UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

DETECCIÓN DE PATRONES DE COMPORTAMIENTO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS EN EXPEDIENTES CLÍNICOS DE PACIENTES PREDIABÉTICOS

Trabajo recepcional bajo la modalidad de Tesis
Que para obtener el grado de

Maestro en Administración de Tecnologías de la Información

Presenta

L.S.C. Henry Jesús Hernández Gómez

Directores

M.I.S. Laura Beatriz Vidal Turrubiates

M.I.S. Homero Alpuín Jiménez

42

Cuerpo Académico:

Sistemas Inteligentes

Líneas de Generación y Aplicación del Conocimiento

Optimización, Clasificación y Minería de datos

Cunduacán, Tabasco Diciembre 2013

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

DETECCIÓN DE PATRONES DE COMPORTAMIENTO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS EN EXPEDIENTES CLÍNICOS DE PACIENTES PREDIABÉTICOS

Trabajo recepcional bajo la modalidad de Tesis
Que para obtener el grado de

Maestro en Administración de Tecnologías de la Información

Presenta

L.S.C. Henry Jesús Hernández Gómez

Directores

M.I.S. Laura Beatriz Vidal Turrubiates

M.I.S. Homero Alpuín Jiménez

Comisión Revisores

M.S.I. Ninfa Urania García Ulín
M.C. Carlos Arturo Custodio Izquierdo
M.C. Guillermo de los Santos Torres

Cunduacán, Tabasco Diciembre 2013

Carta de Autorización

A quien corresponda:

El que suscribe, autoriza por medio del presente escrito a la Universidad Juárez Autónoma de Tabasco para que utilice tanto física como digitalmente la tesis de grado denominada **"DETECCIÓN DE PATRONES DE COMPORTAMIENTO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS EN EXPEDIENTES CLÍNICOS DE PACIENTES PREDIABÉTICOS"** , de la cual soy autor y titular de los Derechos de Autor.

La finalidad del uso por parte de la Universidad Juárez Autónoma de Tabasco de la tesis antes, mencionada, será única y exclusivamente para difusión, educación y sin fines de lucho; autorización que se hace de manera enunciativa más no limitativa para subirla a la Red Abierta de Bibliotecas Digitales (RABID) y a cualquier otra red académica con las que la Universidad tenga relación institucional.

Por lo antes manifestado, libero a la Universidad Juárez Autónoma de Tabasco de cualquier reclamación legal que pudiera ejercer respecto al uso y manipulación de la tesis mencionada y para los fines estipulados en éste documento.

Se firma la presente autorización en la ciudad de Cunduacán, Tabasco a los 04 días del mes de Diciembre del año 2013.

Autorizo

L.S.C. Henry Jesús Hernández Gómez

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

Oficio No. 2274/12/DAIS/D
Septiembre 18 de 2012

MIS. Laura Beatriz Vidal Turrubiates
Profesora-Investigadora
P r e s e n t e

De acuerdo al artículo 44 fracción 3 del Reglamento de Estudios de Posgrado, de la Universidad Juárez Autónoma de Tabasco, me permito informar a Usted, que ha sido designada Directora del trabajo de Tesis titulado **"Detección de Patrones de Comportamiento utilizando Técnicas de Minería de Datos en Expedientes Clínicos de Pacientes Prediabéticos"**, a realizar por el **C. Henry Jesús Hernández Gómez,** para obtener el grado de Maestro en Administración de Tecnologías de la Información.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

Universidad Juárez Autónoma de Tabasco
División Académica de Informática y Sistemas

Lic. Eduardo Cruces Gutiérrez
Director

c.c.p. Lic. Martha Patricia Silva Payró.- Coordinadora de Investigación y Posgrado.
Archivo.
Consecutivo.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

Oficio No. 2275/12/DAIS/D
Septiembre 18 de 2012

MIS. Homero Alpuin Jiménez
Profesor-Investigador
P r e s e n t e

De acuerdo al artículo 44 fracción 3 del Reglamento de Estudios de Posgrado, de la Universidad Juárez Autónoma de Tabasco, me permito informar a Usted, que ha sido designado Director del trabajo de Tesis titulado **"Detección de Patrones de Comportamiento utilizando Técnicas de Minería de Datos en Expedientes Clínicos de Pacientes Prediabéticos"**, a realizar por el **C. Henry Jesús Hernández Gómez,** para obtener el grado de Maestro en Administración de Tecnologías de la Información.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

Universidad Juárez Autónoma de Tabasco
División Académica de Informática y Sistemas

Lic. Eduardo Cruces Gutiérrez
Director

c.c.p. Lic. Martha Patricia Silva Payró.- Coordinadora de Investigación y Posgrado.
Archivo
Consecutivo.

Miembro CUMEX desde 2008
Consorcio de Universidades Mexicanas

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690. Cunduacán, Tabasco, México.
E-mail: direccion.dais@ujat.mx
Teléfonos: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

En la Universidad Juárez Autónoma de Tabasco, de acuerdo al Reglamento de Estudios de Posgrado vigente, se revisó el trabajo de investigación titulado "DETECCIÓN DE PATRONES DE COMPORTAMIENTO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS EN EXPEDIENTES CLÍNICOS DE PACIENTES PREDIABÉTICOS", realizado por el C. Henry Jesús Hernández Gómez, para obtener el Grado de Maestro en Administración de Tecnologías de la Información bajo la modalidad de Tesis.

Los integrantes del jurado, después de revisar el trabajo, lo declararon aceptado. Firmando la presente a los 2 del mes de Diciembre de 2013.

M.C. Carlos Arturo Custodio Izquierdo
Profesor-Investigador

M.C. Guillermo de los Santos Torres
Profesor-Investigador

M.S.I. Ninfa Urania García Ulín
Profesor-Investigador

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

55 ANIVERSARIO UJAT

DAIS 11111000011

DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS

"2013, CENTENARIO LUCTUOSO DE FRANCISCO I. MADERO Y JOSÉ MARÍA PINO SUÁREZ"

Oficio No.2853/13/DAIS/D
Diciembre 4 de 2013

LSC. Henry Jesús Hernández Gómez
P R E S E N T E

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo de investigación **"Detección de Patrones de Comportamiento utilizando técnicas de minería de datos en expedientes clínicos de pacientes prediabéticos"**, para presentar Examen de Grado de la Maestría en Administración de Tecnologías de la Información, bajo la modalidad de Tesis.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

MATI. Eduardo Cruces Gutiérrez
Director

c.c.p. Lic. Martha Patricia Silva Payró.- Coordinadora de Investigación y Posgrado.
Archivo.
Consecutivo.

Miembro CUMEX desde 2008
Consorcio de Universidades Mexicanas
UNA ALIANZA DE CALIDAD POR LA EDUCACIÓN SUPERIOR

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690. Cunduacán, Tabasco, México.
E-mail: direccion.dais@ujat.mx
Teléfonos: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870

A Dios

Por ser ese amigo que brida el apoyo y la fuerza para atravesar cada obstáculo en los tiempos difíciles, por enseñarme con tu palabra el camino y la verdad, por fortalecer mi espíritu e iluminar mi mente.

A mis padres

Por su amor, atención y rectitud que hicieron de mi una persona de bien, con la idea de enfrentar los desafíos de la vida sin tener miedo. Gracias por ser los mejores padres.

A mis hermanos

Gracias por estar en cada instante de mi vida, por entusiasmarme a seguir adelante. Por estar siempre dispuestos a escuchar todas y cada una de mis vivencias en esta etapa de mi vida.

24

Yo juro que vale más ser de baja condición y codearse alegremente con gentes humildes, que no encontrarse muy encumbrado, con una resplandeciente pesadumbre y llevar una dorada tristeza.

William Shakespeare

Agradecimientos

Mis más sinceros agradecimientos:

A mis directores de tesis :

M.I.S. Laura Beatriz Vidal Turrubiates y M.I.S. Homero Alpuín Jiménez los cuales admiro y respeto por su gran capacidad y entusiasmo. Sus conocimientos, recomendaciones y tiempo han sido piezas fundamentales para la elaboración de esta investigación. “Gracias”

A mis profesores de asignatura:

M.A.S.I. Arturo Corona Ferreira, M.T.E. Oscar Alberto González González, Dr. Miguel Wister Ovando, Dr. Pablo Payró Campos, Dra. Marbella Aracely Gómez Lemus, M.I.S. Laura Beatriz Vidal Turrubiates y M.I.S. Homero Alpuín Jiménez, Gracias por compartir su conocimiento y experiencias en las asignaturas impartidas.

A mis revisores de tesis:

M.S.I. Ninfa Urania García Ulín, M.C. Carlos Arturo Custodio Izquierdo y M.C. Guillermo de los Santos Torres. Por su acertadas y valiosas aportaciones, las cuales enriquecieron el contenido de este trabajo recepcional.

Índice General

Índice de ilustraciones xiv
Índice de tablas xvi
Capítulo I. Introducción 18
1.1 Antecedentes 18
1.2 Planteamiento del problema 20
1.3 Objetivos 21
1.3.1 Objetivo general 21
1.3.2 Objetivos específicos 21
1.4 Justificación 22
1.5 Delimitación 23
1.5.1 Alcances 23
1.5.2 Limitaciones 23
1.6 Metodología a utilizar 24
1.6.1 Metodología de la investigación 24
1.6.2 Instrumentos para la recolección de los datos 25
1.6.3 Fuentes de investigación 25
1.6.4 Metodología de minería de datos 26
Capítulo II. Marco teórico 27
2.1 Marco referencial 27
2.2 Marco legal 29
2.3 Marco conceptual 30
2.3.1 Bases de datos 30
2.3.2 Análisis de información 31
2.3.3 Aprendizaje automático (machine learning) 32
2.3.4 Minería de datos 32
2.3.5 Extracción del conocimiento con minería de datos 35
2.3.6 Modelos de minería de datos 37

2.3.7 Metodología de minería de datos 41
2.4 Marco tecnológico 49
2.4.1 Software de minería de datos WEKA 49
2.4.2 Software estadístico PSPP 50
2.4.3 Hoja de cálculo 51
Capítulo III. Aplicación de la metodología y desarrollo 52
3.1 Elección del grupo objeto 52
3.2 Descripción del universo de estudio 52
3.3 Modelo de referencia CRISP-DM 53
3.4 Análisis del problema o comprensión del negocio 54
3.4.1 Evaluación de la situación actual 55
3.4.2 Plan del proyecto 56
3.5 Comprensión de los datos 57
3.5.1 Reporte de recolección de los datos 57
3.5.2 Reporte de descripción inicial de los datos 58
3.5.3 Reporte de exploración de los datos 63
3.5.4 Verificación de la calidad de los datos 66
3.6 Preparación de los datos 67
3.6.1 Datos seleccionados 67
3.6.2 Limpieza de los datos 73
3.6.3 Construcción e integración de los datos 74
3.7 Modelado 74
3.7.1 Selección de las técnicas de modelado 74
3.7.2 Generación del plan de prueba 75
3.7.3 Construcción de los modelos 76
Capítulo IV. Resultados 80
4.1 Descripción de modelos 80
4.1.1 Clasificación variable a302c 80
4.1.2 Modelo de clasificación de las variables a310c, a310g y a310f 82

4.1.3 Clustering (Base de datos Adultos 20 años o más) 87
4.1.4 Clustering (Base de datos Glucosa y Lípidos) 99
4.2 Evaluación de los resultados 104
4.3 Evaluación del grupo objeto 105
Capítulo V. Conclusiones y trabajos futuros 108
5.1 Conclusiones 108
5.2 Trabajos Futuros 109
Bibliografía 110
Glosario 113
Anexos 114
Anexo A. Buffer de algoritmo J48 en BD adultos 20 años o mas 115
Anexo B Buffer de algoritmo J48 a la variable a310c en BD adultos 20 años o mas. 117
Anexo C Buffer de algoritmo J48 a la variable a310f en BD adultos 20años o mas. 118
Anexo D Buffer de algoritmo J48 a la variable a310g en DB adultos 20años o mas. 119
Anexo E Buffer de algoritmo SimpleKmeans BD adultos 20años o mas. 120
Anexo F Buffer de algoritmo SimpleKmeans BD Glucosa y Lípidos 121
Anexo G Reporte de Resultados 122

Índice de ilustraciones

Ilustración 1. Metodología CRISP-DM 26
Ilustración 2. Etapas de un proceso de minería de datos....34
Ilustración 3. Etapas de un proceso KDD 35
Ilustración 4. Proceso de minería de datos....37
Ilustración 5. Fases de comprensión del negocio CRISP-DM 2000 43
Ilustración 6. Fases de comprensión de los datos CRISP-DM 2000....44
Ilustración 7. Fases de la preparación de los datos CRISP-DM 2000....45
Ilustración 8. Fase del modelado CRISP-DM 2000....46
Ilustración 9. Fase de evaluación CRISP-DM 2000 47
Ilustración 10. Fase de implementación CRISP-DM 2000....48
Ilustración 11. Pantalla inicial de WEKA 50
Ilustración 12. Fases de metodología CRISP-DM 54
Ilustración 13. Pantalla de acceso a bases de datos ENSANUT 57
Ilustración 14. Pantalla de SPSS con la base de datos Adultos 20 años o más....58
Ilustración 15. Pantalla de SPSS con la base de datos Glucosa y Lípidos....58
Ilustración 16. Gráficas de las variables de la base de datos Adultos 20 años o más. 64
Ilustración 17. Gráfica de las variables de la base de datos Glucosa y Lípidos....66
Ilustración 18. Resultado de la evaluación del atributo Edad 71
Ilustración 19. Resultado de la evaluación del atributo Sexo 71
Ilustración 20. Resultado de la evaluación del atributo Confdiab....72
Ilustración 21. Resultados iniciales de aplicar técnicas de minería de datos 75
Ilustración 22. Árbol de decisión de la evaluación del atributo a302c 76
Ilustración 23. Modelo de clasificación 77
Ilustración 24. Modelo de clasificación para variables a310c, a310f y a310g....78
Ilustración 25. Modelo de agrupación para Base de datos Adultos, Glucosa y lípidos 79

Ilustración 26. Árbol de decisión de la evaluación del atributo a302c 80
Ilustración 27. Buffer de la precisión del algoritmo J48 81
Ilustración 28. Árbol de decisión generado de la evaluación de la variable a310c.....82
Ilustración 29. Interpretación del árbol de decisión de la variable a310c 83
Ilustración 30. Árbol de decisión generado de la evaluación de la variable a310g 84
Ilustración 31. Interpretación del árbol de decisión de la variable a310g.....84
Ilustración 32. Buffer de la precisión del algoritmo J48 de la variable a310g.....85
Ilustración 33. Buffer de la precisión del algoritmo J48 de la variable a310f 86
Ilustración 34. Interpretación del árbol de decisión de la variable a310f 86
Ilustración 35. Buffer de la aplicación del algoritmo SimpleKmeans 88
Ilustración 36. Buffer de la clasificación de los datos en función del atributo a301 89
Ilustración 37. Visualización de las agrupación en función del atributo a301.....90
Ilustración 38. Conjunto Clúster 1 95
Ilustración 39. Conjunto Clúster 2 96
Ilustración 40. Conjunto Clúster 3 97
Ilustración 41. Conjunto Clúster 4 98
Ilustración 42. Buffer de los clustering creados por Weka 100
Ilustración 43. Porcentaje de cada agrupación.....101
Ilustración 44. Visualización de las variables CONFDIAB y COL.....101
Ilustración 45. Gráfica de la distribución de valores.....103

Índice de tablas

Tabla 1. Descripción inicial de los datos(Base de datos glucosa y lípidos).....59
Tabla 2. Descripción inicial de los datos (Base de datos Adultos 20 años o más) 1/3 60
Tabla 3. Descripción inicial de los datos (Base de datos Adultos 20 años o más) 2/3 61
Tabla 4. Descripción inicial de los datos (Base de datos Adultos 20 años o más) 3/3 62
Tabla 5. Exploración de los datos (Base de datos Adultos 20 años o más) variable a301.....63
Tabla 6. Exploración de los datos (Base de datos Adultos 20 años o más) variable a302c.....63
Tabla 7. Exploración de los datos (Base de datos Adultos 20 años o más) variable a303.....63
Tabla 8. Exploración de los datos (Base de datos glucosa y lípidos) variable Sexo.....64
Tabla 9. Exploración de los datos (Base de datos glucosa y lípidos) variable Confdiab.....65
Tabla 10. Exploración de los datos variables Edad, Glucosa, Colesterol 65
Tabla 11. Exploración de los datos (Base de datos glucosa y lípidos).....65
Tabla 12. Resultados de las evaluaciones del atributo a301(1/2) 67
Tabla 13. Resultados de las evaluaciones del atributo a301(2/2) 68
Tabla 14. Cálculo de la media aritmética.....68
Tabla 15. Atributos seleccionados 69
Tabla 16. Resultados de la evaluación de las variables Edad, Sexo, Confidiab(1/2).....72
Tabla 17. Resultados de la evaluación de las variables Edad, Sexo, Confidiab(2/2).....73
Tabla 18. Atributos a evaluar de la base datos glucosa y lípidos.....87
Tabla 19. Descripción de elementos de diabéticos y no diabéticos clúster 1.....91
Tabla 20. Descripción de elementos de diabéticos y no diabéticos clúster 2.....92
Tabla 21. Descripción de elementos de diabéticos y no diabéticos clúster 3.....93
Tabla 22. Descripción de elementos de diabéticos y no diabéticos clúster 4.....94
Tabla 23. Variables Base de datos Glucosa y Lípidos.....99
Tabla 24. Características de diabéticos y no diabéticos clúster 1 102

Resumen

El objetivo del presente trabajo fue obtener patrones de comportamiento de los expedientes clínicos de pacientes pre-diabéticos, utilizando técnicas de minería de datos, como apoyo a la toma de decisiones para el control de la diabetes. Para el logro de las metas trazadas se aplicó software de minería de datos Weka, que se caracteriza por tener las funciones necesarias que permitieron realizar transformaciones sobre los datos, así como tareas selección de atributos, clasificación, clustering para la extracción del conocimiento en las bases de datos. Las bases de datos sobre las que se trabajó son Adultos 20 años o más de la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2012, Glucosa y Lípidos de ENSANUT 2006. La metodología de minería de datos que se utilizó es CRISP-DM (Cross Industry Standard Process for Data Mining), las fases que la integran son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación, estas permitieron obtener un proceso de minería de datos. Según el problema propuesto y el objetivo planteado, la investigación se encuadró bajo un enfoque cualitativo que se basa en la recolección de datos sin medición numérica para descubrir o afinar preguntas de investigación, así como en el método hermenéutico digital y etnográfico que comparten la finalidad de interpretar y comprender datos, que permitan descubrir contextos nuevos en los grandes volúmenes de información. Los patrones de comportamiento encontrados en las bases de datos se pueden ver en el anexo G. El impacto de esta investigación se ve reflejado en la extracción de conocimiento en grandes almacenes de datos como son los expedientes clínicos, así mismo permite la clasificación, provee un mayor grado de certeza, ahorra tiempo y recursos siendo estos los mas importantes. La forma de innovar utilizando herramientas de minería de datos esta en la parte de que además de conocer las características de un solo individuo, permite también conocer las de una población determinada, así como generar arboles de decisión respecto a datos de control de la enfermedad. El beneficio social esta la cultura de la prevención, así como control de la diabetes.

Capítulo I. Introducción

Existen ciertas normas para el desarrollo de un proyecto que facilitan el manejo y la organización de la información, así como la descripción de las razones de realizar la presente investigación, así como la problemática planteada y las estrategias propuestas para la solución de la misma.

A continuación se describen los antecedes, el planteamiento del problema, los objetivos, la justificación, la delimitación y la metodología utilizada.

1.1 Antecedentes

En la última década el avance tecnológico ha sido muy grande en diferentes áreas las cuales permiten tener computadoras con mayor capacidad de procesamiento y almacenamiento, así como una mejor tasa de transferencia de datos. Estos avances llevaron a una mayor generación y recolección de información. (Ángeles & Santillán, 2004).

Sin duda alguna todos los días se genera información de diferentes maneras esto puede ser al revisar el correo electrónico, a la llegada al trabajo, al reservar una mesa en un restaurant o al asistir a una consulta médica. Los motivos de almacenar toda la información generada son porque al ser analizada permite optimizar, predecir, agrupar, clasificar o tomar decisiones dentro de las organizaciones.

Con el propósito de mejorar el análisis de la información surgen nuevas tecnologías como es la minería de datos que es un conjunto de técnicas utilizadas para encontrar en los datos contextos nuevos, ocultos o inesperados según Césari, (2002). La minería de datos, tiene muchas aplicaciones en las empresas, gobiernos, universidades, hospitales y diversas organizaciones.

En la medicina no es la excepción el utilizar minería de datos, permite realizar análisis de factores de riesgo en distintas patologías como es el Mal de Parkinson o la Osteoartritis en las que se usan técnicas de clúster de minería de datos, al igual que en farmacología en la que se utiliza dichas técnicas para descubrir nuevas propiedades de moléculas conocidas (Limite, 2012). También a través de ella se puede identificar el número de personas con dicho padecimiento, así como la agrupación de acuerdo a patrones de cada paciente.

Existen investigaciones realizadas por instituciones educativas que sirven como referencia para la elaboración de esta investigación, una de ellas fue realizada por la Escuela de Ingeniería de Antioquia-Universidad CES, Medellín, Colombia en la que fueron utilizados métodos de clasificación para identificar lesiones en piel a partir de espectros de reflexión difusa. Otra investigación realizada con la misma perspectiva de detección de patrones tuvo origen en la Universidad de Valencia, España titulada clasificación de estirpes histológicas de tumores de partes blandas mediante reconocimiento de patrones a partir de imágenes de RM (Resonancia Magnética).

En la actualidad en los hospitales pertenecientes al Sector Salud la forma de analizar la información es utilizando consultas a la base de datos lo cual impide tener un nuevo conocimiento de la información almacenada. De esta manera surge la necesidad de aplicar minería de datos utilizando diferentes técnicas aplicables a expedientes clínicos de pacientes pre-diabéticos lo que permitirá generar nuevo conocimiento representado por patrones de comportamiento que sirva a los médicos como apoyo a la toma de decisiones.

1.2 Planteamiento del problema

En México, la diabetes la presentan 10.1 millones de personas y se estima que para el 2030 sea de 16.4 millones (IDF, 2011). En México la diabetes ocupa el primer lugar de defunciones por año, tanto en hombres como en mujeres muestra una tendencia ascendente en ambos sexos con más de 70 mil muertes y 400,000 casos nuevos anuales (Norma Oficial Mexicana, 2010).

En Tabasco, de acuerdo con la última Encuesta Nacional de Salud y Nutrición 2006 (ENSANUT) la prevalencia de la diabetes en adultos de 20 años o más es de 6.2%, siendo más elevada en mujeres 7.3% que en hombres 5.1% (ENSANUT, 2006).

Se calcula que en el 2004 fallecieron 3,4 millones de personas como consecuencias del exceso de azúcar en la sangre, más del 80% de las muertes por diabetes se registran en países de ingresos bajos, medios y casi la mitad de esas muertes corresponden a personas de menos de 70 años, y un 55% a mujeres. La OMS prevé que las muertes por diabetes se multipliquen por dos entre 2005 y 2030 (OMS, 2011).

Hoy en día, la Secretaría Salud (SS) tiene la necesidad de realizar la recopilación de información a través de encuestas aplicadas a los pacientes, para evaluar las condiciones de salud que presenta la población a nivel nacional. La finalidad de la encuesta fue obtener de manera probabilística la cobertura de programas de salud en áreas básicas como las inmunizaciones, la atención a padecimientos crónicos, así como sobre los retos en salud como el control de la hipertensión arterial, el sobrepeso, la obesidad y la diabetes, esta ultima según la Organización Mundial de la Salud es una patología la cual está presente en más de 346 millones de personas en el mundo (OMS, 2011).

En busca de atender el reto que tiene la SS en materia de diabetes, requiere información de los patrones de comportamiento de resultados probabilístico obtenidos de las encuestas, por lo cual, surgió la necesidad de conocer el estado actual y el comportamiento de la información a través de la extracción del conocimiento mediante el uso de herramientas de TI.

La información probabilística de los Pre-diabéticos almacenada en las bases de datos de los últimos diez años, ha servido a los expertos en tomas de decisiones para analizar y evaluar los resultados de incidencia en la diabetes. Actualmente, estos expertos tienen la necesidad de aplicar herramientas innovadoras TI a las probabilidades de más de diez años, para conocer contextos ocultos de los datos que permitan planear estratégicamente medidas preventivas para la Secretaria de Salud Estatal y Nacional, mediante técnicas de minería de datos.

Partiendo de lo anterior, se plantea la siguiente pregunta de investigación:

¿Qué técnicas de minería de datos, permitirán obtener los patrones de comportamiento en los grandes volúmenes de información de pacientes Pre-diabéticos?

1.3 Objetivos

Ya planteado el problema de investigación es necesario buscar una solución a través de unos objetivos, estos son mencionados a continuación:

1.3.1 Objetivo general

Obtener patrones de comportamiento utilizando técnicas de minería de datos, como apoyo a la toma de decisiones para el control de la diabetes en los Hospitales del Sector Salud del Estado de Tabasco.

1.3.2 Objetivos específicos

- Preparar las bases de datos haciendo limpieza de los datos y seleccionar las técnicas de minería de datos.
- Seleccionar las variables más representativas del conjunto de datos a través del software de minería de datos.
- Aplicar diversas técnicas de minería de datos a los expedientes clínicos de pacientes prediabéticos con el propósito de obtener patrones de comportamiento.

- Analizar los datos obtenidos y proporcionar resultados a través de un reporte para que los directivos de la Institución conozcan el comportamiento de la información y así realicen toma decisiones o planeación de estrategias en el combate de la diabetes.

1.4 Justificación

Los avances en las tecnologías de la información TI han ido incrementando en los últimos 15 años, todo esto con el objeto de buscar soluciones más factibles que propicien el desarrollo social, económico y cultural de las personas.

Hoy el uso de herramientas de minería de datos en la medicina tiene gran impacto debido a que permite detectar fácilmente patrones de comportamiento, siendo la técnica más eficiente que el análisis dirigido a la verificación, cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada (García, Miguel, García, & Polo, 1997). Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los grandes volúmenes de datos (Pinto & Torres, 2006).

El aplicar minería de datos en los expedientes clínicos de pacientes prediabéticos con la finalidad de obtener patrones de comportamiento y relaciones entre variables, logrará innovar la forma de analizar la información. Además tendrá el beneficio de detectar pacientes con factores de riesgo de desarrollar diabetes. Otro beneficio de contar con nuevo conocimiento representado por patrones es que permitirá tomar de decisiones a través de la formulación de estrategias para la aplicación de diversas medidas preventivas orientadas a mejorar la salud de los pacientes.

El impacto social que aporta esta investigación es que puede servir como apoyo en el combate del aumento de la diabetes y ligado a ello contribuiría con dos beneficios muy grandes que son el disminuir la tasa de mortalidad y la reducción de gastos en caminados a tratamientos de enfermos por diabetes, todo esto a un largo plazo. Sin duda alguna esta investigación da la pauta a que se aplique herramientas de minería de datos en otras áreas de la medicina buscando el bienestar de la humanidad.

1.5 Delimitación

Una vez realizada la justificación de la investigación se hace mención de los alcances y de las posibles limitaciones que puedan ocurrir durante el desarrollo.

1.5.1 Alcances

- Aplicar técnicas de minería de datos (Clasificación, Agrupación) como apoyo en el proceso de análisis de los datos de los expedientes clínicos de pacientes prediabéticos.
- Generar patrones y tendencias de comportamiento de los expedientes clínicos de pacientes prediabéticos.
- Proporcionar un informe de los resultados obtenidos de la minería de datos, que sirvan como sustento a la toma de decisiones.

1.5.2 Limitaciones

- El presente trabajo de investigación solo se planteó en Secretaria de Salud del estado de Tabasco, como caso de estudio.
- La investigación solo está enfocada en la detección de patrones de comportamiento de expedientes clínicos de pacientes prediabéticos.
- La fuente de bases de datos es la Encuesta Nacional de Salud y Nutrición 2006 y 2012 (ENSANUT).
- Para la investigación se utilizó el software de minería de datos Weka (Waikato Learning Environmen).

1.6 Metodología a utilizar

La metodología es la guía para realizar las tareas indicadas de una investigación; en esta sección se describe la metodología de investigación y del desarrollo de la minería de datos que fueron utilizados en el desarrollo del trabajo recepcional.

1.6.1 Metodología de la investigación

Según el problema propuesto y el objetivo planteado, la investigación se encuadro bajo un enfoque cualitativo con un método hermenéutico digital y etnográfico.

Hernández, R., Fernández, C. y Baptista, P. (1991) menciona que un enfoque cualitativo se basa en la recolección de datos sin medición numérica para descubrir o afinar preguntas de investigación y puede o no probar hipótesis en su proceso de interpretación. De acuerdo a la definición la investigación es cualitativa, porque permitió conocer contextos y comportamientos predominantes a través de la descripción exacta de las acciones, objetos y procesos.

Para Arráez M. et al (2006) la hermenéutica es una disciplina que se dedica a interpretar y develar el sentido de los mensajes haciendo que su comprensión sea posible. La investigación se enmarco dentro del tipo hermenéutico porque determinadamente estuvo enfocado a la interpretación y comprensión de datos digitales, para descubrir contextos nuevos en los grandes volúmenes de información.

De la misma forma se inserto en el método etnográfico que según Galeano M. (2003) se concibe como la descripción, registro sistemático y análisis de un campo de la realidad social especifico, de una escena cultural, de patrones de interacción social. La investigación se enfatizó en este método por que se realizó la recolección de todo tipo de datos accesibles para poder mostrar conocimiento sobre el comportamiento de los datos de pacientes prediabéticos.

Para darle validez se realizó el proceso de **triangulación metodológica** utilizando hermenéutica digital y etnográfica. Para aportar confiabilidad al estudio se realizó la **triangulación de los de datos** obtenidos de manuales, médicos internistas y resultados de la base de datos.

Se determinaron los sujetos mas adecuados para el estudio, usando el muestreo por segmentación.

El *universo o población de estudio* de esta investigación lo constituyen los pacientes prediabéticos.

1.6.2 Instrumentos para la recolección de los datos

Para analizar el problema planteado se utilizó como instrumento para la recolección de los datos la entrevista, que es una técnica de acceso a la información muy empleada en la evaluación. El tipo de entrevista será semiestructurada la cual se encuadra en el método hermenéutico. Su característica esencial es la flexibilidad y es utilizada para realizar exploraciones, para recopilar información previa y estudiar posibilidades de intervención en contextos.

1.6.3 Fuentes de investigación

Para esta investigación se utilizaron fuentes primarias y secundarias. Las fuentes primarias fueron la información digital de pacientes prediabéticos. Las fuentes secundarias son las que nos proporcionaron información abreviada ejemplo de ello son normas, resúmenes.

1.6.4 Metodología de minería de datos

La metodología de minería de datos utilizada fue CRISP-DM (*Cross Industry Standard Process for Data Mining*, por sus siglas en ingles), por tener las características de flexibilidad y personalización, necesarias para obtener un proceso de minería de datos.

Esta metodología de minería de datos está constituida por seis fases:

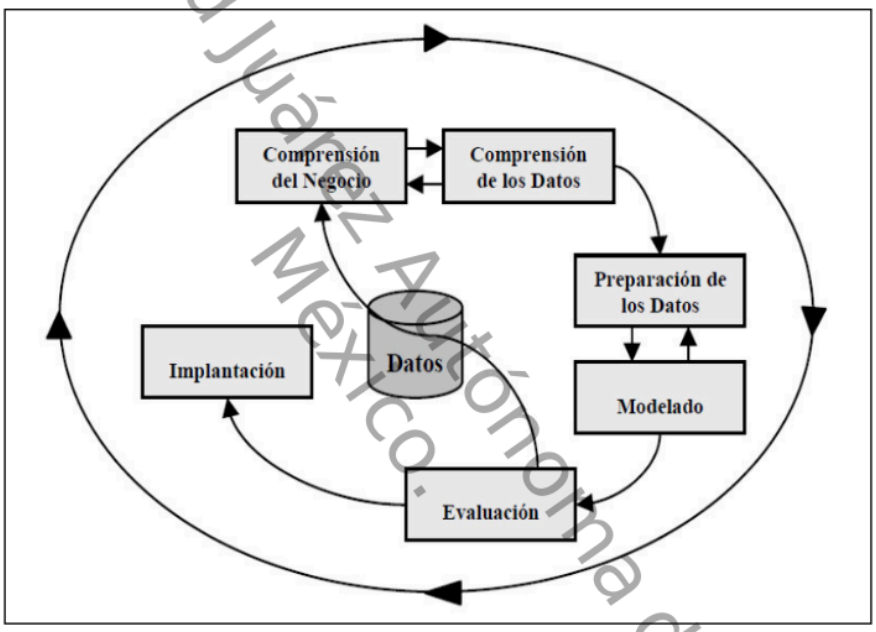


Ilustración 1. Metodología CRISP-DM
(Arancibia, 2010)

Capítulo II. Marco teórico

En este capitulo es parte fundamental dentro del desarrollo de la investigación, permite sustentar teóricamente los conceptos relacionados con la misma. Este integrado por marco referencial, marco legal, marco conceptual y marco tecnológico.

2.1 Marco referencial

Sin duda alguna un punto esencial del marco teórico (que tiene como finalidad fundamentar una investigación), lo constituye el marco referencial, el cual tiene como propósito realizar una revisión de las investigaciones que se han efectuado en el tema que se aborda, por lo que a continuación se presentan algunas de las investigaciones más relevantes que fueron consultadas, no sin antes destacar primero que en el proceso de extracción de nuevo conocimiento a través de herramientas automáticas es algo que se ha venido realizando desde hace años en el ámbito empresarial ya sea para predecir, clasificar y asociar un sinfín de sucesos, pero con el paso del tiempo, se han ido aumentando y mejorando dichas herramientas, incrementando enormemente su potencial. Lo que permite abarcar diferentes rubros en la sociedad. Actualmente existen proyectos enfocados a la identificación de nuevo conocimiento en diferentes áreas de investigación, los cuales son tomados como base para obtener puntos de vistas diferentes, buscando mejorar e innovar.

Flores (2009), en su tesis de maestría, "Detección de patrones de daños y averías en la industria automotriz", menciona que la minería de datos es, en realidad, una prolongación de una práctica estadística de larga tradición, la de Análisis de Datos. Existiendo además, una aportación propia de técnicas específicas de Inteligencia Artificial, en particular sobre la integración de los algoritmos, la automatización del proceso y la optimización del coste.

"Al iniciar un proyecto de Minería de datos (Data Mining), la deducción, el análisis y el modelado de los requisitos del usuario (proceso de Ingeniería de Requisitos), constituyen actividades relevantes para el éxito del proyecto.

Sin embargo estas actividades, normalmente son las menos exploradas debido a la inexistencia de técnicas, procedimientos o métodos ad-hoc para estos propósitos."(Gallardo, 2009).

Para Dueñas (2009), en su tesis de maestría "Minería de datos espaciales en búsqueda de la verdadera información" menciona que la minería de datos permite descubrir el conocimiento no sólo en el ámbito organizacional, sino sustentar investigaciones en la rama biológica.

Las herramientas de minería de datos se pueden aplicar a los datos clínicos para proporcionar conocimiento predictivo y descriptivo acerca de los pacientes individuales, grupos específicos de pacientes, y de una población local, regional o nacional en su conjunto(Robbins & Chiesa, 2011).

46

La importancia del impacto de una herramienta de explotación de datos en la organización y la inversión que la misma debe hacer en términos económicos, sino también en las necesidades propias del negocio (César, 2002).

CRISP-DM, es una metodología para el desarrollo de proyectos de minería de datos que se ha convertido en un estándar. El consorcio CRISP-DM, responsable de esta metodología, está integrado por importantes empresas europeas y estadounidenses que poseen una amplia experiencia en proyectos de análisis de datos relacionados con muy diversos campos de la industria, (Fernández, Jiménez, González, Ávila, 2007).

El tomar como base proyectos ya desarrollados, que aplican técnicas de minería de datos proporciona diferentes perspectivas metodologías, así como conocer las fases, herramientas, características que integran un proyecto de minería de datos.

Para los investigadores de estas tesis todos los objetivos que se plantearon dentro de su investigación se reflejaron en sus conclusiones como metas logradas, lo que garantiza que el aplicar técnicas de minería de datos a expedientes clínicos de pacientes prediabéticos para detección patrones de comportamiento será viable.

2.2 Marco legal

En este apartado se menciona la licencia de uso con la que se debe de contar para el proyecto de investigación.

La herramienta de minería de datos que se utilizó es WEKA y SPSS de cual no se requiere un contrato de licencia para su uso. Debido a que está regido por la ley GPL o Licencia Pública General de GNU. Todos los derechos garantizados por esta Licencia se otorgan como copyright del Programa, y se proporcionan de manera irrevocable siempre y cuando se cumplan las condiciones establecidas.

Algunos de los aspectos que contemplan la versión 3 de la ley GPL:

- Las diversas formas en que alguna persona podría quitar libertades a los usuarios.
- Prohibir el uso de software libre en sistemas que utilizan la llamada "gestión digital de derechos" o DRM, sistema criticado por la comunidad del software libre.
- Resolver ambigüedades y aumentar la compatibilidad de GPLv3 con otras licencias.
- Incluir cláusulas que defiendan a la comunidad de software libre del uso indebido de las patentes de software.

El software de hoja de cálculo utilizado fue Excel, del cual se requiere un contrato de licencia para su uso. Este contrato define una serie de parámetros que delimitan el derecho del usuario hacia el software como lo menciona el artículo 106 de la Ley Federal del Autor de los Estados Unidos Mexicanos. Este software fue utilizado en un equipo de cómputo de la Universidad Juárez Autónoma de Tabasco el cual cuenta con licencia educativa.

Derechos de autor sobre la proyecto de minería de datos

El proyecto de minería de datos, estuvo regido por las leyes de derecho de autor estipuladas en la Ley Federal de Derecho de Autor del software WEKA.

2.3 Marco conceptual

En este apartado se hacen mención de todos los conceptos relacionados con el campo de estudio de la investigación.

2.3.1 Bases de datos

Para Sierra(2012), base de datos es un sistema informático a modo de almacén. En este almacén se guardan grandes volúmenes de información.

Según Elmasri & Navathe (2005), define a una base de datos (BD) como un conjunto de datos lógicamente coherentes con cierto significado inherente. Las principales características de las bases de datos son la independencia física y lógica de los datos, redundancia mínima, integración de datos. Las bases de datos se ha convertido en una herramienta esencial que permiten guardar grandes cantidades de información, y son aprovechada en múltiples áreas como los son: banca, líneas aéreas, universidades, finanzas, hospitales, ventas entre otras.

Existen diferentes tipos de bases de datos:

- Las bases de datos relacionales es una colección de relaciones (tablas) donde cada tabla consta de un conjunto de atributos y puede tener numerosas tuplas, registros o filas.
- Las bases de datos espaciales contienen información relacionada con el espacio físico en un sentido amplio (una ciudad, una región montañosa, un atlas cerebral).
- Las bases de datos temporales almacenan datos que incluyen muchos atributos relacionados con el tiempo o en el que éste sea muy relevante.
- Las bases de datos documentales contienen descripciones para los objetos que pueden ir desde las simples palabras clave a los resúmenes.
- Las bases de datos multimedia almacenan imágenes, audio y video. Soportan objetos de gran tamaño ya que, por ejemplo, los videos pueden necesitar varios gigabytes de capacidad para su almacenamiento.

2.3.2 Análisis de información

Para Rabinowitz & Fawcett (2011), analizar información incluye examinarla de maneras que muestran las relaciones, patrones, tendencias que puedan ser encontradas.

El análisis de información forma parte del proceso de adquisición y apropiación de los conocimientos latentes acumulados en distintas fuentes de información. El análisis busca identificar la información "útil", es decir, aquella que interesa al usuario, a partir de una gran cantidad de datos (Sarduy, 2007).

Según Gómez e Iglesias (2004), el análisis de información es una forma de investigación, cuyo objetivo es la captación, evaluación, selección y síntesis de los mensajes subyacentes en el contenido de los documentos, a partir del análisis de sus significados, a la luz de un problema determinado. Así, contribuye a la toma de decisiones, al cambio en el curso de las acciones y de las estrategias.

Según Fromenta, A. (2011) los pasos básicos para el análisis de información son:

1. Identificar las necesidades del usuario o de la comunidad de usuario.
2. Seleccionar la información requerida.
3. Valorar la calidad del documento y los datos que brinda.
4. Interpretar los datos en correspondencia con la finalidad de la adquisición del documento.

El uso de análisis de información en un principio se realizaba de forma manual, empleando técnicas estadísticas. Sin embargo, actualmente esta forma de análisis resulta inviable por la gran cantidad de datos que puede contener una BD moderna, además de que existen una gran cantidad de formatos para los datos, como tablas (bases de datos relacionales), secuencias, grafos , imágenes, audio, lo cual aumenta la complejidad de un análisis manual.

Con la evolución tecnológica el proceso de analizar información se relacionado con lo que se denomina actualmente knowledge discovery in databases (KDD) y se apoyado con el desarrollo de herramientas tecnológicas capaces de realizar dicha tarea.

2.3.3 Aprendizaje automático (machine learning)

El aprendizaje automático esta constituido por algoritmos que son métodos que dado un conjunto de ejemplos de entrenamiento infieren un modelo de las categorías en las que se agrupan los datos, de tal forma que se pueda asignar a nuevos ejemplos una o más categorías de manera automática mediante analogía de patrones en dicho modelo. Serrano, J, Tomečková, M. & Zvárová, J. (2006).

Según Guerra, A. (2004) El aprendizaje automático es la capacidad de un sistema de computo de aprender a partir de una experiencia E, con respecto a alguna clase de tarea T y una medida de desempeño P, si el desempeño del programa al realizar las tareas T, mejorando con la experiencia E, de acuerdo a la medida P.

Los algoritmos de aprendizaje automático pueden clasificarse en dos grandes categorías: métodos de caja negra (o sin modelo), tales como redes neuronales o los métodos bayesianos, y métodos orientados al conocimiento, tales como los que generan árboles de decisión, reglas de asociación, o reglas de decisión.

2.3.4 Minería de datos

Según Hernando, R. (2004), minería de datos (Data Mining) es un conjunto de técnicas y procesos de análisis de datos que permiten extraer información de bases de datos y almacenes de datos mediante la búsqueda automatizada de patrones y relaciones.

La minería de datos es una nueva tecnología de manejo y análisis de información que aprovecha la capacidad existente hoy día de procesamiento, almacenamiento y transmisión de datos a gran velocidad y bajo costo. Permite encontrar el conocimiento contenido en las inmensas montañas de información para luego tomar decisiones mejor fundamentadas para el futuro de una organización.(Cervantes, López, & Gayosso, 2010).

Para Angeles & Santillán (2004), la minería de datos es el proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación.

2.3.4.1 Tareas de la minería de datos

Según Hernández, J., Ramírez, M. & Ferri, C. (2004) En la minería de datos se aprecian tipos de tareas, que pueden considerarse como un tipo de problema el cual pueden ser resultó a través de algún algoritmo de minería de datos. Cada tarea tiene sus propios requisitos, y que el tipo de información diferir mucho de la obtenida con otra.

Los algoritmos pueden ser catalogados como predictivas o descriptivas. Entre los predictivos encontramos la clasificación y la regresión, mientras que el agrupamiento (clustering), las reglas de asociación y las correlaciones son tareas descriptivas.

Uno de los algoritmos mas utilizados es la clasificación. En aquí la instancia pertenece a una clase, la cual se indica mediante el valor de un atributo que llama clase de la instancia. Este atributo puede tomar diferentes valores discretos, donde cada uno corresponde a una clase y el resto se utiliza para predecir la clase.

La agrupación (clustering) es una tarea descriptiva que ayuda a obtener grupos naturales a partir de los datos. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos del grupo minimizando la similitud entre los distintos grupos.

Las reglas de asociación son consideradas como tareas descriptivas su función es identificar relaciones no explicitas entre los atributos categóricos. Las asociaciones no implican una relación causa-efecto, es decir, puede no existir causa para que los datos estén asociados.

Todas estas tareas ayudan a explorar los datos que se encuentran en las profundidades de las bases de datos. Podemos decir que hurgar y mover a menudo implica el descubrimiento de resultados valiosos e inesperados.

2.3.4.2 Etapas de un proceso de minería de datos

En minería cada proyecto es único. Sin embargo, en términos generales, el proceso se compone de cuatro etapas principales que se muestran en la Ilustración 2, que a consideración (Maneiro, 2008), se interpretan de la siguiente manera :

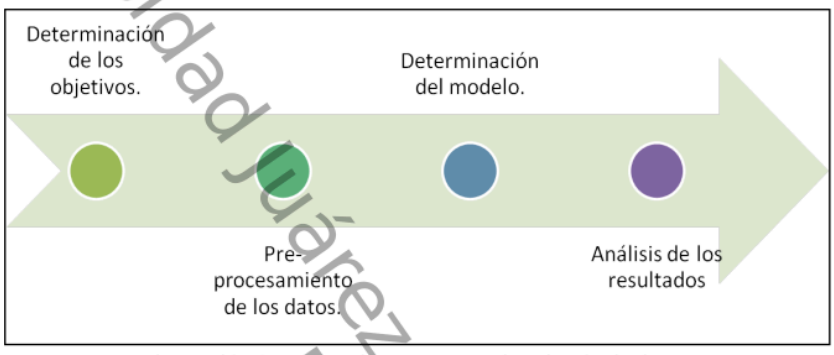


Ilustración 2. Etapas de un proceso de minería de datos
(Maneiro, 2008. Interpretación del autor)

1. **Determinación de los objetivos.-** Trata de la delimitación de los objetivos que el cliente desea.

2. **Pre-procesamiento de los datos**.- Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Es la etapa que consume más de la mitad del tiempo del proyecto.

3. **Determinación del modelo.-** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

4. **Análisis de los resultados.-** Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

2.3.5 Extracción del conocimiento con minería de datos

Según Fayyad et al. (1996), extracción del conocimiento en bases de datos es un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos.

El proceso de minería de datos pertenece a un esquema amplio denominado "Descubrimiento de conocimiento en bases de datos" (*KDD, del Inglés Discovery from Databases)* (Tabuenca, 2011). Este proceso nace como interfaz y se nutre de diferentes disciplinas como la estadística, sistemas de información, bases de datos, inteligencia artificial, visualización computacional, aprendizaje automático entre otras.

El uso de extracción de conocimiento en bases de datos médica con fines de apoyo a la toma de decisiones se ha vuelto al indispensable debido a que permite identificar terapias médicas para diferentes enfermedades, asociar síntomas y clasificar patologías, así como múltiples acciones que se deseen realizar.

2.3.5.1 Fases KDD (Knowledge Discovery from Databases)

El proceso de extracción del conocimiento está constituido por etapas las cuales se muestran en la ilustración 3.

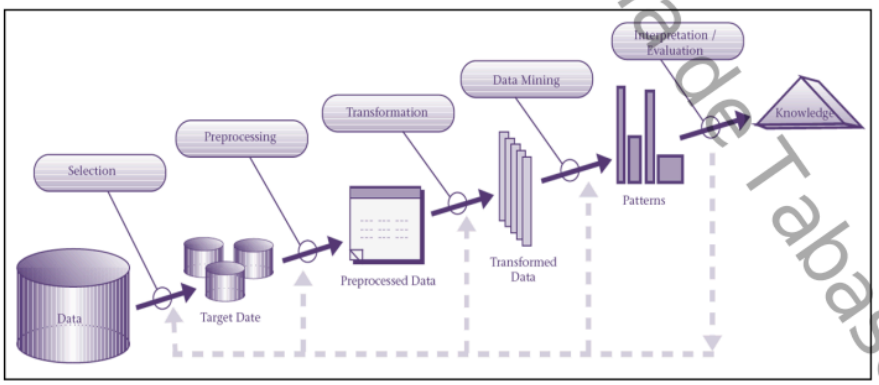


Ilustración 3. Etapas de un proceso KDD
(Tabuenca, 2010)

1. **Selección del conjunto de datos.-** El primer paso en el proceso de extracción de conocimiento a partir de datos es precisamente reconocer y reunir los datos con los que se va a trabajar permite desarrollar un entendimiento sobre el dominio de aplicación, identificando el conocimiento previo relevante e identificando el objetivo principal a alcanzar desde el punto del vista del usuario.

2. **Pre-procesamiento.-** Este proceso se tiene que limpiar la fuente de datos para evitar problemas comunes, pueden contener ambigüedades, ruido o, simplemente, no estar en el formato adecuado para su posterior procesamiento.

3. **Transformación.-** Transformación a un formato específico de la herramienta, por ejemplo fuzzificación, normalización o rotación. La transformación de datos engloba, en realidad, cualquier proceso que modifique la forma de los datos. Prácticamente todos los procesos de preparación de datos involucran algún tipo de transformación. Existen distintas operaciones que transforman atributos, algunas transforman un conjunto de atributos en otros, o bien derivan nuevos atributos, o bien cambian el tipo (mediante numerización o discretización) o el rango (mediante escalado).

4. **Minería de Datos.-** Todas las etapas anteriores son necesarias para que la aplicación de un algoritmo de minería sea exitoso y se puede sacar el conocimiento implícito en los datos que interesan.

   El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar 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.

5. **Análisis e Interpretación.-** En esta etapa se estudia, interpreta y evalúa el modelo de conocimiento generado por el algoritmo de minería de datos. Así como en la mayoría de los casos se evalúa la calidad de las hipótesis de la manera más exacta posible.

2.3.5.2 Extracción de patrones

Como se menciono anteriormente el KDD tiene como objetivo encontrar patrones en los datos que sean validos, novedosos y comprensibles para ser convertidos en conocimiento. De igual forma se describieron las etapas que lo integran, una de ellas es el proceso de minería de datos que puede ser descrito como la vista minable que serian los datos entrantes, así como conocimiento previo, criterios de calidad. Como segunda tarea las técnicas de minería de datos que funcionan como filtro de datos que permite obtener patrones de comportamiento. La ilustración 4 muestra gráficamente la descripción.

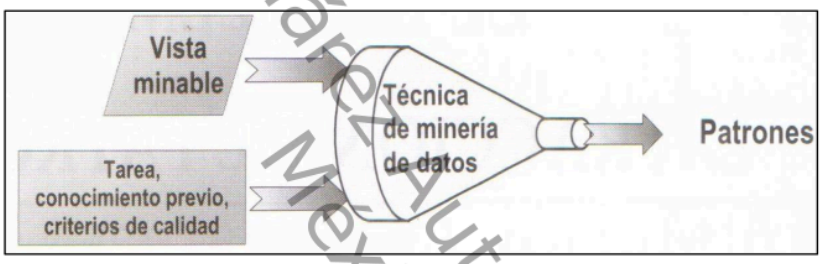


Ilustración 4. Proceso de minería de datos
(Hernández, J., Ramírez, M. & Ferri, C. 2004)

2.3.6 Modelos de minería de datos

Según Hernando, R. (2004), los modelos de minería de datos están clasificados en:

- **De verificación.-** El usuario solicita que se verifique cierta hipótesis, cuando se le responde puede refinar su pregunta, y así sucesivamente.
- **De descubrimiento.-** Con este método se descubre nueva información que no estaba previamente en el almacén de datos (o, en su caso, en las bases de datos. No necesita intervención por parte del usuario. Se buscan patrones en los datos, o bien elementos fuera de la norma.

- **Predictivo.-** Se realizan predicciones sobre el comportamiento futuro de variables a partir de los patrones existentes en los datos. El usuario indica sobre que variable quiere obtener la predicción.

2.3.6.1 Técnicas de minería de datos

Las técnicas de minería de datos se clasifican en dos categorías: Supervisado o predictivo y No supervisado o de descubrimiento del conocimiento (García et al., 1997).

- **Supervisado o predictivo.-** Predicen el valor de un atributo de una conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.

  A continuación se describen algunas técnicas de este tipo:

  - **Arboles de decisión.-** Su representación es en forma de árbol en donde cada nodo es una decisión, los cuales a su vez generan reglas para la clasificación de un conjunto de datos. Los árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación.

    - **Algoritmo C4.5 en WEKA (J48)**

    La clase en la que se implementa el algoritmo C4.5 en la herramienta WEKA es weka.classifers.j48.J48.java. Algunas propiedades concretas de la implementación son las siguientes:

    En primer lugar, en cuanto a los tipos de atributos admitidos, estos pueden ser simbólicos y numéricos. Se permiten ejemplos con faltas en dichos atributos, tanto en el momento de entrenamiento como en la predicción de dicho ejemplo. En cuanto a la clase, ésta debe ser simbólica.

- El algoritmo no posibilita la generación de reglas de clasificación a partir de árbol de decisión.

  - **Inducción neuronal.-** Las técnicas de inducción son capaces de trabajar con datos incompletos e incluso paradójicos, que dependiendo del problema puede resultar una ventaja o un inconveniente.
  - **Regresión.-** El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo, o sobre un conjunto de variables.

    Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

  - **Series de tiempo.-** Las series de tiempo en la minería de datos permiten buscar patrones a partir de grandes cantidades de datos. Algunas de sus variables están en función del tiempo. Esta técnica se utiliza a partir del comportamiento histórico de los datos, que permite modelar los componentes básicos de la serie, y así se logra hacer predicciones

- **No supervisados o descubrimiento de conocimiento.-** Estos se utilizan cuando una aplicación no están madura para realizar tareas de predicción, la función primordial es que descubre patrones y tendencias de los datos actuales

  A continuación se describen algunas técnicas de este tipo:

  - **Agrupación.-** Agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases. Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas.

Este método debido a su naturaleza flexible se puede combinar fácilmente con otro tipo de técnica de minería de datos, dando como resultado un sistema híbrido. Un problema relacionado con el análisis de clúster es la selección de factores en tareas de clasificación, debido a que no todas las variables tienen la misma importancia a la hora de agrupar los objetos.

- **Algoritmo SimpleKmeans**

El algoritmo de k-medias se encuentra implementado en la clase weka.clusterers.SimpleKMeans.java. se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado cuantos clusters se van a crear, éste es el parámetro k, para lo cual se seleccionan k elementos aleatoriamente, que representaran el centro o media de cada clúster. A continuación cada una de las instancias, ejemplos, es asignada al centro del clúster más cercano de acuerdo con la distancia Euclidea que le separa de él. Para cada uno de los clusters así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos clusters. Finalmente se repite el proceso completo con los nuevos centros de los clusters. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración.

- **Asociaciones.-** Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta. Ejemplo: en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.
- **Reglas Generales.-** Patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

- **Correlaciones y factorizaciones.-** son enfocados exclusivamente en los atributos numéricos. Puede existir relaciones bidireccionales o no orientadas para ver si existe algún tipo de orientación (causa/ efecto) se pude utilizar modelos de regresión.

- **Detección de valores e instancias anómalas.-** el objetivo es encontrar aquellas instancias que no son similares a ninguna de las otras instancias. La manera de abordar el problema es generalmente la de agrupación los ejemplos y ver aquellas instancias que se quedan desplazadas de los grupos mayoritarios.

2.3.7 Metodología de minería de datos

En el desarrollo un proyecto existe la necesidad de adquirir un forma de trabajo; un metodología que permita comprender y seguir paso a paso cada una de las fases, una metodología que explique cuando se debe hacer cada actividad y la razón.

2.3.7.1 CRISP-DM (Cross Industry Standard Process for Data Mining)

El inicio de CRISP-DM fue en 1999 cuando un importante consocio de empresa europeas proponen a parir de diferentes versiones de KDD (Knowledge Discovery in Databases) el desarrollo de una guía de referencia de libre distribución denomina CRISP-DM [Cross Industry Standard Process for Data Mining] (Arancibia, 2010).

Este modelo se compone de 6 niveles de abstracción organizados de forma jerárquica. El modelo de referencia CRISP-DM proporciona una descripción del ciclo de vida del proyecto de minería de datos. Es la guía de referencia con más amplitud utilizada en el desarrollo de proyectos de Data Mining.

Su característica es el movimientos adelante y atrás entre cada fase es necesario. Debido a que el resultado de cada etapa da la pauta a que fase o tarea de una fase tiene que se realizado después.

Una de sus propiedades es cíclica lo que se ve reflejado en el círculo externo, la secuencia de las fases no es rígida.

A continuación se describen cada fase en la que esta divida la modelo de referencia CRISP-DM:

1. **Comprensión del negocio o análisis del problema**

La primera etapa de la metodología permite la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional. En esta fase es muy importante tener la capacidad de convertir el conocimiento adquirido en un problema de minería de datos, así como desarrollar los pasos preliminares para lograr los objetivos del negocio con herramientas de minería de datos.

La fase comprensión del negocio es sin duda la mas importante, esta constituida por las siguientes tareas, que se muestra en la ilustración 5.

- **Evaluación de la situación.-** Se realiza búsqueda detalla de los recursos, restricciones, suposiciones y otros factores al definir el plan de proyecto. Se realiza una descripción de los datos y conocimientos previos existentes.

- **Determinación de los objetivos.-** Es indispensable examinar los objetivos del negocio, para abordar aquellas cuestiones que son relevantes y beneficias para organización. En esta etapa se conoce el negocio, en el caso de que el conocimiento no se suficiente se tendrá que entrevistar e involucrar a los expertos en el dominio del negocio.

- **Establecimiento de los objetivos de minería.-** Los objetivos deben ser establecidos en términos de minería de datos, y no en de negocio.

- **Generación del plan de proyecto.-** El plan de proyecto es salida principal debe incluir todas las etapas a desarrollar en el proyecto.

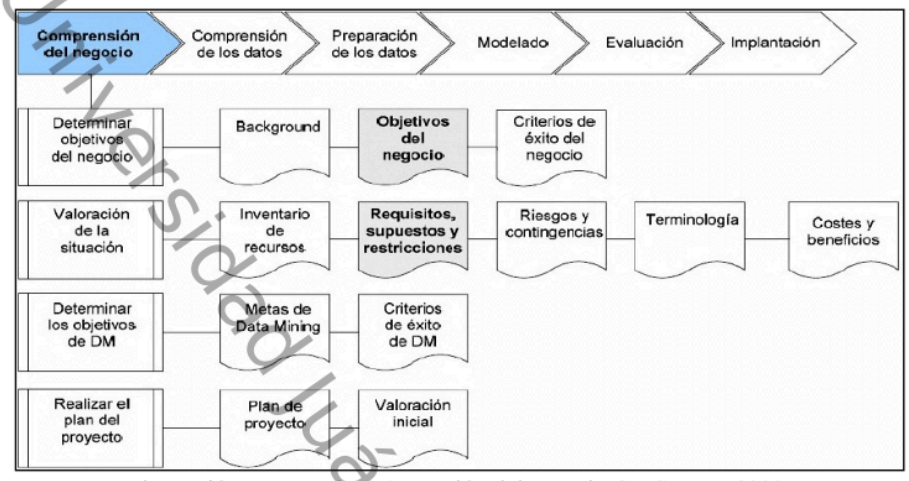


Ilustración 5. Fases de comprensión del negocio CRISP-DM 2000
(Arancibia, 2010)

2. Fase de compresión de los datos

En esta fase se hace el entendimiento de datos y todas aquellas actividades relacionadas con la limpieza de datos, identificación de problemas relacionados con la adquisición de los datos, procedimientos para determinar la calidad de datos con el objetivo de establecer el primer contacto con el problema. En esta etapa se empiezan a establecer los subconjuntos iniciales de datos que puedan contener la información que se esta buscando. Las tareas con las que cuenta esta fase se describen a continuación y muestran gráficamente en la ilustración 6.

- **Recolección inicial de los datos.-** Se realiza una especificación de los datos obtenidos, la forma que se adquirieron y algunos de los problemas encontrados. Esta fase ayuda a las observaciones futuras para la investigación.
- **Descripción de los datos.-** Se describen los datos que han sido adquiridos, incluyendo el formato, la cantidad de datos (el numero de registros y campos).

- **Exploración de los datos.-** Se describen los resultados de las tarea, este análisis directamente pueden dirigir los objetivos de la minería de datos. También se realiza la preparación de datos necesarios para análisis futuros.

- **Verificación de la calidad de los datos.-** Se describen los resultados de la verificación de la calidad de datos, así como los problemas que presentan y las soluciones que se tomaron.

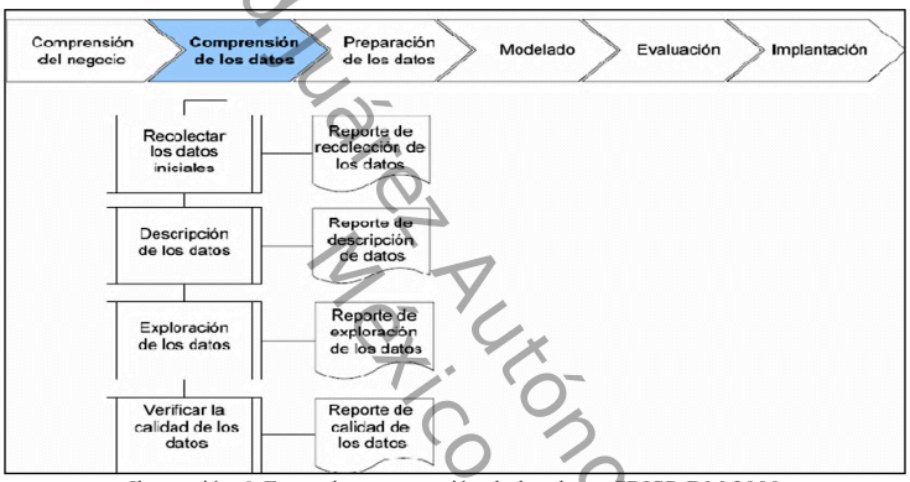


Ilustración 6. Fases de comprensión de los datos CRISP-DM 2000
(Arancibia, 2010)

3. Fase preparación de los datos

Después de la recolección de los datos se prosigue a la preparación para adaptarlos a las técnicas de minería de datos que se utilizaran 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. Estas se muestran en la ilustración 7.

- **Seleccionar los datos.-** Se describen los datos seleccionados para la minería de datos, se realiza una especificación de los métodos utilizados para su elección. También se menciona los datos que serán incluidos o excluidos y los motivos para estas decisiones.
- **Limpieza de los datos.-** Se describe las acciones que fueron tomadas para dirigir los problemas de calidad de los datos. La transformación que sufrieron los datos y el impacto.
- **Construcción de los datos.-** Se describe como que procedimientos fueron pasando los datos para llegar a ser una vista minable. También puede incluir como construcción de operaciones de preparación de datos tales como la producción de atributos derivados o el ingreso de nuevos registros, o la transformación de valores para atributos existentes.

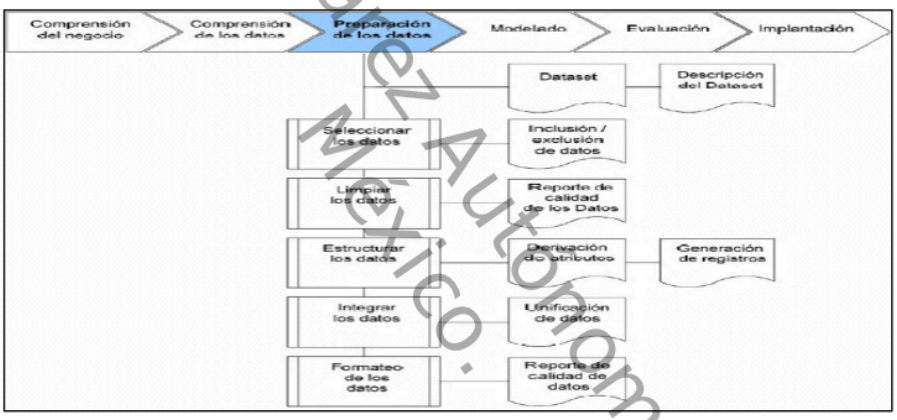


Ilustración 7. Fases de la preparación de los datos CRISP-DM 2000
(Arancibia, 2010)

4. Fase de modelado

En esta fase se hace la elección de las técnicas de modelado que será usada en los datos, se debe tener seleccionada una herramienta de minería de datos para identificarla dentro de la misma, así mismo en esta fase se estudian y ajustan parámetros con los valores corrector para le proyecto. Las técnicas se eligen de acuerdo a criterios como: ser apropiada al problema, disponer de datos adecuados, cumplir los requisitos del problema, disponer del tiempo adecuado para obtener un modelo, tener conocimiento de la técnica.

Esta fase también está constituida por tareas como: selección de la técnica de modelado, generación del plan de prueba, evaluación del modelo, construcción del Modelo. El orden de ejecución se describen a continuación y puede ser visto claramente la ilustración 8.

- **Seleccionar técnicas de modelado.-** Se debe seleccionar las técnicas que se usaran. Es la descripción de los algoritmos, por ejemplo, un árbol de decisión o el clustering (agrupación).

- **Generar el plan de prueba.-** Se describe el plan que se llevara para el entrenamiento de los datos, la prueba de los modelos. También en el se ejecutan algoritmos que permitan explorar previamente el conjunto de datos.

- **Construcción de los modelos.-** En función del plan de prueba se construyen esquemas que proporcionen de forma grafica y descriptiva como se evaluaran finalmente las bases de datos.

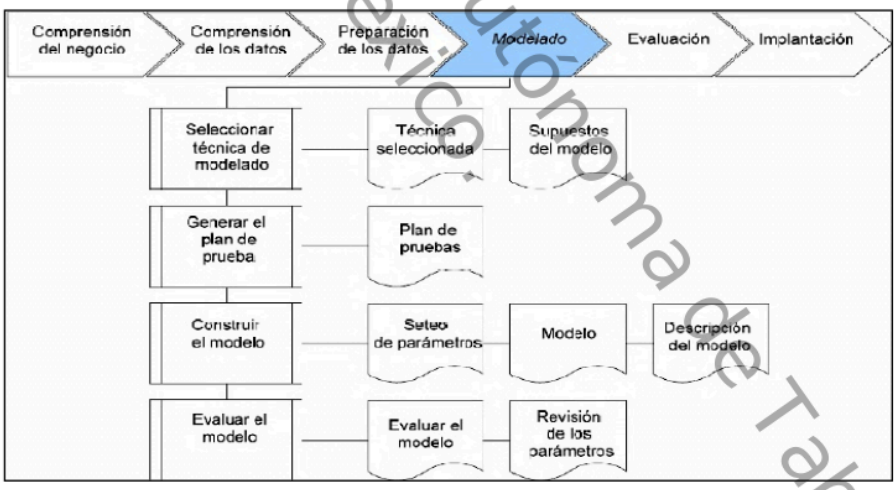


Ilustración 8. Fase del modelado CRISP-DM 2000
(Arancibia, 2010)

5. Fase de evaluación

En esta etapa se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. De la misma manera se debe considerarse, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Las tareas involucradas son proceso de revisión, evaluación de los resultados son mostradas en la ilustración 9.

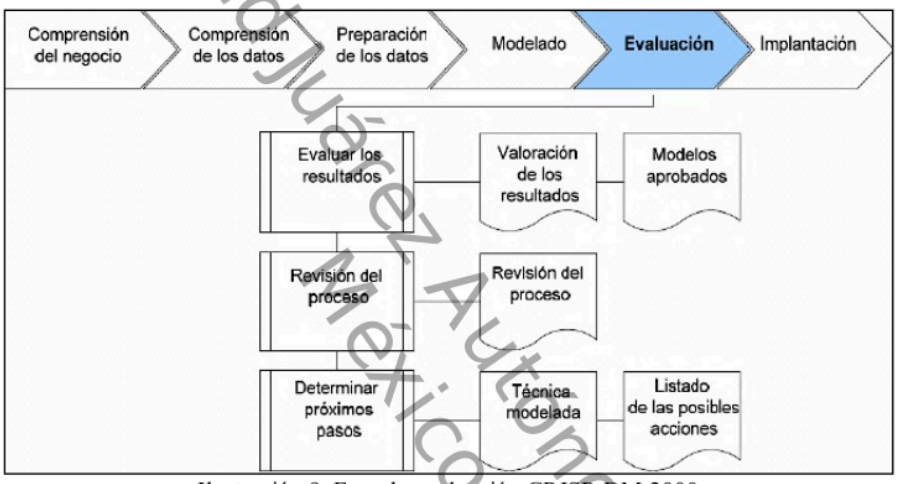


Ilustración 9. Fase de evaluación CRISP-DM 2000
(Arancibia, 2010)

6. Fase de implementación

En la fase anterior se tratan factores como la exactitud y la generalidad del modelo. Ya en esta fase el modelo está construido y validado, se transforma el conocimiento obteniendo en acciones dentro del proceso de negocio. El analista puede realizar recomendaciones basadas en la observación del modelo y sus resultados. También, la evaluación comprueba otros resultados generados por la minería de datos. Los resultados de la minería de datos implican modelos que irremediablemente son relacionados con los objetivos del negocio y todos los otros descubrimientos que nos son relacionados con los objetivos del mismo, pero con la cualidad que pueden revelar desafíos adicionales, información, o insinuaciones para futuras direcciones.

Las tareas que se ejecutan en esta fase son: plan de implantación, informe final que se muestran en la ilustración 10.

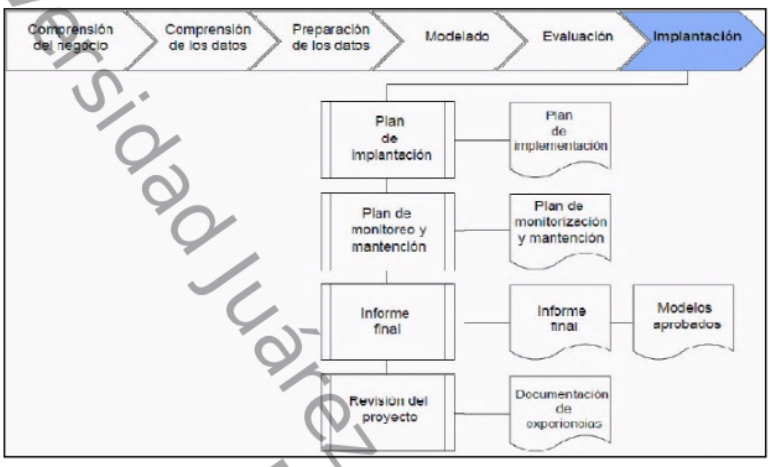


Ilustración 10. Fase de implementación CRISP-DM 2000
(Arancibia, 2010)

2.4 Marco tecnológico

En esta sección se mencionan las herramienta de software de minería de datos, estadísticos y hoja de calculo que fueron utilizadas para el desarrollo del proyecto.

2.4.1 Software de minería de datos WEKA

Weka (Waikato Environment for Knowledge Analysis)

Weka es una plataforma de software para el aprendizaje automático y minería de datos desarrollado por la universidad de Waikato (Nueva Zelanda) escrito en lenguaje java, este software tiene una extensa colección de algoritmos de Máquinas de conocimiento; útiles para ser aplicados sobre datos mediante los interfaces que ofrece o para embeberlos dentro de cualquier aplicación. (Weka, 2013)

Weka contiene las herramientas necesarias que permiten realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. Está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla. (Morate, 2011). Weka está disponible libremente bajo la licencia pública general de GNU tiene la propiedad de ser muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma, así mismo es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

Las características principales son:

- **Pre-procesamiento:** selección de atributos, discretización, tratamiento de valores desconocidos y transformación de datos.
- **Modelos de aprendizaje:** arboles de decisión, tablas de decisión, reglas de asociación, métodos de agrupación.
- **Visualización:** la interfaz grafica se compone de diversos entornos que permiten controlar todas las operaciones de modelos de aprendizaje.

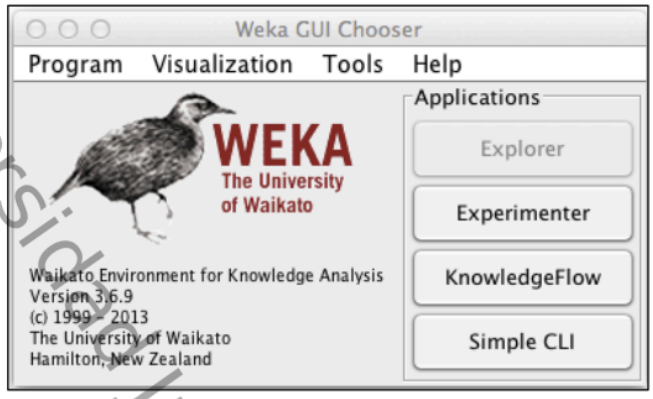


Ilustración 11. Pantalla inicial de WEKA

2.4.2 Software estadístico PSPP

PSPP es un programa para el análisis estadístico de los datos incluidos en una muestra. Funciona como reemplazo libre para el programa propietario SPSS y muy similar aunque con algunas excepciones. (GNU PSPP, 2013)

No tiene límites artificiales en el número de casos o variables que se pueden utilizar. PSPP puede realizar estadística descriptiva, pruebas t, ANOVA, regresión lineal y logística, análisis de conglomerados, análisis factorial, pruebas no paramétricas y más. Puede ser utilizado con su interfaz gráfica o los comandos de sintaxis más tradicionales.

Las características principales son:

- Soporta más de 1 mil millones de casos.
- Soporta más de 1000 millones variables.
- Los archivos de sintaxis y los datos son compatibles con los de SPSS.
- Selección de texto, PostScript, pdf , opendocument formatos de salida o HTML.
- Inter-opera con Gnumeric , LibreOffice , OpenOffice.Org y otro software libre.
- Fácil importación de datos desde hojas de cálculo, archivos de texto y fuentes de bases de datos.

PSPP está disponible libremente bajo la licencia pública general de GNU, es multiplataforma se ejecuta en muchas computadoras diferentes y muchos sistemas operativos, es de muy fácil manejo ya que su interfaz grafica muestra todas herramientas de manera que el usuario las identifique con facilidad.

2.4.3 Hoja de cálculo

Excel es un software de hoja de calculo que posee herramientas de visualización de datos que ayudan a realizar un seguimiento y resaltar tendencias en los datos. Excel es parte de la paquetería de Office de Microsoft.

Sus funciones principales es de escribir e importar datos, administración de hojas y libros, ordenar y filtrar datos imprimir, resaltar datos, Automatizar tareas mediante macros, tablas dinámicas, compartir y colaborar, creación de formulas, tablas, diagramas y gráficos. (Microsoft-Soporte 2013)

Lo que respecta a formatos de archivo y compatibilidad tiene una gran numero de extensiones en las que un archivo de hoja de calculo puede ser exportando. Uno de ellos es el formato de archivo * CSV que es un archivo de texto que utiliza comas para separar valores en las celdas.

Otro formato es el *.prn texto delimitado por espacios es un archivo de texto que usa espacios para separar los valores de las celdas. No se conservan las propiedades de celda, fórmulas, gráfico y otros formatos. Formato de intercambio de datos (.dif) es formato de archivo de texto que se puede usar para intercambiar datos con otras aplicaciones de hojas de cálculo. No se conservan las propiedades de celda, fórmulas, gráfico y otros formatos.

Capítulo III. Aplicación de la metodología y desarrollo

La metodología de investigación es la base para la solución de cualquier hipótesis o problema. En este apartado se hace mención del tipo de técnica de muestreo aplicada, descripción del universo de estudio y el desarrollo del modelo de minería de datos CRISP-DM.

3.1 Elección del grupo objeto

La elección del grupo objeto se realizó considerando la accesibilidad para proporcionar la información, así como la especialización del médico, permitiendo tener una mejor comprensión del problema.

El muestreo en investigaciones cualitativas se debe elegir la menor cantidad de sujetos que proporcionen la mayor cantidad de datos ricos para el estudio (Patton, 2002). De acuerdo a lo mencionado se tomo una muestra cualitativa de cuatro médicos internistas quienes fueron los colaboradores de la investigación.

3.2 Descripción del universo de estudio

El universo o población de estudio de esta investigación lo constituyen los pacientes prediabéticos. Las bases de datos contempladas para esta investigación fueron ENSANUT2006 y 2012.

3.3 Modelo de referencia CRISP-DM

Introducción

El desarrollo de proyectos de minería de datos, requiere la aplicación de una metodología que permita adaptar, guiar y orientar los objetivos del plan, de acuerdo a un conjunto de procedimientos, técnicas y una gran cantidad de información que permite alcanzar estrategias y facilidades en el desarrollo de un proceso de minería.

El modelo de proceso CRISP-DM para minería de datos suministra una delineación del ciclo de vida del proyecto de minería de datos. Esta posee las fases de un proyecto, sus tareas y las relaciones entre tareas. Las fases que la conforman este modelo son las siguientes:

- *Comprensión del negocio o análisis del problema:* en esta fase se hace la identificación de las expectativas y requerimientos. Permite tener una visión general del problema y realizar un plan preliminar en el cual se pueda dar solución y alcanzar el objetivo planteado.

- *Comprensión de los datos:* Se realiza la recolección de los datos, lo cual permite establecer un contacto inicial con el problema. Esta fase tiene como finalidad evitar problemas inesperados durante la fase de preparación de los datos.

- *Preparación de los datos:* En esta fase se llevan acabo todas las tareas (Selección, limpieza y transformación de los datos) con el objetivo de construir el conjunto final de datos que serán utilizados en la herramienta de minería de datos.

- *Modelado:* Se realiza la selección de las técnicas de minería de datos que mas se apropien y que en un final permitan alcanzar el objetivo del proceso de minería de datos. También se realiza generación de pruebas, evaluación del modelo y construcción del mismo.

- *Evaluación:* En la fase de evaluación, se evalúa el modelo teniendo en cuentas los criterios de éxito del problema. Aquí se realiza la documentación y presentación de los resultados de manera comprensible para lograr una incremento del conocimiento del que se tiene.
- *Implementación:* Se trata de explotar la potencialidad de los modelos, intégralos en los proceso de toma de decisiones en la organización y difundir informes del conocimiento extraído.

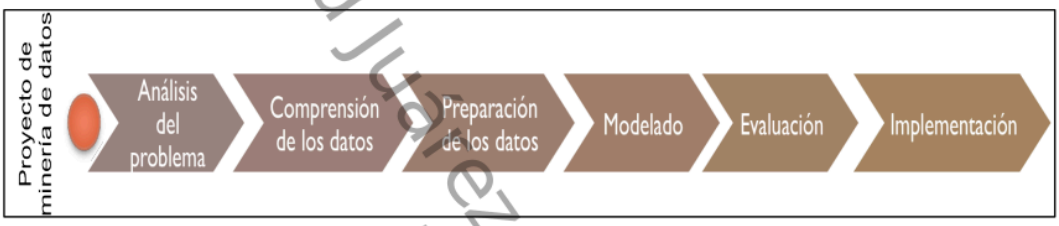


Ilustración 12. Fases de metodología CRISP-DM
(BI Analytics)

3.4 Análisis del problema o comprensión del negocio

En la actualidad las enfermedades no trasmisibles como la diabetes tiene gran prevalencia en México. Según datos de la Encuesta Nacional de Salud y Nutrición 2012 (ENSANUT) la razón de adultos con diagnostico médico previo de diabetes es de 9.2% en comparación con ENSANUT 2006 que era de 7%, lo que muestra que este tipo de padecimientos va en aumento (ENSANUT, 2012).

La Encuesta Nacional de Salud y Nutrición 2012 específica que la prevalencia mas alta de edades que presentan diabetes esta en un rango de 60 a 79 años, siendo mayor en mujeres con un 26.3% y 24.15% para los hombres. (ENSANUT 2012). En el estado de Tabasco la prevalencia de adultos con diagnostico médico previo de diabetes esta entre 9.3 a 10.1% (ENSANUT, 2012).

En busca de soluciones para el Sistema Nacional de Salud, surge lo que es el programa de acción especifica que tiene como objetivo promover y fomentar la generación de información y conocimiento relevantes respecto a tecnologías para la salud.

La Evaluación de Tecnologías en Salud es una actividad compleja, que comprende conocimientos y prácticas provenientes de diversas áreas a saber: investigación básica y aplicada, epidemiología, ingeniería etc. con el propósito fundamental de apoyar la toma de decisiones (Secretaria de Salud, 2012).

El Plan Nacional de Desarrollo 2013-2018 establece como línea de acción el Instrumentar mecanismos que permitan homologar la calidad técnica e interpersonal de los servicios de salud. Así como instrumentar acciones para la prevención y control del sobrepeso, obesidad y diabetes (Plan Nacional de Desarrollo, 2013).

Buscando aportar soluciones tecnológicas que permitan conocer el comportamiento que tienen los grandes conjuntos de datos almacenados en los sistemas de información de la secretaria de salud surgió este proyecto de dominado "Detección de patrones de comportamiento utilizando técnicas de minería de datos en expedientes clínicos de pacientes pre-diabéticos".

Para el cumplimiento de la investigación se estableció el objetivo que fue obtener patrones de comportamiento utilizando técnicas de minería de datos, como apoyo a la toma de decisiones para el control de la diabetes.

3.4.1 Evaluación de la situación actual

Se contó con información especifica de la Encuesta Nacional de Salud y Nutrición que contiene parámetros que permiten identificar características que ayudaron a describir el estado de salud de la población.

Todas las bases de datos están regladas y con la descripción documental de los atributos de la encuesta. Actualmente las bases de datos se encuentra disponibles en la página *Web* de ENSANUT.

Recursos de software y Hardware

Para el desarrollo del proyecto se contó con:

- Software de minería de datos (WEKA).
- Software de Hoja de calculo (Excel).
- Software PSPP.
- Equipo para el entrenamiento de los modelos de datos.
- Acceso irrestricto a la base de datos ENSANUT.

Fuentes de datos y conocimiento

Se contó con la siguiente información:

A. Base de datos adultos contiene datos correspondientes a personas de 20 a 59 años con información vinculada a depresión, Diabetes, Hipertensión, antecedentes de enfermedades, Sobrepeso y obesidad, salud reproductiva, vacunación, accidentes, agresión y violencia.

B. Base de datos adultos contiene datos correspondientes a datos clínicos de personas.

3.4.2 Plan del proyecto

Para cubrir con toda la investigación se elaboró un plan de proyecto que permitió identificar y gestionar la ejecución de las tareas que se debían seguir y poder llevar a cabo el proceso de minería de datos. A continuación se hace mención de cada una de ellas:

- Etapa 1. Se recolectaron los datos para el proceso de minería de datos.
- Etapa 2. Se preparó la información haciendo limpieza de los datos.
- Etapa 3. Se aplicaron algoritmos de selección de atributos para identificar variables.
- Etapa 4.Se Seleccionaron las técnicas de minería de datos mas se ajustaron a el objetivo de proceso de minería de datos.

- Etapa 5. Se aplicó técnicas de minería de datos a los data sets resultado de la limpieza de datos.
- Etapa 6. Se analizaron los resultados obtenidos de cada técnica de minería de datos.
- Etapa 7. Se redactó el reporte final de los resultados obtenidos.

3.5 Comprensión de los datos

3.5.1 Reporte de recolección de los datos

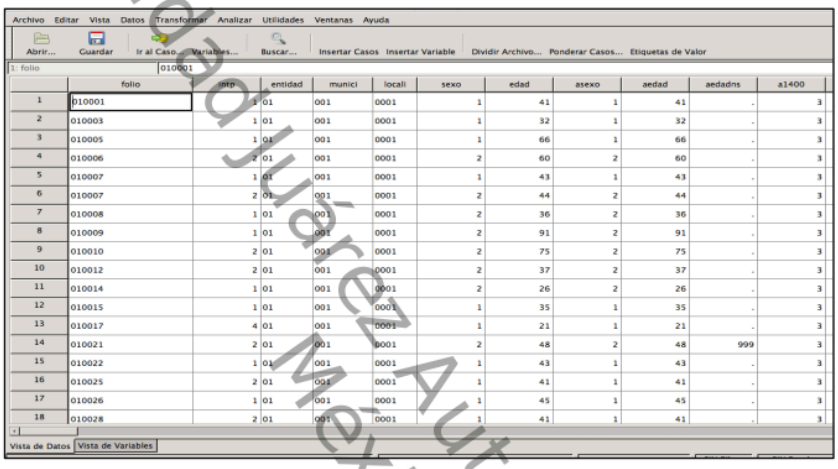
La recolección de la información se realizó de la Encuesta Nacional de Salud y Nutrición 2006 y 2012 que es información de un conjunto de datos fidedignos que la Secretaria de Salud que da a conocer sobre las condiciones de salud de la población en México.

Estos datos fueron recabados de 50,280 hogares donde fueron aplicados 96,031 cuestionarios individuales en los diferentes grupos de edad, así como 14104 cuestionarios de utilizadores de servicios de salud por toda la republica Mexicana. Para la obtención de la información se realizó un registro previo en la página ENSANUT y poder tener acceso a las diferentes bases de datos.

Ilustración 13. Pantalla de acceso a bases de datos ENSANUT (ENSANUT, 2012)

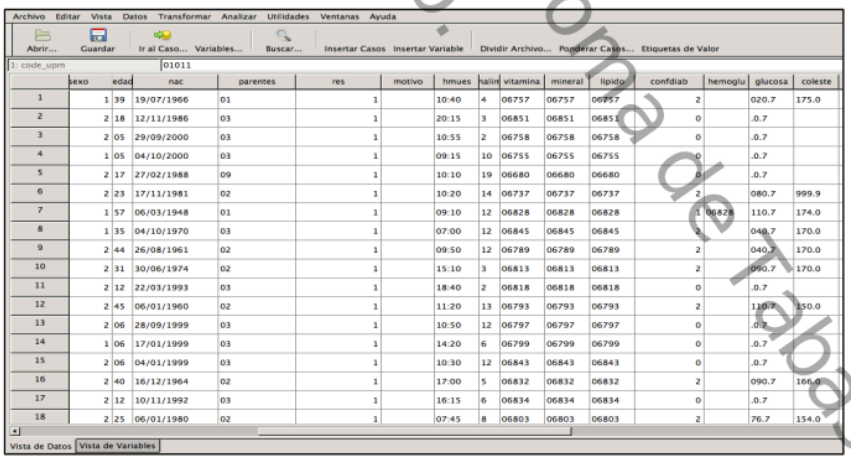
3.5.2 Reporte de descripción inicial de los datos

Las bases de datos seleccionadas para ser analizadas a través minería de datos son: glucosa y lípidos que es donde se representa los valores de un expediente clínico. Así como adultos de 20 años o más. Los almacenes de datos se encontraron en formato .SAV de SPSS.



	folio	tipo	entidad	munici	locali	sexo	edad	asexo	aedad	aedadns	a1400
1	010001	1	01	001	0001	1	41	1	41	.	3
2	010003	1	01	001	0001	1	32	1	32	.	3
3	010005	1	01	001	0001	1	66	1	66	.	3
4	010006	2	01	001	0001	2	60	2	60	.	3
5	010007	1	01	001	0001	1	43	1	43	.	3
6	010007	2	01	001	0001	2	44	2	44	.	3
7	010008	1	01	001	0001	2	36	2	36	.	3
8	010009	1	01	001	0001	2	91	2	91	.	3
9	010010	2	01	001	0001	2	75	2	75	.	3
10	010012	2	01	001	0001	2	37	2	37	.	3
11	010014	1	01	001	0001	2	26	2	26	.	3
12	010015	1	01	001	0001	1	35	1	35	.	3
13	010017	4	01	001	0001	1	21	1	21	.	3
14	010021	2	01	001	0001	2	48	2	48	999	3
15	010022	1	01	001	0001	1	43	1	43	.	3
16	010025	2	01	001	0001	1	41	1	41	.	3
17	010026	1	01	001	0001	1	45	1	45	.	3
18	010028	2	01	001	0001	1	41	1	41	.	3

Ilustración 14. Pantalla de SPSS con la base de datos Adultos 20 años o más



	sexo	edad	nac	parentes	res	motivo	hmues	halim	vitamina	mineral	lipido	confdiab	hemoglu	glucosa	coleste
1	1	39	19/07/1966	01	1		10:40	4	06757	06757	06757	2		020.7	175.0
2	2	18	12/11/1986	03	1		20:15	3	06851	06851	06851	0		.0.7	
3	2	05	29/09/2000	03	1		10:55	2	06758	06758	06758	0		.0.7	
4	1	05	04/10/2000	03	1		09:15	10	06755	06755	06755	0		.0.7	
5	2	17	27/02/1988	09	1		10:10	19	06680	06680	06680	0		.0.7	
6	2	23	17/11/1981	02	1		10:20	14	06737	06737	06737	2		080.7	999.9
7	1	57	06/03/1948	01	1		09:10	12	06828	06828	06828	1	06828	110.7	174.0
8	1	35	04/10/1970	03	1		07:00	12	06845	06845	06845	2		040.7	170.0
9	2	44	26/08/1961	02	1		09:50	12	06789	06789	06789	2		040.7	170.0
10	2	31	30/06/1974	02	1		15:10	3	06813	06813	06813	2		090.7	170.0
11	2	12	22/03/1993	03	1		18:40	2	06818	06818	06818	0		.0.7	
12	2	45	06/01/1960	02	1		11:20	13	06793	06793	06793	2		110.7	150.0
13	2	06	28/09/1999	03	1		10:50	12	06797	06797	06797	0		.0.7	
14	1	06	17/01/1999	03	1		14:20	6	06799	06799	06799	0		.0.7	
15	2	06	04/01/1999	03	1		10:30	12	06843	06843	06843	0		.0.7	
16	2	40	16/12/1964	02	1		17:00	5	06832	06832	06832	2		090.7	166.0
17	2	12	10/11/1992	03	1		16:15	6	06834	06834	06834	0		.0.7	
18	2	25	06/01/1980	02	1		07:45	8	06803	06803	06803	2		76.7	154.0

Ilustración 15. Pantalla de SPSS con la base de datos Glucosa y Lípidos

La primera base de datos seleccionada es glucosa y lípidos se encuentra conformada por 22,815 instancias y 36 atributos iniciales. Esta base de datos muestra 12 atributos que son pruebas rápidas de sangre y el resto de atributos son de pruebas realizadas en laboratorio. A continuación pueden ser vistos en la tabla 1.

Variable	Descripción
ent	Código de la entidad
mun	Código del municipio
Sexo	Sexo
Edad	Edad
Parentes	Parentesco con el jefe del hogar
Res	Resultado de sangre venosa
halim	Hora del ultimo alimento
CONFDIAB	Confirmación de persona diabética
Glucosa	Glucosa
Colesterol	Colesterol
Res2	Resultado de la Hemoglobina
Hemo	Hemoglobina
Variables de determinación de laboratorio	
Col	mg/dL Colesterol total determinado mediante inmunoanalizador
CRP	mg/L Proteína C reactiva determinada mediante nefelómetro
GLU	mg/dL Glucosa determinada mediante analizador clínico
HBGlu	Porcentaje de Hemoglobina glicosilada determinada mediante analizador clínico
HCy	micromoles/L Homocisteina determinada por HPLC
HDLC	mg/dL Colesterol de alta densidad determinado mediante analizador clínico
Ins	Micro U/ml de Insulina determinada mediante inmunoanalizador
LDL	Lipoproteínas de Baja Densidad determinadas mediante analizador clínico
Trig	Triglicéridos determinados mediante analizador clínico
VLDL	Lipoproteínas de Muy Baja Densidad determinadas mediante analizador clínico

Tabla 1. Descripción inicial de los datos(Base de datos glucosa y lípidos)

La segunda base de datos seleccionada esta conformada por 46277 instancias y 163 atributos, a continuación en la tabla 2 se especifica algunos de los atributos que corresponden al área de diabetes:

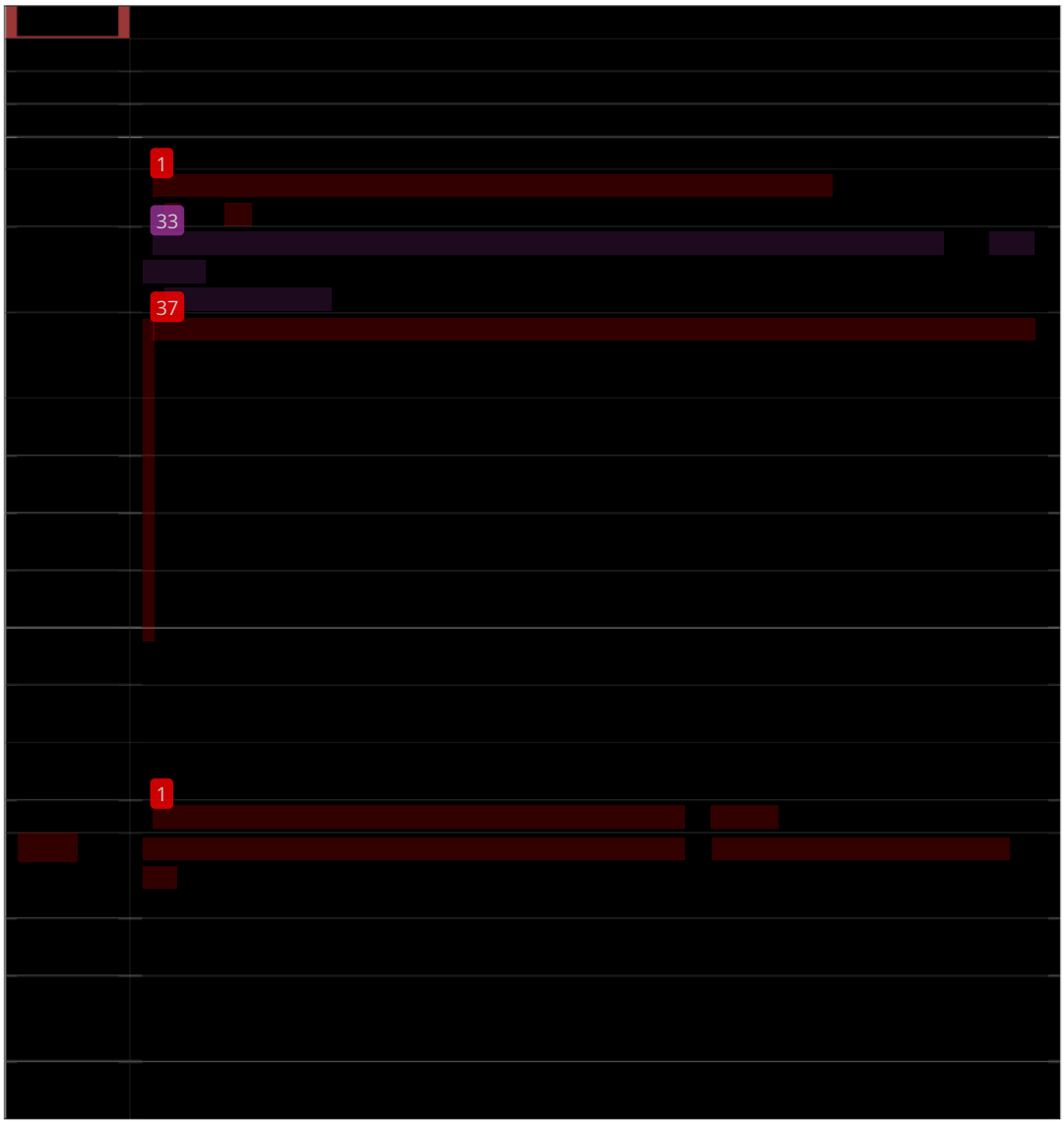


Tabla 2. Descripción inicial de los datos (Base de datos Adultos 20 años o más) 1/3

Variable	Descripción
a306k	¿En que institución se atiende para controlar su diabetes? Otro lugar 1.Si 2.No
a306l	¿En que institución se atiende para controlar su diabetes? NS/NR 1.Si 2.No
a306esp	Especifique
a307	¿Actualmente toma pastillas o le aplican insulina para controlar su azúcar? 1.Si solo insulina 2.Si solo pastillas 3.Ambas 4.Ninguna
a308a	¿Cuantas veces y con que frecuencia se aplica la insulina? FRECUENCIA 1.Diario 2.A la semana 3.NS/NR
a308b	¿Cuántas veces y con que frecuencia se aplica la insulina?
a309a	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Plan de alimentación (Dieta) 1.Si 2.No
a309b	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Si, Realiza algún plan de ejercicio físico 1.Si 2.No
a309c	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Si, Homeopatía (chochos) 1.Si 2.No
a309d	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Si, Herbolaria 1.Si 2.No
a309e	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Si, Medicina alternativa 1.Si 2.No
a309f	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? 1.Si 2.No
a310a	Durante los últimos 12 meses ?Que exámenes le hizo u orden? su médico para vigilar su azúcar? Tiras reactivas en orina 1.Si 2.No
a310b	Durante los últimos 12 meses ?Que exámenes le hizo u orden? su médico para vigilar su azúcar? Tiras reactivas en sangre 1.Si 2.No
a310c	Durante los últimos 12 meses ?Qué exámenes le hizo u orden? su médico para vigilar su azúcar? Examen general de orina 1.Si 2.No
a310d	Durante los últimos 12 meses ?Qué exámenes le hizo u orden? su médico para vigilar su azúcar? Determinación de glucosa en sangre venosa 1.Si 2.No
a310e	Durante los últimos 12 meses ?Qué exámenes le hizo u orden? su médico para vigilar su azúcar? Determinación de hemoglobina glucosilada 1.Si 2.No
a310f	Durante los últimos 12 meses ?Qué exámenes le hizo u orden? su médico para vigilar su azúcar? Examen para medir el nivel de proteínas en la orina (Microalbuminuria) 1.Si 2.No
a310g	Durante los últimos 12 meses ?Que exámenes le hizo u orden? su médico para vigilar su azúcar? Auto monitoreó 1.Si 2.No

Tabla 3. Descripción inicial de los datos (Base de datos Adultos 20 años o más) 2/3

Variable	Descripción
a310h	Durante los últimos 12 meses ?Que exámenes le hizo u orden? su médico para vigilar su azúcar? Ninguno 1.Si 2.No
a311e1	Durante los últimos 12 meses ?Cuántas veces se realizó esta prueba? (Tiras reactivas en orina) 1. Si 2. No
a311e2	Durante los últimos 12 meses ¿Cuantas veces se realizó esta prueba? (Tiras reactivas en sangre)
a311e3	Durante los últimos 12 meses ¿Cuantas veces se realizó esta prueba? (Examen general de orina)
a311e4	Durante los últimos 12 meses ¿Cuantas veces se realizó esta prueba? (Determinación de glucosa en sangre venosa)
a311e5	Durante los últimos 12 meses ¿Cuántas veces se realizó esta prueba? (Determinación de hemoglobina glucosilada)
a311e6	Durante los últimos 12 meses ¿Cuantas veces se realizó esta prueba? (Examen para medir el nivel de proteínas en la orina (Microalbuminuria))
a311e7	Durante los últimos 12 meses ¿Cuantas veces se realizó esta prueba? (Automonitoreo)
a312a	Debido a la diabetes, durante los últimos 12 meses ¿que medidas preventivas ha seguido para evitar complicaciones? Revisión oftalmológica (no lentes) 1.Si 2.No
a312b	Debido a la diabetes, durante los últimos 12 meses que medidas preventivas ha seguido para evitar complicaciones? Toma una aspirina diario 1.Si 2.No
a312c	Debido a la diabetes, durante los últimos 12 meses ¿que medidas preventivas ha seguido para evitar complicaciones? Revisión de pies 1.Si 2.No
a312d	Debido a la diabetes, durante los últimos 12 meses ¿que medidas preventivas ha seguido para evitar complicaciones? Examen general de orina y micro albuminuria 1.Si 2.No
a312e	Debido a la diabetes, durante los últimos 12 meses ¿que medidas preventivas ha seguido para evitar complicaciones? No realiza ninguna medida preventiva 1.Si 2.No
a312f	Debido a la diabetes, durante los últimos 12 meses ¿que medidas preventivas ha seguido para evitar complicaciones? Otro 1.Si 2.No
a312esp	Especifique
a313a	¿Debido a la diabetes ha tenido ulceras en piernas o pies que tarden en sanar más de 4 semanas? 1.Si 2.No
a313b	¿Debido a la diabetes le han amputado alguna parte del cuerpo? 1.Si 2.No
a313c	¿Debido a la diabetes le ha disminuido su visión? 1.Si 2.No
a313d	¿Debido a la diabetes ha sufrido dago en la retina? 1.Si 2.No
a313e	¿Debido a la diabetes ha perdido la vista? 1.Si 2.No
a313f	¿Debido a la diabetes ha perdido la vista? 1.Si 2.No
a313g	¿Debido a la diabetes ha sufrido de un infarto? 1.Si 2.No
a313h	¿Debido a la diabetes sufrió de un coma diabético? 1.Si 2.No
a313i	¿Debido a la diabetes sufre ardor, dolor o perdida de la sensibilidad en la planta de los pies? 1.Si 2.No
a401	¿Algún médico le ha dicho que tiene la presión alta o hipertensión? 1.Si 2.No

Tabla 4. Descripción inicial de los datos (Base de datos Adultos 20 años o más) 3/3

3.5.3 Reporte de exploración de los datos

Las bases de datos seleccionadas una vez de su recolección, se les realizó el proceso de exploración de los datos para identificar propiedades como inconsistencia y calidad de los datos. De este proceso se muestra la siguientes tablas con la descripción de algunas de las variables mas importantes para el estudio. El proceso de selección de variables se detalla en la fase de preparación de los datos.

Variable a301	**¿Algún médico le ha dicho que tiene diabetes o el azúcar alta en la**		
	Valores	Si	4490
		No	41787

Tabla 5. Exploración de los datos (Base de datos Adultos 20 años o más) variable a301

Variable a302c	**¿Hace cuánto tiempo le dijo su médico que tenia diabetes o el**		
	Valores	Meses	0 a 794
		Años	0 a 61

Tabla 6. Exploración de los datos (Base de datos Adultos 20 años o más) variable a302c

Variable a303	**¿Algún médico le ha diagnosticado diabetes durante el embarazo?**		
	Valores	No especifico	43510
		No	2478
		Si	200

Tabla 7. Exploración de los datos (Base de datos Adultos 20 años o más) variable a303

En la ilustración 16 se muestra las gráficas generadas por weka, de la distribución de cada uno de las variables.

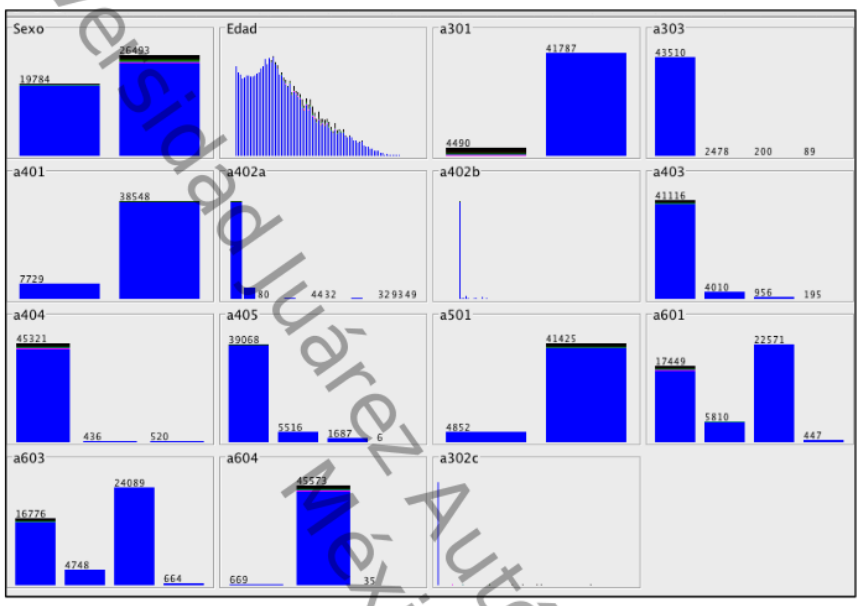


Ilustración 16. Gráficas de las variables de la base de datos Adultos 20 años o más.

La base de datos glucosa y lípidos esta dividida en dos secciones una con variables de determinación por pruebas rápidas y la segunda sección con variables de determinación de pruebas de la laboratorio. A continuación se muestra en tablas los valores que toman los datos de la primera sección.

Variable Sexo		
Valores	1.- Hombre	8984
	2.- Mujer	13831

Tabla 8. Exploración de los datos (Base de datos glucosa y lípidos) variable Sexo

Variable Confdiab		
Valores	0.- Sin especifica	9956
	1.- Si	1904
	2.- No	10952
	3.- Si, embarazo	3

Tabla 9. Exploración de los datos (Base de datos glucosa y lípidos) variable Confdiab

Variables	Valores
Edad	0 a 99 años
Glucosa	0.7 a 999.7
Colesterol	5. a 999

Tabla 10. Exploración de los datos variables Edad, Glucosa, Colesterol

En la tabla 11 se muestra los valores de la sección de valores de determinación por pruebas de laboratorio.

Variables	Valores
Col	10.42 a 1022.188
Glu	8.15 a 737.7
HBGlu	4.9 a 90.6
HCy	2.02 a 55.83
HDLC	2.241 a 144.586
Ins	0 a 354
Trig	3.291 a 2139.321
VLDL	0.658 a 427.864

Tabla 11. Exploración de los datos (Base de datos glucosa y lípidos)

En la ilustración 17 se muestra las gráficas generadas por weka, de la distribución de cada uno de las variables.

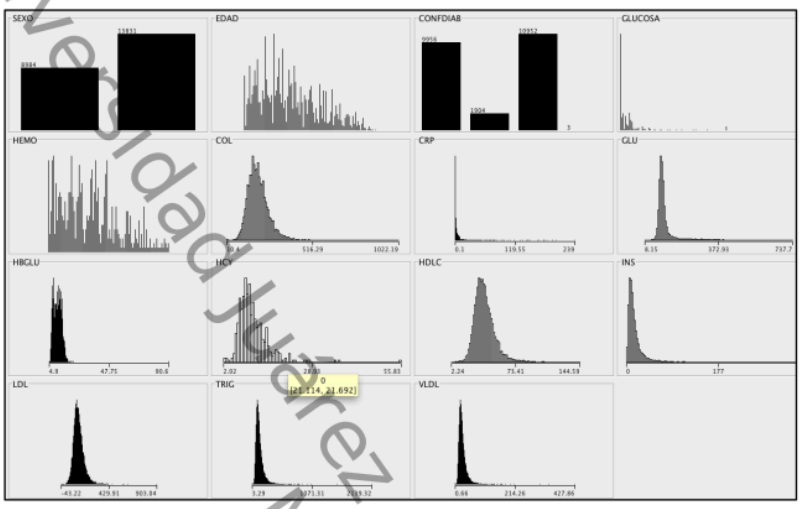


Ilustración 17. Gráfica de las variables de la base de datos Glucosa y Lípidos

3.5.4 Verificación de la calidad de los datos

Posteriormente de explorar los datos se realizó la verificación, para poder determinar que la información este completa o eliminar la existe redundancia. Se realizó a la base adultos la verificación de los datos y se encontró que existía duplicidad de información de las variables Sexo, Edad en las atributos Asexo y Aedad, debido a esto se procedió a eliminar las ultimas dos. Por otra parte el atributo Aedadns no almacena ningún valor en la base de datos de igual forma se realizó la eliminación.

La verificación que se realizó en la base de datos glucosa y lípidos permitió identificar variables como condigo_pum, folio, int, res, res2, vitamina, minerales, que no son atributos relevantes para la investigación, por tal motivo se realizó la eliminación. Después de realizar los procedimientos de recolección, exploración y verificación de los datos se determinó que los datos son apropiados. Cubrieron con los requerimientos necesarios para la obtención de los resultados y lograr los objetivos que es conocer cuales son los patrones de comportamiento de los pacientes pre-diabéticos.

3.6 Preparación de los datos

3.6.1 Datos seleccionados

3.6.1.1 Selección de variables : Base de datos Adultos

Para identificar la relevancia de atributos y realizar una selección, se utilizó la opción Select Attributes. Esta opción supervisada de Weka se divide en dos componentes principales :

- **Método de evaluación:** su función principal es determinar la calidad de subconjunto de atributos.
- **Método de búsqueda:** su función es realizar la búsqueda de conjuntos de atributos.

Seguidamente de conocer su utilidad se realizó como primer acción la carga de los datos, posteriormente se eligió seis métodos de búsqueda, aunado a un método de evaluación. Para identificar las variables más representativas dentro del conjunto de datos, se tomó como punto de partida la variable a301 (¿Algún médico le ha diagnosticado diabetes?) de la sección de diabetes del cuestionario aplicado para la recolección de información. El atributo seleccionado fue evaluado por los seis métodos elegidos obteniendo los siguientes resultados:

Atributo evaluado A301	
Atribute Evaluator: **CfsSubsetEval** Search Method: **Bestfirt- D1- N5**	a302a, a303, a305, a306l, a307, a309a, a309b, a309c, a309d, a309e, a310a, a310b, a310c, a310d ,a310e, a310f, a310g, a310h, a312a, a312b, a312c, a312d, a312e, a312f, a313a, a313b
Atribute Evaluator:**ConsistencySubsetEval** Search Method: **GreedyStepwise-T-1**	a302
Atribute Evaluator: **ClassifierSubsetEval** Search Method: **Genetic Search**	Municipio, a14a08, a14a11, a310c, a311e7, a313d, a406b

Tabla 12. Resultados de las evaluaciones del atributo a301(1/2)

Atributo evaluado A301	
Atribute Evaluator: **FilteredSubsetEval** Search Method:**GreedyStepwise**	a309c, a309e, a310f, a310g, a312b, a313b, a313f, a313g, a313h
Atribute Evaluator: WrapperSubsetEval Search Method: Genetic Search	Municipio, a14a08, a14a11, a310c, a311e7, a313d, a406b
Atribute Evaluator: **ConsistencySubsetEval** Search Method: **Bestfirt- D1-N5**	a302a

Tabla 13. Resultados de las evaluaciones del atributo a301(2/2)

Después de aplicar los métodos y tabular todas la variables que fueron mostradas. Se determinó la media aritmética con respecto al total de atributos dividido entre el número de métodos aplicados, como se muestra en la tabla 14.

Variable a301	Numero de atributos
Atribute Evaluator: CfsSubsetEval **Search Method:** Bestfirt- D1- N5	26
Atribute Evaluator: ConsistencySubsetEval **Search Method:** GreedyStepwise-T-1	1
Atribute Evaluator: ClassifierSubsetEval **Search Method:** Genetic Search	7
Atribute Evaluator: FilteredSubsetEval **Search Method:**GreedyStepwise	9
Atribute Evaluator: WrapperSubsetEval **Search Method:** Genetic Search	7
Atribute Evaluator: ConsistencySubsetEval **Search Method:** Bestfirt- D1-N5	1
Total atributos	51
Media aritmética	$\frac{51}{6} = 8.33$

Tabla 14. Cálculo de la media aritmética

Teniendo la media aritmética podemos identificar que el numero de atributos elementales para la interpretación es de 8. La selección de las variables fue de acuerdo al termino estadístico moda que es entendido como el dato con más presencia en el conjunto.

Se seleccionaron las variables que estuvieron más prevalencia, en este caso la moda más alta fue de tres veces, seguidamente se identificaron las variables con presencia menor a tres y el numero de variables que presentaron la misma característica fueron diez. Para poder elegir las variables se determinó por la relevantes para la investigación.

Atributos mas relevantes	
a302c	¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre? MESES 1. Menos de un mes 2. NS/ NR
a309c	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Si, Homeopatía (chochos) 1.Si 2.No
a309e	¿Actualmente lleva algún otro tratamiento para controlar su azúcar? Si, Medicina alternativa 1.Si 2.No
a312b	Debido a la diabetes, durante los últimos 12 meses que medidas preventivas ha seguido para evitar complicaciones? Toma una aspirina diario 1.Si 2.No
a310c	Durante los últimos 12 meses ?Qué exámenes le hizo u orden? su médico para vigilar su azúcar? Examen general de orina 1.Si 2.No
a310f	Durante los últimos 12 meses ?Qué exámenes le hizo u orden? su médico para vigilar su azúcar? Examen para medir el nivel de proteínas en la orina (microalbuminuria) 1.Si 2.No
a310g	Durante los últimos 12 meses ?Que exámenes le hizo u orden? su médico para vigilar su azúcar? Auto monitoreó 1.Si 2.No
a406b	¿En que institución se atiende su presión alta? ISSSTE

Tabla 15. Atributos seleccionados

Analizando los resultados obtenidos a presencia del atributo a302 estuvo sobre tres métodos. En el caso de las variable 309c tuvo presencia sobre dos métodos, por otra parte el atributo 309e que corresponde a mismo conjunto de respuestas de la variable a309 se vio mostrada sobre los mismos métodos de la variable 309c al igual que el atributo a312b.

Por otra parte las variable a310c tuvieron prevalencia en tres métodos pero solo en dos métodos de la variable a302. Las demás variables estuvieron bajo uno o dos métodos de evaluación. Teniendo en cuenta que los métodos y los evaluadores que se aplicaron a las 163 atributos presentes en la base de datos. Los resultados se concentraron principalmente sobre tiempos de diagnóstico de diabetes o azúcar alta, plan de control del azúcar, exámenes de vigilancia de diabetes, lugar de atención.

3.6.1.2 **Selección de variables: Base de datos Glucosa y Lípidos**

A las variables de la base de datos Glucosa y Lípidos se aplicó un método Ranker aunado a un atributo evaluador InfoGainAttributeEval, este permitió crear un ranking de las variables. Para poder realizar este procedimiento se eligieron tres atributos que serian evaluados con respecto a los demás. El resultado de la evaluación y las variables seleccionadas se muestra en la tabla 16 y 17. Este procedimiento fue aplicado a todas las variables debido a que todas son de gran relevancia para la investigación.

La primera evaluación fue en función de variable edad, el despliegue coloca a la glucosa en primer lugar, seguido de confirmado diabético, colesterol en la tercera y CRP en la cuarta posición. En la ilustración 18 se muestra gráficamente los resultados.

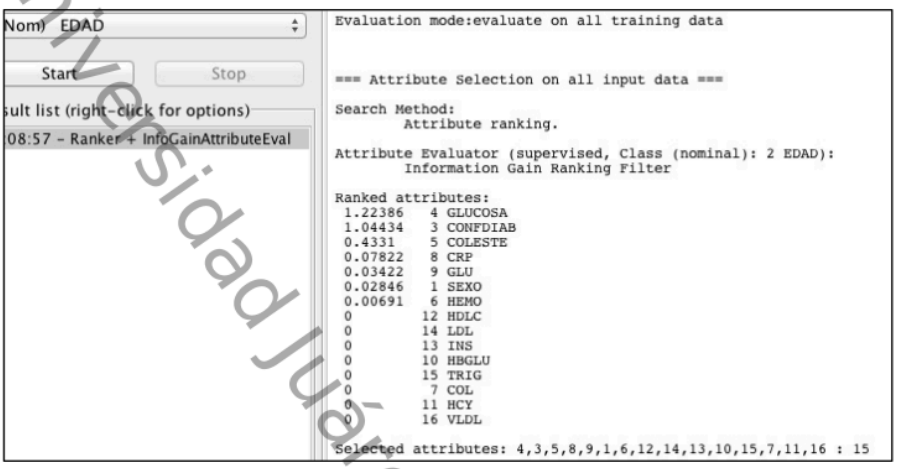


Ilustración 18. Resultado de la evaluación del atributo Edad

La segunda evaluación fue en función de variable SEXO, el despliegue colocan a la edad en primer lugar, seguido de glucosa, confirmado diabético en la tercera y CRP en la cuarta posición. En la ilustración 19 se muestra gráficamente los resultados.

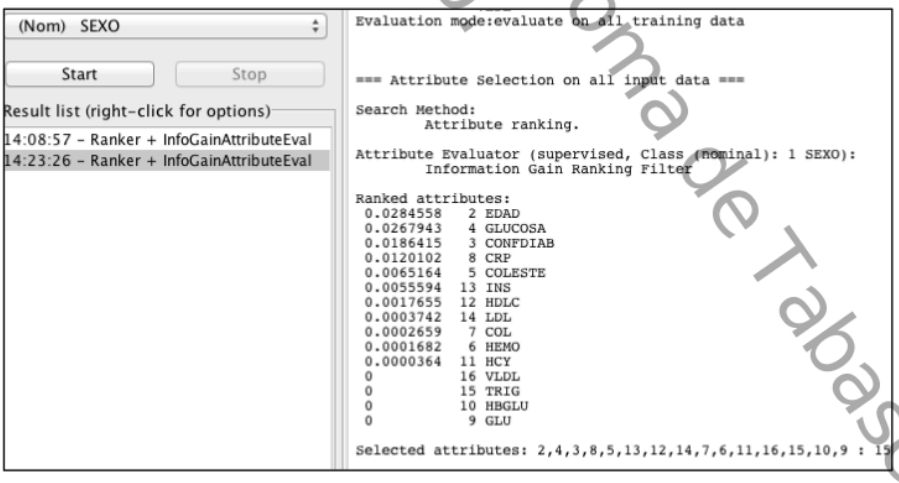


Ilustración 19. Resultado de la evaluación del atributo Sexo

La tercera evaluación fue en función de variable CONFIDIAB, el despliegue colocan a la glucosa en primer lugar, seguido de edad, CRP y Glucosa en la cuarta posición. En la ilustración 20 se muestra gráficamente los resultados.

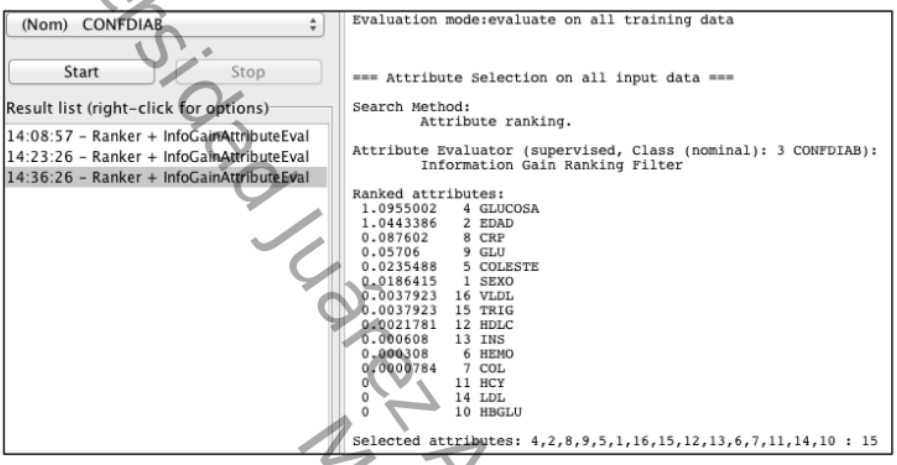


Ilustración 20. Resultado de la evaluación del atributo Confdiab

Teniendo las evaluaciones se realizó el agrupamiento para identificar cuales fueran la posiciones de los demás atributos que no se describieron en cada evaluación. Esto se muestra en la tabla 16 y 17.

	EDAD	SEXO	CONFIDIAB
1	GLUCOSA	EDAD	GLUCOSA
2	CONFDIAB	GLUCOSA	EDAD
3	COLESTE	CONFDIAB	CRP
4	CRP	CRP	GLU
5	GLU	COLESTE	COLESTE
6	SEXO	INS	SEXO
7	HEMO	HDLC	VLDL
8	HDLC	LDL	TRIG

Tabla 16. Resultados de la evaluación de las variables Edad, Sexo, Confidiab(1/2)

	EDAD	SEXO	CONFIDIAB
9	LDL	COL	HDLC
10	INS	HEMO	INS
11	HBGLU	HCY	HEMO
12	TRIG	VLDL	COL
13	COL	TRIG	HCY
14	HCY	HBGLU	LDL
15	VLDL	GLU	HBGLU

Tabla 17. Resultados de la evaluación de las variables Edad, Sexo, Confidiab(2/2)

Efectuando un análisis de los resultados se pudo identificar que el atributo con mayor importancia fue glucosa que estuvo reflejado en primer lugar en dos evaluaciones, seguido de Edad y Confidiab que fueron reflejados en un nivel de importancia 2 y 3. En la quinta posición se vio reflejado el atributo Coleste, seguido de Sexo en sexta posición de prioridad.

3.6.2 Limpieza de los datos

La prime base de datos denominada Adultos contiene información correspondiente a Depresión, Diabetes , Hipertensión, antecedentes de enfermedades, Sobrepeso y obesidad, salud reproductiva, vacunación, , accidentes, agresión y violencia. Pero solo 4 de los temas tiene relevancia para esta investigación. Por tal razón se procedió a eliminar de la base de datos sin afectar el resto de la información. De la eliminación de los datos irrelevantes quedaron los 163 atributos con los que se ha venido trabajando hasta este momento.

Por otra parte la segunda base de datos denominada Glucosa y Lípidos contiene 36 atributos inicialmente, se le aplico limpieza de los datos identificando los atributos como folio_v, código upm, folio entre otros. Estos fueron eliminados sin afectar la naturaleza de los demás.

3.6.3 Construcción e integración de los datos

Los datos extraídos de las bases de datos se encontraron en formato *.SAV de SPSS, estos fueron pasados a una hoja de calculo donde se guardo en la extensión de archivo *.CSV (Valores separados por comas) compatible con el software de minería de datos.

Después de la limpieza y transformación de los datos se garantizo que el formato de datos fuese compatible con las diferentes técnicas de minería de datos.

El orden en que los atributos quedaron después de la limpieza, no causo inconveniente a los procedimientos de construcción e integración de los datos. Tampoco al software de entrenamiento de datos ya que este no requiere una posición especifica de algún atributo.

3.7 Modelado

3.7.1 Selección de las técnicas de modelado

La técnicas de minería de datos que fueron elegidas para la investigación están en función de los criterios iniciales de descubrimiento de patrones:

Los árboles de decisión, que pertenecen a la categoría de los algoritmos predictivos, ayudaron a generar estructuras representadas en conjunto de disposiciones, que permiten predecir el comportamiento de los datos. El algoritmo que se utilizó en la parte de clasificadores es J48, buscando realizar una clasificación de los datos en forma de árbol. Todo esto aplicando un atributo evaluador con respecto al conjunto de datos.

Por otra parte la agrupación(*clúster*) que se ubica en el módulo de las técnicas de descubrimiento de conocimiento, permite aglomerar casos o variables en función del parecido o similitud que existe entre ellos. Para llevar a cabo este procedimiento se planteó utilizar el algoritmo *SimpleKMeans*.

3.7.2 Generación del plan de prueba

Realizada la selección de las técnicas de minería de datos se ejecuto el plan de prueba. Para ellos se utilizó los archivos de datos generados en la etapa de construcción e integración, seguidamente se procedió a explorar los datos en el software para poder ubicar los algoritmos que se habían determinado para la investigación.

Las pruebas ejecutadas permitieron experimentar algunos algoritmos y poder determinar que tan factible era utilizarlos sobre los datos. Durante la ejecución de las pruebas se obtuvieron resultados iniciales que posteriormente fueron considerados como hallazgos de comportamiento de los datos. La ejecución de los algoritmos buscando permitió encontrar un primer hallazgo de comportamiento, esto fue descubierto utilizando el algoritmo J48 de clasificación, con la variable a302c.

Los resultados encontrados con el algoritmo J48 se describen detalladamente en el capitulo 4 y de la misma manera se muestra gráficamente el árbol resultante. En la ilustración 21 se muestra buffer y en la ilustración 22 el árbol de decisión .

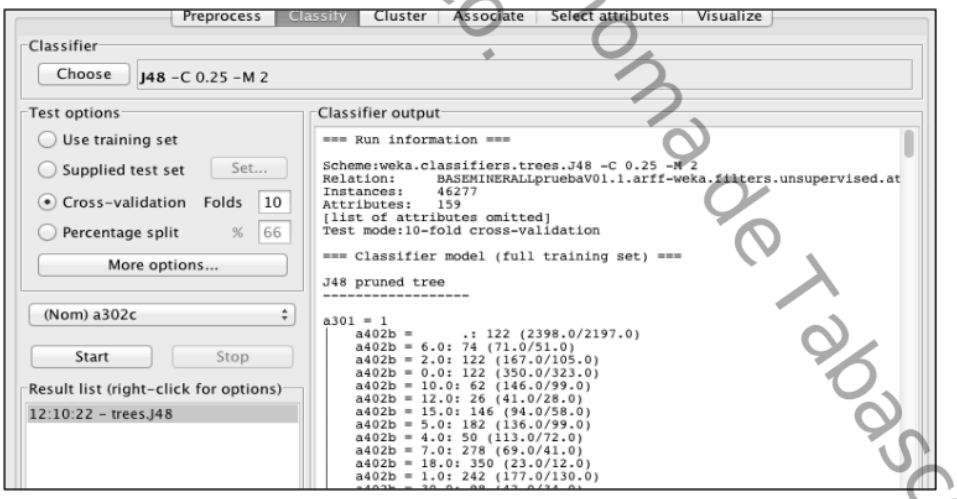


Ilustración 21. Resultados iniciales de aplicar técnicas de minería de datos

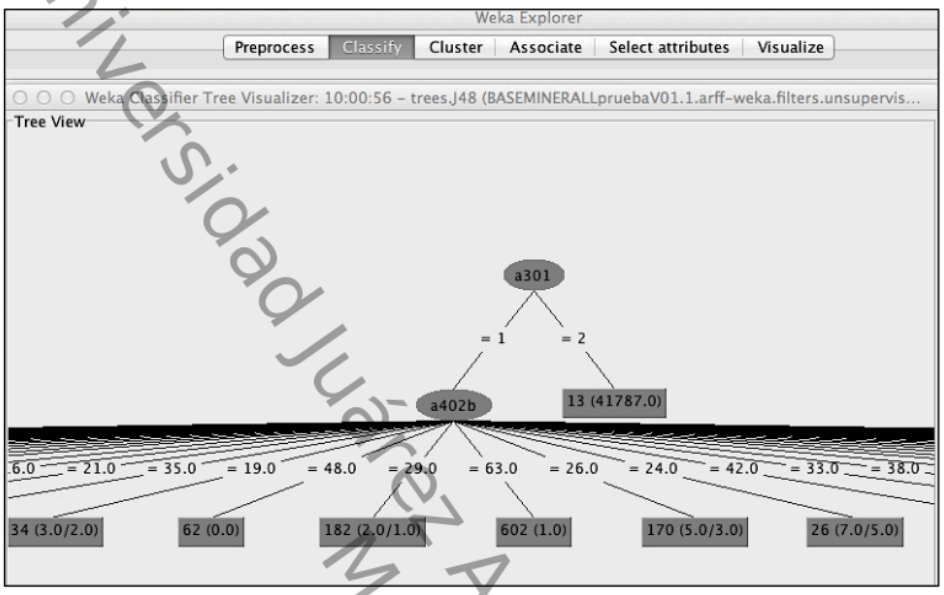


Ilustración 22. Árbol de decisión de la evaluación del atributo a302c

Por ultimo un aspecto que se describió en la construcción e integración de los datos y que es importante recalcar es que el software de minería de datos no requiere ningún orden de los atributos evaluados para realizar la minería de datos, únicamente se tiene que cumplir con el tipo de dato para que los algoritmos puede ejecutarse y obtener resultados.

La ejecución del plan de prueba ayudo a identificar atributos que no son relevantes ser evaluados pero si necesarios para obtener resultados en la investigación.

3.7.3 Construcción de los modelos

El plan previo a esta fase permitió conocer el comportamiento de los datos inicialmente, esto fue de utilidad para construir los modelos de clasificación, agrupación y asociación que dieron óptimos resultado y así una mejor interpretación de la información.

Esta fase es sin duda alguna la más importan debido a que se realiza el plan de como deben ejecutarse los algoritmos, que permitan obtener resultados finales que serán mostrados en los informes. Se identifico las variables que fueron trabajadas en el plan de prueba y se procedió a crear cada modelo, para entender como están formados los modelos se realizó una descripción tanto de las variables como de los gráficos creados para conocer como están constituidos los modelos de datos de la investigación.

3.7.3.1 Modelo de clasificación variable a302c

Para realizar la clasificación se elaboró un procedimiento que seria seguido como modelo en el que se representa la forma en que los datos fueron evaluados por el software de minería.

Primeramente se aprecia que la base de datos adultos fue sometida a evaluación, posteriormente se específica el tipo de tarea que se le realizó, seguido de la variable utilizada que fue a302c y de la misma forma se aprecia el algoritmo aplicado.

- **a302c** ¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre?.

El resultado un árbol de decisión en forma de conjunto de condiciones organizadas de forma de estructura jerárquica. En la ilustración 23 muestra gráficamente el modelo.

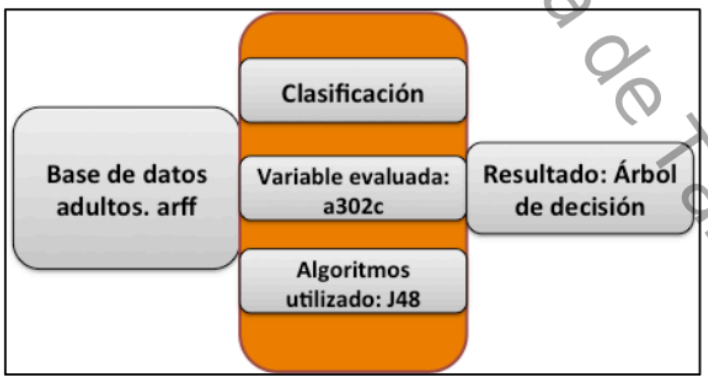


Ilustración 23. Modelo de clasificación

3.7.3.2 Modelo de clasificación de variables a310c, a310f, a310g

Se realizó un modelo de clasificación para identificar el comportamiento de las variables a310c, a310f y a301g, estas son las que describen el tipo de estudio clínico o control que se realizan las personas con diabetes. A continuación se describe cada una de ellas y seguido se aprecia la ilustración 24 donde se ve gráficamente el modelo.

- **a310c** (Durante los últimos 12 meses ¿Qué exámenes le hizo u ordeno su médico para vigilar su azúcar? Examen general de orina), así como la variable
- **a310f** (Durante los últimos 12 meses ¿Qué exámenes le hizo u ordeno su médico para vigilar su azúcar? Microalbuminuria)
- **a310g** (Durante los últimos 12 meses ¿Qué exámenes le hizo u ordeno su médico para vigilar su azúcar? Automonitoreo)

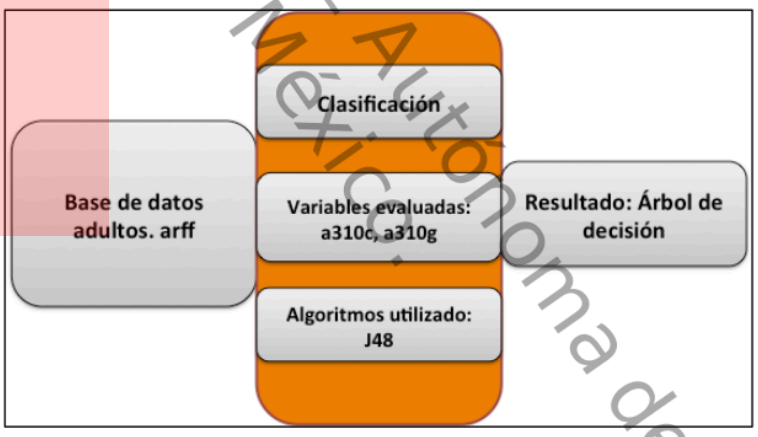


Ilustración 24. Modelo de clasificación para variables a310c, a310f y a310g

3.7.3.3 Modelo de clustering o agrupación

Siguiendo la mismas características del modelo de clasificación, se construyo el esquema para clustering. Se realizó nuevamente la carga de la base de datos adultos, y se realizaron las agrupaciones en función del atributo a301 para la base de datos Adultos de 20 años o más.

El algoritmo utilizado fue *SimpleKmeans* que permitió buscar casos similares en el conjunto de datos. En la ilustración 25 se muestra gráficamente.

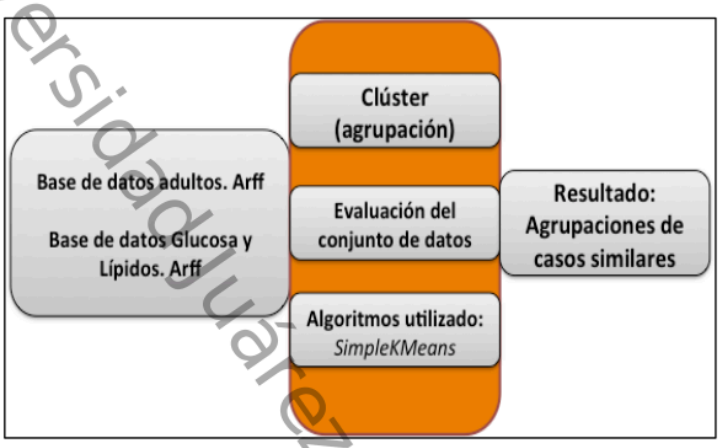


Ilustración 25. Modelo de agrupación para Base de datos Adultos, Glucosa y lípidos

Capítulo IV. Resultados

En este capitulo se describen los hallazgos encontrados de la aplicación de la técnicas de minería de datos.

4.1 Descripción de modelos

4.1.1 Clasificación variable a302c

En este modelo, el cual se especificó anteriormente, se trabajó sobre todo el conjunto de datos, evaluado en función del atributo a302c (¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre?). Se obtuvo como resultado un árbol de decisión el cual se muestra en la ilustración 26.

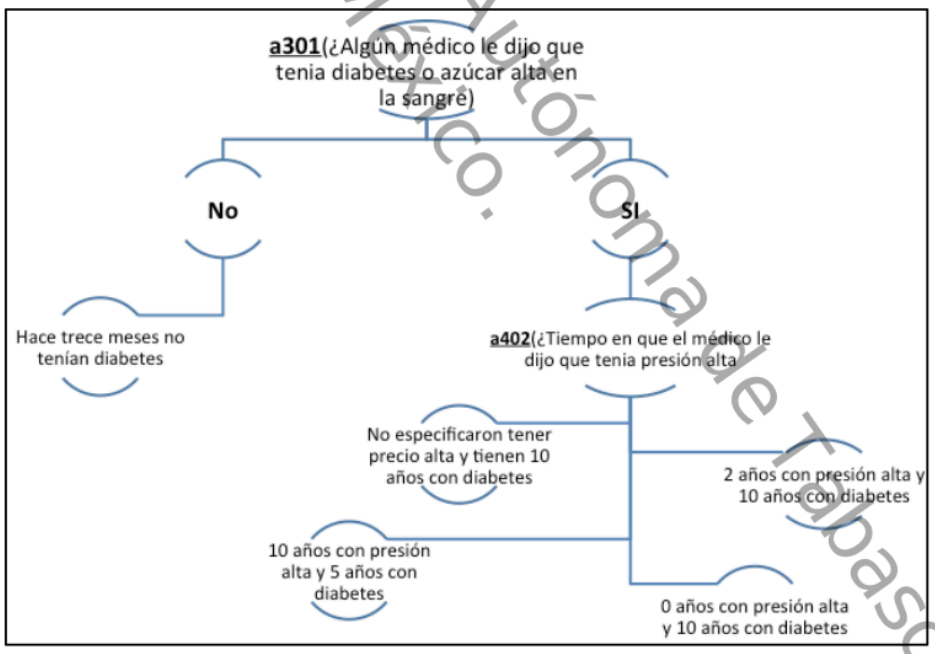


Ilustración 26. Árbol de decisión de la evaluación del atributo a302c

El algoritmo J48, que fue aplicado a los datos del archivo adultos. Arff, presentó una precisión del 91.942% de las instancias clasificadas correctamente y un error relativo del 8.058%. A continuación se muestra en la ilustración 27

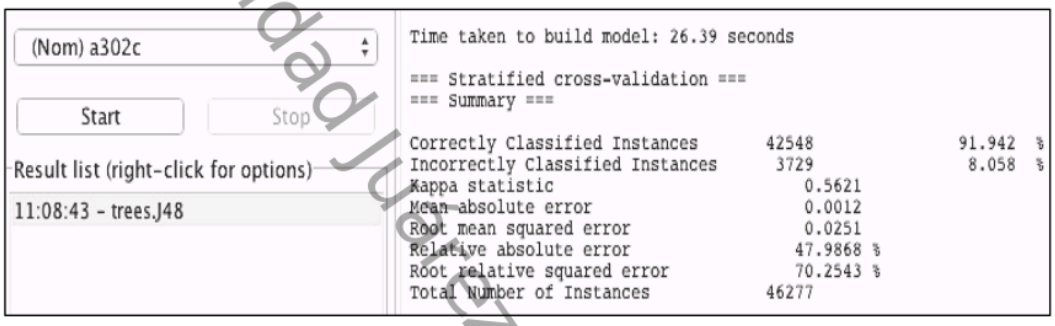


Ilustración 27. Buffer de la precisión del algoritmo J48

Como se puede observar en la ilustración 26 de la página 80 , el árbol de decisión muestra una clasificación del comportamiento de los datos. Para poder realizar una descripción detallada de lo visualizado en árbol se tomó una parte del mismo sin que se perdieran los resultados.

Examinando el árbol se pudo describir que existen personas a las que un médico les ha dicho que Si tienen diabetes o azúcar alta en la sangre, estas tienen una relación con el atributo (a402b) que representa el tiempo en que algún médico les dijo que tenían presión alta. Observando la expansión del atributo a402 se descubrió que:

- Existen personas con diabetes desde hace 10 años pero desconocen si tienen presión alta.
- Hay personas que tienen presión alta desde hace 6 años, llevan 10 años con diabetes.
- Quienes llevan 2 años con presión alta, tienen diabetes desde 10 años.
- Se descubrió que quienes asumen que tienen 10 años con presión alta, tiene 5 años con diabetes .
- Las personas que no tienen presión alta, si presentan diabetes desde hace 10 años.

En base al análisis que se hizo al árbol de decisión, se puede expresar que la presión arterial no es una detonante de la diabetes pero existe un prevalencia alta de personas con diabetes que presentan tensión arterial alta.

4.1.2 Modelo de clasificación de las variables a310c, a310g y a310f

La clasificación que se realizó utilizando el algoritmos J48 a las variables a310c, a310f y a310g permitió obtener arboles de decisión con el numero de veces que las personas con diabetes se realiza determinado examen clínico o automonitoreo en los últimos 12 meses.

Descripción de las variables

- **a310c** (Durante los últimos 12 meses ¿Qué exámenes le hizo u ordeno su médico para vigilar su azúcar? Examen general de orina), así como la variable
- **a310g** (Durante los últimos 12 meses ¿Qué exámenes le hizo u ordeno su médico para vigilar su azúcar? Automonitoreo.)
- **a310f** (Durante los últimos 12 meses ¿Qué exámenes le hizo u ordeno su médico para vigilar su azúcar? Microalbuminuria)

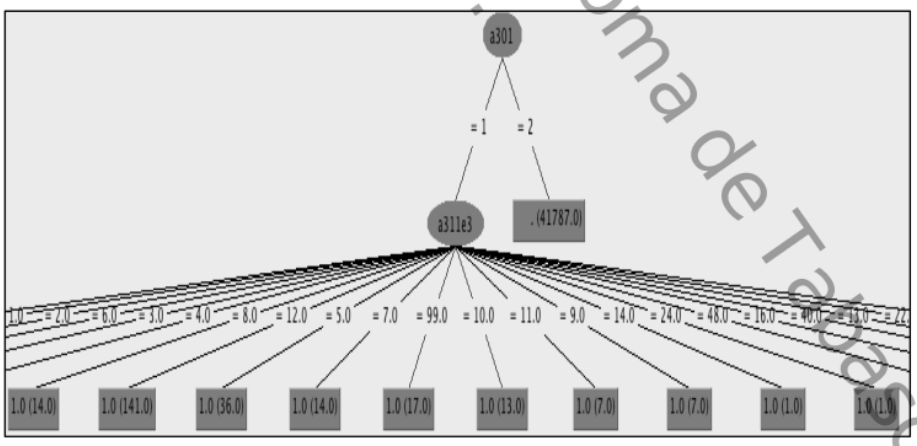


Ilustración 28. Árbol de decisión generado de la evaluación de la variable a310c

Para hacerlo mas comprensible el árbol de decisión y describir las variables que lo integran se construyo un representación donde se visualiza completamente. Los resultados que se pueden apreciar son la frecuencia de veces en que se realizan el examen general de orina. En la ilustración 29 muestra el árbol de decisión.

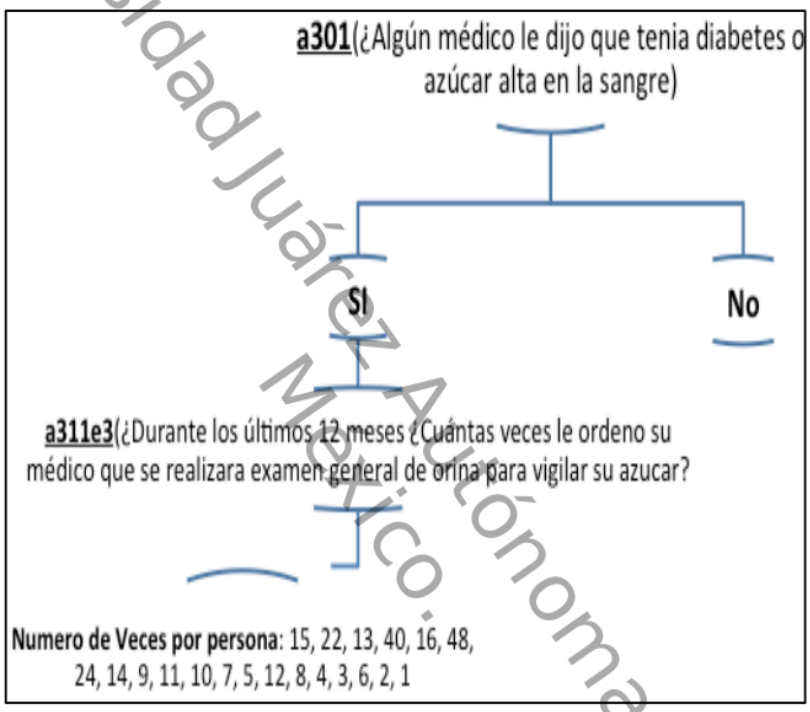


Ilustración 29. Interpretación del árbol de decisión de la variable a310c

Teniendo las frecuencias en que las personas se realizan la prueba general de orina, se determinó el promedio para los 12 meses. Obteniendo el siguiente resultado: 13.5 veces en 12 meses, lo que se puede decir que las personas con diabetes se realizan una vez por mes el examen general de orina para vigilar su azúcar.

La segunda evaluación del atributo a310g se obtuvo el siguiente árbol de decisión :

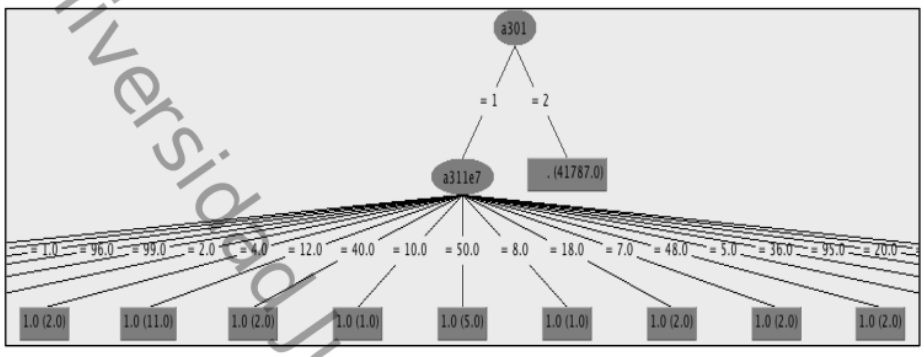


Ilustración 30. Árbol de decisión generado de la evaluación de la variable a310g

Para describir las variables que lo integran se construyo un diagrama donde se visualiza completamente, los resultados del numero de veces que una persona con diabetes se realiza automonitoreo de su azúcar en la sangre. En la ilustración 31 se muestra de forma descriptiva cada una de las variables .

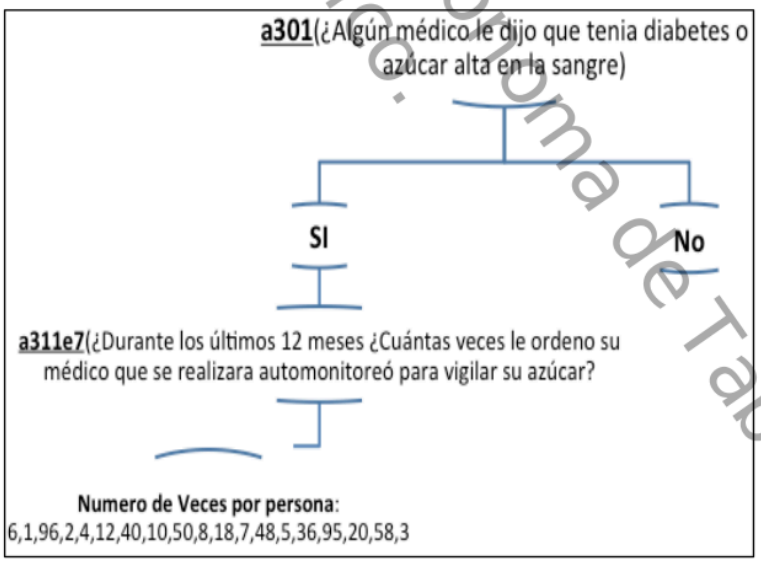


Ilustración 31. Interpretación del árbol de decisión de la variable a310g

Identificando las frecuencias en que las personas se realizan automonitoreo, se determinó el promedio para los 12 meses. Obteniendo el siguiente resultado: 27.31 veces en 12 meses, lo que se puede decir que las personas con diabetes se realizan dos vez por mes automonitoreo para vigilar su azúcar.

El algoritmo J48, que fue aplicado a la variables a310g , presentó una precisión del 99.9849% de las instancias clasificadas correctamente y un error relativo del 0.0151%.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       46270               99.9849 %
Incorrectly Classified Instances         7                0.0151 %
Kappa statistic                          0.9991
Mean absolute error                      0.0001
Root mean squared error                  0.0099
Relative absolute error                  0.0851 %
Root relative squared error              4.1007 %
Total Number of Instances            46277
```

Ilustración 32. Buffer de la precisión del algoritmo J48 de la variable a310g

Del atributo a310f al igual que las demás evaluaciones se obtuvo un árbol de decisión que se puede ver en la ilustración 34 con la frecuencia de veces que las personas se realizan el examen de Microalbuminuria para vigilar su azúcar en la sangre, con base en esto se determinó el promedio para 12 meses.

Se obtuvo un promedio de 7.64 veces cada 12 meses, lo que se puede decir que las personas con diabetes se realizan una prueba de control cada 2 meses.

El algoritmo aplicado a la variables a310f fue J48, presento un precisión del 99.9892 de las instancias clasificadas correctamente y un error relativo del 0.0108%.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       46272               99.9892 %
Incorrectly Classified Instances         5                0.0108 %
Kappa statistic                          0.9994
Mean absolute error                      0.0001
Root mean squared error                  0.0081
Relative absolute error                  0.0589 %
Root relative squared error              3.3648 %
Total Number of Instances            46277
```

Ilustración 33. Buffer de la precisión del algoritmo J48 de la variable a310f

Árbol de decisión de personas que realizan monitoreo de azúcar en la sangre a través de Microalbuminuria.

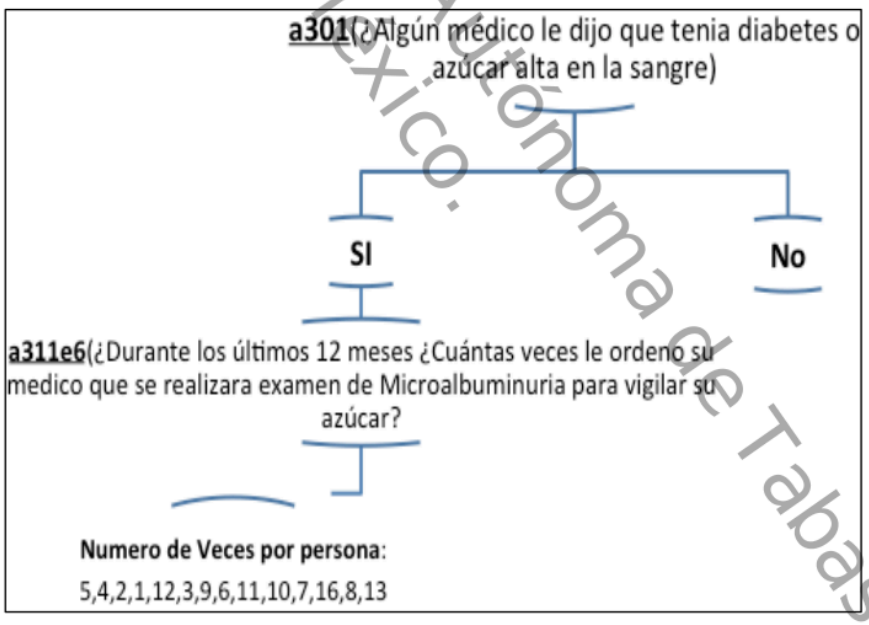


Ilustración 34. Interpretación del árbol de decisión de la variable a310f

4.1.3 Clustering (Base de datos Adultos 20 años o más)

El análisis de clustering se realizó sobre las variables que se muestra en la tabla 18:

Atributos a evaluar	
Sexo	Genero de la persona: 1 Masculino, Femenino
Edad	Edad de la persona
a401	¿Algún médico le ha dicho que tiene la presión alta o hipertensión? 1.Si 2.No
a501	¿Ha tenido alguna vez un dolor fuerte en el pecho, con falta de aire o gran malestar que durara media hora o más? 1. Si 2. No
a502b	¿Le ha dicho el médico que usted tiene o tuvo angina de pecho? 1.Si 2.No
a502c	¿Le ha dicho el médico que usted tiene o tuvo insuficiencia cardiaca? 1.Si 2.No
a502d	¿Le ha dicho el médico que usted tiene o tuvo otra enfermedad del corazón? 1.Si 2.No
a601	¿Alguna vez le han medido el colesterol en la sangre? 1. Sí y lo encontraron normal 2. Sí y lo encontraron alto 3.No 9. NS/NR
a603	¿Alguna vez le han medido los triglicéridos en la sangre? 1. Sí y lo encontraron normal 2. Sí y lo encontraron alto 3.No 9. NS/NR
a604	Alguna vez le ha dicho su médico que tuvo una embolia o un infarto cerebral 1.Si 2.No
a302c	¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre? MESES

Tabla 18. Atributos a evaluar de la base datos glucosa y lípidos

La agrupación permitió conocer las características dentro de cada grupo de variables. A continuación en la ilustración 35 se muestran los resultados obtenidos por el algoritmo *SimpleKmeans*.

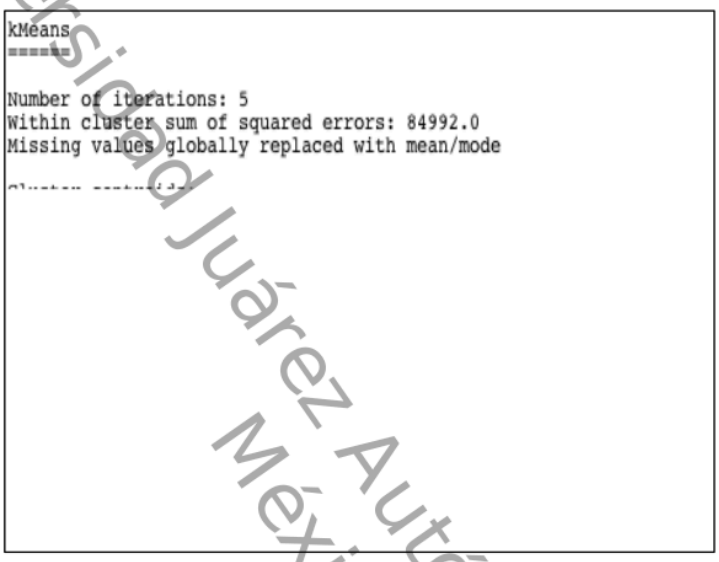


Ilustración 35. Buffer de la aplicación del algoritmo SimpleKmeans

Descripción de las agrupaciones:

- **Clúster 0.-** Es prevalecido por Mujeres en edad de 36 años, ningún médico le ha dicho que tiene presión alta, no tienen diabetes y por consecuente no presentado ningún padecimiento relacionado con la patología. Aquí se agrupan 16,162 de los registros del conjunto de datos que corresponde al 35%.
- **Clúster 1.-** Conformado en su mayoría por Mujeres en edad de 29 años, no presentan diabetes, tampoco presentan presión alta o colesterol. Este lo integran el 695 de los registros que equivale al 2%.

- **Clúster 2.-** Al igual que el clúster 0 predominan las mujeres en edad de 38 años, estas han presentado colesterol alto, así como triglicéridos en limite establecido para considerarlo como normal. El porcentaje comprendido en este clúster es de 38% que corresponde a un total de 17,370 de registros.

- **Clúster 3.-** Se constituye por hombres en edad de 20 años, no presentan diabetes, tampoco padecen colesterol o triglicéridos. Es 26% que conforma un total de 12,050 de registros

Sobre las agrupaciones realizadas se puede expresar que en las mujeres en edad 38 años en adelante empiezan a presentar elementos que pueden llevar a detonar la diabetes. Mientras que los hombres en la misma edad no presentan ningún criterio que lo lleve a considerarlos como prediabéticos.

Dentro de la misma conglomeración se realizó la agrupación de la variable a301 (¿Algún médico le dijo que tenia diabetes o azúcar alta en la sangre?), obteniendo que en el clúster 0 hay 1,506 personas que tienen azúcar alta. Mientras que en el clúster 1 hay 3 personas, por ultimo el clúster 3 muestra que hay 838 personas con el azúcar alta en la sangre. Todo esto puede visualizarse en ilustración 36.

```
Class attribute: a301
Classes to Clusters:

    0     1     2     3  <-- assigned to cluster
 1506     3  2133   848 | 1
14656   692 15237 11202 | 2
```

Ilustración 36. Buffer de la clasificación de los datos en función del atributo a301

Posteriormente a la agrupación del atributo a301 en cada clúster, se realizó la visualización de los grupos, que puede ser visto en la ilustración 37.

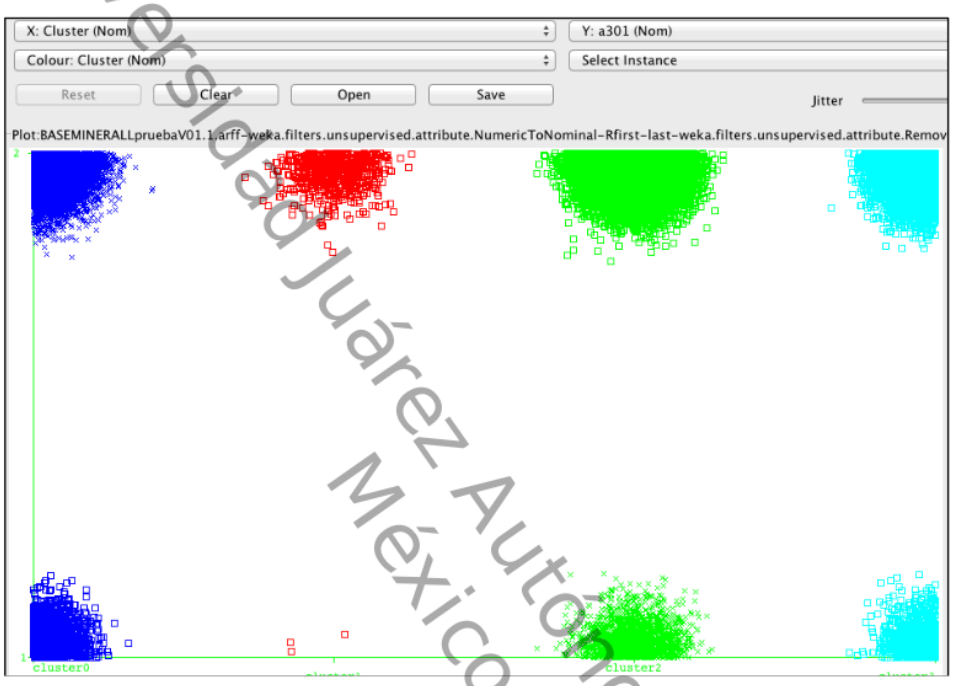


Ilustración 37. Visualización de las agrupación en función del atributo a301

Describiendo la visualización de como se representan los grupos en el eje X y Y del cuadrante del plano cartesiano. En el eje X se puede ver las agrupaciones que realizó el software de minería de datos, mientras que en el eje Y se encuentran los valores de las variable a301 ¿Algún médico le ha dicho que tiene diabetes o el azúcar alta en la sangre?.

En total se puede visualizar ocho clúster de los cuatro que se hicieron inicialmente, debido a que se evaluó con respecto a la variables a301 que tiene dos valores (Si o No). Cada grupo tiene una característica que lo hace fijarse en una posición del plano. Estos son descritos en las siguientes tablas y narrativas.

Se eligieron de cada clúster cuatro participantes divididos en 2, unos con diabetes y otros sin la patología, en las tablas 19, 20, 21 y 22 se describen las características cada agrupación.

Clúster 1				
Variable				
a301 (¿Algún médico le ha dicho que tiene diabetes o azúcar alta en la sangre?)	**No**	**No**	**Si**	**Si**
Sexo	Mujer	Mujer	Mujer	Mujer
Edad	32	21	47	53
a401 (¿Algún médico le ha dicho que tiene la presión alta o hipertensión?)	No	No	Si	Si
a501 (¿Ha tenido alguna vez un dolor fuerte en el pecho, con falta de aire o gran malestar que durara media hora o más?)	No	No	No	No
a502b (¿Le ha dicho el médico que usted tiene o tuvo angina de pecho?)	No	No	No	No
a502c (¿Le ha dicho el médico que usted tiene o tuvo insuficiencia cardiaca?)	No	No	No	No
a502d (¿Le ha dicho el médico que usted tiene o tuvo otra enfermedad del corazón?)	No	No	No	No
a601 (¿Alguna vez le han medido el colesterol en la sangre?)	No	No	No	No
a603(¿Alguna vez le han medido los triglicéridos en la sangre?)	No	No	Si, lo encontraron altos	No
a604 (¿Alguna vez le ha dicho su médico que tuvo una embolia o un infarto cerebral?)	No	No	No	No
a302c ¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre? MESES	13 meses	13 meses	12 años	13 años

Tabla 19. Descripción de elementos de diabéticos y no diabéticos clúster 1

De acuerdo a los datos representados anteriormente podemos decir que las mujeres en edad de 45 a 60 años son diabéticas, empiezan a manifestar otros tipos padecimientos como presión arterial alta, triglicéridos altos. Por otra parte quienes no son diabéticas el rango de edad esta entre 21 a 35 años muestran ciertas características como el colesterol alto. Estas particulares se pueden empezar a considerar como factores de riesgo que lleven a la detonación de la diabetes.

En el clúster 2 podemos ver al igual que los demás una separación de grupos de quien no y quien si presenta diabetes. Las características se muestran en la siguiente tabla.

Clúster 2				
Variable				
a301 (¿Algún médico le ha dicho que tiene diabetes o azúcar alta en la sangre?)	**No**	**No**	**Si**	**Si**
Sexo	Mujer	Mujer	Mujer	Mujer
Edad	28	27	29	29
a401 (¿Algún médico le ha dicho que tiene la presión alta o hipertensión?)	No	No	No	No
a501 (¿Ha tenido alguna vez un dolor fuerte en el pecho, con falta de aire o gran malestar que durara media hora o más?)	No	No	No	No
a502b (¿Le ha dicho el médico que usted tiene o tuvo angina de pecho?)	No	No	No	No
a502c (¿Le ha dicho el médico que usted tiene o tuvo insuficiencia cardiaca?)	No	No	No	No
a502d (¿Le ha dicho el médico que usted tiene o tuvo otra enfermedad del corazón?)	No	No	No	No
a601 (¿Alguna vez le han medido el colesterol en la sangre?)	No	Si, lo encontraron alto	No	No
a603(¿Alguna vez le han medido los triglicéridos en la sangre?)	No	Si, lo encontraron alto	No	Si, lo encontraron alto
a604 (¿Alguna vez le ha dicho su médico que tuvo una embolia o un infarto cerebral?)	No	No	No	No
a302c ¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre? MESES	13 meses	13 meses	12 meses	12meses

Tabla 20. Descripción de elementos de diabéticos y no diabéticos clúster 2

Las características que muestran las personas agrupadas en el clúster 2 son las siguientes: Las Mujeres en edad de 29 años presentan diabetes, así como triglicéridos altos. Quienes no son diabéticos presentan dos elementos (Colesterol y triglicéridos altos) que pueda ser considerado como factor de riesgo en la presencia de la diabetes.

A continuación se puede ver las características del clúster 3.

Clúster 3				
Variable				
a301 (¿Algún médico le ha dicho que tiene diabetes o azúcar alta en la sangre?)	**No**	**No**	**Si**	**Si**
Sexo	Mujer	Mujer	Mujer	Mujer
Edad	69	44	61	56
a401 (¿Algún médico le ha dicho que tiene la presión alta o hipertensión?)	No	Si	No	Si
a501 (¿Ha tenido alguna vez un dolor fuerte en el pecho, con falta de aire o gran malestar que durara media hora o más?)	No	No	No	No
a502b (¿Le ha dicho el médico que usted tiene o tuvo angina de pecho?)	No	No	No	No
a502c (¿Le ha dicho el médico que usted tiene o tuvo insuficiencia cardiaca?)	No	No	No	No
a502d (¿Le ha dicho el médico que usted tiene o tuvo otra enfermedad del corazón?)	No	No	No	No
a601 (¿Alguna vez le han medido el colesterol en la sangre?)	No	Si, lo encontraron normal	Si, lo encontraron normal	Si, lo encontraron normal
a603(¿Alguna vez le han medido los triglicéridos en la sangre?)	Si, lo encontraron normal	Si, lo encontraron normal	Si, lo encontraron normal	Si, lo encontraron normal
a604 (¿Alguna vez le ha dicho su médico que tuvo una embolia o un infarto cerebral?)	No	No	No	No
a302c ¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre? MESES	13 meses	13 meses	2.1 años	10.1 años

Tabla 21. Descripción de elementos de diabéticos y no diabéticos clúster 3

Las características de los participantes elegidos que si presentan diabetes se encuentran en un rango de edad de 55 a 65 años, su presión alta es alta en uno de los casos, triglicéridos y colesterol están en el rango normal. Los participantes que representan a quienes no tiene diabetes tienen criterios como edad entre 44 y 69 años, presentan triglicéridos en estado normal, así como un nivel de colesterol normal.

En el clúster 4 podemos ver al igual que los demás una clasificación de quien no y quien si presenta diabetes. Las características se muestran en la siguiente tabla.

Clúster 4				
Variable				
a301 (¿Algún médico le ha dicho que tiene diabetes o azúcar alta en la sangre?)	**No**	**No**	**Si**	**Si**
Sexo	Hombre	Hombre	Hombre	Hombre
Edad	49	46	66	51
a401 (¿Algún médico le ha dicho que tiene la presión alta o hipertensión?)	No	No	No	No
a501 (¿Ha tenido alguna vez un dolor fuerte en el pecho, con falta de aire o gran malestar que durara media hora o más?)	No	No	No	No
a502b (¿Le ha dicho el médico que usted tiene o tuvo angina de pecho?)	No	No	No	No
a502c (¿Le ha dicho el médico que usted tiene o tuvo insuficiencia cardiaca?)	No	No	No	No
a502d (¿Le ha dicho el médico que usted tiene o tuvo otra enfermedad del corazón?)	No	No	No	No
a601 (¿Alguna vez le han medido el colesterol en la sangre?)	No	Si, lo encontraron normal	No	Si, lo encontraron alto
a603(¿Alguna vez le han medido los triglicéridos en la sangre?)	No	No	No	Si, lo encontraron alto
a604 (¿Alguna vez le ha dicho su médico que tuvo una embolia o un infarto cerebral?)	No	No	No	No
a302c ¿Hace cuánto tiempo le dijo su médico por primera vez que tenia diabetes o el azúcar alta en la sangre? MESES	13 meses	13 meses	17.1 años	13.1 años

Tabla 22. Descripción de elementos de diabéticos y no diabéticos clúster 4

Los elegidos del clúster se caracterizaron por ser hombres en su totalidad, en primer instancia se describen a los que presentan diabetes estos tienen presión arterial, triglicérido y colesterol en un nivel alto . La edad de quienes poseen diabetes esta entre 50 a 66 años. Describiendo a quienes no presentan diabetes su edad esta entre 45 a 49 años, su colesterol esta en estado normal.

4.1.3.1 Vista de los clúster en conjuntos (Base de datos adultos 20 años o más)

Las personas elegidas tiene ciertas características ya sea si son o no diabéticos, cada agrupación se represento en 1 conjunto, donde se puede conocer a través de la intercesión las variables que comparten. El sujeto seleccionado se le asigno un numero en este caso del 1 al 4 siendo utilizada esta forma para todos los clustering.

- **Conjunto del clúster 1**

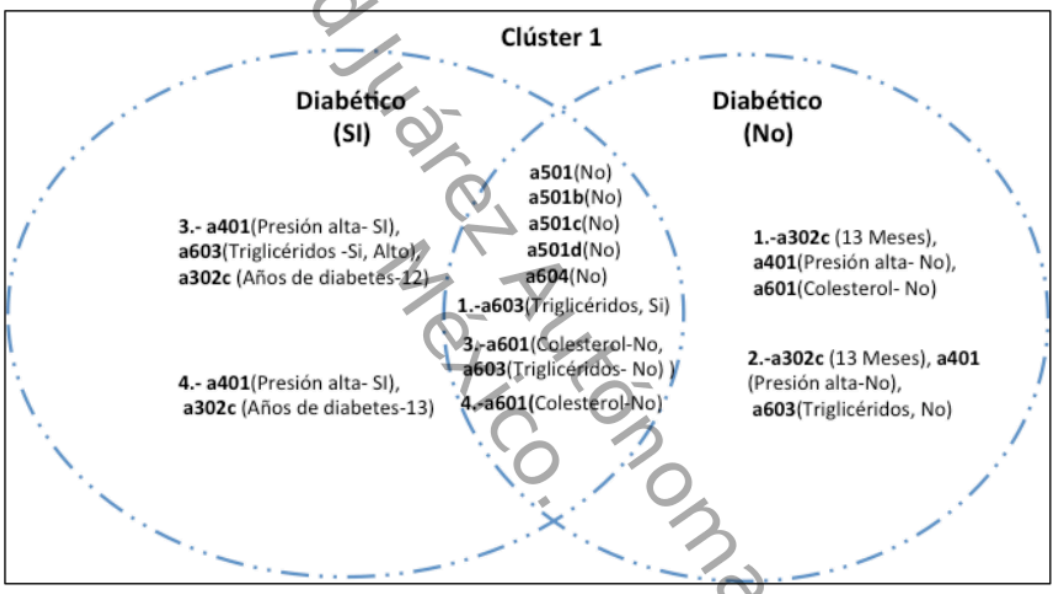


Ilustración 38. Conjunto Clúster 1

En la ilustración 38 se puede ver la intersección donde las personas diabéticas y no diabéticas comparten el atributo a 501, 501b 501c, 501d y 604. Otras variables que también están incluidas en la mediación son: a603 que representa los triglicéridos, en este caso es una variable del conjunto de los diabéticos pero la persona 1 de los no diabéticos presenta esta características. Por otra parte la persona con el numero 3 tiene algunas variables de los no diabéticos, no tiene colesterol, ni triglicéridos altos. Finalmente la persona 4 solo tiene relación con la variable a601 de los no diabéticos que es no tener colesterol alto.

- **Conjunto del clúster 2**

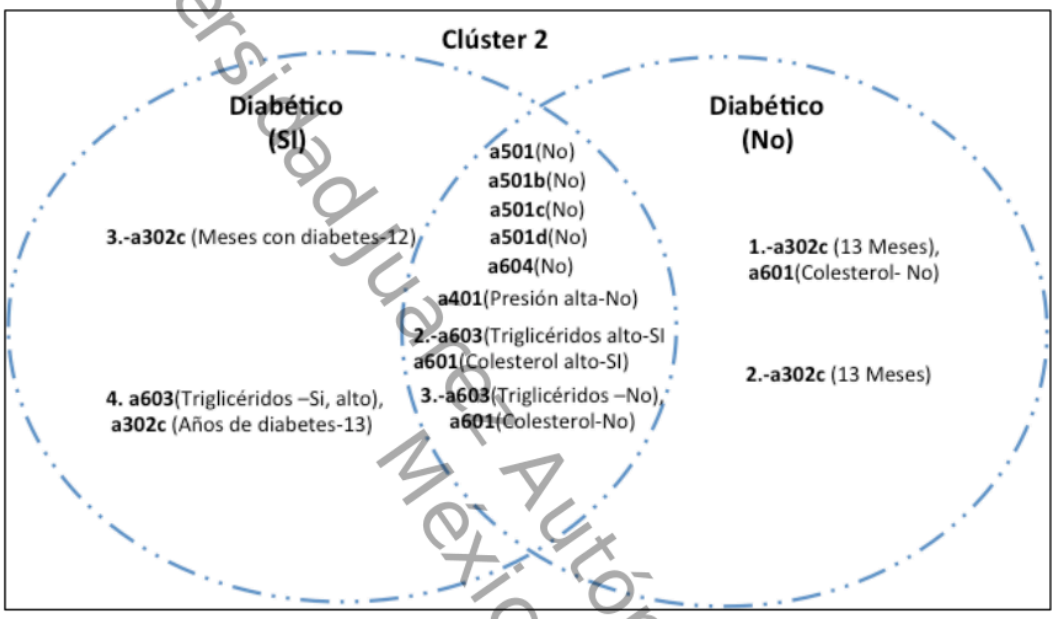


Ilustración 39. Conjunto Clúster 2

En la intersección de los conjuntos comparten el atributo a 501, 501b 501c, 501d, 604 y a401. Otras variables que también están incluidas en la mediación son: a603 que representa los triglicéridos, en este caso es una variable del conjunto de los diabéticos pero la persona 2 de los no diabéticos presenta esta particularidad. Por ultimo la persona con el numero 3 tiene algunas variables de quienes no tienen la patología, entre ellas esta no tener colesterol, ni triglicéridos altos. En los conjuntos en la intercesión se identifico que los no diabéticos y los que si de el clúster 2, no tienen presión alta o hipertensión.

- **Conjunto del clúster 3**

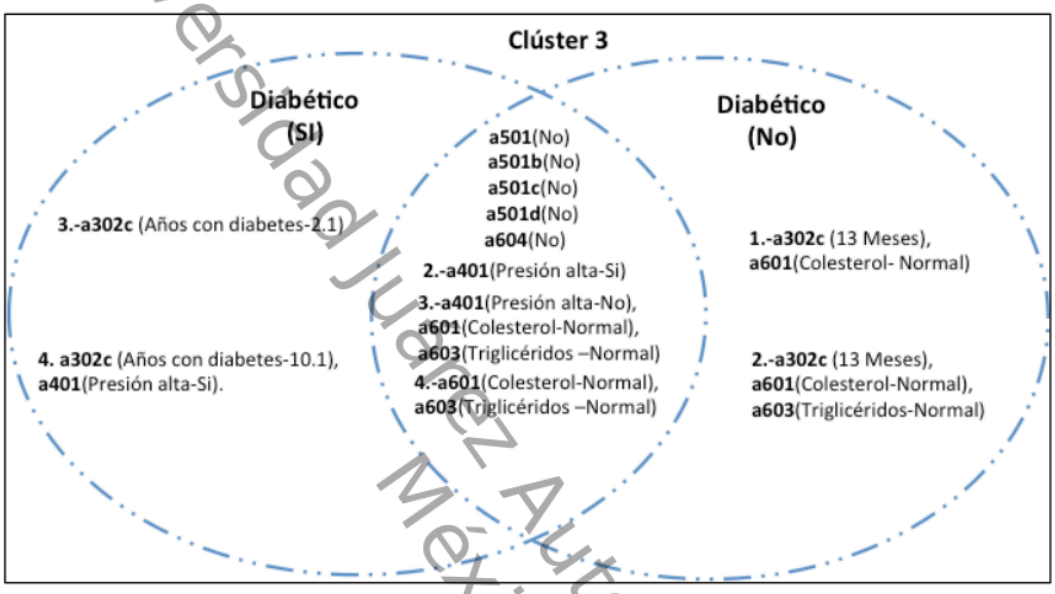


Ilustración 40. Conjunto Clúster 3

Los conjuntos comparten los atributos a501, 501b 501c, 501d, 604 al igual que las anteriores representaciones. Las persona 2 tiene una variable de los diabéticos que es la presión arterial alta, la persona 3 tiene dos variables características de las personas no diabéticas estas son: no presenta presión alta y su colesterol y triglicéridos están normales. Por último la persona 4 que presenta dos variables de los no diabéticos como son colesterol y triglicéridos normales.

- **Conjunto del clúster 4**

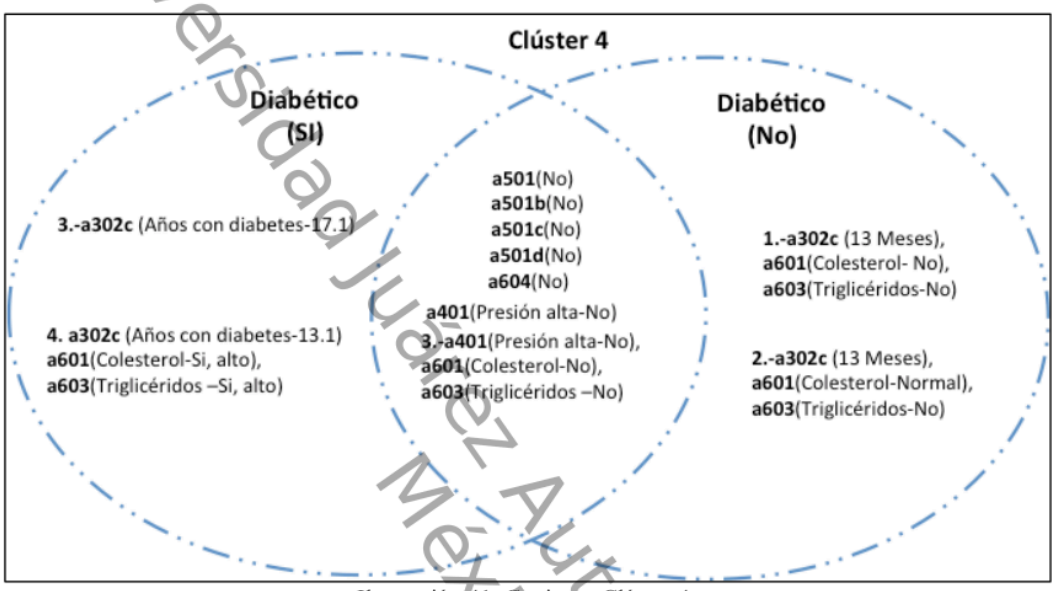


Ilustración 41. Conjunto Clúster 4

Los atributos a501, 501b 501c, 501d, 604 y 401 son visualizados en la intersección. Por su parte la persona 3 presenta tres variables de los no diabéticos estas son no tener colesterol, triglicéridos y presión alta.

Cabe mencionar que cada intercesión en los conjuntos presento variable similares aunque fue las personas seleccionadas las que presentaron ciertas características.

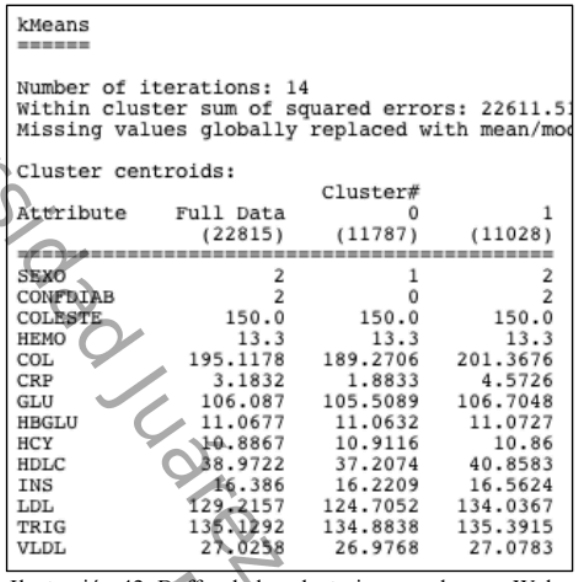
4.1.4 Clustering (Base de datos Glucosa y Lípidos)

El análisis de clustering se realizó sobre las variables que se muestra en la tabla 23:

	Variable	Descripción
1	**Sexo**	Sexo (Masculino, Femenino)
2	**CONFDIAB**	Confirmación de persona diabética
3	**COLESTE**	Cantidad de colesterol
4	**CRP**	mg/dL Proteína C reactiva determinada mediante nefelómetro
5	**GLU**	mg/dL Glucosa determinada mediante analizador clínico
6	**HEMO**	Cantidad de hemoglobina
7	**HDLC**	mg/dL Colesterol de alta densidad determinado mediante analizador clínico
8	**LDL**	Lipoproteínas de Baja Densidad determinadas mediante analizador clínico
9	**INS**	microU/ml de Insulina determinada mediante inmunoanalizador
10	**HBGLU**	Porcentaje de Hemoglobina glicosilada determinada mediante analizador clínico
11	**TRIG**	Triglicéridos determinados mediante analizador clínico
12	**COL**	mg/dL Colesterol total determinado mediante inmunoanalizador
13	**HCY**	micromoles/dL Homocisteína determinada por HPLC
14	**VLDL**	Lipoproteínas de Muy Baja Densidad determinadas mediante analizador clínico

Tabla 23. Variables Base de datos Glucosa y Lípidos

En la ilustración 42 se muestran las dos agrupaciones obtenidas de la base de datos glucosa y lípidos utilizando el algoritmo *SimpleKmeans*.



```
kMeans
======

Number of iterations: 14
Within cluster sum of squared errors: 22611.5
Missing values globally replaced with mean/mo

Cluster centroids:
                          Cluster#
Attribute    Full Data          0          1
               (22815)    (11787)    (11028)
=============================================
SEXO                2          1          2
CONFDIAB            2          0          2
COLESTE         150.0      150.0      150.0
HEMO             13.3       13.3       13.3
COL           195.1178   189.2706   201.3676
CRP             3.1832     1.8833     4.5726
GLU            106.087   105.5089   106.7048
HBGLU          11.0677    11.0632    11.0727
HCY            10.8867    10.9116      10.86
HDLC           38.9722    37.2074    40.8583
INS             16.386    16.2209    16.5624
LDL            129.2157   124.7052   134.0367
TRIG           135.1292   134.8838   135.3915
VLDL            27.0258    26.9768    27.0783
```

Ilustración 42. Buffer de los clustering creados por Weka

Descripción de las agrupaciones:

- **Clúster 0.-** Esta agrupación es prevalecida por hombres quienes no están confirmados como diabéticos, su colesterol esta en 150 mg/dL, la hemoglobina es de 13.3 mg/dL, el colesterol determinado por inmunizador es de 189.2706 mg/dL, la proteína C reactiva determinada mediante nefelómetro es de 1.8833, la glucosa es 105.5089, el Colesterol de alta densidad es de 37.2074 mg/dL. Aquí se encuentran el 52 % de los registros.

- **Clúster 1.-** Conformado por mujeres no confirmadas como diabéticas, su colesterol esta en 150 mg/dL, la hemoglobina es de 13.3 mg/dL, el colesterol determinado por inmunizador es de 201.3676 mg/dL, la proteína C reactiva determinada mediante nefelómetro es de 4.5726, la glucosa es 106.7048 mg/dL, el colesterol de alta densidad es de 40.8583 mg/dL. En esta agrupación están el 48% de los datos.

De los clustering se puede decir que tienen características similares en el colesterol y hemoglobina, para las demás variables los datos son variables.

Por otra parte se puede especificar que en el clúster 0 hay 11,787 personas. Mientras que en el clúster 1 hay 11,028 personas. Todo esto puede visualizarse en ilustración 43

```
Clustered Instances

0       11787 ( 52%)
1       11028 ( 48%)
```

Ilustración 43. Porcentaje de cada agrupación

Después de obtener los clustering , se realizó la visualización de las variables Confidiab(Confirmado diabético) , COL(Colesterol) que puede ser visto en la ilustración 44.

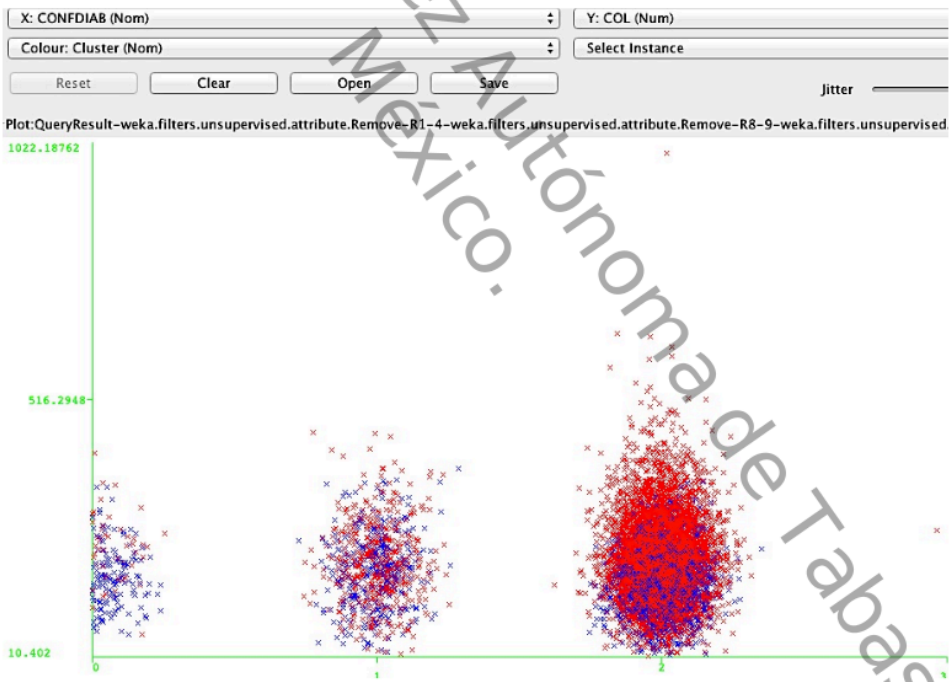


Ilustración 44. Visualización de las variables CONFDIAB y COL

Describiendo la visualización de como se representan las variables en el eje X y Y del cuadrante del plano cartesiano. En el eje X se encuentra la variable Confidiab como se puede ver en la ilustración 44. aquí están agrupados de acuerdo a los valores (0,1,2,3) del atributo, mientras que en el eje Y se encuentran los valores de las variable COL.

De acuerdo a la visualización, se tomaron 4 personas una de cada grupo para realiza la caracterización. Esto se muestra en la tabla 24.

Clúster 1				
Variable				
CONFDIAB (Confirmación de persona diabético (a))	**No especifico**	**Si**	**No**	**Si en el embarazo**
Sexo (Hombre, Mujer)	Hombre	Mujer	Hombre	Mujer
TRIG (Triglicéridos determinados mediante analizador clínico)	128.21	83.59	140.70	104.43
COL (mg/dL Colesterol total determinado mediante inmunoanalizador)	179.63	257.64	476.08	220.44
GLU mg/dL Glucosa determinada mediante analizados clínico	115.69	378.89	126.40	263.10
HDLC (mg/dL Colesterol de alta densidad determinado mediante analizado clínico.	33.71	67.694	79.43	46.31
LDL (Lipoproteínas de baja densidad determinadas mediante analizador clínico.	120.27	173.22	368.51	153.23
INS (microU/ml de Insulina determinada mediante inmunoanalizador)	14.9	6.0	33.40	6.5
HBGLU (Porcentaje de Hemoglobina glicosilada determinada mediante analizador clínico)	-	13.4	-	12.6
COLESTE (Cantidad de colesterol)	-	223.0	196.00	180.00

Tabla 24. Características de diabéticos y no diabéticos clúster 1

Las características encontradas dentro de los grupos son que quien no especifico que saber si tiene o no la patología sus niveles de triglicéridos oscilan en 128.21 mg/dl, el colesterol es de 179.63 mg/dl, la glucosa 115.69 mg/dl y su insulina es de 14.0 microU/ml.

Quienes especificaron tener diabetes sus niveles de triglicéridos fluctúan en 83.59 mg/dl, el colesterol es de 257.64 mg/dl, la glucosa 378.89 mg/dl y su insulina es de 6.0 microU/ml. Por otra parte las mujeres que en su embarazo fueron confirmadas con diabetes cuentan con las siguientes características: triglicéridos oscilan en 104.43 mg/dl, el colesterol es de 220.44 mg/dl, la glucosa 263.10 mg/dl y su insulina es de 6.5 microU/ml. En la ilustración 45 se muestra gráficamente la distribución de valores

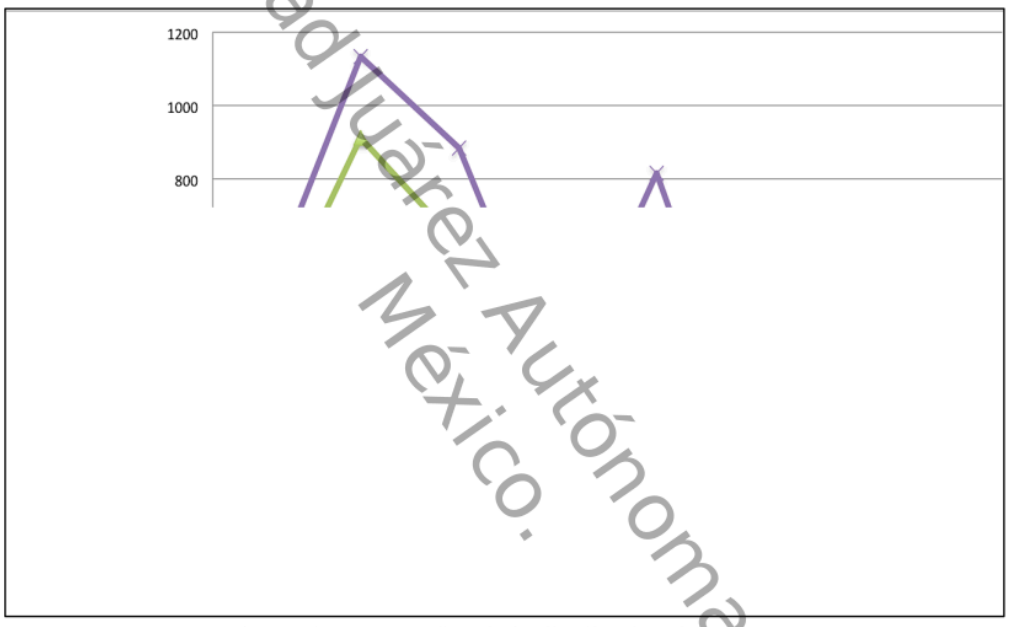


Ilustración 45. Gráfica de la distribución de valores

Se puede observar que los valores de la glucosa están a un nivel mayor 130 mg/dl cuando las personas presentan diabetes, así mismo el colesterol presentan en un nivel mayor a los 200 mg/dl. Otro aspecto por mencionar es que cuando una persona presenta diabetes la insulina esta presente en menor cantidad.

Por otra parte quienes no tienen diabetes presentan sus nivel de colesterol y triglicéridos altos, mucho mas que los que tienen desarrollada la patología, esto puede considerarse como factor de riesgo para presentar la enfermedad.

4.2 Evaluación de los resultados

En este parte del documento se describen de forma general resultados obtenidos de la minería de datos en función de los objetivos de la investigación.

Cada clustering o árbol de decisión realizado a través de técnicas de minería de datos contiene información detallada sobre algunas características y relaciones de los pacientes diabéticos y no diabéticos. A continuación se realizara la descripción del comportamiento de la información al ser utilizadas herramientas de explotación de datos.

El árbol de decisión creado a través de la evaluación de la variable a302c (¿Hace cuánto tiempo le dijo su médico por primera vez que tenía diabetes o el azúcar alta en la sangre?) se caracterizo por mostrar si existe dependencia en tener o no diabetes y el tiempo que las personas empieza a presentar problemas de presión arterial alta.

Algunos de los hallazgos que se pueden mencionar son que existen personas con 10 años con diabetes y no presentan presión alta, por otra parte hay quienes su presión alta ha sido alta desde hace 10 años y llevan menor tiempo en que les diagnosticaron diabetes.

La clasificación de las variables a310c, a310g y a310f permitió conocer la cantidad de veces que las personas realizan pruebas de rutinarias para llevar el seguimiento de la diabetes unas vez enterado de tener la patología.

Por otra parte los clustering o agrupaciones realizadas permitieron hacer la base de datos en grupos, conocer el numero de personas que están en ellas y cuales son las características que presentan cada individuo del mismo. Para la base de datos (Adultos de 20 años o más) grupos fueron evaluados en función de la variable a301, permitiendo segmentar los cuatro grupos en ocho agrupaciones, divididos en quienes si presentan diabetes y quienes no.

4.3 Evaluación del grupo objeto

Con la necesidad de asegurar la certeza del conocimiento encontrado en las bases de datos se redacto un reporte del conocimiento encontrado (anexo G) que fue evaluado por los colaboradores (Médicos Internistas), para obtener su comentarios se les dio un cuestionario. A continuación se presentan las preguntas realizadas y las respuestas de los médicos.

Cuestionario

1.- Los resultados mostrados ¿Qué tan acertados son respecto a criterios de edad, seco y otras características de los prediabéticos y diabéticos?

Médico 1.- Son acertados, pero tendríamos que checar otros factores de riesgo.

Médico 2.- Considerando un 80% de certeza, con respecto a criterios y sexo.

Médico 3.- Es aceptable la distribución respecto a los grupos, edad y sexo.

Medico 4.- Adecuados. Tienen compatibilidad con los pacientes que se ven al día y se toman en cuenta criterios de laboratorio conforme a edad así como resultados del paciente.

2.-¿Como evaluaría usted los resultados encontrados con las técnicas de minería de datos?

Médico 1.- Muy buenos, pero como mencione en la pregunta anterior has ocasiones que hay que considerar otros factores como la ingesta de alimentos, obesidad etc.

Médico 2.- Si hay concordancia con la realidad en cuanto a la edad del diagnóstico y la sintomatología.

Médico 3.- Los resultados me parecen correctos, ya que por estudios el grupo de edad de personas diabéticas prevalecen en ese grupo de edad y sexo. En mi opinión se puede buscar otra forma de presentar la información encontrada.

Medico 4.- Son buenos, pero dependería de los valores de referencia que se tengan o que se obtengan. Colesterol <150 mg/dl, Glucosa<100 mg/dl.

3.- ¿En que otras áreas medicas cree usted que pueda ser de ayuda aplicar minería de datos?

Médico 1.-

- Epidemiología

Médico 2.-

- Pacientes con obesidad
- Enfermedades crónicas como IRC

Médico 3.-

- Enfermedades crónico degenerativas
- Epidemiologia
- Accidentes

Médico 4.-

- Epidemiologia
- Ginecología
- Prevención médica

4.- ¿De acuerdo a los resultados mostrados, cree usted que utilizar herramientas tecnológicas como minería de datos, pueda servir como instrumento de apoyo para tomar decisiones y planear estrategias en el combate de la diabetes?

Médico 1.- Si, Se conocería cuales personas están en riesgo de presentar o no la diabetes y podríamos disminuir esos riesgos y por tanto la diabetes.

Médico 2.- Si, Como herramienta tecnológica es de gran utilidad ya que disminuiría los tiempos en el proceso de la información, además seria muy útil en estadísticas y toma de decisiones. Muy practico en realidad.

Médico 3.- Si, Se sintetiza y facilita a la hora de tabular toda la información, hay menos forma de equivocarse, se ahorra tiempo y recursos, estos últimos creo que son lo mas importante para un buen estudio.

Médico 4.- Si, Se puede tomar laboratorios en cuanto se empieza a sospechar del diagnostico probable. Con la diabetes hemoglobina, colesterol, glucosa principalmente, urea etc. Y así administrar o ajustar el tratamiento adecuado.

5.- ¿Qué tan innovador le es a usted que se estén buscando nuevas formas de interpretar los grandes volúmenes de información médica?

Médico 1.- De alguna manera muy bueno, para conocer el riesgo no de personas sino de una población.

Médico 2.- Considero de gran importancia ya que reduciría el rango de tiempo, evitar retraso en los diagnósticos y atención oportuna y pronta en estos pacientes

Médico 3.- Es aceptable, ojala se tenga competencia para ayudar a la población, ya que la medicina preventiva es el primer paso para todas las enfermedades.

Medico 4.- Bien, seria un buen método de utilidad para interpretar la información.

Los descubrimientos del proceso de minería de datos son los siguientes:

- Existen personas con diabetes desde hace 10 años pero desconocen si tienen presión alta.
- Hay personas que tienen presión alta desde hace 6 años, y llevan 10 años con diabetes.
- Se descubrió que quienes asumen que tienen 10 años con presión alta, tiene 5 años con diabetes.
- El promedio de exámenes de general de orina es de 13.5 veces en 12 meses.
- El automonitoreo es realizado 27.31 veces en 12 meses.
- Las personas no diabéticas están en un rango de edad de 21 a 35 años pero empiezan mostrar factores de riesgo como el colesterol alto.
- Las mujeres en edad de 45 a 60 años son diabéticas, empiezan a manifestar otros tipos de padecimientos como presión alta y triglicéridos.
- Los hombres en edad de 55 a 65 años, son diabéticos aunque no manifiestan ningún otro padecimiento.

Capítulo V. Conclusiones y trabajos futuros

Este capítulo tiene la finalidad de mostrar las conclusiones que se llegaron después de desarrollar metodológicamente la investigación, así como los trabajos futuros del mismo por lo que a continuación se detallan cada uno de ellos.

5.1 Conclusiones

Cabe mencionar dos aspectos que intermedian en la conclusión de este trabajo: el primero es dar respuesta a la pregunta de investigación y el segundo es el cumplimiento del objetivo planteado.

Para concluir con el primer aspecto es dar la respuesta a la pregunta de investigación:

¿Qué técnicas de minería de datos, permitirán obtener los patrones de comportamiento en los grandes volúmenes de información de pacientes Pre-diabéticos?

Las técnicas que permitieron obtener patrones de comportamiento fueron los árboles de decisión específicamente la técnica J48 que están en la categoría de los algoritmos supervisados o predictivos, así como la técnica de clustering (agrupación) en la que se utilizó el algoritmo SimpleKmeans que esta en la categoría de los no supervisados o descubrimiento de conocimiento que permitió determinar las características por grupo o individuo.

El objetivo de esta investigación se cumple para el caso de estudio de detección de patrones de comportamiento de pacientes Pre-diabéticos y Diabéticos, esto porque permitió descubrir y conocer sus características como edad, sexo, niveles de colesterol, triglicéridos, glucosa e insulina representadas a través de los arboles de decisión y clusters mostrados por la herramienta de minería de datos.

Los patrones de comportamiento encontrados (Anexo G) a través de la herramienta de minería de datos (Weka) permite llegar a la conclusión que el utilizar tecnologías dedicada a la extracción del conocimiento es una solución de mucha utilidad para descubrir contextos ocultos en los grande volúmenes de información.

Se concluye que el modelo CRISP-DM fue de mucha ayuda debido a que permitió orientar los objetivos del plan del proyecto, ya que suministra un delineación de un ciclo de vida para minería de datos. Integrado por actividades y tareas propias de la evolución del proyecto.

Sin duda alguna la solución tecnológica presentada en este documento, puede funcionar de manera útil en la medicina, como apoyo en verificación de riesgos que presentan cierto grupo de personas con enfermedades no trasmisibles como la diabetes. También puede ser utilizado como punto de partida para tomar decisiones y formular estrategias de prevención y control rutinario de diabetes.

5.2 Trabajos Futuros

Tomando en cuenta las respuestas de la pregunta numero tres del cuestionario aplicado al grupo objeto. Esta investigación da pauta para nuevas investigaciones :

- Utilización de herramienta de minería de datos para la extracción del conocimientos en datos médicos del padecimiento crónico IRC.
- Minería de datos en bases de datos epidemiológicas
- Análisis de clúster de datos de pacientes del área de Ginecología
- Evaluación de bases de datos de medicina familia para la prevención médica
- Detección de patrones de comportamiento en bases de datos de pre-diabéticos y diabéticos realizando agrupaciones por año.
- Evaluar técnicas de minería para la detección de patrones de comportamiento de bases de datos de pacientes con cáncer de mama.
- Análisis de clúster de datos de pacientes con cardiopatías.

Bibliografía

Ángeles, M. I., & Santillán, A. M. (2004). Minería de datos: Concepto, características, estructura y aplicaciones. Recuperado de http://www.ejournal.unam.mx/rca/190/RCA19007.pdf

Arancibia, J. A. G. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. Recuperado de http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

BI Analytics. Inteligencia de negocios [BI Services]- Construcción de modelos de minería de datos [BIA DataMining]. Recuperado de http://bianalytics.biz/index.php/bi-analytics-inteligencia-de-negocios-bi-services/75-construccion-mineria-datos

Cervantes, A. M., López, V. G., & Gayosso, G. Y. (2010). Minería de Datos. Recuperado de http://www.ingenieria.buap.mx/DOCUMENTOS/REVISTA/REV_11/art_4.pdf

Césari, M. (2002). Minería de Datos, 1-42. Recuperado de http://ai.frm.utn.edu.ar/micesari/files/01_Matilde_Mineria_datos.pdf

Dueñas, M. X. (2009). Minería de datos espaciales en búsqueda de la verdadera información *. *Energy*, *13*(1), 137-156. Recuperado de http://redalyc.uaemex.mx/pdf/477/47711998007.pdf

Elmasri, R., & Navathe, S. B. (2005). *Sistemas de bases de datos (Conceptos fundamentales)*. (2° ed.) (A. WESLEY, Ed.)

ENSANUT (2012). Encuesta Nacional de Salud y Nutrición- Resultados Nacionales. Recuperado de http://ensanut.insp.mx/doctos/ENSANUT2012_Sint_Ejec-24oct.pdf

Fayyad U., Piatesky-Shapiro G., Smyth P. (1996) Advances in Knowledge Discovery and Data Mining, (3° ed)Madrid:MIT Press.

Fernández, O., Jiménez, G., González, B. E., & Ávila, J. (2007) Aplicación de minería de datos al sistema cubano de Farmacolovigilancia. Recuperado de http://scielo.sld.cu/pdf/far/v41n3/far03307.pdf

Flores, H. D. (2009). *Detección de Patrones de Daños y Averías en la Industria Automotriz*. Universidad Tecnológica Nacional de Buenos Aires. Recuperado de http://www.unla.edu.ar/sistemas/gisi/tesis/flores-tesisdemagister.pdf

Gallardo, J. A. (2009). *Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM)*. Universidad Politécnica de Madrid. Recuperado de http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

García, M., Miguel, L. A., Garcia, F. J., & Polo, M. J. (1997). *Construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software* (pp. 1-14). Recuperado de http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf

GNU Operating System. GNU PSPP. Recuperado de http://www.gnu.org/software/pspp/

Gómez, A. M. M., & Iglesias, M. E. D. (2004). Análisis documental y de información: dos componentes de un mismo proceso. Recuperado de http://bvs.sld.cu/revistas/aci/vol12_2_04/aci11204.htm

Hernández, R., Collado, C.F, & Baptista, L. (2003). Metodología de la Investigación. (3° ed.) McGraw-Hill.

Hernández, J., Ramírez, M. & Ferri, C. (2004). Introducción a la Minería de Datos, (2° ed.) Ed. Pretice Hall

Limite, C. (2012). Minería de Datos. Retrieved May 22, 2012, Recuperado de http://ciclolimite.com/mineria-de-datos/

Lozaya, S. V. M., Jiménez, R. Y. R., & Flores, E. karina C. (2010). ITSON | Paradigma de Investigación Cuantitativa | Características de la Investigación Cuantitativa. Recuperado de http://biblioteca.itson.mx/oa/educacion/oa3/paradigmas_investigacion_cuantitativa/p3.htm

Maneiro, M. Y. (2008). "Minería de Datos." Recuperado de http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosYany2008.pdf

Morate, D. G. (2011). Manual de Weka. Recuperado de http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf

32

Mohammad N, Namakforoosh(2000) Metodología de la investigación. (2ª ed.). (Versión Digital) Recuperado el 9 de mayo del 2012 de http://books.googl.co.ve/books?id=ZEJ7-0hmvhwC

Microsoft- Soporte. ¿Que es Excel?. Recuperado de http://office.microsoft.com/es-mx/excel-help/que-es-excel-HA010265948.aspx

Méndez, P. D., & Rodríguez, A. D. (2009). *VIABILIDAD PARA PROYECTOS QUE UTILIZAN LA METODOLOGÍA P 3 TQ Laboratorio de Sistemas Inteligentes Índice*. Universidad de Buenos Aires. Recuperado de http://laboratorios.fi.uba.ar/lsi/mendez-rodriguez-trabajoprofesional.pdf

Organización Mundial de la Salud (2011) Diabetes. Recuperado de http://www.who.int/mediacentre/factsheets/fs312/es/index.html

Patton, M.Q. (2002). Qualitative research and evaluation methods. 2a. ed. New York: SAGE.

Gobierno de la Republica (2013). Plan Nacional de Desarrollo 2013-2018. Recuperado de http://pnd.gob.mx

Pinto, M. A., & Torres, I. H. (2006). *Metodología de Evaluación de Técnicas de la Minería de Datos Aplicadas a Datos Biológicos*. Benemerita Universidad Autonoma de Puebla. Recuperado de http://perseo.cs.buap.mx/bellatrix/tesis/TES219.pdf

Rabinowitz, P., & Fawcett, S. (2011). Recolectar y analizar informacion. Recuperado de http://ctb.ku.edu/es/tablecontents/capitulo37_seccion5_seccion_principal.aspx

Robbins, D. E., & Chiesa, M. (2011). Clinical applications of data mining. In V. P. Gurupur & S. Suh (Eds.), *Clinical Applications of Data Mining*. Springer.

Ruiz, K. (2010). Instrumentos para recopilación de datos. Recuperado de http://cienciassocialeskathy.obolog.com/instrumentos-recopilacion-datos-608877

Sierra, M. (2012). ¿QUÉ ES UNA BASE DE DATOS Y CUÁLES SON LOS PRINCIPALES TIPOS? EJEMPLOS: MYSQL, SQLSERVER, ORACLE, POSTGRESQL, INFORMIX… (DV00204A). Recuperado de http://www.aprenderaprogramar.com/index.php?option=com_attachments&task=download&id=500

Tabuenca, J. G. (2010). Introducción a la Minería de Datos. Recuperado de http://tornillosygenes.com/2011/05/19/introduccion-a-la-mineria-de-datos/

Weka (2013) . Data Mining Software in Java. Recuperado de http://www.cs.waikato.ac.nz/ml/weka/

Glosario

ENSANUT.- Encuesta Nacional de Salud y Nutrición.

OMS.- Organización Mundial de la Salud.

CRISP-DM.- Cross Industry Standard Process for Data Mining) Procedimiento Industrial Estándar para realizar Minería de Datos.

Patrón.- tipo de tema de sucesos u objetos recurrentes.

Comportamiento.- es la manera de proceder que tienen las personas u organismos, en relación con su entorno o mundo de estímulos.

Diabetes.- Enfermedad crónica e irreversible del metabolismo en la que se produce un exceso de glucosa o azúcar en la sangre y en la orina; es debida a una disminución de la secreción de la hormona insulina o a una deficiencia de su acción.

KDD.- Descubrimiento de Conocimiento en bases de datos.

PSPP.- es un una aplicación de software libre para el análisis de datos.

* **CSV.-** son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas.

***.arff.-**Formato de archivo de Weka y se pueden utilizar para tareas tales como la agrupación de datos y la regresión.

J48.- es un algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan.

SimpleKmeans.- es un método de agrupamiento, que tiene como objetivo la [partición [de un conjunto]] n en k grupos en el que cada observación pertenece al grupo más cercano a la media.

Anexos

Anexo A. Buffer de algoritmo J48 en BD adultos 20 años o mas

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: BASEMINERALLpruebaV01.1.arff-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 46277
Attributes: 159
[list of attributes omitted]
Test mode:10-fold cross-validation

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

J48 pruned tree

---

a301 = 1
| a402b = .: 122 (2398.0/2197.0)
| a402b = 6.0: 74 (71.0/51.0)
| a402b = 2.0: 122 (167.0/105.0)
| a402b = 0.0: 122 (350.0/323.0)
| a402b = 10.0: 62 (146.0/99.0)
| a402b = 12.0: 26 (41.0/28.0)
| a402b = 15.0: 146 (94.0/58.0)
| a402b = 5.0: 182 (136.0/99.0)
| a402b = 4.0: 50 (113.0/72.0)
| a402b = 7.0: 278 (69.0/41.0)
| a402b = 18.0: 350 (23.0/12.0)
| a402b = 1.0: 242 (177.0/130.0)
| a402b = 30.0: 98 (42.0/34.0)
| a402b = 3.0: 158 (155.0/108.0)
| a402b = 41.0: 110 (4.0/3.0)
| a402b = 8.0: 170 (77.0/49.0)
| a402b = 20.0: 206 (99.0/70.0)
| a402b = 99.0: 310 (34.0/27.0)
| a402b = 13.0: 338 (26.0/12.0)
| a402b = 17.0: 194 (20.0/12.0)
| a402b = 32.0: 26 (4.0/3.0)
| a402b = 11.0: 266 (24.0/17.0)
| a402b = 25.0: 422 (34.0/25.0)
| a402b = 44.0: 62 (0.0)
| a402b = 50.0: 74 (6.0/5.0)
| a402b = 27.0: 470 (5.0/3.0)
| a402b = 23.0: 242 (7.0/4.0)
| a402b = 14.0: 134 (23.0/13.0)
| a402b = 40.0: 86 (12.0/8.0)
| a402b = 22.0: 110 (7.0/4.0)
| a402b = 28.0: 314 (14.0/11.0)
| a402b = 9.0: 254 (27.0/12.0)
| a402b = 16.0: 374 (13.0/9.0)
| a402b = 21.0: 410 (9.0/7.0)
| a402b = 35.0: 158 (9.0/7.0)
| a402b = 19.0: 234 (3.0/2.0)
| a402b = 48.0: 62 (0.0)
| a402b = 29.0: 182 (2.0/1.0)
| a402b = 63.0: 602 (1.0)
| a402b = 26.0: 170 (5.0/3.0)
| a402b = 24.0: 26 (7.0/5.0)
| a402b = 42.0: 62 (3.0/2.0)
| a402b = 33.0: 50 (1.0)
| a402b = 38.0: 182 (2.0/1.0)
| a402b = 36.0: 134 (6.0/5.0)
| a402b = 45.0: 338 (2.0/1.0)
| a402b = 75.0: 494 (1.0)

```
|   a402b = 52.0: 62 (0.0)
|   a402b = 67.0: 62 (0.0)
|   a402b = 39.0: 26 (1.0)
|   a402b = 55.0: 62 (0.0)
|   a402b = 34.0: 62 (5.0/4.0)
|   a402b = 70.0: 590 (1.0)
|   a402b = 65.0: 62 (0.0)
|   a402b = 83.0: 230 (1.0)
|   a402b = 61.0: 62 (0.0)
|   a402b = 49.0: 62 (0.0)
|   a402b = 53.0: 122 (2.0/1.0)
|   a402b = 31.0: 410 (1.0)
|   a402b = 51.0: 194 (2.0/1.0)
|   a402b = 37.0: 74 (3.0/2.0)
|   a402b = 66.0: 310 (1.0)
|   a402b = 43.0: 62 (0.0)
|   a402b = 68.0: 62 (0.0)
|   a402b = 73.0: 50 (1.0)
|   a402b = 71.0: 62 (0.0)
|   a402b = 54.0: 62 (0.0)
|   a402b = 46.0: 62 (0.0)
|   a402b = 47.0: 290 (1.0)
|   a402b = 79.0: 206 (1.0)
|   a402b = 60.0: 62 (0.0)
|   a402b = 69.0: 62 (0.0)
|   a402b = 84.0: 62 (0.0)
|   a402b = 59.0: 62 (0.0)
|   a402b = 78.0: 62 (0.0)
|   a402b = 82.0: 422 (1.0)
a301 = 2: 13 (41787.0)

Number of Leaves  :  77
Size of the tree :  79


Time taken to build model: 27.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       42548
91.942 %
Incorrectly Classified Instances      3729
8.058 %
Kappa statistic                  0.5621
Mean absolute error              0.0012
Root mean squared error          0.0251
Relative absolute error         47.9868 %
Root relative squared error     70.2543
%
Total Number of Instances        46277
```

Anexo B Buffer de algoritmo J48 a la variable a310c en BD adultos 20 años o mas.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: BASEMINERALLpruebaV01.1.arff-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 46277
Attributes: 159
[list of attributes omitted]
Test mode:10-fold cross-validation
== Classifier model (full training set) ==

J48 pruned tree

---

a301 = 1
| a311e3 = .: 0.0 (3005.0)
| a311e3 = 1.0: 1.0 (313.0)
| a311e3 = 2.0: 1.0 (336.0)
| a311e3 = 6.0: 1.0 (132.0)
| a311e3 = 3.0: 1.0 (258.0)
| a311e3 = 4.0: 1.0 (189.0)
| a311e3 = 8.0: 1.0 (14.0)
| a311e3 = 12.0: 1.0 (141.0)
| a311e3 = 5.0: 1.0 (36.0)
| a311e3 = 7.0: 1.0 (14.0)
| a311e3 = 99.0: 1.0 (17.0)
| a311e3 = 10.0: 1.0 (13.0)
| a311e3 = 11.0: 1.0 (7.0)
| a311e3 = 9.0: 1.0 (7.0)
| a311e3 = 14.0: 1.0 (1.0)
| a311e3 = 24.0: 1.0 (1.0)
| a311e3 = 48.0: 1.0 (1.0)
| a311e3 = 16.0: 1.0 (1.0)
| a311e3 = 40.0: 1.0 (1.0)
| a311e3 = 13.0: 1.0 (1.0)
| a311e3 = 22.0: 1.0 (1.0)
| a311e3 = 15.0: 1.0 (1.0)
a301 = 2: . (41787.0)

Number of Leaves : 23

Size of the tree : 25

Time taken to build model: 0.79 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	46269
99.9827 %	
Incorrectly Classified Instances	8
0.0173 %	
Kappa statistic	0.999
Mean absolute error	0.0001
Root mean squared error	0.0072
Relative absolute error	0.0645 %
Root relative squared error	2.9376 %
Total Number of Instances	46277

Anexo C Buffer de algoritmo J48 a la variable a310f en BD adultos 20años o mas.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: BASEMINERALLpruebaV01.1.arff-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 46277
Attributes: 159
[list of attributes omitted]
Test mode:10-fold cross-validation

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

J48 pruned tree

---

a301 = 1
| a311e6 = .: 0.0 (4311.0)
| a311e6 = 5.0: 1.0 (4.0)
| a311e6 = 4.0: 1.0 (28.0)
| a311e6 = 2.0: 1.0 (39.0)
| a311e6 = 1.0: 1.0 (36.0)
| a311e6 = 12.0: 1.0 (19.0)
| a311e6 = 3.0: 1.0 (18.0)
| a311e6 = 99.0: 1.0 (5.0)
| a311e6 = 9.0: 1.0 (3.0)
| a311e6 = 6.0: 1.0 (20.0)
| a311e6 = 11.0: 1.0 (1.0)
| a311e6 = 10.0: 1.0 (1.0)
| a311e6 = 7.0: 1.0 (1.0)
| a311e6 = 16.0: 1.0 (1.0)
| a311e6 = 8.0: 1.0 (2.0)
| a311e6 = 13.0: 1.0 (1.0)
a301 = 2: . (41787.0)

Number of Leaves : 17

Size of the tree : 19

Time taken to build model: 0.72 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	46272 99.9892 %
Incorrectly Classified Instances	5 0.0108 %
Kappa statistic	0.9994
Mean absolute error	0.0001
Root mean squared error	0.0081
Relative absolute error	0.0589 %
Root relative squared error	3.3648 %
Total Number of Instances	46277

Anexo D Buffer de algoritmo J48 a la variable a310g en DB adultos 20años o mas.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: BASEMINERALLpruebaV01.1.arff-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 46277
Attributes: 159
[list of attributes omitted]
Test mode:10-fold cross-validation

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

J48 pruned tree

---

a301 = 1
| a311e7 = .: 0.0 (4434.0)
| a311e7 = 6.0: 1.0 (6.0)
| a311e7 = 1.0: 1.0 (5.0)
| a311e7 = 96.0: 1.0 (1.0)
| a311e7 = 99.0: 1.0 (3.0)
| a311e7 = 2.0: 1.0 (5.0)
| a311e7 = 4.0: 1.0 (2.0)
| a311e7 = 12.0: 1.0 (11.0)
| a311e7 = 40.0: 1.0 (2.0)
| a311e7 = 10.0: 1.0 (1.0)
| a311e7 = 50.0: 1.0 (5.0)
| a311e7 = 8.0: 1.0 (1.0)
| a311e7 = 18.0: 1.0 (2.0)
| a311e7 = 7.0: 1.0 (2.0)
| a311e7 = 48.0: 1.0 (2.0)
| a311e7 = 5.0: 1.0 (1.0)
| a311e7 = 36.0: 1.0 (2.0)
| a311e7 = 95.0: 1.0 (1.0)
| a311e7 = 20.0: 1.0 (2.0)
| a311e7 = 58.0: 1.0 (1.0)
| a311e7 = 3.0: 1.0 (1.0)
a301 = 2: . (41787.0)

Number of Leaves : 22

Size of the tree : 24
Time taken to build model: 0.6 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	46270	99.9849 %
Incorrectly Classified Instances	7	0.0151 %
Kappa statistic	0.9991	
Mean absolute error	0.0001	
Root mean squared error	0.0099	
Relative absolute error	0.0851 %	
Root relative squared error	4.1007 %	
Total Number of Instances	46277	

Anexo E Buffer de algoritmo SimpleKmeans BD adultos 20años o mas.

Test mode:evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 5
Within cluster sum of squared errors: 89482.0
Missing values globally replaced with mean/mode
Cluster centroids: Cluster#

Attribute	Full Data	0	1	2	3
	(46277)	(16162)	(695)	(17370)	(12050)
Sexo	2	2	2	2	1
Edad	38	36	29	38	20
a301	2	2	2	2	2
a401	2	2	2	2	2
a501	2	2	2	2	2
a502b	2	2	2	2	2
a502c	2	2	2	2	2
a502d	2	2	2	2	2
a601	3	3	3	1	3
a603	3	3	3	1	3
a604	2	2	2	2	2
a302c	13	13	13	13	13

Time taken to build model (full training data) : 0.88 seconds
=== Model and evaluation on training set ===

Clustered Instances
0 16162 (35%)
1 695 (2%)
2 17370 (38%)
3 12050 (26%)
Class attribute: a301

Classes to Clusters:
0 1 2 3 <-- assigned to cluster
1506 3 2133 848 | 1
14656 692 15237 11202 | 2

Cluster 0 <-- 2
Cluster 1 <-- No class
Cluster 2 <-- 1
Cluster 3 <-- No class

Anexo F Buffer de algoritmo SimpleKmeans BD Glucosa y Lípidos

Test mode:evaluate on training data

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

Number of iterations: 14
Within cluster sum of squared errors: 22611.51676882856
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(22815)	(11787)	(11028)
SEXO	2	1	2
CONFDIAB	2	0	2
COLESTE	150.0	150.0	150.0
HEMO	13.3	13.3	13.3
COL	195.1178	189.2706	201.3676
CRP	3.1832	1.8833	4.5726
GLU	106.087	105.5089	106.7048
HBGLU	11.0677	11.0632	11.0727
HCY	10.8867	10.9116	10.86
HDLC	38.9722	37.2074	40.8583
INS	16.386	16.2209	16.5624
LDL	129.2157	124.7052	134.0367
TRIG	135.1292	134.8838	135.3915
VLDL	27.0258	26.9768	27.0783

Time taken to build model (full training data) : 1.47 seconds
=== Model and evaluation on training set ===
Clustered Instances

0 11787 (52%)
1 11028 (48%)

Algoritmo de clasificación

2

-
-
-
-
-
-
-

- Por lo general el examen de Microalbuminuria es realizado en un promedio de 7.64 veces en 12 meses, lo que se puede decir que las personas con diabetes se realizan una prueba de control cada 2 meses.

Clustering (Agrupación)

De la aplicación del algoritmo de agrupación se obtuvieron grupos (Clúster). A continuación se describen algunas características que presentan los grupos:

- **Clúster 0.-** Es prevalecido por Mujeres en edad de 38 años, ningún médico le ha dicho que tiene presión alta, no tienen diabetes y por consecuente no presentado ningún padecimiento relacionado con la patología.
- **Clúster 1.-** Conformado en su mayoría por Mujeres en edad de 36 años, no presentan diabetes, tampoco presentan presión alta o colesterol.
- **Clúster 2.-** Al igual que el clúster 0 predominan las mujeres en edad de 38 años, estas han presentado colesterol alto, así como triglicéridos en limite establecido para considerarlo como normal.
- **Clúster 3.-** Se constituye por hombres en edad de 20 años, no presentan diabetes, tampoco padecen colesterol o triglicéridos.

Sobre las agrupaciones realizadas se puede expresar que en algunas mujeres en edad 38 años en adelante empiezan a presentar elementos que pueden llevar a detonar la diabetes. Mientras que los hombres en la misma edad no presentan ningún criterio que lo lleve a considerarlos en riesgo de desarrollo de la patología. De estas agrupaciones se eligieron aleatoriamente a 4 personas, 2 con diabetes y 2 sin la patología, a continuación se describen las características que presentan los pares de personas:

Clúster 1. Las mujeres en edad de 45 a 60 años son diabéticas, empiezan a manifestar otros tipos padecimientos como presión arterial alta, triglicéridos altos. Por otra parte quienes no son diabéticas el rango de edad esta entre 21 a 35 años muestran ciertas características como el colesterol alto. Estas particulares se pueden empezar a considerar como factores de riesgo que lleven a la detonación de la diabetes.

Clúster 2. Las Mujeres en edad de 29 años presentan diabetes, así como triglicéridos altos. Quienes no son diabéticos presentan dos elementos (Colesterol y triglicéridos altos) que pueda ser considerado como factor de riesgo en la presencia de la diabetes.

Clúster 3. Las características de los participantes elegidos que si presentan diabetes se encuentran en un rango de edad de 55 a 65 años, su presión alta es alta en uno de los casos, triglicéridos y colesterol están en el rango normal. Los participantes que representan a quienes no tiene diabetes tienen criterios como edad entre 44 y 69 años, presentan triglicéridos en estado normal, así como un nivel de colesterol normal.

Clúster 4. Los elegidos del clúster se caracterizaron por ser hombres en su totalidad, en primer instancia se describen a los que presentan diabetes estos tienen presión arterial, triglicérido y colesterol en un nivel alto . La edad de quienes poseen diabetes esta entre 50 a 66 años. Describiendo a quienes no presentan diabetes su edad esta entre 45 a 49 años, su colesterol esta en estado normal.

Clúster (Agrupación) Base de datos Glucosa y Lípidos

De la aplicación del algoritmo de agrupación se obtuvieron grupos (Clúster). A continuación se describen algunas características que presentan los grupos :

Clúster 0. Esta agrupación es prevalecida por hombres quienes no están confirmados como diabéticos, su colesterol esta en 150 mg/dL, la hemoglobina es de 13.3 mg/dL, el colesterol determinado por inmunizador es de 189.2706 mg/dL, la proteína C reactiva determinada mediante nefelómetro es de 1.8833, la glucosa es 105.5089, el Colesterol de alta densidad es de 37.2074 mg/dL. Aquí se encuentran el 52 % de los registros

Clúster 1. Conformado por mujeres no confirmadas como diabéticas, su colesterol esta en 150 mg/dL, la hemoglobina es de 13.3 mg/dL, el colesterol determinado por inmunizador es de 201.3676 mg/dL, la proteína C reactiva determinada mediante nefelómetro es de 4.5726, la glucosa es 106.7048 mg/dL, el colesterol de alta densidad es de 40.8583 mg/dL. En esta agrupación están el 48% de los datos.

Se realizó un proceso de visualización de las variables Confirmado Diabético y Colesterol, se obteniendo lo siguiente:

Las características encontradas dentro de los grupos son que quien no especifico que saber si tiene o no la patología sus niveles de triglicéridos oscilan en 128.21 mg/dl, el colesterol es de 179.63 mg/dl, la glucosa 115.69 mg/dl y su insulina es de 14.0 microU/ml. Quienes especificaron tener diabetes sus niveles de triglicéridos fluctúan en 83.59 mg/dl, el colesterol es de 257.64 mg/dl, la glucosa 378.89 mg/dl y su insulina es de 6.0 microU/ml. Por otra parte las mujeres que en su embarazo fueron confirmadas con diabéticas cuentan con las siguientes características: triglicéridos oscilan en 104.43 mg/dl, el colesterol es de 220.44 mg/dl, la glucosa 263.10 mg/dl y su insulina es de 6.5 microU/ml.

DETECCIÓN DE PATRONES DE COMPORTAMIENTO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS EN EXPEDIENTES CLÍNICOS DE PACIENTES PREDIABÉTICOS

ORIGINALITY REPORT

17%
SIMILARITY INDEX

PRIMARY SOURCES

1	ensanut.insp.mx Internet	771 words — 4%
2	repositorioinstitucional.buap.mx Internet	220 words — 1%
3	posgrado.frba.utn.edu.ar Internet	195 words — 1%
4	repositorio.uci.cu Internet	179 words — 1%
5	pag.org.mx Internet	165 words — 1%
6	core.ac.uk Internet	157 words — 1%
7	vbook.pub Internet	152 words — 1%
8	id.scribd.com Internet	97 words — < 1%
9	www.slideshare.net Internet	96 words — < 1%
10	www.bibliomatica.blogspot.com Internet	

		92 words — < 1%
11	es.slideshare.net Internet	82 words — < 1%
12	sedici.unlp.edu.ar Internet	73 words — < 1%
13	idoc.pub Internet	63 words — < 1%
14	chivasraulv.blogspot.com Internet	62 words — < 1%
15	scielo.sld.cu Internet	61 words — < 1%
16	fc5scrim.blogspot.com Internet	60 words — < 1%
17	www.coursehero.com Internet	59 words — < 1%
18	www.clubensayos.com Internet	57 words — < 1%
19	hdl.handle.net Internet	55 words — < 1%
20	openaccess.uoc.edu Internet	50 words — < 1%
21	posgrado.lapaz.tecnm.mx Internet	36 words — < 1%
22	repositorio.unap.edu.pe Internet	33 words — < 1%
23	fcasua.contad.unam.mx	

	Internet	32 words — < 1%
24	frasesdelavida.com Internet	32 words — < 1%
25	www.buenastareas.com Internet	32 words — < 1%
26	repositorio.ug.edu.ec Internet	31 words — < 1%
27	repositorio.uncp.edu.pe Internet	30 words — < 1%
28	www.scribd.com Internet	27 words — < 1%
29	biblioteca.unet.edu.ve Internet	26 words — < 1%
30	e-archivo.uc3m.es Internet	26 words — < 1%
31	ktisis.cut.ac.cy Internet	25 words — < 1%
32	mriuc.bc.uc.edu.ve Internet	23 words — < 1%
33	www.insp.mx Internet	23 words — < 1%
34	documentop.com Internet	22 words — < 1%
35	repositorio.unapiquitos.edu.pe Internet	22 words — < 1%
36	etd.aau.edu.et	

	Internet	21 words — < 1%
37	repositorio.tec.mx Internet	20 words — < 1%
38	ntnuopen.ntnu.no Internet	19 words — < 1%
39	www.slideserve.com Internet	19 words — < 1%
40	patientsafetymovement.org Internet	18 words — < 1%
41	bcdigi.unse.edu.ar:8080 Internet	17 words — < 1%
42	docslide.us Internet	17 words — < 1%
43	dspace.utpl.edu.ec Internet	17 words — < 1%
44	repositorio.espe.edu.ec Internet	17 words — < 1%
45	scielosp.org Internet	16 words — < 1%
46	www.itba.edu.ar Internet	16 words — < 1%
47	revistas.elpoli.edu.co Internet	15 words — < 1%
48	tesis.ucsm.edu.pe Internet	15 words — < 1%
49	uvadoc.uva.es	

Internet

15 words — < 1%

50 www.monografias.com
Internet

15 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 15 WORDS