BENEMÉRITA UNIVERSIDAD AUTONOMA DE PUEBLA

TESIS PROFESIONAL

QUE PARA OBTENER EL GRADO DE:

LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN

"SISTEMA DE APOYO A LA TOMA DE DECISIONES UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA EL CASO DE DIABETES EN JUCHITÁN, OAXACA "

PRESENTA

JOSÉ CASTILLEJOS LÓPEZ

ASESOR

DRA. MARÍA JOSEFA SOMODEVILLA GARCÍA

PUEBLA, MEXICO **AGOSTO 2013**

AGRADECIMIENTOS

….. Antes que nada a Dios por permitirme lograr mis metas, y por llenar mi vida de dicha, salud y sobre todo por las bendiciones que me ha otorgado.

….. A mis padres a quienes agradezco de corazón el esfuerzo que han realizado para que yo logre esta meta, pero sobre todo por su amor, cariño y comprensión, en todo momento los llevo conmigo en las enseñanzas que me han brindado, Los Amo.

….. A mis hermanos Jesús y Rocio, por el gran apoyo que me brindan, sé que estarán ahí siempre que los necesite, y también por ser mis primeros amigos, Los quiero mucho.

….. A mis abuelos Consuelo Dehesa, José López (+), Hilda Fuentes, Martin Castillejos, por la confianza, el apoyo incondicional y sobre todo por creer en mí.

….. A mis amigos que me han apoyado en las buenas y en las malas, a ellos que me brindan su confianza, amistad, y sobre todo por aconsejarme en los momentos que más se necesitan.

…A mí cuñado Mario por facilitarme la información utilizada en la investigación, sin esos datos esto no pudiera haber sido posible.

….. A mi asesora, Dra. María J. Somodevilla Garcia, por haber confiado en mí, por brindarme el apoyo incondicional para la elaboración de este trabajo, por su paciencia, sus consejos y sobre todo por haber compartido conmigo sus conocimientos.

Por eso y más GRACIAS a todos…..

"Muchos de los fracasos de la vida son de personas

que no se dieron cuenta cuán cerca

estaban del éxito cuando se dieron por vencidos."

Thomas A. Edison

RESUMEN

En la actualidad, son diversas las enfermedades que encabezan las listas de índices de mortalidad en el mundo. Por lo cual ya son muchas las instituciones que buscan, una opción viable, para realizar actividades que ayuden al control y prevención de ellas.

Una de ellas es la Diabetes Mellitus(DM),la cual es una enfermedad producida por una alteración del metabolismo, caracterizada por un aumento de la cantidad de glucosa en la sangre y por la aparición de complicaciones microvasculares y cardiovasculares. Estas complicaciones incrementan sustancialmente los daños en otros órganos (riñones, ojos, corazón, nervios periféricos) y la mortalidad asociada con la enfermedad y reduce la calidad de vida de las personas.

En el presente trabajo de tesis se plantea el desarrollo de unas vistas minables, con la finalidad de organizar la información relativa a diabetes en la zona de Juchitán, Oax. , para su posterior explotación utilizando minería de datos en WEKA. El análisis de la información en su conjunto se realizará siguiendo la metodología de desarrollo KDD[1] (***K****nowledge****D****iscovery in* ***D****atabases*).

El objetivo de este proyecto es proveer una herramienta de software, que pueda dar apoyo a los usuarios, al momento de tomar decisiones.

[1]*KDD (KnowledgeDiscovery in Database)* es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos [3].

ÍNDICE

ÍNDICE DE TABLAS.

ÍNDICE DE FIGURAS

CAPÍTULO 1

INTRODUCCIÓN

A lo largo de las tres últimas décadas, muchas organizaciones han generado una gran cantidad de datos automatizados en forma de archivos y bases de datos. Para procesar estos datos, disponemos de la tecnología de bases de datos que proporciona lenguajes de consulta como *SQL*[1] [6].

Gran parte de esa información es histórica, lo que representa transacciones que se han producido. Este tipo de información es de gran utilidad, ya que ayuda a explicar el pasado, entender el presente y predecir la información futura. De modo resumido puede considerarse la Minería de Datos (*Data Mining*, DM en inglés) como un proceso de descubrimiento de nuevas técnicas y significativas relaciones, patrones y técnicas al examinar grandes cantidades de datos [7]. El descubrimiento de los patrones en grandes volúmenes de datos es un proceso que se conoce como *KDD* (*Knowledge Discovery in Databases*)[2]. Es un proceso iterativo e interactivo que comprende seis fases, entre las que se cuenta la Minería de Datos.

La Minería de Datos se realiza por lo general con ciertos objetivos o para determinadas aplicaciones finales. Estos objetivos pueden clasificarse en los siguientes tipos: Predicción, Identificación, Clasificación y Optimización.

En el presente trabajo de tesis se plantea realizar estudios con la información relativa a diabetes en la zona de Juchitán, Oaxaca, para su explotación utilizando minería de datos en WEKA. El análisis de la información en su conjunto se realizara siguiendo la metodología de desarrollo KDD[2].

[1] LENGUAJE DE CONSULTA ESTRUCTURADO (*SQL por sus siglas en inglés*) [6].

[2] *KDD (KnowledgeDiscovery in Database) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos [3].*

1.1 PLANTEAMIENTO DE LA INVESTIGACIÓN

En esta sección se precisa el problema de investigación a resolver, se definen los objetivos del proyecto, al mismo tiempo se plantea la propuesta de solución y por último se describe la organización de la tesis.

1.1.1 PROBLEMA A RESOLVER

Para abordar el problema de prevención y/o tratamiento de la Diabetes en Juchitán, Oax. Haremos uso de los datos, otorgados por el Instituto Mexicano del Seguro Social (IMSS) y recursos del Portal Web del Instituto Nacional de Estadística Geográfica e Informática (INEGI). Debido al gran volumen de información se aplicará un proceso de extracción de conocimientos, con el objetivo de sentar bases para la toma de decisiones relativas al control de Diabetes en el área de influencia anteriormente especificada.

1.1.2 OBJETIVOS DE LA INVESTIGACIÓN

El objetivo general de la tesis es el siguiente:

Desarrollar un sistema para la toma de decisiones para analizar el caso de diabetes en la zona Juchitán, Oax, utilizando la metodología KDD.

Los objetivos particulares se enlistan a continuación:

1. Preprocesar la información proveniente de diferentes fuentes para crear la vista minable.
2. Utilización de técnicas de minería de datos para descubrimiento de patrones en las vistas minables para la predicción de futuros casos de diabetes.
3. Realizar cálculos de estimadores estadísticos acerca de la población en cuanto a los casos reales de la enfermedad en la ciudad de Juchitán Oaxaca.

1.1.3 JUSTIFICACIÓN DE LA INVESTIGACIÓN

Los criterios mostrados a continuación justifican la investigación en la que se fundamenta este trabajo de tesis:

- **DIABETES UN PROBLEMA MAYOR:** La obesidad, el sedentarismo tienen mucho que ver, en que la población sufra de diabetes, un trastorno en el que cuerpo humano no produce la cantidad de insulina necesaria para procesar los alimentos. Las personas que la padecen son propensos a desarrollar enfermedades como: **insuficiencia renal, ceguera y problemas cardiovasculares [8].**
- Durante las últimas décadas, el almacenamiento, organización y recuperación de la información, se han facilitado gracias a que los sistemas de bases de datos han sido automatizados, así como también para encontrar significado que tienen los datos almacenados en los grandes volúmenes de información. Esto ayuda a la toma de decisiones ya que permite analizar y explorar las bases de datos disponibles.
- Todas estas exigencias han hecho que surjan las nuevas generaciones de herramientas y técnicas para el soporte de la extracción de conocimiento, desde información disponible y que se denominan Minería de Datos.

1.2 PRESENTACIÓN DE LA SOLUCIÓN

La propuesta de solución al problema definido en la sección 1.1.1 y los productos desarrollados son comentados en la siguiente subsección.

1.2.1 PROPUESTA DE SOLUCIÓN

"Se propone el estudio de los datos obtenidos del IMSS y del INEGI para la creación de vistas minables, las cuales posteriormente serán explotadas con las técnicas adecuadas de Minería de Datos".

Para llevar a cabo la propuesta de solución se hará uso de Weka (*Waikato Environment for Knowledge Analysis*), el cual es un entorno para análisis del conocimiento desarrollado por la Universidad de Waikato, Nueva Zelandia. WEKA, es un conocido software para aprendizaje automático y Minería de Datos escrito en Java y de distribución libre, bajo licencia de GNU-GPL[3].

[3] Disponible en :http://es.wikipedia.org/wiki/Licencia_pública_general_de_GNU

1.3 APORTACIONES DE LA INVESTIGACIÓN.

Las aportaciones se derivan de la comparación de las técnicas de Minería de Datos, aplicadas sobre las vistas minables, enriqueciéndose por medio de la visualización de ejemplos, y de los alcances de trabajos similares presentados en el Capítulo 2, “Estado del Arte”.

1.4 ORGANIZACIÓN DE LA TESIS.

Este trabajo se organiza en capítulos, siendo distribuidos de la siguiente manera:

- CAPÍTULO 1: Introducción. En este capítulo se presenta el planteamiento del problema, los objetivos de la investigación, la propuesta de la solución y las aportaciones.

- CAPÍTULO 2: Estado del arte. En esta sección se detallan algunos trabajos de investigación sobre la Diabetes.

- CAPÍTULO 3: Marco teórico. En este capítulo, se profundiza sobre los conocimientos disponibles acerca de las Técnicas de la Minería de Datos. Así mismo se profundiza en los fundamentos teóricos para el sustento formal de la tesis.

- CAPÍTULO 4: Análisis y Diseño de vistas minables. Se muestra el preprocesamiento de los datos, para después hacer un análisis de los datos de diabetes en Juchitán, Oax.

- CAPÍTULO 5: Resultados. En este capítulo se presentan los resultados obtenidos por la aplicación de las técnicas de minería de datos aplicados a las vistas minables mostradas en el capítulo IV.

- CAPÍTULO 6: Conclusiones y Trabajo a Futuro. En esta sección se detallan las conclusiones de la investigación, con base a la experiencia adquirida en el desarrollo de la tesis.
- Por último se enlistan las referencias y material consultado.

CAPÍTULO 2

ESTADO DEL ARTE

La diabetes mellitus clínica representa un síndrome con metabolismo alterado, hiperglucemia debida a deficiencia absoluta de la secreción de insulina o a la reproducción de su eficacia biológica, o ambas.

El problema de la diabetes mellitus ha puesto en alerta, a las instituciones de salud a nivel mundial, ya que es una enfermedad que tiene grandes índices de mortalidad. Tan solo en México, la Diabetes Mellitus ocupa la segunda causa de muertes solo debajo de las enfermedades del corazón. Es tan grave el problema que ya hay diversos actos por parte de las instituciones no solo nacionales si no mundiales para prevenir y para controlar esta enfermedad.[2]

En Junio de 1.997, tras un acuerdo formulado por un Comité de expertos de la ADA y de la OMS, se propone una nueva clasificación de la diabetes, así como nuevos métodos de cribado y de diagnóstico. En esta nueva clasificación, se eliminan los términos de insulina-dependiente y no-insulinodependiente y se introducen los términos de diabetes tipo 1 y 2 (con números arábigos para evitar confusiones).La nueva clasificación queda de la siguiente manera:

1- Diabetes Mellitus tipo 1
   ..diabetes mediada por procesos autoinmunes

   ..diabetes idiopática

2- Diabetes Mellitus tipo 2
3- Alteración del metabolismo de la glucosa
4- Diabetes gestacional
5- Otros tipos de diabetes

2.1 CLASIFICACIÓN DE LA DIABETES

2.1.1-DIABETES MELLITUS TIPO 1

Diabetes mediada por procesos autoinmunes

Representa la mayoría de los casos de la diabetes tipo 1 y es debida a una destrucción autoinmune la célula beta pancreática.

Aunque puede ocurrir a cualquier edad, lo más frecuente es que aparezca en la infancia o adolescencia y se suele aparecer de forma brusca, siendo frecuente la cetoacidosis. Habitualmente el peso puede ser normal o por debajo de lo normal pero la obesidad no debe excluir el diagnóstico.

Estos pacientes pueden presentar otras enfermedades autoinmunes como la enfermedad de Graves, tiroiditis de Hasimoto, enfermedad de Adisson, vitiligo y anemia perniciosa.

Diabetes idiomática

Es de etiología desconocida y tiene un fuerte factor hereditario, no hay fenómenos autoinmunes y no se asocia al HLA..

Estos individuos pueden tener cetoacidosis y presentar diversos grados de deficiencia insulínica. La necesidad absoluta de insulina puede aparecer y desaparecer.

2.1.2-DIABETES MELLITUS TIPO 2

Su comienzo suele ser en la vida adulta y se caracteriza por una resistencia insulínica asociada con frecuencia a un déficit relativo a la insulina. Representa el 90-95 % de los pacientes con diabetes mellitas.

Estos pacientes suelen ser obesos y su comienzo normalmente es insidioso siendo raros lo episodios de cetoacidosis, aunque puede aparecer en situaciones de stress o infección.

El riesgo de aparición de este tipo de diabetes, aumenta con la edad, el peso y la falta de actividad física y es más frecuente en mujeres con diabetes gestacional y en individuos con hipertensión y dislipemia. No precisan insulina para mantener la vida aunque pueden requerirla para conseguir el control glucémico.

Aunque se sabe que tiene una fuerte predisposición genética, este factor no está claramente definido.

2.1.3-ALTERACIÓN DEL METABOLISMO DE LA GLUCOSA

Se incluyen dos categorías que se consideran factores de riesgo para futura diabetes y enfermedad cardiovascular:

Glucemia basal alterada (IFG: Impaired Fasting Glucose). Nueva categoría incluida en la clasificación de la diabetes. Cuando la glucemia basal es >= a 110 mg/dl y < de 126 mg/dl.

Tolerancia alterada a la glucosa (TAG o IGT. Imaireg Glucosa Tolerante).

2.1.4- DIABETES GESTACIONAL

Ocurre en el 2-5% de todos los embarazos. Comienza o se diagnostica por primera vez en el embarazo. Estas mujeres tienen a corto, medio o largo plazo, mayor riesgo de desarrollar DM2.

Teniendo en cuenta que esta es una enfermedad de las más tratadas en México y a nivel internacional, se cuenta con grandes cantidades de información almacenada; información histórica que nos indica, el avance o retroceso de la enfermedad. La minería de datos nos puede dar, una forma de tratar esos grandes volúmenes almacenados, para que podamos encontrar patrones y así nos podría facilitar la toma de decisiones.

Existen ya diversos métodos para predecir la diabetes, además de los que son pruebas médicas, existen algunos que son estadísticos, algunos de los cuales serán tratados en el transcurso de este capítulo. [2]

TIPO	DESCRIPCIÓN
Diabetes Mellitus tipo 1	Caracterizada por destrucción de la célula beta, que habitualmente lleva a déficit absoluto de insulina. Hay dos formas: ○ Diabetes Mellitus mediada por procesos inmunes. La destrucción de la célula beta resulta de un proceso autoinmune ○ Diabetes Mellitus idiopática: etiología desconocida
Diabetes Mellitus tipo 2	Caracterizada por resistencia insulínica, que habitualmente se acompaña de un déficit relativo de insulina. Puede variar desde resistencia insulínica predominante con deficit relativo de insulina a déficit insulínico predominante con alguna resistencia insulínica.
Alteraciones del metabolismo de la glucosa	Es un estado metabólico intermedio entre la normalidad y la diabetes. Es factor de riesgo para diabetes y enfermedad cardiovascular. ○ Glucemia Basal Alterada: Glucemia plasmática basal por encima de los valores normales y menor que el valor diagnóstico de Diabetes ○ Tolerancia alterada a la Glucosa: Glucemia plasmática mayor que los valores normales y menor que los diagnóstico de diabetes tras Sobrecarga de 75 gramos de glucosa.
Diabetes Gestacional	Sin cambios en la definición
Otros tipos específicos	Diabetes causada por otras etiologías identificables: 5. Defectos genéticos en la función de la célula beta 6. Defectos genéticos en la acción de la insulina 7. Enfermedades del páncreas exocrino 8. Endocrinopatías 9. Fármacos y drogas 10. Infección 11. Formas raras de diabetes relacionadas con procesos inmunes 12. Otros síndromes genéticos

Tabla 2.1.- Clasificación de los síndromes de diabetes mellitus idiopática. [3]

Cabe mencionar que los datos en la actualidad se consideran como un valioso recurso que debe ser transformado en información. Si la información es precisa y oportuna, es probable que su uso desencadene acciones que mejoren la posición competitiva de la compañía y genere riqueza [9].

2.2 PROYECTOS PARA LA PREVENCIÓN Y CONTROL DE LA DIABETES.

En esta sección mencionaremos el arduo trabajo que han realizado universidades, instituciones de salud tanto nacionales como internacionales, para poder dar una solución a la diabetes, ya sea con proyectos que ayudan a la detección de la enfermedad como en planeación para llevar un control.

2.2.1 PROYECTO MEXRISC.

En el año 1989 el Centro de Estudios en Diabetes A.C y el Instituto Nacional de Salud Pública arrancaron el Estudio de la Diabetes de la Ciudad de México, a más de 20 años de ello, se han generado datos muy valiosos para la salud de los mexicanos.

La Benemérita Universidad Autónoma de Puebla y el Instituto Nacional de Salud Pública se unen al llamado internacional de poner a la diabetes en el foco de atención y tomar el control, con el Observatorio MexRisc [4].

El observatorio MexRisc entrelaza resultados de investigaciones en diabetes y las tecnologías de la información para:

- Transmitir la buena noticia de que la diabetes es prevenible.
- Que existe una forma rápida, sencilla y gratuita de calcular el riesgo de padecer diabetes a futuro.
- Y una vez que se conoce el riesgo, orientar en como modificar de manera personalizada la forma de vida para retrasar la aparición de la diabetes.

La plataforma desarrollada por investigadores de la Facultad de Ciencias de la Computación, con apoyo del Instituto Nacional de Salud Pública (INSP) y el Conacyt, consiste en una prueba metodológica de diagnóstico alternativo**,** para conocer el nivel riesgo de padecer diabetes.

El proyecto consiste en hacer una encuesta, la cual pide información que ayudaría a calcular el riesgo de padecer la enfermedad. La encuesta consta de 8 sencillas preguntas con las cuales se podrá se podrá diagnosticar a cualquier persona si es propensa de padecer Diabetes.

Las preguntas cubren temas sobre los factores riesgos como lo son:

- Edad
- Sobrepeso y obesidad
- Circunferencia de la cintura
- Inactividad física.

Las preguntas están basadas en el proyecto FindRisk (FINnish Diabetes Risk Score), este proyecto maneja puntajes para cada pregunta, que han sido probados, en varios países ya, tales como: Finlandia y España. Y ahora con el laboratorio MexRisc, poner a prueba a la población Mexicana.

2.2.2 PROYECTO EXTRACCIÓN DE REGLAS USANDO PROGRAMACIÓN GENÉTICA COMO BASE PARA UN SISTEMA DE SOPORTE A LA TOMA DE DECISIONES CLÍNICAS.

La toma de decisiones se da en todo tipo de procesos. En la práctica clínica es especialmente complejo tomar una decisión. Los sistemas computacionales que apoyan a la toma de decisiones en la salud (HDSS por sus siglas en inglés) han ido evolucionando desde sistemas que dan soporte a los diferentes procesos que utiliza un hospital hasta los que apoyan la toma de decisiones clínicas del médico con conocimiento nuevo [15].

El propósito de este proyecto, es el descubrimiento de conocimiento nuevo sobre un paciente denominado síndrome metabólico, aplicando un enfoque socio-técnico que guía todo el proceso de descubrimiento de conocimiento. El producto de este análisis se debe obtener una clasificación de riesgo del síndrome metabólico que permita al médico, ubicar el tratamiento inclusive en etapas tempranas del padecimiento [15].

2.3 COMPARATIVA ENTRE LOS PROYECTOS PRESENTADOS EN LAS DOS SECCIONES ANTERIORES.

Entre los dos proyectos anteriores hay una gran diferencia pero también hay gran similitud, una similitud es que ninguno de los dos pretende sustituir una prueba médica.

Una diferencia entre ellos que es bastante notoria, es el hecho que MexRisc no usa una técnica computacional, solo se basa en una serie de puntajes que están establecidos en un proyecto que se titula FindRisk.

Los puntajes manejados en el proyecto, ya fueron utilizados, para predecir diabetes en población de diferentes países.

Por otro lado el proyecto nombrado en la sección 2.1.2, maneja técnicas computacionales propias de la minería de datos, no está enfocado en su totalidad a la diabetes, pero es una de las cuantas enfermedades de las cuales analizaron.

2.4 TRABAJOS RECIENTES DE DIABETES EN MÉXICO.

Debido a que esta enfermedad, tiene un gran avance entre la población, se han tomado diversas medidas, y se han llevado a cabo diversas actividades, tanto preventivas como de tratamiento para la misma.

A pesar de las iniciativas por parte del gobierno, por reducir la obesidad, podemos notar que esta está fuera de control debido a que la Organización Mundial de la Salud (OMS), indica que México ocupa el primer sitio en obesidad infantil y segundo en adulta. Esto es preocupante ya que un gran número de personas con sobrepeso tienden a desarrollar diabetes.

Recientemente, el Instituto de Fisiología Celular (IFC) y el Programa Universitario de Investigación en Salud (PUIS) de la Universidad Nacional Autónoma de México (UNAM) presentaron el libro Advances in Obesity–Diabetes Research at UNAM, en el cual se indagan los orígenes, desarrollo y nuevos abordajes para el tratamiento temprano de ambos padecimientos [10].

2.5 DIABETES Y LA TOMA DE DECISIONES.

Los sistemas de toma de decisiones en el área médica, son gran ayuda, ya que ahora nos permiten realizar un estudio rápido, en los pacientes y así poder hacer una predicción de ciertas enfermedades.

El IMSS ha puesto desde hace algunos años, la encuesta de diabéticos que se realiza anualmente, esto para llevar el control de los pacientes que tienen esta

enfermedad, y así poder llevar un plan para prevención y tratamiento de dicha enfermedad. El resultado de esta encuesta resulta en una gran cantidad de información que se almacena en bases de datos para llevar el control del número de diabéticos y realizar estadísticas. [11]

Es por eso que la toma de decisiones tiene un papel importante en este caso, ya que se puede llevar a cabo el tratamiento de ese gran volumen de datos, y así poder tomar decisiones adecuadas respecto a los pacientes. La minería de datos aprenderá y descubrirá patrones en la información pasada y presente y podrá predecir el futuro.

2.6 CONCLUSIÓN

La diabetes es una enfermedad, que cada vez se va acrecentando, es por ello, que muchas instituciones, están tomando medidas preventivas, con el fin de reducir el número de nuevos casos.

Hoy en día, podemos encontrar mucha información sobre ella, es tan grave que ya se ha proclamado el Día Internacional de la Diabetes, marcado el día 14 de noviembre. En esta fecha, se dan a conocer estadísticas, y programas de apoyo para llevar el control de ella.

En el año 2011 en el día internacional de la diabetes, la BUAP y la SSP dieron a conocer el observatorio MexRisc, del cual hablamos en la sección 2.1.1.

CAPÍTULO 3

MARCO TEÓRICO

3.1 MINERÍA DE DATOS

El instituto SAS describe el concepto de *Data Mining* como el proceso de Seleccionar (*Selecting*), Explorar (*Exploring*), Modificar (*Modifyning*), Modelizar (*Modeling*) y Valorar (*Assessment*) grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores [7].

Al momento de pensar en minería de datos, surgen diversas cuestiones, como cuales son las diferencias que realizan otras disciplinas con el proceso de *Data Mining*. Unas pequeñas diferencias es que en los sistemas tradicionales existen hipótesis o modelos previos, una vez que se tenga formulada la hipótesis, se analiza de manera empírica de acuerdo a la información que se tiene en la base de datos, y las respuestas obtenidas se interpretan como respuestas de la hipótesis.

En *Witten & Frank* 2000 [1] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Como su nombre lo indica, la minería de datos tiene relación con la extracción o descubrimiento de nueva información en términos de patrones o reglas a partir de grandes cantidades de datos [6].

3.2 RELACIÓN DE LA MINERÍA DE DATOS CON OTRAS DISCIPLINAS.

La minería de datos es un campo multidisciplinario, que integra elementos de las siguientes áreas del conocimiento, como se muestra e la figura 3.1:

1.- Sistemas de información / bases de datos: Tecnologías de bases de datos y bodegas de datos, maneras eficientes de almacenar, accesar y manipular datos.

2.- Estadística, aprendizaje automático / IA (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos.

3.- Reconocimiento de patrones: Desarrollo de herramientas de clasificación.

4.- Visualización de datos: interfaz entre humanos y datos, y entre humanos y patrones.

5.- Computación paralela / distribuida: cómputo de alto desempeño, mejora de desempeño de algoritmos debido a su complejidad y a la cantidad de datos.

6.- Interfaces de lenguaje natural a bases de datos.

La minería de datos se puede aplicar a una gran variedad de contextos de toma de decisiones de negocios.

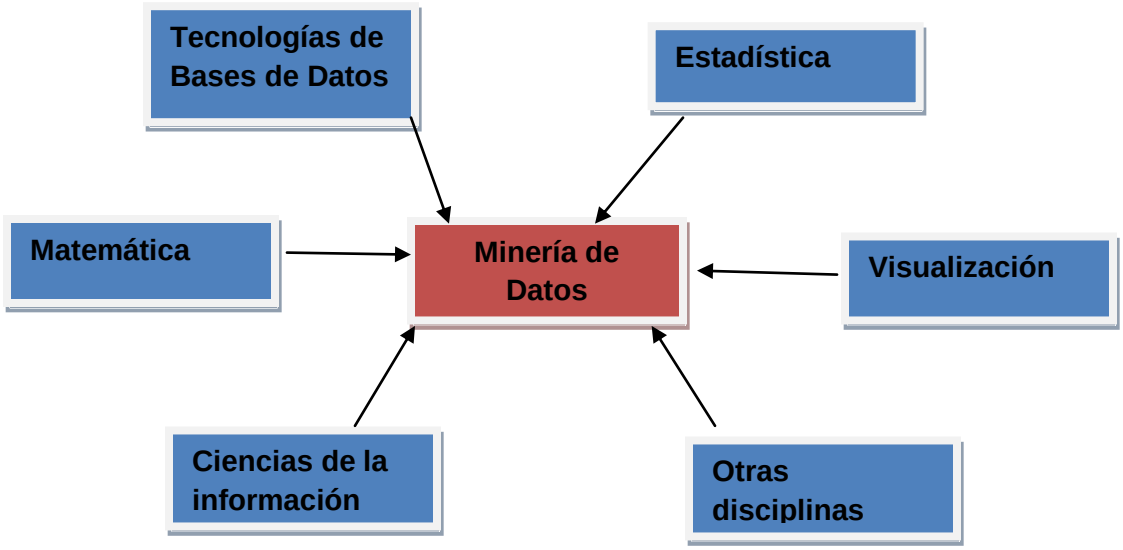

Figura 3.1 Minería de Datos relacionada con Otras disciplinas

3.3 METODOLOGÍA KDD

Este es un campo de rápido crecimiento ya que la demanda creciente hacia las herramientas, que ayuden a comprender información no clara en grandes volúmenes de datos. Estos datos diariamente se van generando por diversas instituciones como bancos,

tiendas minoritarias, empresas de seguros y en la WWW. Esto se debe al gran auge en el uso de las computadoras, códigos de barras, cámaras digitales entre otras cosas.

Actualmente hay una gran cantidad de información guardada en Bases de datos, hojas de cálculo, entre otros repositorios de datos, que están disponibles pero no son muy entendibles.

En este capítulo se explica el modelo KDD, mostraremos los seis pasos que conlleva este modelo, para permitir el fortalecimiento del conocimiento y entablar una comunicación entre diversas herramientas, como son Minería de datos, bases de datos y algunos otros repositorios de información.

3.3.1 EL PROCESO KDD Y EL PROCESO DE MINERÍA DE DATOS.

El descubrimiento de conocimiento en bases de datos, abreviado normalmente como KDD, es por lo regular un tema más amplio que la minería de datos, el proceso de descubrimiento del conocimiento comprende seis fases de los cuales hablaremos en este tema.

El objetivo de seguir y diseñar un proceso KDD es, el seguir una serie de pasos que nos faciliten el procesado de datos. Este modelo nos lleva a una planeación y con ello a una reducción de costos al detallar los procesos que se realizan en cada paso. Los seis pasos del proceso se describen a continuación:

1. **Comprendiendo el dominio del problema.** En este paso se trabaja en conjunto con los expertos del área para definir el problema y determinar los objetivos del proyecto., se aprende sobre temas actuales que se relacionen al problema y se aprenden sobre soluciones.

2. **Compresión de los datos.** En este paso, se hace la recolección de los datos, y es momento de tomar la decisión de que datos nos son útiles, incluyendo su formato y tamaños. Si el conocimiento existe ciertos datos los podemos clasificar como más importantes. Posteriormente se debe verificar la utilidad de los datos de acuerdo a los objetivos del KDD. Los datos deben ser verificados en cuanto a

integridad, redundancia, falta de valores, la plausibilidad del valor de los atributos y cuestiones similares.

3. **Preparación de los datos.** Este es el punto crucial para que el proyecto funcione, ya que el descubrimiento del conocimiento depende de este proceso, por lo general este paso conlleva la mitad del esfuerzo de todo el proyecto.

La MD es la parte central del proceso KDD, en la cual se buscan o encuentran patrones de interés para el usuario. Los patrones descubiertos pueden ser subgrafos, reglas de asociación, árboles de clasificación, una red neuronal entrenada, entre otros [13] .

1. **Evaluación del conocimiento descubierto.** Aquí es donde los expertos analizan los resultados, para verificar que verdaderamente sean resultados novedosos e interesantes. En todo proceso KDD se pueden volver a realizar análisis de los datos, para así poder identificar alternativas y poder mejorar los resultados.

2. **Uso del conocimiento descubierto.** Es aquí donde intervienen los dueños de las bases de datos, ya que le corresponde a ellos, llevar a cabo la aplicación del conocimiento descubierto, para ello se requiere de una planeación de donde y como se debe aplicar. El área de aplicación debe ampliarse a otros dominios dentro de la organización, también se debe crear un plan para vigilar dicho conocimiento y cabe mencionar que todo el proyecto debe ser documentado.

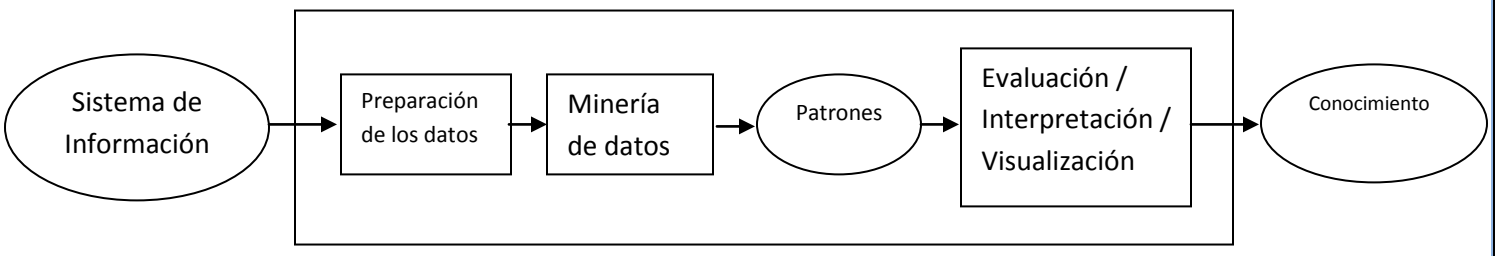

Figura 3.2 Proceso de extracción de conocimiento [1].

Una vez obtenida la definición anterior del proceso KDD (figura 3.2) podemos notar, la estrecha relación que existe entre este y la minería de datos. La MD se refiere a la aplicación de métodos de aprendizajes y estadísticos para la obtención de patrones y modelos, mientras que el KDD es el proceso para la extracción de conocimiento desde las bases de datos.

3.3.2 METODOLOGÍAS ESPECÍFICAS.

El empleo de una metodología bien estructurada y organizada presenta las siguientes ventajas:

- Facilita la planificación y dirección del proyecto.
- Permite realizar un mejor seguimiento del proyecto.
- Facilita el desarrollo de nuevos proyectos de Minería de Datos con características similares.

Entre las metodologías que podemos encontrar actualmente, tenemos las siguientes: CRISP-DM (*Cross Industry Standard Process for Data Mining*), SEMMA(*Sample, Explore, Modify, Model, Asses)*, Metodología de las cinco A's (*Asses, Acces, Analyze, Act, Automate*), Modelo de proceso de Minería de Datos de *Two Crows*, CRITIKAL(*Client-Server Rule Induction Technology for Industrial Knowledge Acquisition from Large Databases*), y Metodología SQL Server-2005, entre otras [14].

Un análisis de las principales características de cada una de estas metodologías permitió concluir que todas comparten la misma esencia: Estructura el KDD en fases similares, que se encuentran interrelacionadas entre sí; y lo describen de forma iterativa e interactiva. Se pudo comprobar que algunas (como SEMMA) se centran más en los aspectos técnicos del proceso; mientras que otras (como CRISP-DM) ven el proyecto de forma global, tomando en cuenta los aspectos relacionados con el campo de aplicación. También se pudo observar la estrecha relación de algunas con las herramientas de desarrollo (SEMMA, "SQL-Server 2005", CRITIKAL) [14].

3.4 TAREAS Y TÉCNICAS DE MINERÍA DE DATOS.

Dentro de las tareas de Minería de Datos existen tipos, cada una de las cuales se pueden considerar como un ejemplo de problema a ser resuelto por un algoritmo de

Minería de Datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra [2].

Como ya lo hemos mencionado anteriormente, las tareas de minería de datos pueden ser predictivas o descriptivas. Entre las tareas predictivas encontramos la clasificación y la regresión, mientras que el *clustering*, las reglas de asociación secuenciales y las correlaciónales son tareas descriptivas.

3.4.1 TAREAS PREDICTIVAS.

La minería de datos puede mostrar cómo se comportaran ciertos atributos en el futuro de cierto conjunto de datos. Entre los ejemplos de predicciones de datos podemos incluir el análisis de transacciones de compra para predecir que compraran los consumidores ante determinados descuentos, que volumen de ventas se generará en una tienda en un periodo determinado y si la eliminación de una línea de productos generará más beneficios. En aplicaciones de ese tipo, la lógica de negocio se utiliza unida a la minería de datos. En un contexto científico, determinados patrones de ondas sísmicas podrían predecir un terremoto con una probabilidad alta [6].

No obstante una vez más podemos analizar estos problemas con más detenimiento para observar que la variable a predecir puede ser una variable categórica (si compra o no un producto) [7]. Esta distinción nos lleva a que podemos ver este tipo de problemas en dos que son:

- Problemas de clasificación: hacen referencia a los problemas en los que la variable a predecir tiene un número finito de valores, esto es la variable es categórica [7].
- Problemas de predicción de valores: se refieren a los problemas en los que la variable a predecir es numérica [7].

Hacer estas distinciones es importante, ya que de acuerdo a la problemática, se debe buscar la técnica adecuada que se utilizará para solucionar el problema.

3.4.2 TAREAS DESCRIPTIVAS.

La descripción es normalmente usada para realizar un análisis preliminar de los datos. Busca derivar descripciones concisas de características de los datos: medias, desviaciones estándares, etc.

La meta principal de todas estas tareas es una descripción del conjunto de datos origen. Pero las tareas descriptivas las podemos dividir en dos más que son:

- Análisis de segmentación: que se refiere a los problemas donde la meta es encontrar grupos homogéneos en la población de objetos origen. A estos problemas también se les denomina problemas de aprendizaje no supervisado o *clustering* [7].
- Análisis de asociaciones: hace referencia a los problemas en los que se persigue obtener relaciones entre los valores de atributos de una base de datos [7].

3.4.3 EJEMPLOS DE TÉCNICAS Y TAREAS QUE REALIZAN.

La tabla 3.1 muestra algunas técnicas de minería de datos y algoritmos asociados a las tareas que realizan.

NOMBRE	PREDICTIVAS		DESCRIPTIVAS		
	CLASIFICACION	REGRESION	AGRUPAMIENTO	REGLA DE ASOCIACION	CORRELACIONES / FACTORIZACIONES
Redes neuronales	X	X	X		
Arboles de decisión ID3, C5.0	X				
K-means			X		
Naive Bayes	X				
Vecino más próximo	X	X	X		
Algoritmos genéticos y evolutivos	X	X	X	X	X

Tabla 3.1 Algunas técnicas de minería de datos

3.5 TÉCNICAS DE MINERÍA DE DATOS UTILIZADAS.

En este punto, se darán a conocer las técnicas de minería de datos, que serán aplicadas, al conjunto de datos, para su análisis, dando una explicación de cada una y en qué consisten.

Las técnicas son implementaciones específicas de los algoritmos que se utilizan para llevar a cabo las operaciones de construcción del modelo [7]. Cabe mencionar que no todos los algoritmos de *data mining* son iguales y cada uno de ellos tendrá una serie de ventajas y desventajas.

Hay que tener muy en cuenta que los algoritmos utilizados en *data mining*, son provenientes de distintas aéreas de investigación, como son: la estadística o la inteligencia artificial.

A continuación se enlistan las técnicas de minería de datos que se pueden aplicar.

- Modelos predictivos: Clasificación: En este tipo de modelos se utiliza aprendizaje supervisado. Se suelen utilizar arboles de decisión, regresiones logísticas y redes neuronales [7].
- Modelo predictivos: predicción de valores: para la predicción de valores se utilizan, junto a los métodos anteriores, la regresión lineal y la no lineal [7].
- Segmentación de bases de datos: *clustering* no jerárquico: se trata de comparar cada registro de la base d datos con todos los segmentos (semillas) creadas por la función. Se mide la distancia del registro de entrada con los segmentos ya creados y se asigna el registro de entrada al segmento correspondiente. Cabe mencionar que el número de clusters se ajusta automáticamente. Este método se conoce como el método de las K-medias.
- Segmentación de las bases de datos: *clustering* jerárquico: este tipo de técnica de minería de datos es apropiado cuando se desconoce y no se tiene información sobre los grupos en los que se clasifican los *clusters*. Suelen usarse algoritmos de tipo jerárquico como los aglomerativos o divisivos. Junto con ellos se utilizan redes neuronales basadas en aprendizaje no supervisado, como por ejemplo los mapas de Kohonen [7].
- Análisis de relaciones: asociaciones: El objetivo de esta técnica de minería de datos es encontrar elementos que implican la presencia de otros elementos dentro de una misma transacción. El resultado de esta técnica son reglas de tipo "*if X then Y*". Lo que hace es contar las ocurrencias de todos los elementos presentes en las transacciones de la base de datos [7].

- Análisis de relaciones: patrones secuenciales: trata de descubrir patrones entre transacciones en las que un conjunto de elementos va seguido de otro conjunto de elementos distanciados por un periodo de tiempo determinado [7].
- Análisis de relaciones: patrones en series temporales: con esta técnica se pretenden descubrir ocurrencias o secuencias similares a una dada en una base de datos que almacene información que represente una serie temporal, como puede ser la evolución de los precios de mercado o datos de telemetría provenientes de algún sensor.

3.5.1 K-MEANS.

El algoritmo K-means, creado por MacQueen en 1967 es el algoritmo de clustering más conocido y utilizado ya que es de muy simple aplicación y eficaz. Sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado número K de clusteres, K determinado a priori.

El nombre de K-means viene porque representa cada uno de los clusters por la media (o media ponderada) de sus puntos, es decir, por su centroide. La representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato ver figura 3.3.

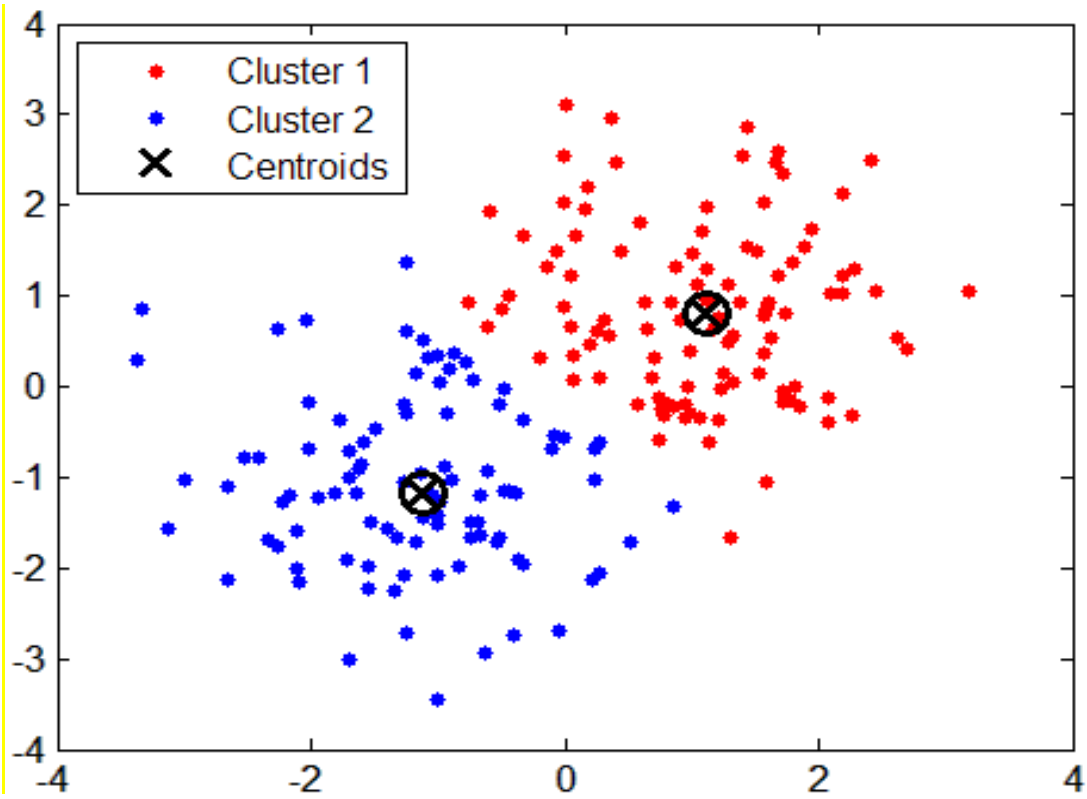

Figura 3.3 Clústeres caracterizados por su centroide.

El algoritmo *K-Means* trata de un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar, sin etiquetar. Es el más popular entre los métodos de agrupamiento denominados "por partición". La idea del *K-Means* es situar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares [12].
Todo ejemplo nuevo, una vez que los prototipos han sido correctamente situados, es comparado con éstos y asociado a aquél que sea el más próximo, en los términos de una distancia previamente elegida. Normalmente se usa la distancia Euclidiana [2].

3.5.2 ARBOLES DE DECISIÓN.

De todos los métodos de aprendizaje, los sistemas basados en árboles de decisión son quizás los más fáciles de utilizar y de entender. Un árbol de decisión es un conjunto de de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar e puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Los árboles de decisión se utilizan desde hace siglos, y son especialmente apropiados para expresar procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, etc. Sus aplicaciones básicamente son la clasificación y la predicción.

Un árbol de decisión lleva a cabo un para alcanzar una decisión. El árbol de decisión suele contener nodos internos, nodos de probabilidad, nodos hojas y arcos (ver figura 3.4). Un nodo interno contiene un las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, este tipo de nodos es redondo, los demás son cuadrados. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada. Los arboles de decisión poseen:

- Ramas: se representan con líneas.
- Nodos de decisión: de ellos salen las ramas de decisión y se representan por un cuadrado.
- Nodos de incertidumbre: de ellos salen las ramas de los eventos y se representan con un círculo.

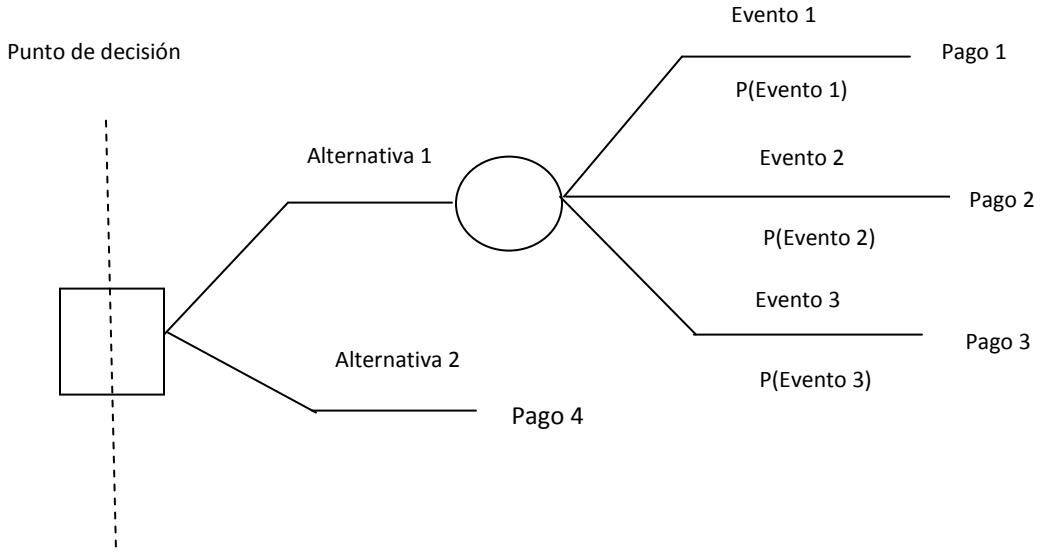

Figura 3.4 Componentes y estructura de árbol de decisión

3.6 ENTORNO DE MINERÍA DE DATOS EN WEKA

Existen diversas herramientas que nos permiten la implementación de MD, como por ejemplo SPSS Clementine la cual es una herramienta integrada de minería de datos, inicialmente de Integral Solutions Limited (ISL) y ahora de SPSS. Clementine incluye distintas herramientas de minería de datos: correlación, reglas de asociación (GRI, a priori), patrones secuenciales (regresión), segmentación (Kohonen, Two-step y k-means), clasificación (redes neuronales, reglas y árboles de decisión). Otra de los entornos de MD es el RapidMiner este es un sistema prototipado para el descubrimiento del conocimiento y MD. Este es un software de tipo Open-Source con licencia GNU GPL, basado en Java, usa el lenguaje de scripting XML para describir los operadores y su configuración. Nosotros para este proyecto nos enfocaremos en el uso de WEKA.

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 [2]:

- Acceso a los datos: Los datos son cargados desde un archivo 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, medias y varianzas de los atributos numéricos, distribución de frecuencias en los atributos nominales, etc.)
- Pre procesado de datos (destacar la gran cantidad de filtros disponibles):
    - Selección de atributos.
    - Desratización.
    - Tratamiento de valores desconocidos.
    - Transformación de atributos numéricos
- Modelos de aprendizaje:
    - Árboles de decisión (J4.8, versión propia del método C4.5).
    - Tablas de decisión.
    - Vecinos más próximos
    - Máquinas de vectores de soporte (método *sequential minimal optimization*).
    - Reglas de asociación (método Apriori).
    - Métodos de agrupamiento (K-medias, EM y Cobweb).
    - Modelos combinados (*bagging, boosting, stacking,* etc).
- Visualización en la figura 3.5 podemos observar la interfaz de inicio de WEKA. (la interfaz gráfica se compone de diversos entornos):

Figura 3.5 Entorno GUI WEKA

- El entorno *Explorer* permite controlar todas las operaciones anteriores (filtrado, selección y especificación del modelo, diseño de experimentos, etc.).
- El entorno consola (CLI) posibilita la invocación textual de las operaciones anteriores. (También es posible acceder directamente a los métodos que implementan dichas tareas e incorporarlos en el código fuente de la aplicación de Minería de Datos que se esté programando.)
- El entorno *Experimenter* facilita el diseño y la realización de experimentos complejos
- El proceso global de Minería de datos en WEKA se acelera considerablemente gracias al entorno *KnowledgeFlow* que, de una forma gráfica y a modo de flujos de operaciones, permite definir la totalidad del proceso (carga de datos, preproceso, obtención de modelos, comprobación y visualización de resultados

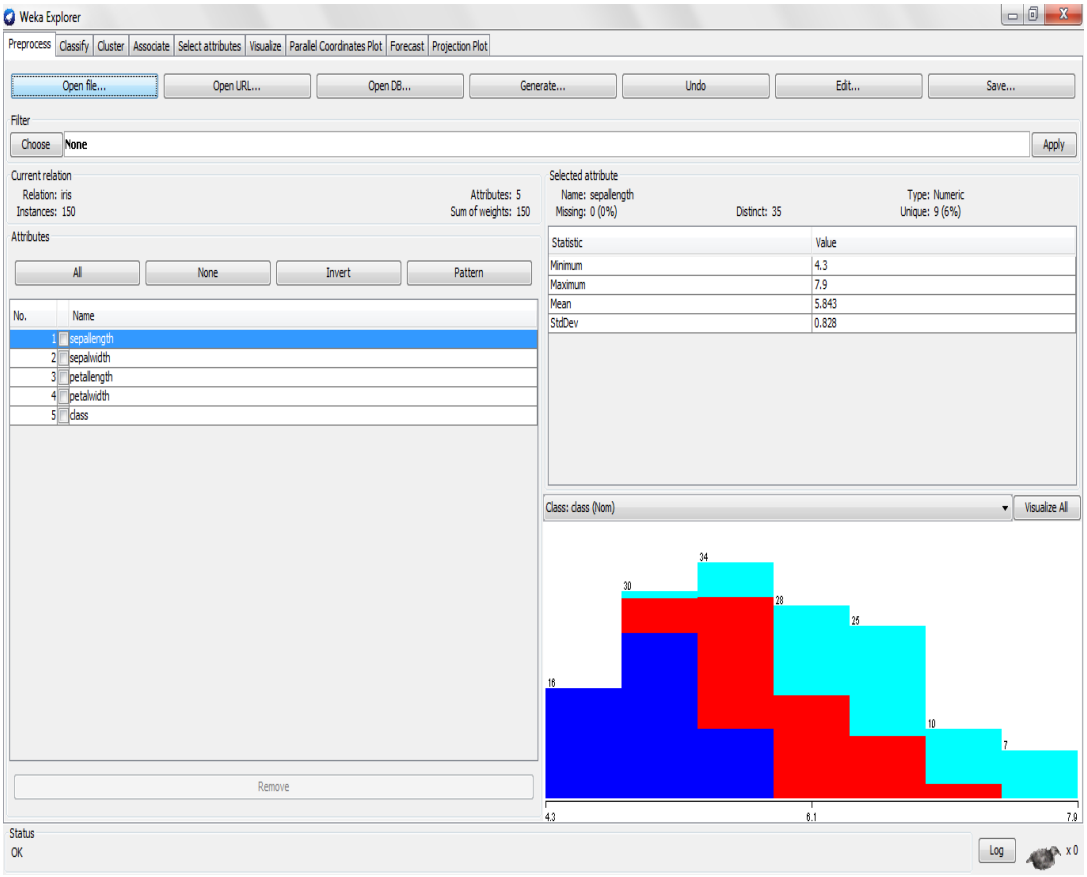

Figura 3.6 Detalle del entorno EXPLORER de WEKA

En la figura 3.6 se puede apreciar parte del entorno de trabajo de la herramienta WEKA. En esa sección podemos seleccionar los datos queremos utilizar para la extracción de conocimiento. También nos permite aplicar filtros, tales como discretización, entre otros procesos para la generación de la vista minable para la posterior aplicación de los clasificadores correspondientes.

3.7 CONCLUSIONES

El marco teórico utilizado en este trabajo de tesis amplía nuestros conocimientos científicos acerca del proceso de minería de datos, para apoyarnos en el planteamiento y solución del problema descrito en la sección 1.1.1. Así también se llevó a cabo la integración de la teoría existente sobre los conocimientos científicos con investigaciones y sus relaciones mutuas.

CAPÍTULO 4

ANÁLISIS Y DISEÑO

El objetivo del presente capitulo es mostrar el análisis y el diseño de las vistas minables a los datos, para el SISTEMA DE APOYO A LA TOMA DE DECISIONES UTILIZANDO TÉCNICAS DE CLASIFICACIÓN PARA EL CASO DE DIABETES EN JUCHITÁN, OAXACA, posteriormente se describe el procedimiento correspondiente a la aplicación de Técnicas de Minería de Datos, en concordancia con lo teóricamente expuesto en el capítulo 3.

4.1 PLANTEAMIENTO DE REQUERIMIENTOS

Tomando en cuenta los problemas identificados en la sección 1.1.1 y los objetivos de la investigación planteados en la sección 1.1.2, los requerimientos que se necesitan siguen la metodología del proceso de KDD expuesto en la sección 3.1.3.1**,** como a continuación se muestra:

1. Preparación de los datos
   - Efectuar un proceso de selección, limpieza y transformación de la información, y así eliminar y corregir los datos incorrectos, con esto se decidirá la estrategia a seguir con los datos incompletos. En esta parte se proyectarán los datos y se consideraran únicamente aquellas variables o atributos que son relevantes, con el objetivo de facilitar la tarea propia de la Minería y para que los resultados de la misma sean útiles.

2. Aplicación de técnicas de minería de datos
   - En esta fase se decidirá cuál es la tarea a realizar (clasificación, agrupar, etc.) y se elegirá el método a utilizar.

3. Fase de evaluación e interpretación
   - Se evaluarán los patrones y se analizarán por los expertos, y si es necesario se realizarán las fases anteriores para una nueva iteración.

4. Fase de difusión y uso
   - Se hará uso del nuevo conocimiento y se hará partícipe de él a todos los posibles usuarios.

A manera de ilustrar los requerimientos anteriores éstos se muestran en la figura 4.1:

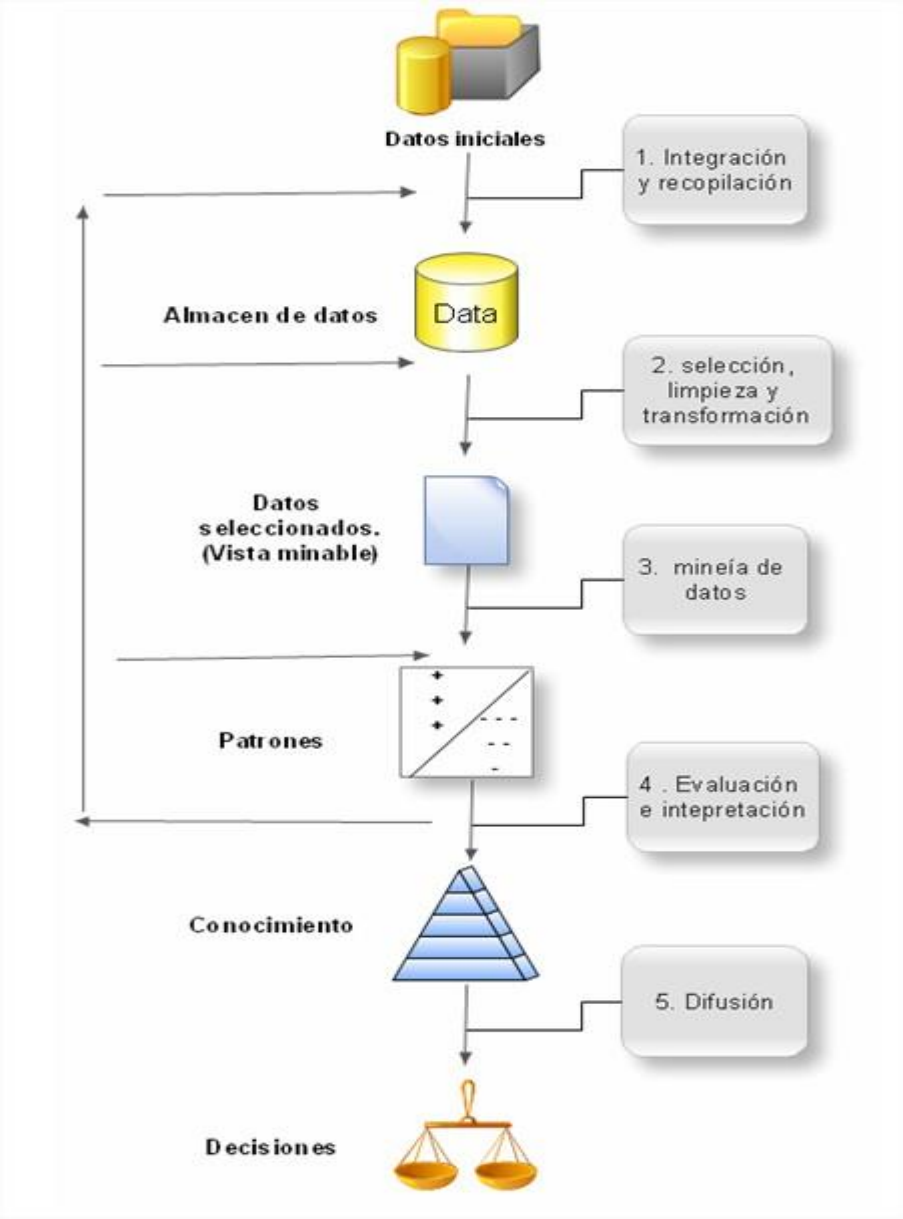

Figura 4.1 Fases del proceso de descubrimiento del conocimiento en Bases de Datos, KDD

4.2 FASE DE INTEGRACIÓN Y RECOPILACIÓN

En esta fase se llevó a cabo la recopilación de los datos, obtenidos por el Instituto Mexicano del Seguro Social (IMSS) y el Instituto Nacional de Estadística Geográfica e Informática (INEGI).

Los datos obtenidos estaban almacenados en una hoja de cálculo bajo el nombre encuesta diabéticos 2012, cuyas columnas eran: Peso, Talla, Sexo, Edad, Población, Ult. Glucosa (última medida de glucosa tomada al paciente), Altura entre otras. A continuación en la tabla 4.1 y 4.2 se muestran algunos ejemplos de los datos obtenidos, así como la nomenclatura utilizada.

Población	Nomenclatura
Juchitán	Juo
Espinal	Esp
La Ventosa	Lve
Unión Hidalgo	Uh
La Mata	Lma
Santa María Xadani	Smx
Monte Grande	Mgd

Tabla 4.1 Nomenclatura de las poblaciones utilizadas en las vistas minables

Últ. Peso	Últ. Glucosa	ALTURA	EDAD	POBLACION	SEXO
122	150	1.74	77	C	M
93	191	1.7	71	D	M
73.5	358	1.59	42	F	F
90	292	1.53	28	A	F
87	183	1.82	57	G	F
76	285	1.76	32	B	F
98.5	309	1.72	43	G	M

Tabla 4.2 Ejemplo de datos obtenidos

4.3 FASE DE SELECCIÓN LIMPIEZA Y TRANSFORMACIÓN

La calidad del conocimiento obtenido depende en gran parte de la calidad de los datos minados, es por ello, que después de haber seleccionado los datos, el siguiente paso en la metodología KDD es preparar el subconjunto de datos que se va a minar, los cuales van a constituir lo que se conoce como vista minable.

En este proceso se eliminaron 10 registros los cuales no nos iban a ser de utilidad para el proceso ya que tenían información perdida o faltante. También se eliminaron valores *outliers* (valores que no se ajustaron al comportamiento general de los datos). Posteriormente se calcularon nuevos atributos que son más relevantes (índice de masa corporal), se rellenaron algunos valores faltantes (*missing values*), se realizó una discretización.

Todas estas actividades descritas anteriormente se realizan con la finalidad de mejorar la eficiencia de la herramienta de minería de datos, mejorar la calidad (precisión) del conocimiento obtenido.

Para la realización de esta tarea, es posible hacerla de diversas maneras entre ellas son hacerlas a mano, utilizar herramientas de proceso.

El procedimiento especificado en la sección 4.3 se realizó utilizando una visualización previa del histograma conseguido con la herramienta *WEKA* (ver figura 4.2).

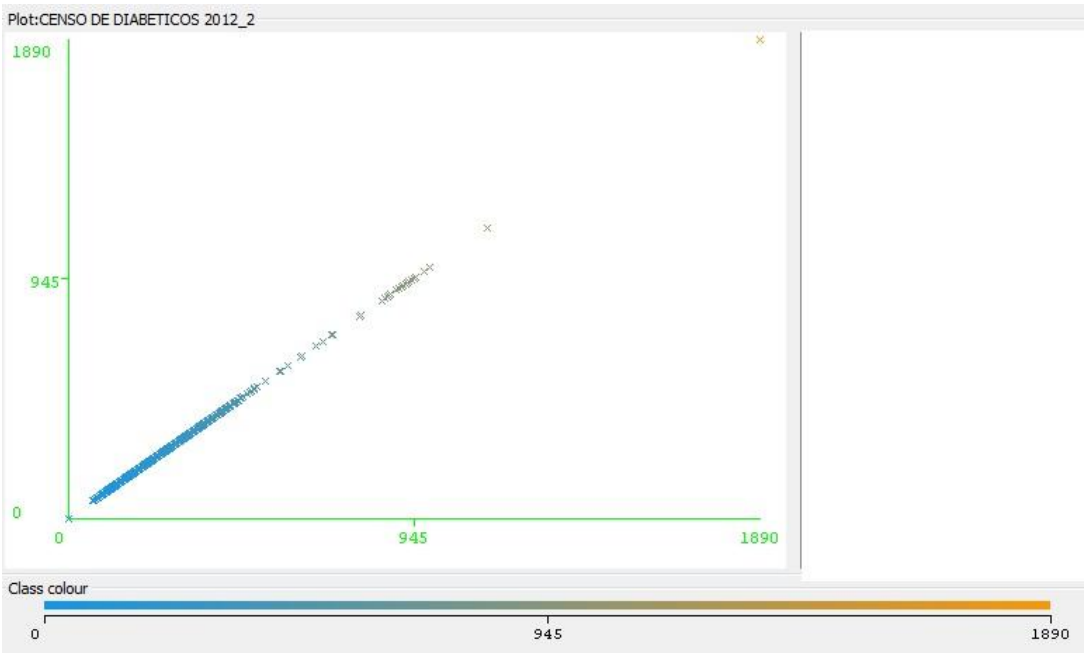

Figura. 4.2: Visualización de datos *outliers* con Weka.

En la figura 4.2 se muestra el histograma del parámetro última Glucosa, donde podemos visualizar datos con valor 0 y llegando a un máximo valor de 1890, este y otros datos como el número de seguro social, el consultorio médico al cual son atendidos, el turno en el cual asisten a consulta fueron eliminados del conjunto ya que no son datos relevantes o son datos que nos arrojan ruido al momento de realizar las pruebas.

Una vez seleccionados los atributos, se procedió a realizar una preparación de los datos, es decir, la construcción automática de nuevos atributos, con objeto de que estos nuevos atributos hagan más fácil el proceso de Minería.

Entre la preparación que se les dio a los datos fue la eliminación de los correspondientes a la diabetes tipo 1 (DID), ya que solo contábamos con 6 datos de un total de 1732 con que cuenta la base de datos original, el cual solo equivale al 0.34% del total de los datos. La eliminación de estos 6 registros nos ayudaría a centrar el estudio en la diabetes tipo II.

Con estos datos que hemos eliminado, hacemos un estimado con respecto a las poblaciones a las cuales los pacientes correspondían dando como resultados los datos presentados en la tabla 4.3:

Población	Sexo	Cantidad en la BD	Población Real	% Diabéticos tipo 1	% Diabéticos tipo respecto al Sexo
El Espinal	F	1	8310	0.01	4278(0.02)
El espinal	M	1	8310	0.01	4031(0.02)
Juchitán	M	1	93038	0.001	47828(0.002)
Juchitán	F	1	93038	0.001	45210(0.002)
Sta. Ma. Xadani	M	1	7781	0.01	3929(0.02)
Union Hidalgo	F	1	13970	0.007	7221(0.01)

Tabla 4.3: Diabéticos tipo 1 eliminados

Otro de los procesos que se realizó sobre los datos fue la discretización. Por ejemplo el atributo de *peso* que fue discretizado (se puede tomar como atributos categóricos) con un número más pequeño de valores.

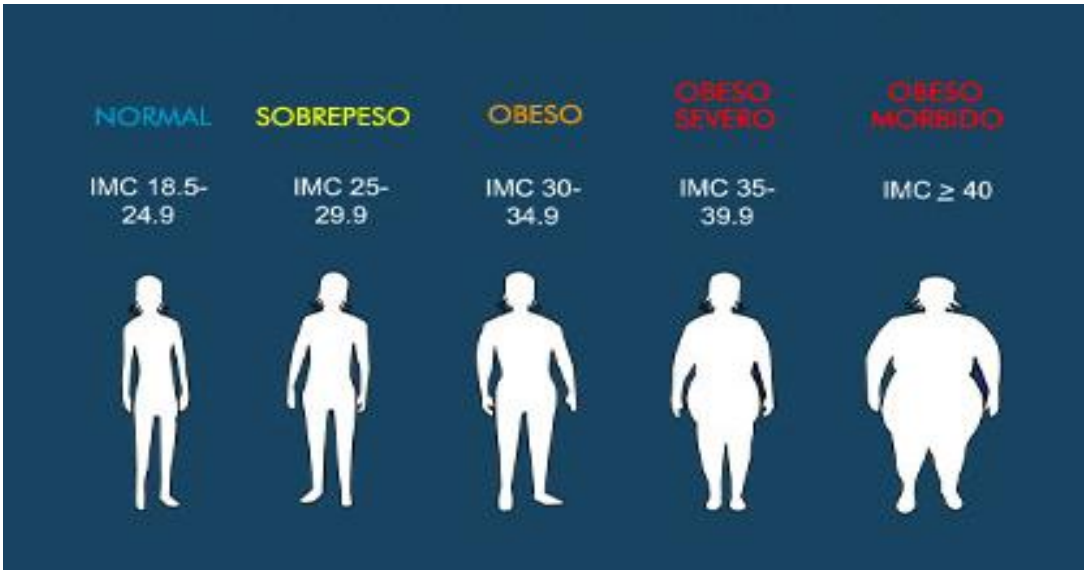

Fig. 4.3 Clasificación de peso de acuerdo al IMC

En la figura 4.3 se muestra la clasificación de acuerdo a grados de obesidad la cual fue base para poder realizar una discretización de los datos, con esto conseguir tener un rango menor de valores a evaluar, a continuación se muestra la clasificación que la OMS nos otorga sobre la obesidad.

Atendiendo al Consenso SEEDO 2000(Sociedad Española para el Estudio de la Obesidad), a los pacientes se les clasifica en función del porcentaje graso corporal, cuando este está por encima del 25% en los varones y del 33% en las mujeres los podemos catalogar como personas obesas. Los valores comprendidos entre el 21 y el 25% en los varones y entre el 31 y el 33% en las mujeres se consideran límites.

La OMS ha propuesto una clasificación del grado de obesidad utilizando el índice ponderado como criterio:

- Peso Normal: IMC 18,5 - 24,9 Kg/m^2
- Sobrepeso: IMC 25 -29 Kg/m^2:
    - Obesidad grado I con IMC 30-34 Kg/m^2
    - Obesidad grado II con IMC 35-39,9 Obesidad grado I con IMC 30-34 Kg/m^2
    - Obesidad grado III con IMC >= 40 Obesidad grado I con IMC 30-34 Kg/m^2

4.3.1 TRATAMIENTO DE DATOS.

En esta sección daremos a conocer, todo el proceso que se le dio a los datos, hasta obtener la información que se uso para las pruebas.

Entre las modificaciones que se realizaron a los datos, se encuentra la discretización de ciertos atributos, separación de columnas, y creación de nuevos campos, así como también la eliminación de datos que no eran de utilidad.

La figura 4.4 muestran parcialmente los datos, que nos fueron proporcionados por la clínica 6 del IMSS, de Juchitán, OAX., provenientes de una encuesta realizada a su población.

Consultorio	Turno	Diabetes	Dx Diabetes	Fecha Registro Diabetes	Último Dx Diabetes	Últ. Dx	Últ. Peso	Últ. P. Sistolica	Últ. P. Diastolica	Últ. Glucosa	Fecha Alta Censo
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	122	210	100	150	06/10/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E149.-DIABETES	M948	93	180	120	191	05/04/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E143.-DIABETES	Z719	73.5	160	100	358	04/04/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E109.-DIABETES MELLITUS INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/06/2006	E119.-DIABETES	Z719	90	130	86	232	01/06/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E125.-DIABETES	Z713	87	140	80	183	20/01/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E125.-DIABETES MELLITUS ASOCIADA CON DESNUTRICION, CON COMPLICACIONES CIRCULATORIAS PERIFERICAS	06/10/2006	E125.-DIABETES	Z719	76	180	90	285	04/09/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E110.-DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMA	01/01/1900	E110.-DIABETES	R42X	98.5	130	90	309	04/07/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	02/10/2009	E148.-DIABETES	Z719	46	140	80	500	07/07/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E117.-DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMPLICACIONES MULTIPLES	01/01/1900	E117.-DIABETES	N19X	71	180	100	213	07/01/2009
01-MEDICINA FAMILIAR	1-Matutino	Si	E125.-DIABETES MELLITUS ASOCIADA CON DESNUTRICION, CON COMPLICACIONES CIRCULATORIAS PERIFERICAS	01/01/1900	E127.-DIABETES	Z719	82	160	90	363	04/09/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E149.-DIABETES	Z719	111	210	90	230	08/05/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E125.-DIABETES MELLITUS ASOCIADA CON DESNUTRICION, CON COMPLICACIONES CIRCULATORIAS PERIFERICAS	01/01/1900	E125.-DIABETES	Z719	88	140	90	204	04/09/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E130.-OTRAS	Z719	80	140	90	250	01/06/2007
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	83	150	90	698	08/05/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E149.-DIABETES MELLITUS NO ESPECIFICADA, SIN MENCION DE COMPLICACION	05/12/2007	E149.-DIABETES	Z719	74	120	80	400	05/11/2007
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	80	130	80	88	11/06/2009
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E125.-DIABETES	Z719	97	140	100	233	16/03/2004
01-MEDICINA FAMILIAR	1-Matutino	Si	E113.-DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMPLICACIONES OFTALMICAS	01/01/1900	E119.-DIABETES	Z719	84	150	80	173	04/08/2008
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E127.-DIABETES	R509	67	150	80	251	04/04/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E117.-DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMPLICACIONES MULTIPLES	01/01/1900	E119.-DIABETES	Z763	93	130	80	325	04/09/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E149.-DIABETES	Z719	110	170	100	154	04/07/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	09/07/2010	E149.-DIABETES	Z719	63	130	80	395	08/12/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	70	130	80	376	05/01/2007
01-MEDICINA FAMILIAR	1-Matutino	Si	E127.-DIABETES MELLITUS ASOCIADA CON DESNUTRICION, CON COMPLICACIONES MULTIPLES	01/01/1900	E127.-DIABETES	Z719	64	180	90	250	04/07/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	04/07/2006	E119.-DIABETES	Z719	98	160	90	192	04/07/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	93.5	130	80	104	04/09/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E117.-DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMPLICACIONES MULTIPLES	01/01/1900	E148.-DIABETES	Z719	75	150	90	500	09/04/2007
01-MEDICINA FAMILIAR	1-Matutino	Si	E127.-DIABETES MELLITUS ASOCIADA CON DESNUTRICION, CON COMPLICACIONES MULTIPLES	01/01/1900	E127.-DIABETES	Z719	104	140	90	137	08/05/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	89	150	100	121	07/08/2006
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	Z719	85.5	120	90	267	01/06/2007
01-MEDICINA FAMILIAR	1-Matutino	Si	E117.-DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMPLICACIONES MULTIPLES	12/08/2011	E117.-DIABETES	Z719	95	140	80	216	12/08/2011
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	03/11/2009	E119.-DIABETES	Z719	75.5	120	70	277	03/11/2009
01-MEDICINA FAMILIAR	1-Matutino	Si	E119.-DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	01/01/1900	E119.-DIABETES	I64X	91	180	90	150	03/11/2010

Figura 4.4 Datos Originales obtenidos en la encuesta de diabetes.

Como se puede observar en la columna Dx.Diabetes, se presentan 3 datos, los cuales son de gran importancia para la investigación por el hecho de contener en un solo campo, el código médico, que se le asigna a las complicaciones médicas que surgen con la diabetes, también tenemos el tipo de diabetes, y en caso de presentar alguna complicación.

En la figura 4.5, se muestran como quedan los datos después de realizar la separación de las columnas y la discretización de datos, para poder obtener los datos finales, con los cuales se realizan las pruebas.

Codigo	Tipo	Complicacion	IMC	POBLACION	SEXO	ALTURA CLASIFIC	EDAD CLASIF	PESO CLASIF	Últ. Glucosa	TALLA EN CM	Últ. P. Sistoli	Últ. P. Diastc
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	40.296	Lvs	M	Media	Anciana	Obesidad1	150	96	210	100
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	32.18	UH	M	Media	Anciana	Obesidad	191	80	180	120
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	29.073	Smx	F	Media	Adulta	Sobre Peso	358	75	160	100
E109	DIABETES MELLITUS INSULIN	SIN MENCION DE COMPLICACIOI	38.447	Juo	F	Media	Adulta	Obesidad	292	115	130	86
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	26.265	MtG	F	Alta	Adulta	Sobre Peso	183	73	140	80
E125	DIABETES MELLITUS ASOCIA	CON COMPLICACIONES CIRCULA	24.535	Esp	F	Media	Adulta	Peso Normal	285	70	180	90
E110	DIABETES MELLITUS NO INSI	CON COMA	33.295	MtG	M	Media	Adulta	Obesidad	309	94	130	90
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	19.396	Juo	F	Media	Adulta	Peso Normal	500	81	140	80
E117	DIABETES MELLITUS NO INSI	CON COMPLICACIONES MULTIPL	31.556	Smx	F	Baja	Adulta	Obesidad	213	103	180	100
E125	DIABETES MELLITUS ASOCIA	CON COMPLICACIONES CIRCULA	24.22	Esp	M	Alta	Anciana	Peso Normal	369	115	160	90
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	38.864	Lvs	M	Media	Adulta	Obesidad	290	117	210	90
E125	DIABETES MELLITUS ASOCIA	CON COMPLICACIONES CIRCULA	27.465	Lvs	M	Media	Adulta	Sobre Peso	204	96	140	90
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	27.359	UH	M	Media	Adulta	Sobre Peso	250	103	140	90
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	30.86	Lvs	F	Media	Adulta	Obesidad	898	109	150	90
E149	DIABETES MELLITUS NO ESP	SIN MENCION DE COMPLICACIOI	25.606	Esp	M	Media	Adulta	Sobre Peso	400	80	120	80
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	26.73	MtG	M	Media	Adulta	Sobre Peso	88	114	130	80
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	33.962	Esp	M	Media	Adulta	Obesidad	293	95	140	100
E113	DIABETES MELLITUS NO INSI	CON COMPLICACIONES OFTALM	27.745	Smx	F	Media	Adulta	Sobre Peso	173	120	150	80
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	22.913	Juo	F	Media	Anciana	Peso Normal	251	95	150	80
E117	DIABETES MELLITUS NO INSI	CON COMPLICACIONES MULTIPL	27.77	Lmt	F	Alta	Adulta	Sobre Peso	325	119	130	80
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	35.511	MtG	M	Media	Adulta	Obesidad	154	120	170	100
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	25.559	MtG	F	Media	Adulta	Sobre Peso	395	116	130	80
E119	DIABETES MELLITUS NO INSI	SIN MENCION DE COMPLICACIOI	30.298	Esp	F	Media	Adulta	Obesidad	376	111	130	80
E127	DIABETES MELLITUS ASOCIA	CON COMPLICACIONES MULTIPL	21.139	Esp	F	Media	Anciana	Peso Normal	250	107	180	90

Fig. 4.5 Datos Finales.

Como se muestra la Figura 4.5, podemos observar algunos de los datos con los cuales se realizan las tareas de minería de datos. Entre las columnas de mayor interés podemos observar que, a diferencia de la Figura 4.4 en la cual se tenían en una sola columna los datos de clave, tipo de diabetes y la complicación, en la nueva representación estos son atributos independientes, de los cuales se pueden obtener mejores resultados a los que se nos otorgaban con los datos originales. También cómo podemos observar en la Figura 4.5 se muestran algunos de los datos que fueron discretizados, durante todo el proceso de selección, limpieza y transformación.

4.4 FASES DE MINERÍA DE DATOS.

El objetivo de esta tarea, es generar nuevo conocimiento que pueda ser utilizado por el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas.

A continuación se describen las decisiones tomadas, para minar el conjunto de datos conseguido en el paso anterior:

1. **Elección de tarea de minería de datos.**

   Describiremos brevemente algunas técnicas aplicadas al conjunto de datos:

   - **K-Means:** fue la primera técnica aplicada, (a partir de un número *k* de *clúster*, obtenidos por medio de una previa visualización de los datos, que sugiere un número *k=3 clústers*, (Figura 4.17)). Dado que el conjunto de datos es numérico, la aplicación de *k-Means* encaja perfectamente. La técnica también permite dividir los datos en grupos teniendo en cuenta el criterio de la distancia Euclidiana.

   - **Arboles de decisión J48 y Random Forest. :** Se aplicaron estos dos tipos de árboles, para ver la diferencia, entre ellos de los cuales, se mostró un resultado distinto, por lo cual posteriormente se realizaran unas comparaciones, entre los dos métodos.

   - **Reglas de Asociación:** Estas al igual que *KMEANS* son utilizadas para la descripción de datos, y así obtener las normas, para las cuales se cumplan, con la diabetes tipo 2.

Las técnicas descritas anteriormente, fueron aplicadas, utilizando la herramienta WEKA, la cual es comúnmente utilizada para minar un conjunto de datos.

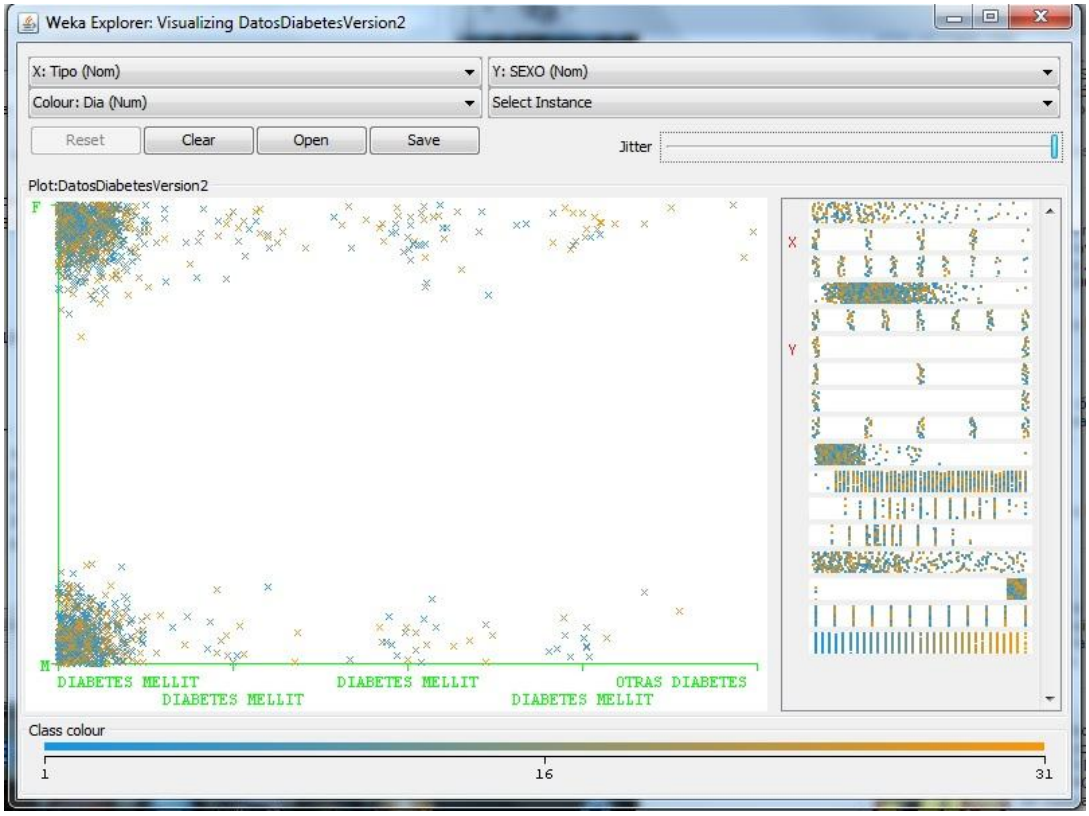

Fig. 4.6 Grupos de tipo de diabetes respecto al Sexo.

En la Fig. 4.6 Se muestra un ejemplo de los grupos que se tienen de diabetes, correspondiente, a su tipo con respecto al sexo, en el lado de las abscisas se tiene la información respecto al sexo, y del lado de las ordenadas se tienen los tipos de diabetes.

4.5.- CONCLUSIONES.

En este capítulo se presentó el diseño, construcción e implantación de las vistas minables de datos para el sistema de Toma de decisiones acerca de la Diabetes, siguiendo los pasos de la metodología KDD. Así como también se describieron las Técnicas de Minería de Datos utilizadas para la explotación de las vistas minables de datos.

CAPITULO 5

RESULTADOS

5.1 INTRODUCCIÓN

En este capítulo se presentan los resultados obtenidos en la aplicación de las técnicas de minería de datos a través de comparaciones para establecer la diferencia entre los datos antes de procesar y después de ser procesados. Dichos resultados se presentan de acuerdo al tipo de tarea de minería aplicada, ya sean estas, descriptivas o predictivas.

En la figura 5.1, se presentan algunas de las tareas más significativas de minería de datos, esto con la finalidad de dejar en claro a qué tipo de tareas cada una de ellas pertenece, además de ofrecer una pequeña explicación del descubrimiento que realizan.

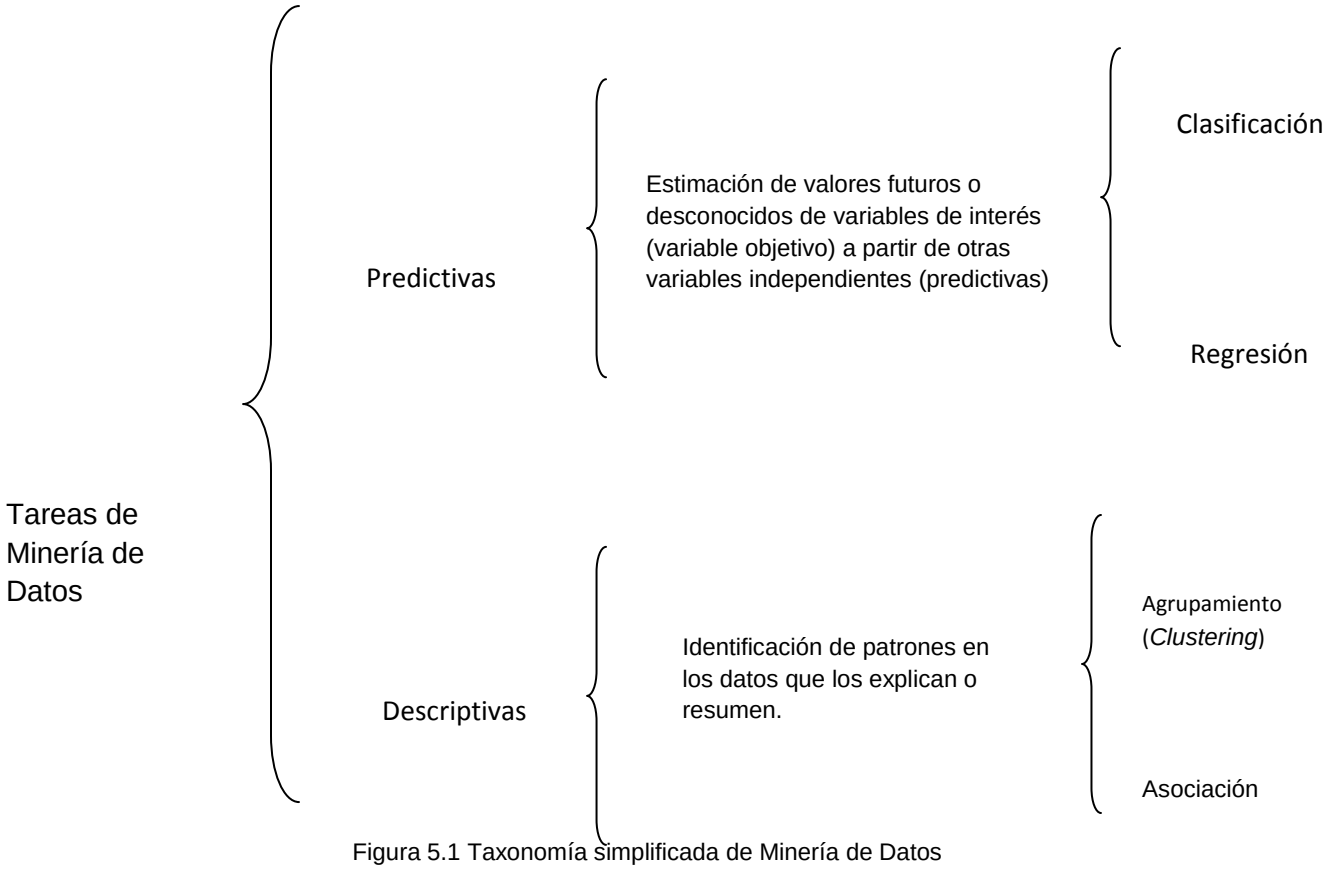

Figura 5.1 Taxonomía simplificada de Minería de Datos

.

5.2 TAREAS PREDICTIVAS.

Como lo explica en la figura 5.1 un modelo predictivo, es aquel que nos ayuda a realizar estimaciones futuras con respecto a los valores almacenados de años anteriores. En esta sección se mostrarán los resultados obtenidos al aplicar técnicas predictivas a los datos originales y también después de su preprocesamiento.

1. Datos antes de ser procesados:

   El primer problema encontrado al intentar aplicar las técnicas de clasificación a los datos sin ser preprocesados. Consiste en que estos no son nominales.

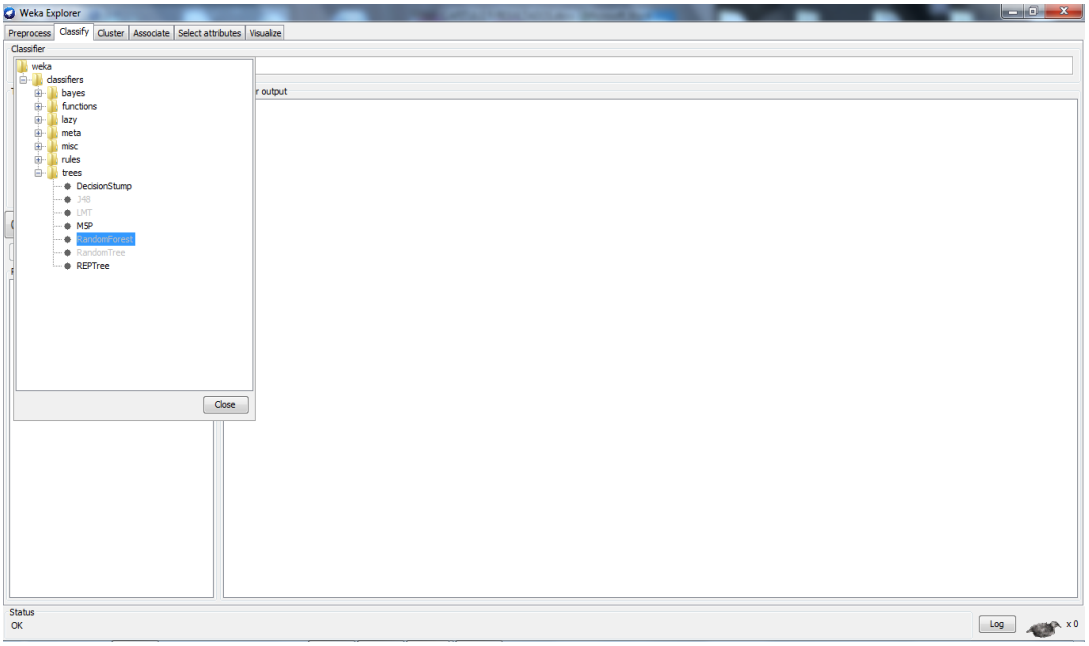

Fig. 5.2 Muestra de problemas al aplicar técnicas de clasificación

Como se muestra en la Figura 5.2 *WEKA* no nos permite seleccionar las técnicas que queremos aplicar ya que la fuente de datos no presenta las características necesarias, para realizar esta tarea.

2. Datos después de ser procesados:

Una vez que tengamos preparados los datos para este proceso, seleccionamos la técnica de minería de datos que deseamos aplicar, en este caso realizamos 3 pruebas, 2 de árboles de decisión (*J48* y *RadomForest*) y *Naive Bayes*, Los resultados se muestran de las figuras 5,3 a la 5.6, de manera separada para cada técnica.

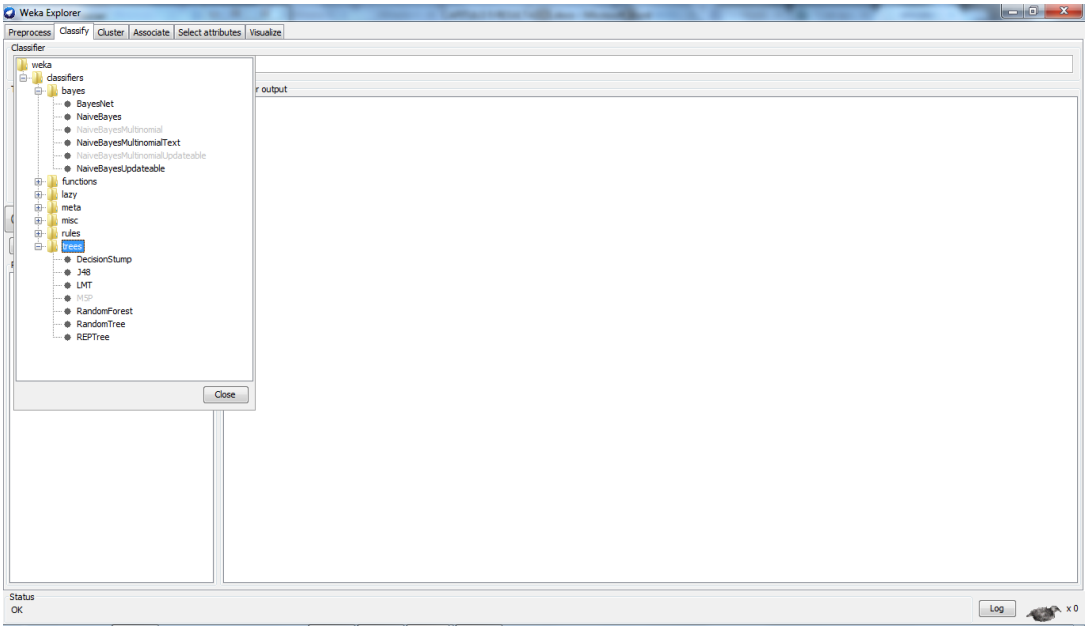

Fig. 5.3 Técnicas activadas después del preprocesamiento.

Como se muestra en la figura 5.3, ya es posible seleccionar las técnicas que se desean utilizar. Ahora procederemos al análisis de los resultados obtenidos con cada una de ellas.

- Técnica *Naive Bayes*:

  Naive Bayes es una técnica estadística asumiendo eventos equiprobables que sirve para realizar clasificación en Minería de Datos. Y en la figura 5.4 mostraremos los resultados obtenidos con dicha técnica.

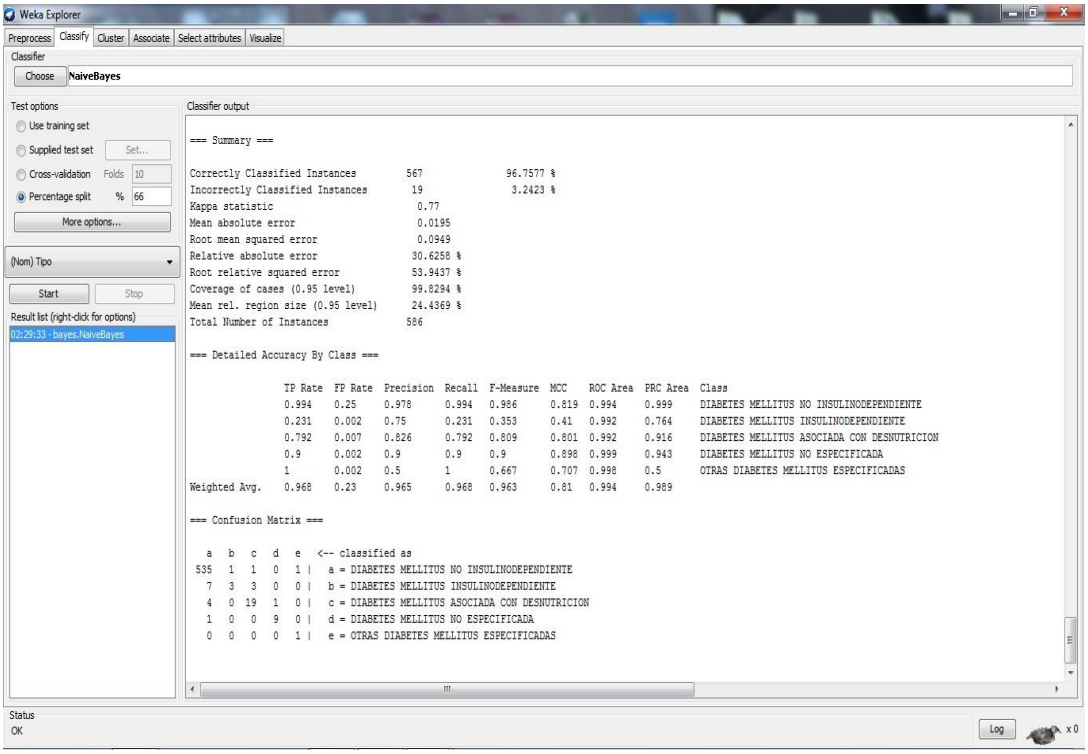

Fig. 5.4 Resultados por Naive Bayes

Como se muestra en la figura 5.4, del 66% de los datos que fueron usados para el entrenamiento, el 96.7577% fue correctamente clasificado. Esté resultado es bastante aceptable, ya que de los 586 datos utilizados, solo 19 no fueron clasificados correctamente. Esos 19 registros mal clasificados representan solo el 3.2423% del total. Cabe mencionar que esta clasificación se llevó a cabo mediante el tipo de diabetes y su complicación en caso de presentar alguna.

- Arboles de decisión *RandomForest*:

  Aplicando esta técnica, podemos observar que tenemos menos datos erróneos en la clasificación con respecto a *NaiveBayes*. A diferencia de los resultados de la figura 5.4, en la figura 5.5 podemos observar que los datos mal clasificados descendieron a un 1.5358% del total.

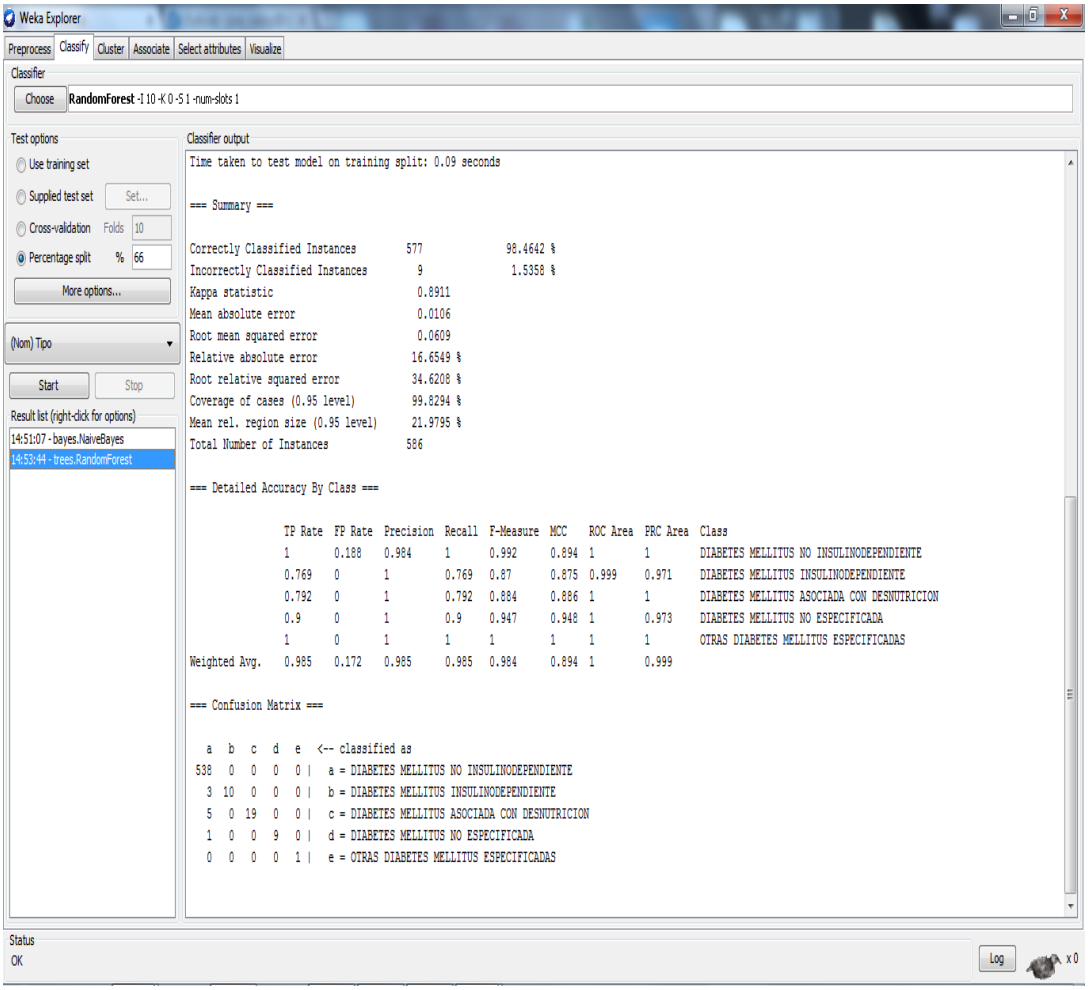

Fig. 5.5 Resultados por *Random Forest*

- Arboles de decisión *J48*:

  Esta técnica es diferente a la anterior, los arboles J48, tratan con valores continuos y utiliza criterios estadísticos para impedir que el árbol se sobre adapte (que crezca demasiado, que se aprenda los datos en vez de generalizar) [17].

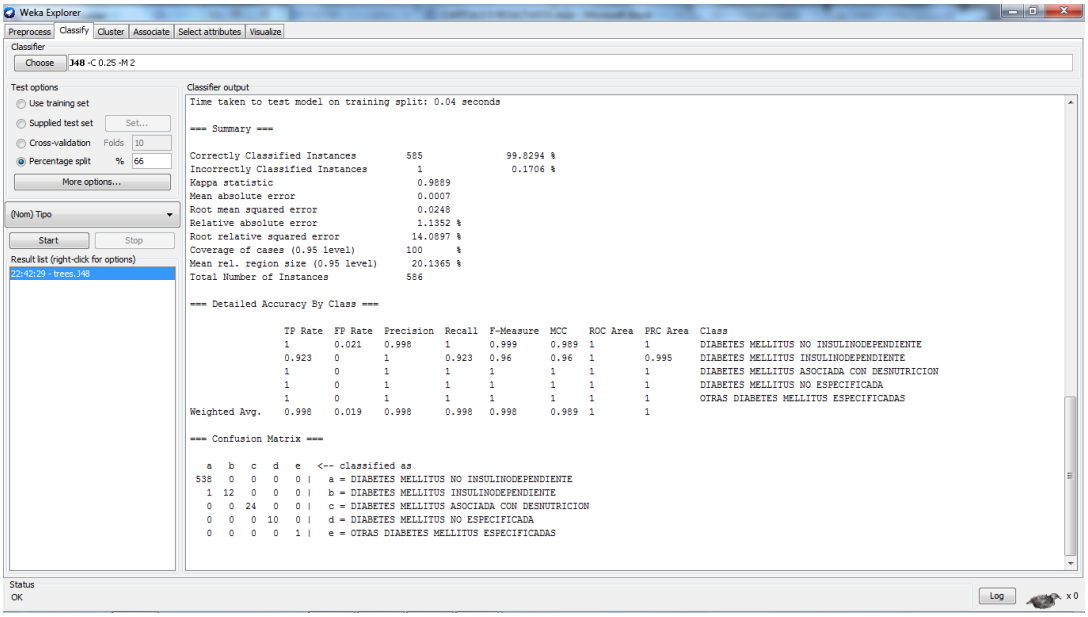

Fig.5.6 Resultados por Arboles J48

Como podemos observar en la figura 5.6 los arboles J48 ofrecieron una mejor clasificación para el conjunto de datos ya que de todos los datos usados para el entrenamiento que fueron 586, solo 1 dato no fue bien clasificado.

5.3 TAREAS DESCRIPTIVAS.

En esta sección se presentan los resultados obtenidos al haber aplicado las tareas descriptivas al conjunto de datos. Se mostraran en dos secciones una con los datos antes de ser tratados y la otra con los datos ya preprocesados.

- **Agrupamiento o *clustering*:** Es un procedimiento de reunión de una serie de vectores según criterios habitualmente basados en distancias. Se trata de disponer los vectores de entrada, de forma que estén más cercanos aquellos que tengan características comunes. [16]**.** A diferencia de las técnicas de clasificación en esta técnica se habla de grupos y no de clases.

1. Datos antes de ser preparados.

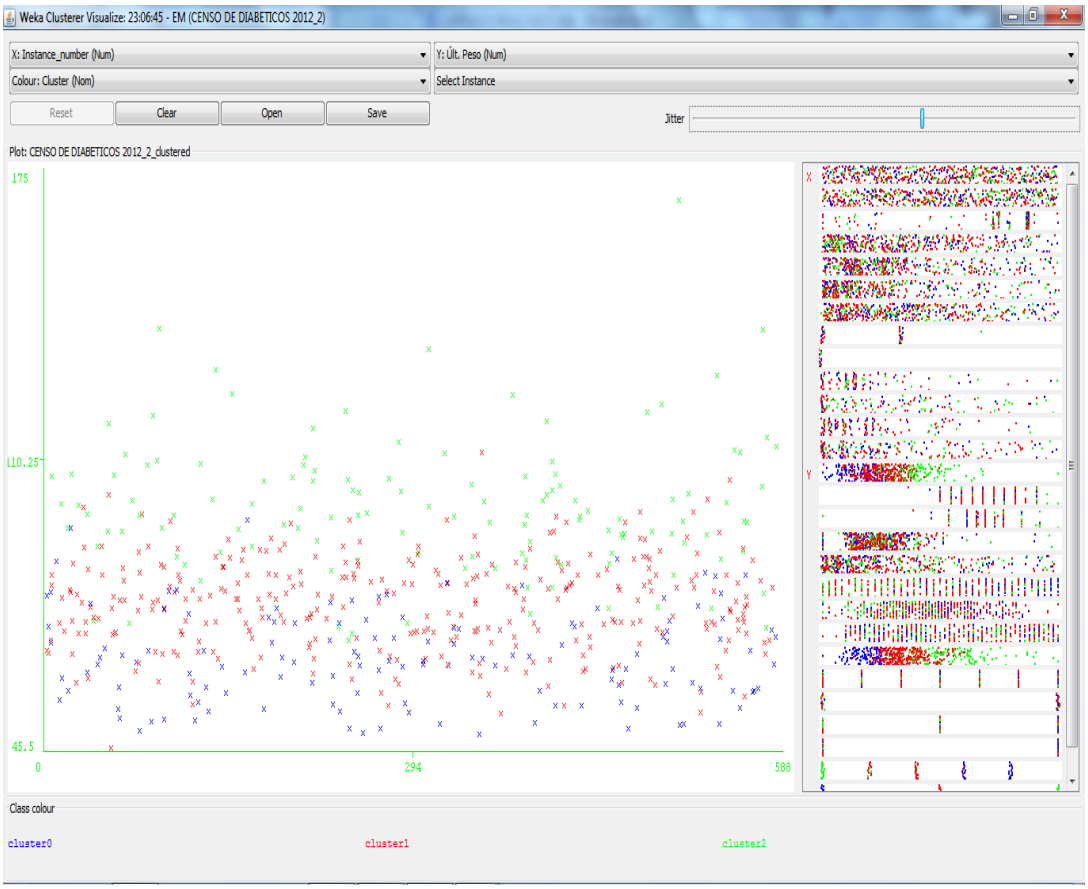

Fig. 5.7 Resultados de *simpleKmeans* a los datos originales

Como se muestra en la figura 5.7 realmente no se notan los *clusters* ni los resultados de forma adecuada. El eje X representa el número de la instancia y el de las Y al *último peso* del paciente. Este resultado es debido a que *último peso* es el único valor en todo el conjunto que WEKA reconocía para arrojar un resultado.

Los datos originales no permitían por si mismos observar gráficamente relaciones importantes entre los atributos de la muestra, ya que en una sola columna se incluía: el tipo de diabetes, la complicación que presentaba y el código que los médicos le asignan de acuerdo a la complicación.

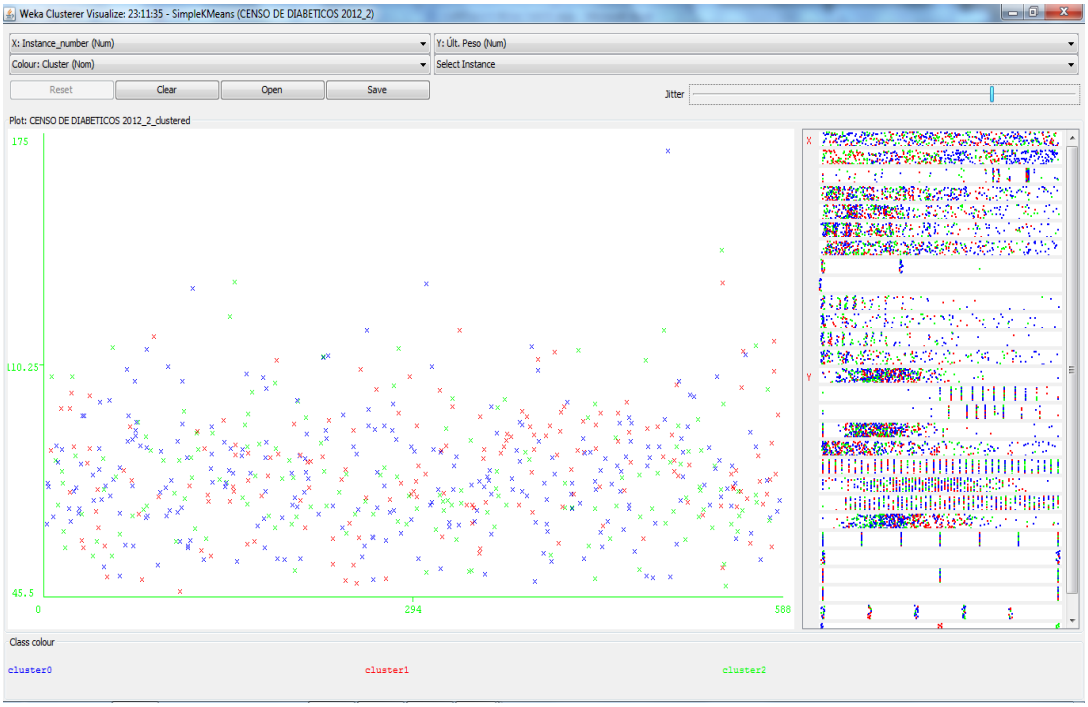

Fig 5.8 Resultados con clusters de tipo simpleKmeans

Como se muestra en la Figura 5.8 tampoco podemos apreciar de manera clara los resultados, así que procederemos a mostrar los resultados con los datos ya preprocesados.

Como se menciona en [18] para conocer el número correcto de *clusters* se realiza una prueba con el método EM con un parámetro cantidad de *clusters* de -1. Esto con la finalidad de obtener el número de K del algoritmo *Kmeans*, donde la K indica el número de clusters que nos serán suficientes para poder realizar las pruebas. En la figura 5.11 podemos observar las reglas de asociación obtenidas durante las pruebas realizadas.

2. Datos después de ser procesados.

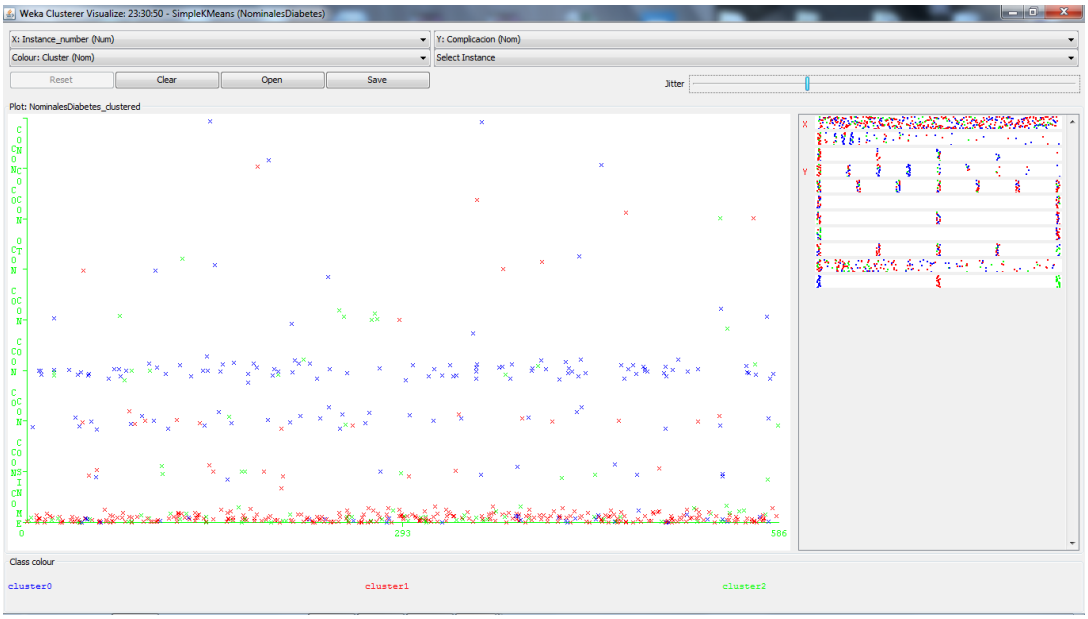

Fig. 5.9 Resultados por simpleKmeans

Como podemos observar en la figura 5.9 se pueden revisar los *clusters* de acuerdo a la complicación que tienen los pacientes, ya que se han separado en columnas distintas al igual que el código que el médico asigna y el tipo de diabetes (de acuerdo a una complicación). Todos los ejemplos corresponden a DM tipo 2, por lo anteriormente mencionado en el capítulo 4, donde se eliminaron los de diabetes tipo 1 por el simple hecho de no tener suficientes datos.

En esta sección se mostrará un conjunto de imágenes que muestran los *clusters* obtenidos a las pruebas realizadas

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2474.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                                                                    Cluster#
Attribute                                                   Full Data                      0                              1                              2
                                                              (1138)                   (584)                          (388)                          (166)
==========================================================================================================================================================
Tipo                          DIABETES MELLITUS NO INSULINODEPENDIENTE DIABETES MELLITUS NO INSULINODEPENDIENTE DIABETES MELLITUS NO INSULINODEPENDIENTE DIABETES MELLITUS NO INSULINODEPENDIENTE
Complicacion                           SIN MENCION DE COMPLICACION         SIN MENCION DE COMPLICACION         SIN MENCION DE COMPLICACION         SIN MENCION DE COMPLICACION
POBLACION                                                      Juo                                 Smx                                  UH                                 Lvs
SEXO                                                             F                                   M                                   F                                   F
ALTURA CLASIFICADA                                           Media                               Media                               Media                               Media
EDAD CLASIFICADA                                            Adulta                              Adulta                              Adulta                             Anciana
PESO CLASIFICADO                                          Obesidad                            Obesidad                         Peso Normal                          Sobre Peso



Time taken to build model (percentage split) : 0.02 seconds

Clustered Instances

0      313 ( 53%)
1      195 ( 33%)
2       79 ( 13%)
```

Fig.5.10 Resultados por simpleKmeans con datos preprocesados.

En la figura 5.10 podemos observar los resultados obtenidos, el conjunto de datos, y los *clusters* resultantes después de haber aplicado las técnicas de agrupamiento. A pesar de tener algo de diferencia respecto al porcentaje, podemos observar entre los *clusters 1 y 2* son muy similares ya que por ejemplo en el *cluster* número 2 los valores son muy parecidos hasta el momento que analizamos la altura, edad y el peso del paciente, agregando también a la lista de diferencias la población.

5.4 REGLAS DE ASOCIACIÓN

Los algoritmos de asociación permiten la búsqueda automática de reglas que relacionan conjuntos de atributos entre sí [18].

```
Apriori
=======

Minimum support: 0.25 (431 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 19

Size of set of large itemsets L(3): 16

Size of set of large itemsets L(4): 4

Best rules found:

 1. Complicacion= SIN MENCION DE COMPLICACION SEXO=M 466 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 450    <conf:(0.97)> lift:(1.05) lev:(0.01) [22] conv:(2.29)
 2. EDAD CLASIFICADA=Adulta PESO CLASIFICADO=Obesidad 455 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 437    <conf:(0.96)> lift:(1.05) lev:(0.01) [19] conv:(2)
 3. Complicacion= SIN MENCION DE COMPLICACION 1124 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 1077    <conf:(0.96)> lift:(1.05) lev:(0.03) [46] conv:(1.95)
 4. PESO CLASIFICADO=Obesidad 577 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 552    <conf:(0.96)> lift:(1.04) lev:(0.01) [23] conv:(1.85)
 5. Complicacion= SIN MENCION DE COMPLICACION EDAD CLASIFICADA=Adulta 865 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 826    <conf:(0.95)> lift:(1.04) lev:(0.02) [33] conv:(1.81)
 6. Complicacion= SIN MENCION DE COMPLICACION ALTURA CLASIFICADA=Media 918 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 875    <conf:(0.95)> lift:(1.04) lev:(0.02) [33] conv:(1.74)
 7. Complicacion= SIN MENCION DE COMPLICACION SEXO=F 658 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 627    <conf:(0.95)> lift:(1.04) lev:(0.01) [23] conv:(1.72)
 8. ALTURA CLASIFICADA=Media PESO CLASIFICADO=Obesidad 508 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 483    <conf:(0.95)> lift:(1.04) lev:(0.01) [17] conv:(1.63)
 9. Complicacion= SIN MENCION DE COMPLICACION SEXO=F EDAD CLASIFICADA=Adulta 496 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 471    <conf:(0.95)> lift:(1.04) lev:(0.01) [16] conv:(1.59)
10. Complicacion= SIN MENCION DE COMPLICACION ALTURA CLASIFICADA=Media EDAD CLASIFICADA=Adulta 713 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 677    <conf:(0.95)> lift:(1.04) lev:(0.01)
11. Complicacion= SIN MENCION DE COMPLICACION SEXO=F ALTURA CLASIFICADA=Media 531 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 503    <conf:(0.95)> lift:(1.03) lev:(0.01) [16] conv:(1.53)
12. SEXO=M EDAD CLASIFICADA=Adulta 542 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 509    <conf:(0.94)> lift:(1.02) lev:(0.01) [12] conv:(1.33)
13. SEXO=M ALTURA CLASIFICADA=Media 595 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 554    <conf:(0.93)> lift:(1.02) lev:(0.01) [8] conv:(1.18)
14. SEXO=M 719 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 668    <conf:(0.93)> lift:(1.01) lev:(0.01) [9] conv:(1.15)
15. EDAD CLASIFICADA=Adulta 1249 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 1151    <conf:(0.92)> lift:(1.01) lev:(0) [6] conv:(1.05)
16. ALTURA CLASIFICADA=Media EDAD CLASIFICADA=Adulta 1031 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 946    <conf:(0.92)> lift:(1) lev:(0) [1] conv:(1)
17. ALTURA CLASIFICADA=Media 1415 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 1295    <conf:(0.92)> lift:(1) lev:(0) [-1] conv:(0.98)
18. SEXO=F 1005 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 913    <conf:(0.91)> lift:(0.99) lev:(0) [-8] conv:(0.9)
19. SEXO=F EDAD CLASIFICADA=Adulta 707 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 642    <conf:(0.91)> lift:(0.99) lev:(0) [-5] conv:(0.89)
20. SEXO=F ALTURA CLASIFICADA=Media 820 ==> Tipo=DIABETES MELLITUS NO INSULINODEPENDIENTE 741    <conf:(0.9)> lift:(0.99) lev:(-0.01) [-10] conv:(0.86)
```

Fig. 5.11 Reglas de Asociación Apriori

El principal algoritmo de asociación implementado en WEKA es el algoritmo "Apriori". Este algoritmo únicamente puede buscar reglas entre atributos simbólicos, razón por la que se requiere haber discretizado todos los atributos numéricos. [18]

De LA figura 5.11 tomaremos las primeras 6 reglas de asociación, ya que estas nos aportan un mayor valor de convicción estando en un rango de 1.74 a 2.29. Lo cual que con estas conformamos los *clusters* que tenemos en la figura 5.10

5.5 CONCLUSIONES

En este capítulo, se mostraron los resultados obtenidos después de haber aplicado las técnicas de minería de datos. Como podemos observar y mencionó en el capítulo 2, el trabajo más arduo de la minería de datos, es la preparación de los datos, ya que sin ese paso, el o los programas que realizan descubrimiento de conocimiento normalmente no pueden realizar su tarea de forma adecuada.

Se aplicaron tres métodos de predicción: *NaiveBayes*, J48 y *RandomForest*. J48 tuvo el mejor rendimiento con un 99.8294%. En cuanto a las técnicas descriptivas se aplicaron simpleKmeans y A priori donde surgieron 3 *clusters* y 20 reglas de asociación.

CAPÍTULO 6

CONCLUSIONES Y TRABAJO A FUTURO

La tesis fue encamida al desarrollo de vistas minables que nos facilitaran el conocimiento de los casos de diabetes que existen en Juchitán Oaxaca, por lo tanto se cumplieron los objetivos planteados. Durante el desarrollo del trabajo de tesis, fueron diversos los obstáculos a los cuales me enfrenté, desde el momento de definir el tema, hasta el comenzar a redactar los resultados obtenidos. Algunos de estos obstáculos fueron, por ejemplo, conseguir la información ya que todos los datos utilizados en el proceso de pruebas, son datos confidenciales. Por este motivo fue algo complicado el hecho de conseguir dicho material, que una vez encontrado, se nos otorgo sin datos personales de los pacientes para no incurrir en violación a su privacidad.

También cabe mencionar que durante el desarrollo de la tesis, se adquirieron diversas habilidades para investigación. Cabe mencionar también que se adquieren conocimientos más amplios sobre los temas que se van desarrollando ya que, en este caso, se pone mucha dedicación al momento de preparar, limpiar, y seleccionar los datos que nos van a ser de utilidad para ir desarrollando el trabajo.

Es factible mencionar que es una gran satisfacción haber desarrollado un trabajo de investigación de este tipo, ya que es un tema bastante alarmante en todo el mundo. Es posible mencionar que existen diversos puntos que podemos tomar como experiencia adquirida, los cuales los enumeraré a continuación:

1. Existe mucha información que se está desaprovechando debido al mal uso que se le da, con esto me refiero a que hay empresas que tienen informes o datos sobre su control los cuales no son procesados para poder conocer el futuro de la empresa, y así generarle mayores beneficios.
2. En la actualidad hay diversas herramientas de minería de datos de distribución gratuita y de paga las cuales contienen métodos que nos facilitan el análisis de los datos.
3. No se necesita de una capacitación exhaustiva para poder hacer uso de estas herramientas, ya que personas ajenas al campo informático podrían hacer uso de tales, con una capacitación básica.
4. El trabajo en conjunto con expertos en el área de la salud ha sido de mucha importancia para el desarrollo de esta tesis sobre todo en el análisis de los resultados.

Como trabajo a futuro se plantea, desarrollar un sistema para la toma de decisiones que incluya el sistema de captura de datos asociados con los pacientes diabéticos y un almacén de datos. Se propone además, el desarrollo de este almacén de datos, que integrará información interna del sistema antes mencionado y otras fuentes externas como lo son: las secretarías de salud estatal y federal, plataforma MexRisc de la BUAP, así como la organización mundial de la salud entre las más significativas.

REFERENCIAS

[1]. Hernández Orallo J, Ramirez Quintana Ma. J, Ferri Ramírez C, (2008), *Introducción a la Minería de Datos..*, 2da. Edición, editorial Prentice Hall.

[2]. GUÍAS ALAD DE DIAGNÓSTICO, CONTROL Y TRATAMIENTO DE LA DIABETES MELLITUS TIPO 2(Asociación Latinoamericana de Diabetes).

[3]. Krupp A. M, Chatton J. M,(1983) *Diagnóstico clínico y tratamiento*.., México DF, Editorial El Manual Moderno, S.A. de C.V.

[4]. http://mexrisc.cs.buap.mx/quienes.php, MexRisc, ¿quiénes somos?.,MexRisc, [última consulta 12 Agosto 2012].

[5]. http://www.fundaciondiabetes.org/findrisk/CampanaOnLine.asp , Findrisk, [última consulta 13 Agosto 2012]

[6]. Ramez E, (2007), *Fundamentos de Sistemas de Bases de Datos..(* Madrid.) Addison Wesley. 5ta Edición,

[7]. Pérez C, Santín D, (2006) *Data Mining Soluciones con Enterprise Miner..(*México). Alfaoemega. Primera edición.

[8].http://www.elmundo.es/elmundosalud/2007/03/01/corazon/1172778362.html , Diabetes un problema mayor de lo esperado, [Ultima consulta 03 septiembre 2012]

[9]. Rob P, Coronel C , (2004) *Sistemas de Bases de Datos diseño, implementación y administración..* , (México), Thomson, 5ta Edición.

[10].http://vivecondiabetes.com/investigacion-en-mexico/noticias/191-recientes-investigaciones-en-diabetes-y-obesidad-de-la-unam-recopiladas-en-libro- Libro de la UNAM, [última consulta 15 de agosto de 2012]

[11] Secretaria de Salud.(2008), Programa de acción especifico 2007-2012, Mexico Df.

[12].J. Moody and C. Darken. *Fast Learning in networks of locally-tuned processing units.*

Neural Computation.., 1(2):281-294, 1989.

[13]. González Bernal J.A., Olmos-Pineda I.,. *Minería de Datos, México*

[14]. Ing. Rolando Acosta Sánchez, Dr. Alejandro Rosete Suárez, Lic. Alfredo Rodríguez Díaz, Ing. Raycos Brito Sarasa. *Empleo de Minería de Datos en la predicción de diabetes. Preprocesado de datos,Habana Cuba.*

[15] Velázquez Montalvo Amado R.,(Enero 2011), *Proyecto extracción de reglas usando programación genética como base para un sistema de soporte a la toma de decisiones clínicas,Maestria,CICESE,Ensenada, Baja California, México*

[16]. López Cumplido A, (Agosto 2010), *Sistema para toma de decisiones acerca del cambio climatico*.Tesis de Licenciatura, BUAP, Puebla,Mexico.

[17].http://ocw.uc3m.es/ingenieria-informatica/herramientas-de-la-inteligencia-artificial/contenidos/transparencias/MDWEBHIA-clase.pdf , Introducción al aprendizaje automático y a la minería de datos con weka herramientas de la inteligencia artificial ingeniería informática.[4 Julio 2013].

[18]. Análisis de Datos en WEKA.

[19]. Nikhil R. Pal and Lakhmi Jain (Eds). Advanced Techniques in Knowledge Discovery and Data Mining. London. Springer-Verlarg.2005. 234 p.

[20]. Riquelme José C., Ruiz Roberto y Gilbert Karina. Minerìa de datos: Conceptos y tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, primavera. 10(029): 11-18, 2006.

[21]. Hernández Orallo J. Parte II: *Almacenes de Datos, transparencias basadas parcialmente en el "tutorial DW" de Matilde Celma* [diapositiva]. Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia. 122 diapositivas, col.