientes que están entre 45 y 65años de edad y el grupo rojo, aunque está muy poco representado por este sexo siempre destacan pacientes mayores de 65 años.
- ✓ Para sexo femenino; en el grupo rojo se encuentran personas mayores de 65 años, en el verde menores de 45 y en el azul aquellas que están entre 45 y 65 años de edad. Es válido destacar que aunque en pequeñas proporciones el grupo rojo está representado también dentro del azul y el verde.

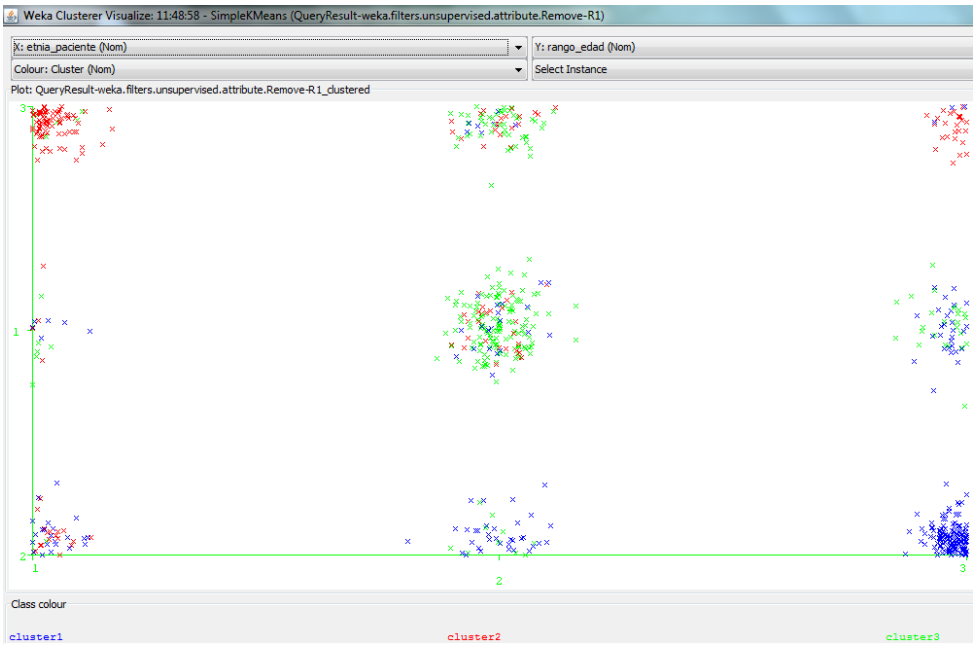

Imagen 19. Distribución de grupos según atributo *etnia_paciente y rango_edad.*

La imagen 19 demuestra que el grupo rojo está mayormente representado por personas mayores de 65 años de edad y que además son de raza blanca; así mismo, aunque en menor proporción, existe representación de personas en este rango de edad de raza negra. En el grupo azul se encuentran los pacientes que están entre 45 y 65 años de edad generalmente de raza negra, aunque existe una buena representación de los mismos que son de raza mestiza. Finalmente, el grupo verde se destaca por agrupar generalmente a las personas de menos de 45 años de edad y de raza mestiza.

Así mismo las gráficas siguientes son analizadas a partir de otras variables, el re sumen de todas ellas se encuentra explícitamente en el epígrafe 3.1.3.2 (Evaluación del modelo generado por Simple K-Means), donde se describen las características específicas de cada grupo obtenido.

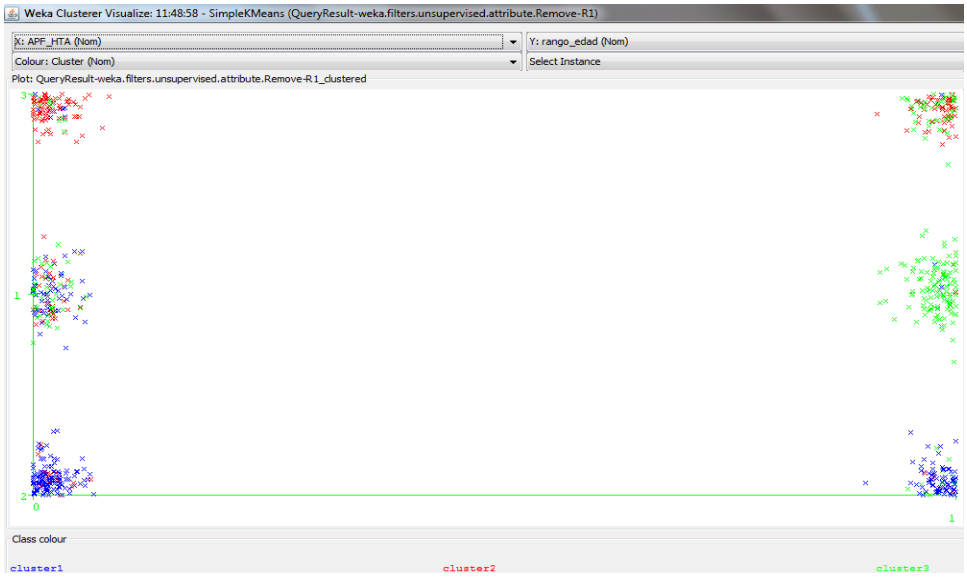

Imagen 20. Distribución de grupos según atributo *APF_HTA y rango_edad.*

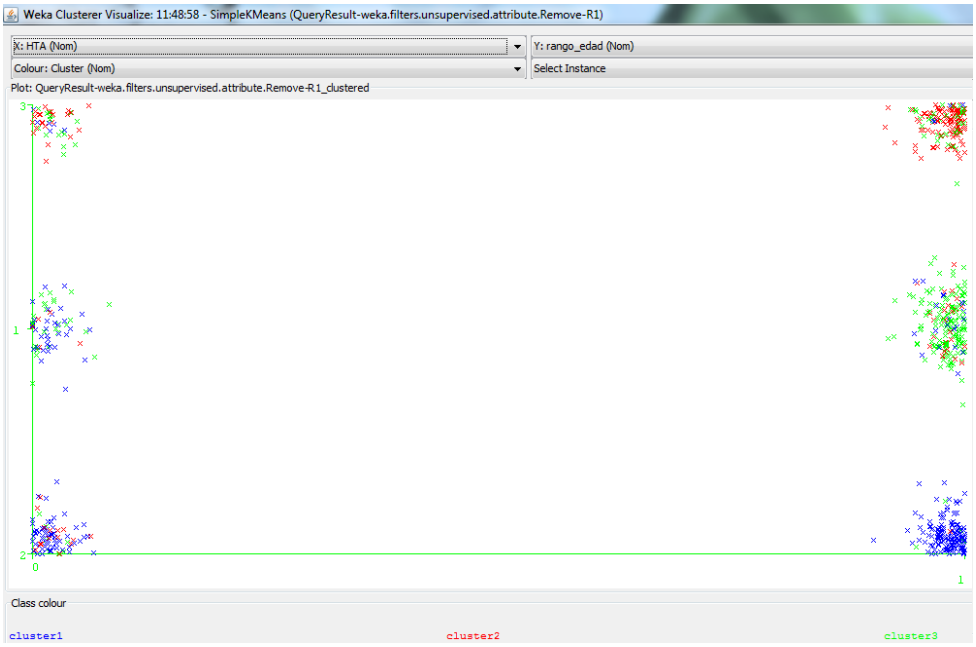

Imagen 21. Distribución de grupos según atributo *HTA y rango_edad.*