EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Aplicación de Minería de Datos para el pronóstico de la evolución de la diabetes en México

presentado por

Ing. Daniel Flores Guerrero

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dr. Joaquín Pérez Ortega

Cuernavaca, Morelos, México. Septiembre de 2019.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Mor., 13/agosto/2019
Oficio No. DCC/078/2019
Asunto: Aceptación de documento de tesis

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **Ing. Daniel Flores Guerrero**, con número de control M17CE031, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado **"Aplicación de Minería de Datos para el pronóstico de la evolución de la diabetes en México"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Joaquín Pérez Ortega
Doctor en Ciencias Computacionales
4795984

REVISOR 1

Dr. Moisés González García
Doctor en Ciencias en la Especialidad de Ingeniería Eléctrica
7501724

REVISOR 2

Dra. Alicia Martínez Rebollar
Doctora en Informática
7399055

C.p. M.E. Guadalupe Garrido Rivera - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Mor., 24/septiembre/2019
No. de Oficio: SAC/268/2019
Asunto: Autorización de impresión de Tesis

ING. DANIEL FLORES GUERRERO
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Aplicación de Minería de Datos para el Pronóstico de la Evolución de la Diabetes en México"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

SEP TecNM
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

C.p. M.E. Guadalupe Garrido Rivera.- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Dedicatoria.

A la hermosa familia que Dios me ha dado, a ustedes, que siempre han confiado en mí.

A mis maravillosos padres, quienes han trabajado arduamente para proveernos de un hogar, estudio y comodidad a mis hermanos y a mí; gracias por ser siempre fuente de inspiración y motivación para salir adelante.

A mis hermanos, quienes siempre me han apoyado y ayudado en cada proyecto de vida. A mis grandiosas sobrinas, quienes me motivan para ser un ejemplo de que cuando algo se quiere no hay obstáculo que lo impida. A mi novia, quien al final sin importar la situación, continúa estando a mi lado.

Agradecimientos

A Dios por brindarme, fortaleza, protección y sabiduría todo este tiempo. No habría adquirido este logro sin su presencia en mi vida.

A mis padres, Mónica y José. A mis hermanos Nora y Pepe. A mi cuñada Lili, a mis hermosas sobrinas Sofi y Mia. Los amo infinitamente, muchas gracias. por ser mi inspiración día a día.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por darme la oportunidad de realizar mis estudios de maestría, uno de mis sueños más grandes.

Al Dr. Joaquín Pérez Ortega, quien nunca dudo de mí, quien me apoyo en innumerables veces y gracias a su amplia experiencia dirigió apropiadamente este trabajo de investigación. Simplemente mi más sincero agradecimiento por su paciencia, tiempo, orientación y su valioso apoyo.

Al comité revisor: Dra. Alicia Martínez Rebollar por el tiempo dedicado y contribuir en la revisión de este trabajo; Dr. Moisés González García por su disposición y atención en cada momento que lo necesité. A la Dra. Leticia Sánchez Lima por su apoyo en los diferentes seminarios realizados en el grupo de investigación y por su valiosa ayuda en la revisión de esta tesis.

A mis buenos amigos, Celia, Oscar, Luis, Neri, Luz; quienes fueron compañeros de varias travesías en esta aventura llamada Maestría. A mis compañeros de laboratorio y de línea de investigación; Gudy, Cubo, Lorenzo, Santiago, Tocayo, Pablo, Marianita y a la MC. Itzel quien me apoyo en varias ocasiones con sus conocimientos.

A mis ex compañeros y amigos de Servicios de Salud de Morelos, quienes fueron parte para que tomara esta decisión de continuar creciendo académicamente, Jimmy, Mario, JC, Sandra, Gus, Blanquita, Gil y en especial a mi gran amigo Fernando Espinoza, quien siempre me apoyo cuando lo necesite, Gracias Amigo.

Al buen Pepe Salazar, quien me guio en momentos difíciles y que gracias a sus buenos consejos logre salir adelante.

Por ultimo gracias al gran ser humano, que sirvió como inspiración, apoyo y equilibrio para superarme día a día durante más de una década de mi vida. ¡Gracias! este logro, también es tuyo Soni.

RESUMEN

En la actualidad, la minería de datos ha mostrado ser una actividad de interés en diferentes dominios de la sociedad, particularmente en el área de la salud, al cual va dirigida esta investigación.

Con este estudio se muestra que es factible desarrollar un prototipo de minería de datos, capaz de realizar proyecciones de tasas de mortalidad por causa E11 (*diabetes mellitus no insulinodependiente*). Para validar el funcionamiento del prototipo se utilizó información proveniente de fuentes oficiales como: el Instituto Nacional de Estadística y Geografía, el Sistema Nacional de Información en Salud, el Consejo Nacional de Población y el Centro Mexicano para la Clasificación de Enfermedades. Para procesar, manipular y extraer la información obtenida se realizó un proceso de minería de datos con la metodología *Cross Industry Standard Process for Data Mining* (CRISP-DM), ya que es un estándar para el desarrollo de proyectos de minería de datos en la industria. Asimismo, esta metodología ayudo a obtener la información histórica de la población y de la tasa de mortalidad de diferentes municipios del país de México por causa Tipo E11. Toda la información obtenida respecto a mortalidad por la enfermedad de *diabetes mellitus* se almaceno en un *Data werehouse.*

Para encontrar los grupos de municipios con mayores tasas de mortalidad en México se utilizó el algoritmo de agrupamiento *K-Means*, debido a que sus resultados son fáciles de interpretar y su implementación computacional es relativamente sencilla. En esta investigación se utilizaron los tres grupos que presentaban mayores tasas de mortalidad en México los cuales se identificaron como C08, C24 y C51. Posteriormente se analizaron meticulosamente los respectivos municipios de cada grupo, cuya finalidad fue localizar los dos municipios con mayor tasa de mortalidad y el municipio que menor tasa presentara en un periodo comprendido entre los años 1998 al 2015.

Una vez identificados los municipios de interés se aplicaron técnicas de regresión polinomial, con la finalidad de obtener la predicción a cinco años de las tasas de moralidad de los municipios, tomando como año de partida el 2015. Tal técnica de regresión se implementó en el lenguaje de programación R.

Los resultados más representativos respecto a la zona metropolitana del país, corresponden a la alcaldía de Venustiano Carranza de la Ciudad de México (CDMX), este muestra el mayor aumento en su tasa de mortalidad para el año 2020, mientras que, para la zona de provincia, el municipio de Orizaba del estado de Veracruz es el que tendrá mayor aumento en su tasa de mortalidad para el 2020. El municipio que presenta el mayor decremento en su tasa de mortalidad dentro de la zona metropolitana es la alcaldía de Miguel Hidalgo de la CDMX y en provincia es el municipio de San Pedro Cholula de estado de Puebla.

Es importante la necesidad de desarrollar nuevos programas que sirvan para la promoción y prevención de la enfermedad en mención, y que a su vez permitan reducir las incidencias de los diferentes tipos de diabetes. Por otra parte, es de suma importancia que las autoridades correspondientes en México continúen contribuyendo de una forma más asertiva a los programas que sirven para el control y tratamiento de la diabetes, lo que serviría para reducir los índices de mortalidad.

Con base en los resultados obtenidos desde punto de vista computacional, los resultados son destacables, tomando en cuenta la calidad de la información que se obtiene después de realizar el proceso de minado de datos con la metodología seleccionada y la utilización de la técnica de predicción. Esta investigación muestra que es factible realizar predicciones del comportamiento de las tasas de mortalidad de la enfermedad *diabetes mellitus* Tipo E11.

ABSTRACT

Nowadays, data mining has shown to be an activity of interest in different domains of society, particularly in the area of health, to which this research is directed.

This study shows that it is feasible to develop a data mining prototype, capable of making projections of mortality rates due to E11 cause (non-insulin dependent diabetes mellitus). To validate the operation of the prototype, information from official sources such as Instituto Nacional de Estadística y Geografía, Sistema Nacional de Información en Salud, Consejo Nacional de Población and Centro Mexicano para la Clasificación de Enfermedades. To process, manipulate and extract the information obtained, a data mining process was carried out with the Cross Industry Standard Process for Data Mining methodology (CRISP-DM), which is a standard for the development of data mining projects in the industry, this methodology helped to obtain the historical information of the population and of the mortality rate of different municipalities of the country of Mexico by Type E11 cause. All information was stored in a data warehouse.

To find the groups of municipalities with the highest mortality rates in Mexico, the K-Means clustering algorithm was used, because its results are easy to interpret and its computational implementation is relatively simple.

In this research, the three groups with the highest mortality rates in Mexico were used, which were identified as C08, C24 and C51. Then, municipalities of each group were analyzed, whose purpose was to locate the two municipalities with the highest mortality rate and the municipality with the lowest rate in a period between 1998 and 2015.

Once the municipalities of interest were identified, polynomial regression techniques were applied, in order to obtain the five-year prediction of the morality rates of the municipalities, taking 2015 as the starting year. For this the programming language R was used.

The most representative results regarding the metropolitan area of the country, correspond to the mayor of Venustiano Carranza of Mexico City, this shows the largest increase in its mortality rate for the year 2020, while, for the province area, the municipality of Orizaba in the state of Veracruz is the one with the highest increase in its mortality rate by 2020.

The municipality that presents the greatest decrease in its mortality rate within the metropolitan area is the mayor of Miguel Hidalgo of the CDMX and in the province is the municipality of San Pedro Cholula of the state of Puebla

The need to develop new programs that serve to promote and prevent the disease is important, on the other hand, it is of the utmost importance that the corresponding authorities in Mexico continue to contribute more assertively to the programs that are used for the control and treatment of diabetes, which would serve to reduce mortality rates.

The results are remarkable, taking into account the quality of the information obtained after performing the data mining process with the selected methodology and the use of the prediction technique.

This research shows that it is feasible to make predictions of the behavior of the mortality rates of diabetes mellitus disease Type E11

Contenido

Página

1. Introducción 1
1.1. Contexto de la investigación 2
1.2 Descripción del problema 2
1.3. Objetivo 3
1.4. Justificación 3
1.5. Impacto computacional 4
1.6. Impacto social 4
1.7. Alcances y limitaciones de la investigación 5
1.7.1. Alcances 5
1.7.2. Limitaciones 5
1.8. Estado del arte 6
1.8.1. Antecedentes en el CENIDET 6
1.8.2. Trabajos relacionados 7
1.9. Organización del documento 10
2. Marco conceptual 11
2.1. Metodología CRISP-DM 12
2.2. Minería de datos 15
2.3. Minería de datos predictiva 15
2.4. Base de datos poblacional 16
2.5. Epidemiologia 16
2.6. Patrones epidemiológicos 16
2.7. Almacén de datos 16
2.8. Tasa de mortalidad por diabetes 17
2.9. Regresión 17
2.9.1. Regresión polinomial 17
3. Desarrollo del prototipo de acuerdo a la metodología CRISP-DM 18
3.1. Minado de datos con la metodología CRISP-DM 19
3.1.1. Comprensión del negocio. 19

3.1.2. Comprensión de los datos 19
3.1.3. Recolección de datos iniciales 20
3.1.4. Descripción de los datos 20
3.1.4.1. Cantidad de datos 20
3.1.5. Calidad de Datos 21
3.1.6. Exploración de los datos 22
3.1.7. Verificación de la calidad de los datos 22
3.1.8. Preparación de datos 23
3.2. Modelado 33
3.2.1. Seleccionar técnica de modelado 34
3.2.2. Generar el plan de prueba 34
3.2.2.1. Objetivo del plan de pruebas 34
3.2.2.2. Ambiente de pruebas 34
3.2.3. Construir el modelo 35
3.2.4. Evaluar el modelo 35
3.3. Despliegue 35
3.3.1. Evaluar resultados 37
4. Patrones obtenidos 38
4.2.1. Grupo C08 39
4.2.2. Grupo C24. 42
4.2.3. Grupo C51 45
5. Conclusiones y trabajos futuros 46
5.1. Conclusiones 47
5.2. Trabajos futuros 48
Referencias 49

Lista de figuras

Página

Figura 1. Descripción del problema, visualización del grupo y tendencia de mortalidad..... 3
Figura 2. Modelo CRISP-DM 12
Figura 3. Fase 1 del estándar CRISP-DM (Comprensión del negocio) 12
Figura 4. Fase 2 del estándar CRISP-DM (Compresión de los datos) 13
Figura 5. Fase 3 del estándar CRISP-DM (Preparación de los datos)..... 13
Figura 6. Fase 4 del estándar CRISP-DM (Modelado) 14
Figura 7. Fase 5 del estándar CRISP-DM (Evaluación) 14
Figura 8. Fase 6 del estándar CRISP-DM (Despliegue) 15
Figura 9. Cantidad de datos obtenido y analizados 21
Figura 10. Metodología de preparación de datos orientada a aplicaciones de epidemiología basada en el modelo CRISP-DM 23
Figura 11. Flujo de trabajo en KNime, para datos de defunciones de los años 2000 al 2015..... 26
Figura 12. Flujo de trabajo KNime, para datos de defunciones 27
Figura 13. Flujo de trabajo de concatenación de identificador 10 28
Figura 14. Flujo de trabajo para excluir municipios con una población menor a 100,000 habitantes. 28
Figura 15. Flujo de trabajo para excluir municipios con población menor a 100,000 habitantes. 29
Figura 16. Estructura Data werehouse 31
Figura 17. Representación del modelo multidimensional 32
Figura 18. Diagrama de flujo de la regresión polinomial..... 33
Figura 19. Esquema del prototipo de minería de datos para el pronóstico de diabetes mellitus 37
Figura 20. Tendencia real y proyección de la tasa de mortalidad Grado 2 - Azcapotzalco 40
Figura 21. Tendencia real y proyección de la tasa de mortalidad Grado 3 - Miguel Hidalgo..... 41
Figura 22. Tendencia real y proyección de la tasa de mortalidad Grado 2 - Matamoros..... 42
Figura 23. Tendencia real y proyección de la tasa de mortalidad Grado 2 - Venustiano Carranza .. 43
Figura 24. Tendencia real y proyección de la tasa de mortalidad Grado 3 - Orizaba 44
Figura 25. Tendencia real y proyección de la tasa de mortalidad Grado 2 – San Pedro Cholula 45

Capítulo 1

1. Introducción

En este capítulo se presenta el panorama general de la presente investigación. Asimismo, se exponen los motivos que provocaron esta investigación. De igual forma se presentan los antecedentes de esta investigación, la descripción del problema, el objetivo general, los resultados esperados entre otros.

1.1. Contexto de la investigación

En la actualidad el proceso de Minería de datos ha sido de interés en diferentes dominios: finanzas, educación, comercio y en especial en el sector de la salud, a este último campo va dirigido este trabajo, porque mediante la exploración de grandes volúmenes de datos es posible extraer información previamente desconocida que, posteriormente se convierte en información altamente útil gracias al proceso de minado de datos.

Para el desarrollo de esta investigación los datos utilizados están directamente relacionados con el área epidemiológica en México. Los datos de mortalidad se obtienen de las actas de defunción, que posteriormente son recopilados por el Sistema Nacional de Información en Salud (SINAIS) [1] y difundidos por el Instituto Nacional de Estadística, Geografía (INEGI) [2].

Este estudio es continuación de una serie de investigaciones previas que forman parte de un proyecto de minería de datos desarrollado en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), orientado al área de salud. Su objetivo principal es encontrar patrones de interés epidemiológico aplicando técnicas de minería de datos en bases de datos poblacionales de mortalidad por alguna enfermedad que se presente en el país. De igual forma, se pretende encontrar tendencias en las tasas de mortalidad por la enfermedad seleccionada.

1.2 Descripción del problema

El problema que se aborda en esta investigación, consiste en determinar mediante un prototipo de minería de datos la evolución de la *diabetes mellitus* en México para el año 2020. En los últimos años la enfermedad llamada *diabetes mellitus* se ha convertido en una de las principales causas de mortalidad en México, de acuerdo con [3], en 2014 el 8.5% de los adultos mayores de 18 años tenían diabetes.

En 2012 el alto nivel de glucosa en sangre fue la causa de otras 2.2 millones de muertes y en 2015, la diabetes fue la causa directa de 1.6 millones de muertes. Por otra parte, de acuerdo con [4], en las últimas tres décadas, la prevalencia de diabetes Tipo 2 ha aumentado dramáticamente en países de todos los niveles de ingresos, por lo que se prevé que en los próximos años aumente la tasa de mortalidad por este Tipo de diabetes.

En la Figura 1, se muestra de manera figurada como pueden ir creciendo las regiones con altas tasas de mortalidad en México. Por ejemplo, para el municipio de Orizaba, se pronostica un crecimiento del 30% en su tasa de mortalidad para el 2020 por causa Tipo E11.

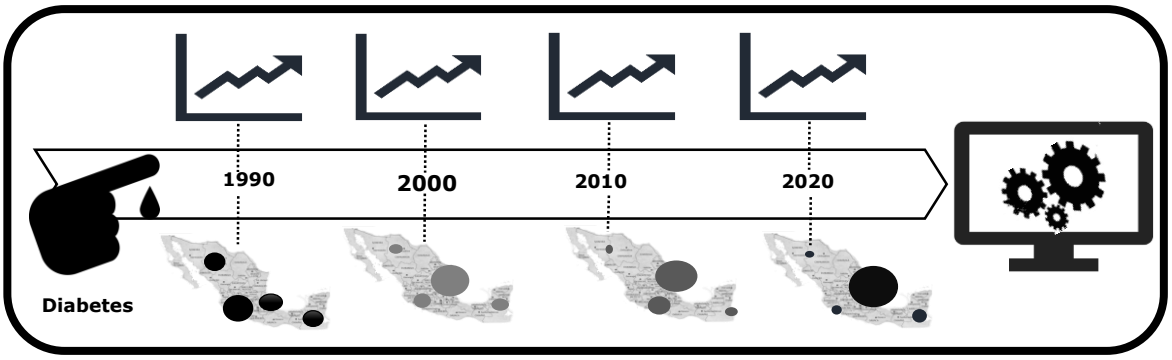

Figura 1. Descripción del problema, visualización del grupo y tendencia de mortalidad.

1.3. Objetivo

Realizar un prototipo de minería de datos, que permita realizar predicciones de las tasas de incidencias por mortalidad de la enfermedad de *diabetes mellitus* en México y que muestre la evolución de la diabetes en regiones de municipios en México, utilizando bases de datos de mortalidad.

1.4. Justificación

El proceso de minería de datos no es problema trivial, al contrario, en computación su aplicación es bastante utilizada en una gran variedad de dominios. Al aplicar minería de datos a varios esquemas de bases de datos, computacionalmente implica: un gran trabajo de procesamiento, de integración de tales esquemas y de aplicación de técnicas de minado, entre otros. El objetivo de la minería de datos es descubrir patrones e información altamente rentable que sirva para toma de decisiones para determinado dominio. En particular, en el CENIDET se ha aplicado en el área de la salud, con resultados bastante alentadores.

Esta investigación computacional permite desarrollar experiencias relacionadas con la integración de información de bases de datos con esquemas diferentes, ya que estas no cuentan con el mismo nombre en sus atributos ni tienen los mismos datos. Al trabajar con bases de datos, obtenidas de fuentes oficiales y a estas aplicarles procesos de minería de datos, permite aportar soluciones a problemas reales y proporcionar información útil para la toma de decisiones por parte de las autoridades de salud. Su finalidad es ayudar en la identificación de aquellas zonas y municipios que requieran atención en cuanto a la promoción y prevención de la enfermedad de *diabetes mellitus*.

Por otra parte, el problema que se aborda en esta investigación, dará pauta para que en futuros trabajos consideren otro tipo de técnicas predictivas para la identificación de zonas y municipios con altas tasas de mortalidad. De igual manera se permitirá la creación de metodologías para trabajos posteriores que aborden problemas reales con grandes cantidades de datos.

1.5. Impacto computacional

Desde el punto de vista computacional, las actividades que constan de limpiar, manejar, manipular y procesar grandes cantidades de datos reportadas por fuentes oficiales relacionados con defunciones por *diabetes mellitus* en México, no es una tarea trivial, debido a que las bases de datos sobre las que se hace minería no están diseñadas para realizar minería de datos, por lo cual es necesario su integración y selección de los datos relevantes para el proceso de minería.

La manipulación de tal información presenta un gran reto, porque en algunos casos no existen datos, o los tipos de datos de las variables no coinciden entre un esquema de base de datos y otro, en varios casos los formatos de la información difieren de un año a otro, en la sección *3.1.8.1 Limpieza de datos* se menciona de forma general lo que se realizó para la preparación de los datos.

Por otro lado, se presenta la complejidad de implementar la técnica predictiva de regresión polinomial, la cual es capaz de trabajar con datos numéricos. También es importante mencionar que es complejo realizar una interpretación del análisis de regresión. Por otra parte, el desarrollar el prototipo en un lenguaje de programación R que está emergiendo en la comunidad científica implico un gran reto. Es cierto que la documentación oficial es rica en cuanto a información para su uso del lenguaje, no es lo mismo para programar la técnica de regresión seleccionada ya que esta no es fácil de implementar. Para ello se tuvieron que utilizar librerías estadísticas propias del lenguaje de R.

1.6. Impacto social

La *diabetes mellitus* es una enfermedad frecuente en el mundo. En México, la mortalidad por dicho padecimiento es de aproximadamente 75 mil muertes por año. Por tal motivo, se posiciona en los primeros lugares de mortalidad a nivel mundial [4].

En varias investigaciones realizadas en el CENIDET se han identificado varias regiones con altas tasas de mortalidad, de igual manera se ha logrado identificar que las tendencias de las tasas irán creciendo anualmente.

De acuerdo con [5] se estima que cada año aproximadamente 400,000 personas contraen tal enfermedad a nivel mundial. Esta enfermedad se ha convertido en un problema a nivel mundial.

Expertos de la Organización Paranamericana de la Salud y de la Organización Mundial de la Salud, señalan que crecerá sustancialmente en las próximas dos décadas. Se estima que solo en América existirán aproximadamente 110 millones de personas con diabetes para el año 2040 [6].

Por los motivos señalados previamente resulta de gran interés para la comunidad científica aplicar procesos de minería de datos, utilizando bases de datos poblacionales reales de defunciones que provengan de fuentes oficiales. Con el objetivo de identificar patrones en el comportamiento epidemiológico en México. De igual forma, es importante la implementación de técnicas predictivas sobre bases de datos de mortalidad, porque el resultado permitirá a las autoridades correspondientes establecer estrategias para la prevención y promoción de la salud. De esta forma, se estaría contribuyendo a tomar medidas que permitan reducir la mortalidad y las defunciones por diabetes.

1.7. Alcances y limitaciones de la investigación

1.7.1. Alcances

- Esta investigación está orientada al uso de bases de datos poblacionales oficiales de México, que se puedan conseguir de forma gratuita. Se utilizará la metodología CRISP-DM [13], para su aplicación al dominio epidemiológico.
- Se utilizará como base un prototipo de sistema minería de datos, al cual se le realizarán las extensiones o modificaciones necesarias.
- La predicción de la evolución de las tasas de mortalidad por *diabetes mellitus* será para cinco años. El prototipo desarrollado será capaz de trabajar con la granularidad de municipios con altas tasas de mortalidad.

1.7.2. Limitaciones

- Se utilizará un algoritmo de agrupamiento tipo K-Means mejorado en el CENIDET
- Se utilizará software y hardware disponible en el CENIDET.

1.8. Estado del arte

1.8.1. Antecedentes en el CENIDET

A continuación, se describen de manera general las investigaciones desarrolladas en CENIDET y las que preceden a la presente tesis.

En la investigación llamada "Aplicación de Minería de datos en el área de salud pública" [7]. El problema que aborda dicha investigación es identificar grupos con una alta tasa de mortalidad por *diabetes mellitus* en México y Estados Unidos. Para detectar estos grupos se usaron bases de datos de fuentes oficiales de los países mencionados.

El autor define que los problemas más grandes que enfrentó en su investigación fueron los siguientes: búsqueda de la información, obtención de las bases de datos de México y Estados Unidos, la preparación de los datos, la creación del almacén de datos, la selección de la técnica de Minería de datos para encontrar patrones y la visualización de estos patrones en un mapa de los países en mención.

Para realizar las tareas anteriormente mencionadas, se utilizó la metodología CRISP-DM. El autor menciona que esta es una de las guías más utilizadas para el desarrollo de proyectos de minería de datos.

El producto de la investigación fue el desarrollo de un prototipo de una aplicación de Minería de datos que posibilita encontrar patrones de interés representados como grupos de población con altas tasas de mortalidad por diabetes en poblaciones de más de cien mil habitantes. El autor considera que es factible encontrar patrones de mortalidad en bases de datos mediante técnicas de Minería de datos, también menciona que le fue posible pintar cada grupo en una imagen del mapa de México y Estados Unidos.

En la tesis de maestría denominada "Metodología de Preparación de datos orientada a aplicaciones de epidemiología basada en el modelo CRISP-DM" [8]. Se propuso una metodología para la fase de preparación de datos, con un nivel de detalle mayor al propuesto en la metodología CRIPS-DM. La metodología propuesta fue validada mediante una aplicación en el área de epidemiologia, con resultados satisfactorios.

En la tesis de maestría "Desarrollo de una aplicación de ciencia de datos" [9]. Se muestra que es factible la asimilación de conceptos de ciencia de datos y aplicarlos a una infraestructura de conocimientos que apoyen el desarrollo de aplicaciones de ciencia de datos. El autor estudio conceptos generales de ciencia de datos, los cuales se validaron mediante un caso práctico, donde se empleó el lenguaje de programación R y la metodología propuesta por IBM llamada "Foundational Methodology for Data Science". Para validar este trabajo se utilizó una aplicación del sector salud.

1.8.2. Trabajos relacionados

En la literatura especializada acerca de la incidencia de mortalidad por alguna enfermedad se observó en los siguientes documentos una relativa similitud con el presente proyecto de investigación. En los cuales se realiza un análisis comparativo de algunos métodos aplicados para el desarrollo de las investigaciones. En la mayoría de los artículos se encontró la utilización de algoritmos de minería de datos, pero estos no cumplían con ciertas condiciones que el presente estudio si pretende cumplir. Ejemplo de esto es la predicción de casos de mortalidad en bases de datos de mortalidad por diabetes.

Los criterios de búsqueda incluyeron artículos de aplicación de técnicas de Minería de datos sobre bases de datos epidemiológicas oficiales. Además, se investigaron trabajos que utilizan algoritmos predictivos para el sector salud. A continuación, se describen brevemente las investigaciones que se consideran más relacionadas con la presente investigación.

En el artículo "An Epidemiological Data Mining Application Based on Census Databases" [10], se desarrolló un proceso de minería de datos especializado, que integra datos de mortalidad tomados de censos oficiales del 2000 y 2010. El objetivo de esta investigación es la generación de patrones de interés basados en el agrupamiento de distritos con altas tasas de mortalidad por diferentes causas de muerte.

Las contribuciones de esta investigación son la implementación de un subsistema de preparación de datos y la integración de un almacén de datos que contiene registros de defunciones, ocurridas en los años antes mencionados, y de las 2049 causas diferentes de muerte registradas en México. Para validar los resultados, se analizaron cuatro causas de muerte relacionadas con cáncer C16 (estómago), C34 (pulmón), *diabetes mellitus* E11 (no insulinodependiente) y E14 (no especificado). Como resultado se logró mostrar patrones de interés. Además, se identificó un aumento en las tasas de mortalidad del año 2010 en comparación con el año 2000.

En el artículo "A Data Mining System for the Generation of Geographical C16 Cancer Patterns" [11], se describen los resultados de un sistema de minería de datos, desarrollado para identificar patrones que sean de interés, utilizando bases de datos de mortalidad por la enfermedad de cáncer. Este trabajo presenta de manera muy particular sus resultados obtenidos, así como la arquitectura innovadora del sistema que implementa, ya que este integra un mapa cartográfico visual, un almacén de datos y un sistema de minería de datos.

El algoritmo K-Means fue utilizado para generar patrones, lo que permitió que este trabajo expresara patrones como grupos de distritos con afinidad en sus parámetros de localización y tasa de mortalidad. Las bases de datos utilizadas fueron obtenidas de instituciones oficiales mexicanas. Como resultado del sistema, se generó un conjunto de patrones de agrupación, que definen la distribución de la mortalidad por cáncer de estómago en los municipios de México. Se identificó una alta tasa de mortalidad en los distritos del sureste de México.

El autor consideró que los patrones generados por el sistema, pueden ser útiles como herramienta para ayudar a los estudios sobre el cáncer y para la toma de decisiones sobre la asignación de recursos que permitan organizar servicios para la prevención y el tratamiento de ese padecimiento.

En el artículo "A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases" [12], se muestra una nueva metodología de preparación de datos orientada al área epidemiológica en la que se han identificado dos conjuntos de tareas. Una es la preparación de datos generales y la segunda es la preparación de datos específicos. Para ambos conjuntos, se utiliza la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) [13]. La principal aportación de esta investigación son catorce tareas especializadas relacionadas a este dominio.

Para la validación de la metodología propuesta, se desarrolló un sistema de minería de datos, el cual se aplicó a bases de datos reales de mortalidad. Los resultados obtenidos fueron alentadores porque se observó que el uso de la metodología redujo algunas de las tareas que consumen mucho tiempo en el tratamiento de la información. El autor destaca que el sistema de minería de datos implementado, mostró hallazgos de patrones desconocidos, estos resultan ser potencialmente útiles para los servicios de salud pública en México. Porque les servirá en los procesos de toma de decisiones al momento de realizar campañas de atención primaria, orientadas a distritos de México con altas tasas de mortalidad. Los servicios de salud podrían tener una nueva perspectiva de cómo estas enfermedades afectan a algunas regiones en particular.

En el artículo "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes" [15] se toma como problema de investigación los crecientes casos de *diabetes mellitus* en el mundo. Los responsables del sector salud han identificado que esta enfermedad se encuentra en focos rojos, porque día a día las incidencias por tal padecimiento van aumentando a un ritmo acelerado. Por tal motivo, la prevención y predicción de esta enfermedad va ganando interés en la comunidad sanitaria. Por esta razón, en el mundo se han propuesto sistemas que emplean técnicas de minería de datos para el apoyo en la toma de decisiones clínicas y para la predicción de la diabetes. En su mayoría estos sistemas son convencionales y únicamente se basan en clasificar o combinar datos.

Recientemente, en diversas investigaciones se esfuerzan por mejorar la precisión de estos sistemas usando clasificadores de conjuntos. Por esta razón, el autor decide utilizar técnicas de conjunto de *adaboost* y *bagging* junto al uso de árboles de decisión J48. Este último lo utiliza como un aprendiz básico. Esta técnica la emplea para clasificar a los pacientes con *diabetes mellitus*, utilizando factores de riesgo por tal padecimiento. Esta clasificación se realiza, a través de tres grupos diferentes de adultos ordinarios, estos son de la Red Canadiense de Vigilancia Centinela de Atención Primaria (CPCSSN).

El resultado experimental muestra que el rendimiento general del método del conjunto de *adaboost* es mejor que el *bagging*, así como el árbol de decisión *J48*. El árbol de decisiones es una de las técnicas de clasificación y predicción más potentes y ampliamente utilizadas. Mediante el uso de *adaboost* y el árbol de decisiones *J48*. Este autor construyó modelos razonables, con un mayor rendimiento para clasificar a los pacientes diabéticos, en tres grupos de edad en la población canadiense. El conjunto de datos utilizado en este estudio se obtuvo de la base de datos de CPCSSN.

En "Intelligible Support Vector Machines for Diagnosis" [14], se desarrolló una herramienta para diagnosticar y pronosticar la diabetes. El autor predice estos diagnósticos utilizando máquinas de vectores de soporte (SVM) y con datos de diabetes de la vida real. Como resultado se logró generar un conjunto de reglas comprensibles para la realización de predicciones.

De igual manera se menciona que la herramienta desarrollada tiene una precisión de predicción y sensibilidad superior a un 90 % en cada rubro que contempló. Además, las reglas extraídas son medicamente validadas y coinciden con resultados de estudios médicos. Esta herramienta está destinada a funcionar como una segunda opinión para el diagnóstico de diabetes en personas con alto riesgo. También se pretende que con el uso de estas reglas se pueda ayudar en la detección de sujetos no diagnosticados. El autor piensa que de seguir actualizando su aplicación disminuiran los costos de la atención medica de esta enfermedad ya que se planea realizar predicciones tempranas para el diagnóstico de tal padecimiento.

En el trabajo "Diabetes Classification using k means" [17], se presenta una encuesta sobre las técnicas que se proponen en el campo de la medicina para comprender qué grupos de edad de personas está siendo afectado por la diabetes. Para este trabajo se empleó un procedimiento de minería de datos, el uso del algoritmo K-Means y una agrupación jerárquica del vecino más cercano, aplicado a bases de datos de diabetes, con la finalidad de poder extraer información importante y convertirla en datos comprensibles para su uso posterior.

La precisión, sensibilidad y especificidad son las diferentes métricas que se evaluaron en esta encuesta. El objetivo de este trabajo consiste en encontrar los efectos de la diabetes en grupos de edad, que posteriormente se evalúa su tasa de supervivencia de una manera eficiente, también en este trabajo se buscó predecir la probabilidad de cómo se ven afectadas las personas de diferentes edades por la enfermedad de la diabetes en función a sus actividades, estilo de vida. De igual forma se intentó averiguar los factores responsables del por qué la persona es diabética. Como resultado de este trabajo se resalta la importancia de implementar técnicas de minería de datos en el campo de la medicina para comprender que grupo de personas se está afectando por la diabetes, así como también la visualización del conocimiento descubierto.

El trabajo “Performance Analysis of Classifier Models to Predict *diabetes mellitus*” [19], ofrece una comparación del rendimiento de los algoritmos que se ocupan para realizar predicciones de diabetes, utilizando técnicas de minería de datos. Esta investigación se derivada de la problemática que implica la diabetes en varios países. El autor menciona que es de suma importancia trabajar para prevenir esta enfermedad en una etapa temprana mediante la predicción de los síntomas. Comparo los clasificadores de aprendizaje automático de los árboles de decisión *J48*, *K-Nearest Neighbors* (KNN), *Random Forest* y *Support Vector Machines* (SVM), para clasificar a los pacientes con *diabetes mellitus*.

Estos enfoques se probaron con muestras de datos descargadas del repositorio de datos de la Universidad de California (UCI). Los algoritmos se han medido en dos conjuntos de casos, el primer conjunto de información contiene datos ruidosos fueron utilizados antes del pre procesamiento. El segundo conjunto contiene datos no ruidosos, mismos que se ocuparon después del pre procesamiento. Se compararon en términos de precisión, sensibilidad y especificidad. Los resultados de la comparación de los cuatro modelos de predicción de *diabetes mellitus* concluyen que el clasificador del árbol de decisiones *J48* logra una precisión mayor al 73.82% que otros tres clasificadores. Este resultado se obtuvo antes de procesar el conjunto de datos. En el otro caso, después de pre procesar el conjunto de datos se obtuvo precisión en comparación con el primer estudio. En el caso de *KNN* y *Random Forest* son mejores que los otros clasificadores, porque proporcionan una precisión cercana al 100%. Con este trabajo se llegó a saber que después de eliminar los datos ruidosos del conjunto de datos, se proporcionará un buen resultado. Este estudio se podrá utilizar para seleccionar el mejor clasificador para predecir la diabetes.

1.9. Organización del documento

El presente documento se encuentra organizado por capítulos de la siguiente manera:

El Capítulo 2 expone el marco teórico, que sirve de base a la presente investigación. En este se desarrollan los conceptos básicos relacionados a minería de datos y algunos otros conceptos relacionados con el dominio de aplicación.

El Capítulo 3 se explica la metodología CRIPS-DM relacionada al dominio epidemiológico. Se describe el proceso de creación del almacén de datos y el esquema general del sistema implementado.

En el Capítulo 4 se presentan los resultados de las pruebas realizadas al almacén de datos y al prototipo con los datos de mortalidad por diabetes entre los años 1990 y 2015 en México.

En el Capítulo 5 se exponen las conclusiones derivadas del desarrollo de la investigación, así como las aportaciones. Además, se proponen temas para trabajos futuros, que sirvan para dar continuidad al tema de investigación.

Capítulo 2

2. Marco conceptual

En este capítulo se explican los distintos conceptos que fueron utilizados en la presente de investigación. Su finalidad es dar a entender de mejor forma el desarrollo de la presente investigación. Se amplían conceptos relacionados con Minería de datos, epidemiologia, algoritmo predictivo, el almacén de datos y el uso de la metodología empleada para el tratamiento de la información.

2.1. Metodología CRISP-DM

Para desarrollar este proyecto de mineria de datos se ocupó la metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM) [13] por sus siglas en ingles. Esta metodología describe las fases normales de un proyecto, las tareas necesarias en cada fase y ofrece una explicación de las relaciones entre las tareas. La Figura 2 muestra el proceso CRISP-DM.

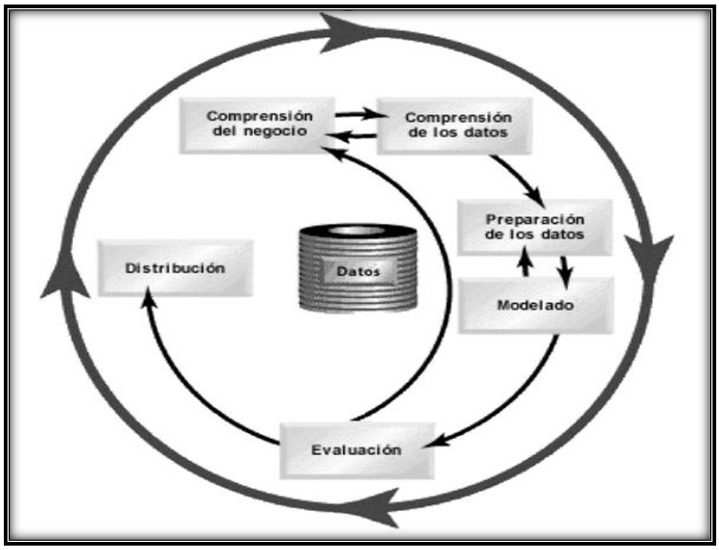

Figura 2. Modelo CRISP-DM

El estándar CRISP-DM consiste en una estructura de seis fases que indican las dependencias más importantes y frecuentes entre fases, las cuales se describen a continuación.

1.- Comprensión del negocio. Esta fase se enfoca en comprender los objetivos del proyecto y los requerimientos necesarios desde la perspectiva del negocio, para después convertir este conocimiento en un problema de minería de datos y de esta manera diseñar un plan para lograr los objetivos del proyecto (Figura 3).

Figura 3. Fase 1 del estándar CRISP-DM (Comprensión del negocio)

2.- Comprensión de los datos. Esta fase comienza con una colección de datos y continúa con actividades que permiten familiarizarse con estos. Además, se identifican problemas en la calidad de los datos. También se detectan subconjuntos para formar hipótesis sobre la información oculta en los datos (Figura 4).

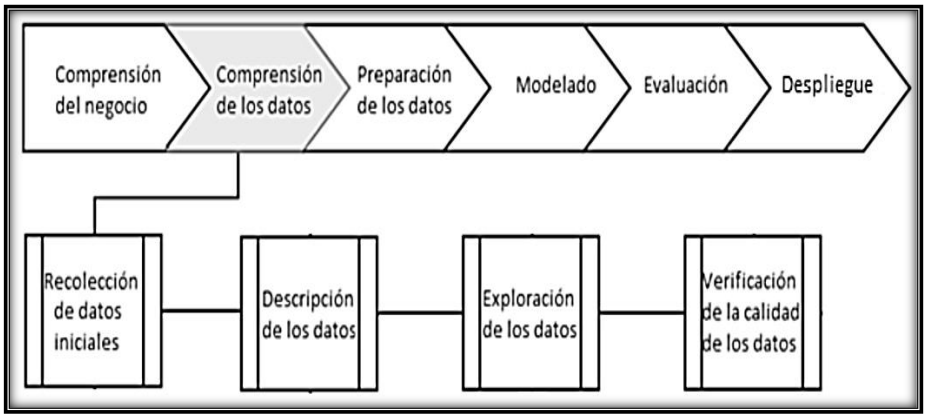

Figura 4. Fase 2 del estándar CRISP-DM (Compresión de los datos)

3.- Preparación de los datos. En esta fase se cubren las actividades necesarias para construir el último conjunto de datos a partir de los datos de inicio. Es probable que las tareas para la preparación de los datos se realicen múltiples veces y no en algún orden prescrito. Las tareas incluyen selección de tablas, columnas, registros y atributos, así como la transformación y limpieza de los datos mediante las herramientas de modelado (Figura 5).

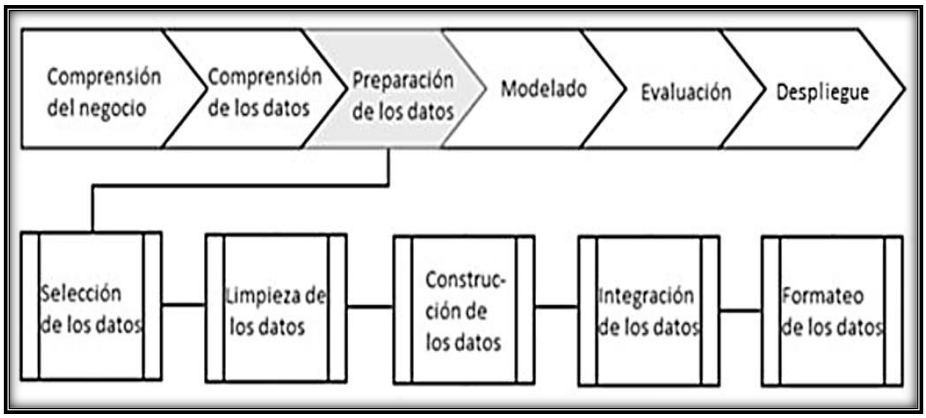

Figura 5. Fase 3 del estándar CRISP-DM (Preparación de los datos)

4.- Modelado. Esta fase se encarga de seleccionar y aplicar diversas técnicas de modelado. Normalmente existen diferentes técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos en la forma de los datos, por tanto, algunas veces es necesario regresar a la preparación de los datos (Figura 6).

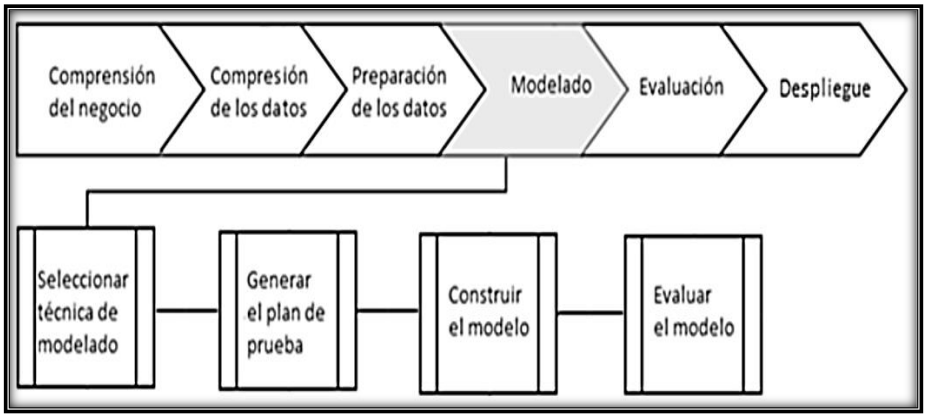

Figura 6. Fase 4 del estándar CRISP-DM (Modelado)

5.- Evaluación. Antes de proceder al despliegue final del modelo, es importante evaluar a fondo y revisar los pasos ejecutados para crearlo, con el fin de que el modelo cumple con los objetivos del negocio. Un objetivo clave es determinar si hay un aspecto importante del negocio que no ha sido considerado con suficiencia. Al final de esta fase, deberá tomarse una decisión sobre el uso de los resultados (Figura 7).

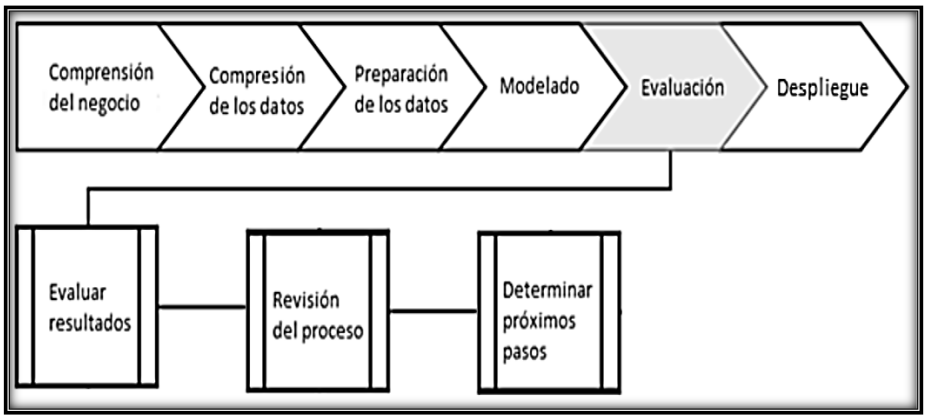

Figura 7. Fase 5 del estándar CRISP-DM (Evaluación)

6.- Despliegue. Se muestra el modelo y la forma en que el cliente puede utilizar el conocimiento generado. Es importante mencionar que el propósito del modelo es incrementar el conocimiento de los datos, por lo que el modelo tendrá que ser organizado y presentado de una forma que el cliente pueda utilizar el conocimiento adquirido (Figura 8).

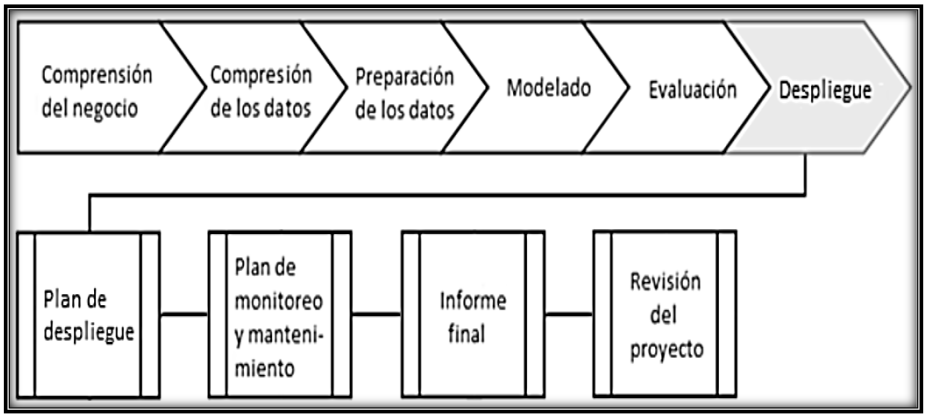

Figura 8. Fase 6 del estándar CRISP-DM (Despliegue)

7.- Documentación de la investigación. Esta actividad no está contemplada dentro del estándar CRISP-DM. Se incluye para establecer la importancia de documentar la investigación realizada.

2.2. Minería de datos

La minería de datos [21] es un proceso iterativo dentro del cual el progreso se define mediante el descubrimiento, utilizando métodos automáticos o manuales. Además, consiste en la búsqueda de información nueva, valiosa y no trivial que se encuentra dentro de grandes volúmenes de datos.

De acuerdo a [22], la minería de datos permite el análisis de conjuntos de datos observacionales usualmente grandes, dentro de los que se pretende encontrar relaciones insospechadas y sirve para resumir datos de forma novedosa, comprensible y útil para el propietario de los datos.

2.3. Minería de datos predictiva

La minería de datos es el proceso de extracción de información de grandes conjuntos de datos para hacer predicciones y estimaciones sobre resultados futuros [23].

2.4. Base de datos poblacional

Las bases de datos poblacionales contienen información minuciosa sobre algunas características de la población. Los datos se examinan y organizan por temas para obtener las estadísticas sociodemográficas de México, las cuales abarcan una gran cantidad de información, entre la que se encuentra: volumen de la población y distribución geográfica de la población, población con discapacidad, tipo, número de discapacidad, enfermedades en una población y mortalidad poblacional [2].

2.5. Epidemiologia

La epidemiología es el estudio de la distribución y los determinantes de estados o eventos (en particular de enfermedades) relacionados con la salud y la aplicación de esos estudios al control de enfermedades y otros problemas de salud. Hay diversos métodos para llevar a cabo investigaciones epidemiológicas. La vigilancia y los estudios descriptivos se pueden utilizar para analizar la distribución y los estudios analíticos permiten analizar los factores determinantes [24].

2.6. Patrones epidemiológicos

Los patrones epidemiológicos se refieren a la forma en que se distribuyen los eventos relacionados con la salud de acuerdo con tiempo, lugar y características de la población. Son utilizados dentro de la epidemiologia descriptiva para ofrecer una descripción detallada de los fenómenos de salud y enfermedad, basados en la observación cuidadosa y en el registro objetivo de hechos [25].

2.7. Almacén de datos

Un almacén de datos es un conjunto de datos históricos, internos o externos, y descriptivos de un contexto o área de estudio. Que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas [26].

2.8. Tasa de mortalidad por diabetes

La tasa de mortalidad por diabetes se define como el número de muertes por causa E11, E12, E14 entre otras, para lugares con una población mayor o igual a 100,000 habitantes. Esta cifra se obtiene de dividir el número de muertes por diabetes en un año entre la población de estudio [27].

2.9. Regresión

El análisis de regresión es una forma de técnica de modelado predictivo que investiga la relación entre una variable dependiente e independiente. La importancia de la relación entre las variables, es para encontrar la mejor línea de ajuste o la ecuación de regresión que se pueda utilizar para realizar predicciones [28].

2.9.1. Regresión polinomial

La regresión polinomial se aplica cuando existe una correlación de datos, pero estos, no parecen tener una relación lineal. Es ahí cuando es factible la aplicación de una regresión polinomial, con la finalidad de que se realice un mejor ajuste de curva o de la ecuación polinómica [28]. La ecuación de un polinomio de grado n es:

$$\gamma = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_n x^n = \sum_{i=0}^{n} \alpha_n$$

$\gamma =$ Valor de la pendiente

$\alpha_0 \; y \; \alpha_1 =$ Ordenada al origen y la pendiente de la línea recta

$\alpha_0 + \alpha_1 x =$ Valor pronosticado de la variable dependiente

Capítulo 3

3. Desarrollo del prototipo de acuerdo a la metodología CRISP-DM

En este capítulo se describe el proceso desarrollado en cada fase de la aplicación de la metodología CRIS-DM. Adicionalmente también se especifican las tareas necesarias para el proceso de minado de los datos relacionados para esta investigación.

3.1. Minado de datos con la metodología CRISP-DM

Para realizar el proceso de minado de datos de la información obtenida del INEGI [2], SINAIS [1], CONAPO [19] y CEMECE [20], de los años 1990 al 2015. Se utilizó la metodología CRISP-DM por su amplio uso en la industria para proyectos de minería de datos. A continuación, se explica cómo se realizó cada fase de la metodología utilizada.

3.1.1. Comprensión del negocio.

Las actividades realizadas en esta fase, consistieron en explorar las expectativas del proyecto basado en estudios previamente realizados en el CENIDET. Se produjo el plan del proyecto tomando como referencia los resultados esperados y el tiempo máximo estipulado para el desarrollo.

En esta parte principalmente se establecieron los objetivos de minería de datos, los cuales debían realizar el agrupamiento de los datos conforme a los municipios con altas tasas de mortalidad relacionada por diabetes Tipo E11, entre los años 1998 al 2015. También se estableció de donde se debía obtener la información y cuáles eran los recursos disponibles.

Además, se estudiaron investigaciones previas a esta, con la finalidad entender como procesar, manipular, exportar la información que se va a minar. En el estudio realizado por [29], se desarrolló un prototipo de un sistema de minería de datos. En él se destaca: que es factible el desarrollo de un prototipo de minería de datos con el objetivo de encontrar regiones de territorio mexicano con alta tasa de mortalidad por diabetes. Principalmente establece que es posible realizar un proyecto de minería de datos partiendo de bases de datos poblaciones para determinar que existe una relación en la distribución de las tasas de mortalidad con respecto al lugar en el que ocurrió el deceso. Asimismo, se estudió la investigación realizada por [9], el cual destaca que es posible realizar predicciones de tasas de mortalidad utilizando una metodología de ciencia de datos llamada Metodología Fundacional para Ciencia de Datos (FMDS, por sus siglas en inglés). Ambas investigaciones sirvieron para comprender el proceso para tratar, manipular y explorar la información con la que se trabajara en esta investigación.

3.1.2. Comprensión de los datos

Esta fase involucró el estudio cercano de las bases de datos disponibles con la finalidad de evitar problemas inesperados durante la fase de preparación de datos. Las actividades llevadas a cabo abarcan desde acceder a los datos y explorarlos, de tal manera que estos queden organizados y así poder determinar la cantidad de los datos.

3.1.3. Recolección de datos iniciales

En esta etapa, es importante destacar que los datos se buscaron en las bases de datos de diferentes fuentes oficiales previamente mencionadas, como son:

Tabla 1. Fuentes oficiales

Institución	Fuentes oficiales de información
INEGI	De [2], se obtuvieron datos de localización geográfica representada en coordenadas en grados decimales, provenientes del Marco de Geoestadica Nacional, así como la población existente en los años 1990, 1995, 2000, 2010 y 2015.
CONAPO	De [19] se obtuvo la proyección de población 2010 al 2030
SINAIS	De [1] se obtuvieron bases de datos de mortalidad de los periodos 1990 al 2015
CEMECE	De [20] se obtuvo la clasificación de Enfermedades del año 2013

Cabe mencionar que esta investigación, basada en el proceso de minería de datos tiene un alto grado de complejidad computacional, porque toda la información con que se cuenta, proviene de diferentes fuentes oficiales, los esquemas de las bases de datos difieren entre una fuente y otra.

3.1.4. Descripción de los datos

En esta actividad se describen las características de los diferentes tipos de datos, como por ejemplo, si es numérico o carácter, su longitud, entre otros.

3.1.4.1. Cantidad de datos

A continuación, en la Figura 9 se describe la cantidad de bases de datos obtenidas de cada fuente de información con los que se desarrolló la presente investigación.

En total se trabajó con 65 bases de datos; 1 del CEMICE, 32 de la CONAPO, 6 del INEGI y 26 del SINAIS. De cada fuente de información variaban los formatos de las bases de datos. Por ejemplo, la información obtenida del SINAIS venía en un formato DBF, para lo cual se tuvo que utilizar el software *DBF Manager* para poder leer cada registro de las 26 bases de datos que se obtuvieron de esta fuente.

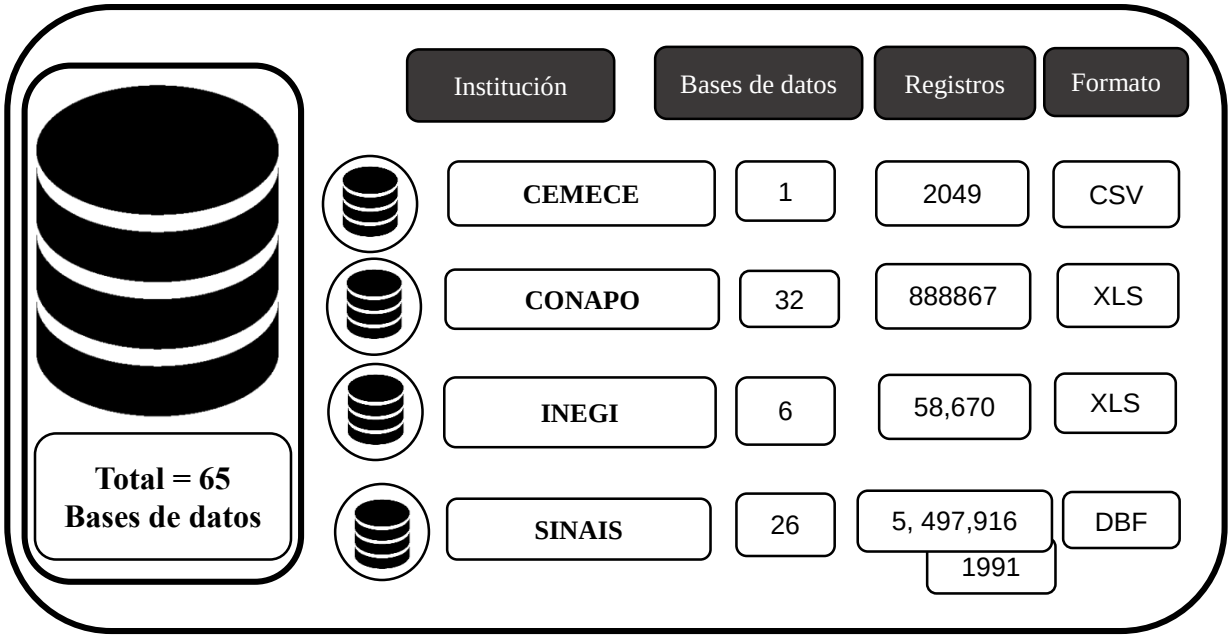

Figura 9. Cantidad de datos obtenido y analizados

3.1.5. Calidad de Datos

Es de suma importancia determinar la calidad de los datos disponibles, porque de esto depende un buen resultado al aplicar la técnica de minería de datos seleccionada. En este apartado se lleva a cabo una descripción detallada de los datos, en la que se incluyen los atributos de las tablas, tipos de valores y esquema de codificación.

Utilizar datos de México facilitó que la información en su mayoría fuera numérica, esto permitía que los datos minados de diferentes años fueran compatibles con el *Data werehouse* existente.

Otro factor que ayudó para lograr la compatibilidad con el almacén de datos, fue que la información de los años, contaba con las mismas variables, solo que, con diferente etiqueta. De tal modo, la actividad de analizar detenidamente los catálogos de las fuentes de información fue de suma importancia.

3.1.6. Exploración de los datos

Para la preparación de los datos fue necesario su exploración, ya que desde aquí se puede observar la calidad de la información recolectada. Esta actividad permitió determinar las tareas que deberían realizarse en la fase de preparación de datos, tal como es la selección de atributos.

Una herramienta altamente eficaz en proyectos de minería de datos es *Knime* [30], la cual es una plataforma modular de exploración y procesamiento de información que permite crear flujos de datos de forma visual e interactiva. Esta herramienta sirvió como apoyo para la exploración de los datos de mortalidad de los años 1990 al 2015.

Esta fase es de suma importancia porque es necesario explorar minuciosamente la información. Se deben buscar errores en los datos, como son los campos vacíos, en esta investigación las tuplas con atributos de datos vacíos fueron excluidas del proceso de minería.

3.1.7. Verificación de la calidad de los datos

Se reconoce que los datos obtenidos no son perfectos. De hecho, en su mayoría contienen errores de codificación, valores perdidos u otro tipo de incoherencias que hacen que los análisis resulten difíciles en algunas ocasiones. Antes de proceder al modelado de los datos disponibles es importante realizar un buen análisis de calidad. Esta es una forma de evitar posibles problemas al aplicar la técnica de minería de datos.

Tomando en cuenta la metodología desarrollada por [12], en la cual se define con un mayor nivel de detalle al que se propone en CRISP-DM para su aplicación al dominio epidemiológico (Figura 10); se definen cinco sub-fases para la fase de preparación de datos. Dichas fases son las siguientes: limpieza, selección, formateo y construcción e integración de datos. Posteriormente son clasificados en dos niveles: preparación de datos general y preparación de datos especifica.

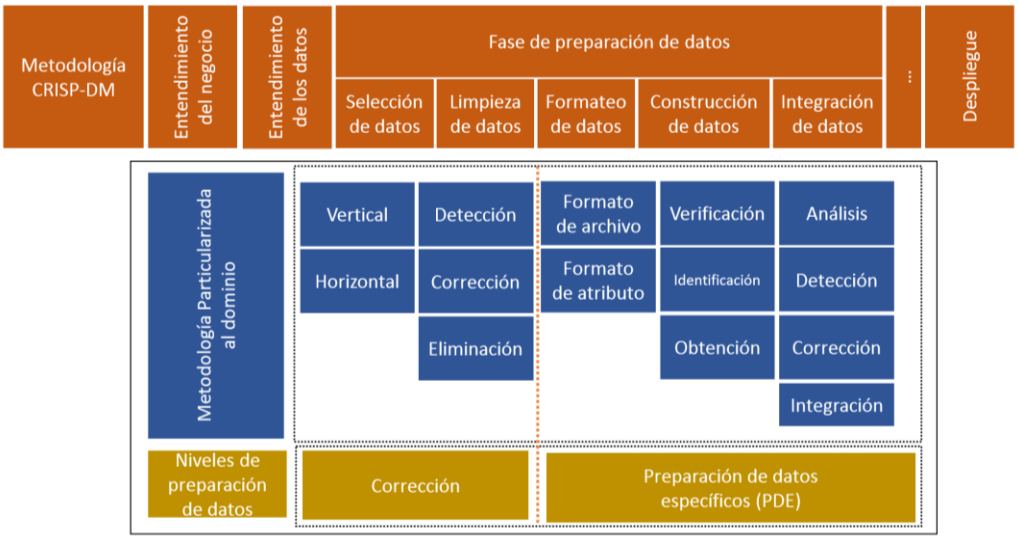

Figura 10. Metodología de preparación de datos orientada a aplicaciones de epidemiología basada en el modelo CRISP-DM

3.1.8. Preparación de datos

Uno de los aspectos más importantes y que más tiempo requiere en un proyecto de minería de datos es la preparación de los datos. Se estima que esta fase suele llevarse más del 70% del tiempo. Sin embargo, dedicar el esfuerzo adecuado a las primeras fases de comprensión del negocio y comprensión de los datos, apoya en gran parte a esta tarea. En esta preparación se pretende seleccionar los datos desde diferentes fuentes de información. De igual forma se deben homogenizar los datos que sean necesarios. Esta actividad consiste en llevar a cabo las tareas que involucran la limpieza y la selección de los datos.

3.1.8.1. Limpieza de datos.

La limpieza de datos implica identificar datos perdidos, errores de datos, incoherencia de codificación y metadatos ausentes o erróneos.
Para tener la calidad de los datos requerido por las técnicas de análisis seleccionadas es necesario la detección de diversos errores. A este nivel ya se tiene identificados cuales son aquellos datos anómalos, aquellos registros con valores que no son de utilidad para la investigación. Después del análisis de los datos, se realizó la detección de algunas anomalías.

En las bases de datos de mortalidad del año 1990 a 1997, solo existe una sola clasificación para la enfermedad de *diabetes mellitus*. En estos años no hay registros de alguna otra clasificación de dicha enfermedad. Por esa razón no existía una clasificación de los tipos E11, E12, E14.

En lo general no se cuenta con datos nulos. Solo se observó que, en algunas bases de datos, en especial las del 1996, 1997 y 1998 existían años anteriores a dichas bases de datos.

3.1.8.2. Selección de Datos

Para seleccionar los datos se realizó una selección de datos horizontales, es decir, elementos o registros que consideran útiles. Asimismo, se hizo una selección de los datos verticalmente que son atributos o columnas de interés.

Lo datos que se utilizaron corresponden a poblaciones de municipios mexicanos que tuvieran una población mayor a 100,000 habitantes de los años 1990 hasta el 2015. Asimismo, se utilizó información referente a los casos de muerte ocurridos en México por diabetes Tipo E11 en el rango de tiempo antes mencionado.

3.1.8.3. Selección vertical

En la selección vertical únicamente se seleccionan aquellos atributos que representan información de interés. Se elaboraron los filtros necesarios con el fin de obtener la información necesaria para el desarrollo del trabajo. A continuación, en las Tablas 2 y 3 se describen los atributos seleccionados de las bases de datos.

Tabla 2. Descripción de los atributos de la tabla poblacional

Atributo	Tipo de dato	Descripción	Observaciones
Clave del municipio (clave)	Numérico	Identificador del estado y del municipio.	La clave está conformada por el ID de la Ent_Regis mas el ID Mun_registro.
Municipio	Carácter	Nombre del municipio.	Nombre oficial del municipio.
Año	Numérico	Año en el que se contabilizo la población.	Sirve para identificar el año de registro.
Población	Numérico	Total de la población por año.	Se tiene el total en este atributo el total de la población por municipio y año.

Tabla 3. Descripción de los atributos de la tabla mortalidad

Atributo	Tipo de dato	Descripción	Observaciones
Clave del municipio (clave)	Numérico	Identificador del estado y del municipio.	La clave está conformada por el ID de la ENT_Regis más el ID Mun_registro.
Causa	Carácter	Nombre del municipio.	Nombre oficial del municipio.
Genero	Numérico	Año en el que se contabilizo la población.	Sirve para identificar el año de registro.
Estado Civil	Numérico	Número de habitantes.	Número de habitantes de acuerdo al año
Lugar de defunción	Numérico	Lugar de la defunción.	El lugar de la defunción está conformada por el ID de la Ent_reg mas el ID Mun_registro.
Escolaridad	Numérico	Indica de forma numérica la escolaridad.	Indica de forma numérica la escolaridad de la persona que falleció, esta esta especificada en el catálogo de la fuente de información.
Ocupación	Numérico	Indica de forma numérica la ocupación.	Indica la ocupación.
Edad	Numérico	Indica la edad .	Indica la edad de forma numérica de la persona que falleció.
Año	Numérico	Indica año de la defunción.	Indica el año de la defunción.

3.1.8.4. Selección horizontal

Para la selección de datos, se utilizó una condición, la cual fue que los municipios seleccionados debían contar con una población mayor a 100,000 habitantes.

3.1.8.5. Inclusión o exclusión de datos

Una vez recopilados los datos se procede a la utilización de la herramienta *Knime* para llevar a cabo las tareas de exploración y manipulación de la información. Tales actividades ayudarán a incluir o excluir información.

En la Figura 11 se muestra el flujo de trabajo que se utilizó para minar información de defunciones del año 2000 al 2015. Se tuvieron que eliminar datos que estaban fuera del rango, por ejemplo: en municipios hubo datos fuera del rango ya que únicamente deben existir valores del 1 al 32, este número sirve para identificar a cada estado del país. Por esta razón se tuvo que filtrar todo número distinto a este rango.

Otras acciones importantes fueron cambiar el tipo de dato de algunas columnas, por ejemplo, el hecho de convertir a cadenas los datos (*string*) y después pasarlos a un tipo numérico (*int*). Al final el flujo de trabajo se obtiene un archivo con extensión .csv para su posterior migración al *Data werehouse*.

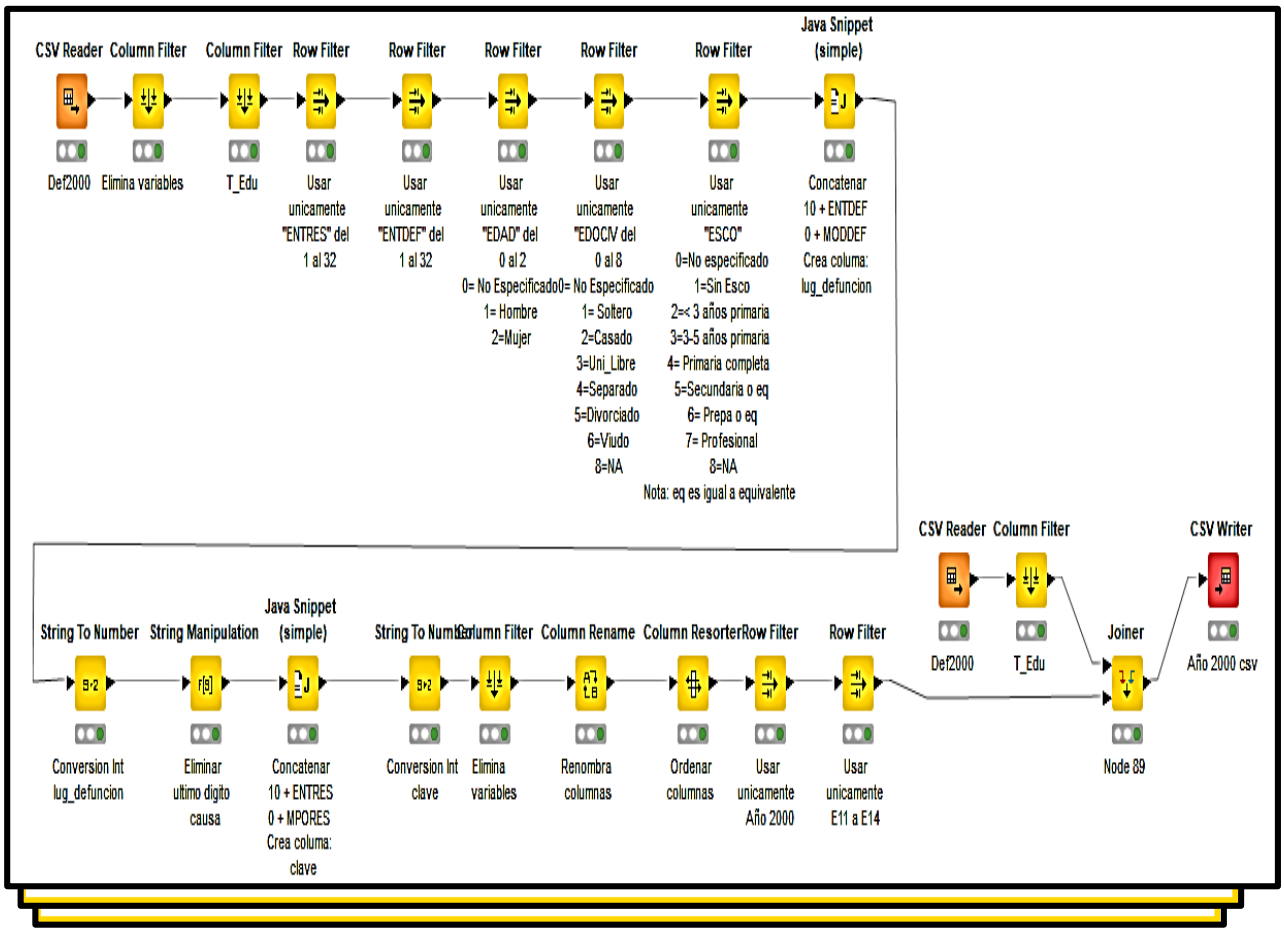

Figura 11. Flujo de trabajo en *KNime*, para datos de defunciones de los años 2000 al 2015

En la Figura 12 se visualiza el flujo de trabajo modificado y ampliado respecto a los años posteriores al 2000, debido a que la información previa al año 2000 contaba con columnas diferentes a las posteriores ese año. Se tuvieron que agregó una mayor cantidad de filtros, para cuadrar la información en la tabla de mortalidad del *Data werehouse*. Este flujo de trabajo se utilizó para los años de 1990 a 1999.

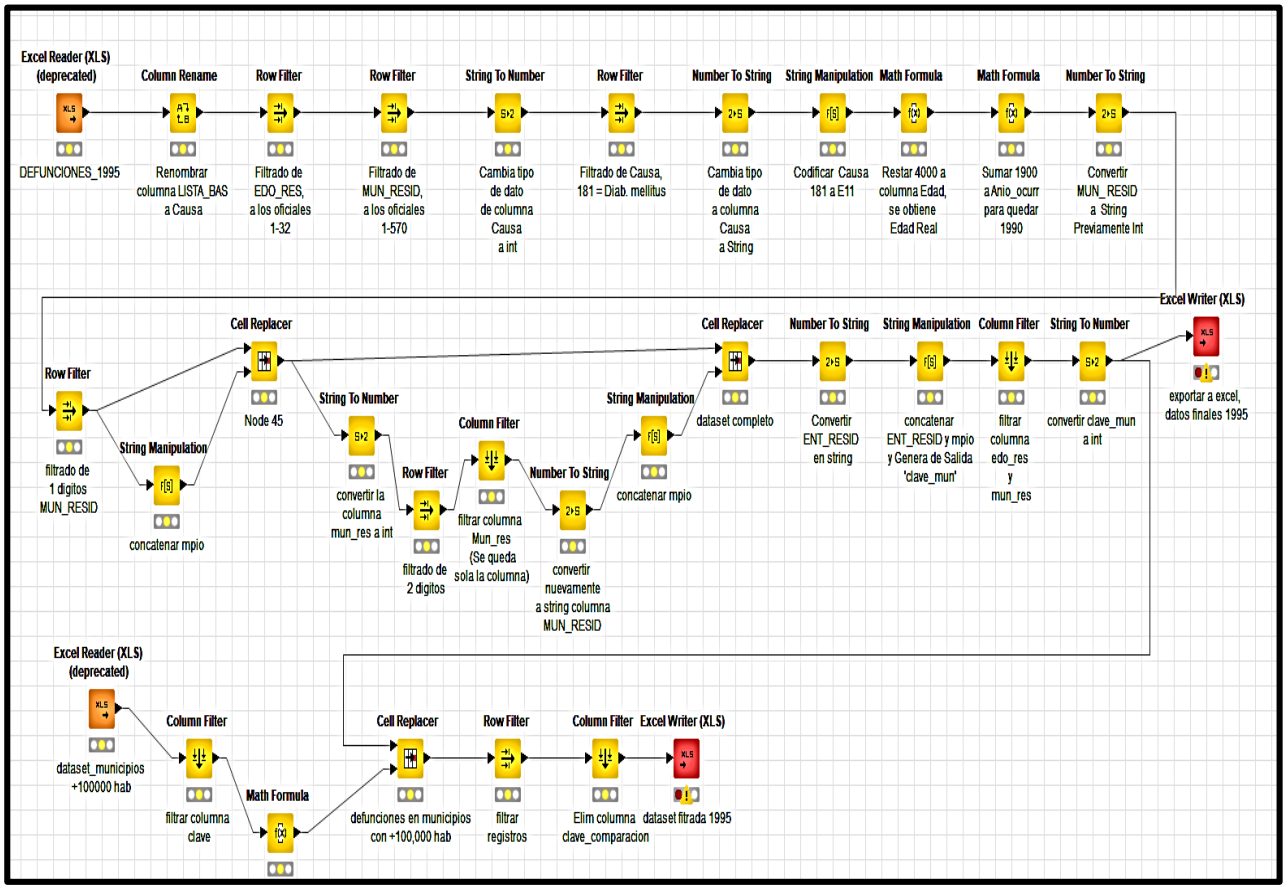

Figura 12. Flujo de trabajo *KNime*, para datos de defunciones

En la Figura 13, se muestra un flujo de trabajo que sirve para concatenar el número 10 a los atributos clave y lugar de defunción para que estos datos posteriormente se puedan integrar a las tablas de población y mortalidad del *Data werehouse.*

El dígito 10 sirve para identificar registros que pertenecen al país de México. Remitiéndose a la tesis de [7], se agregan datos de condados correspondientes al país de Estados Unidos, esos condados se distinguían por tener una clave que inicia con el número 20. Este flujo de trabajo se utilizó para los años 1990 a 1999 y para los años 2000 al 2015 respecto a los datos de población.

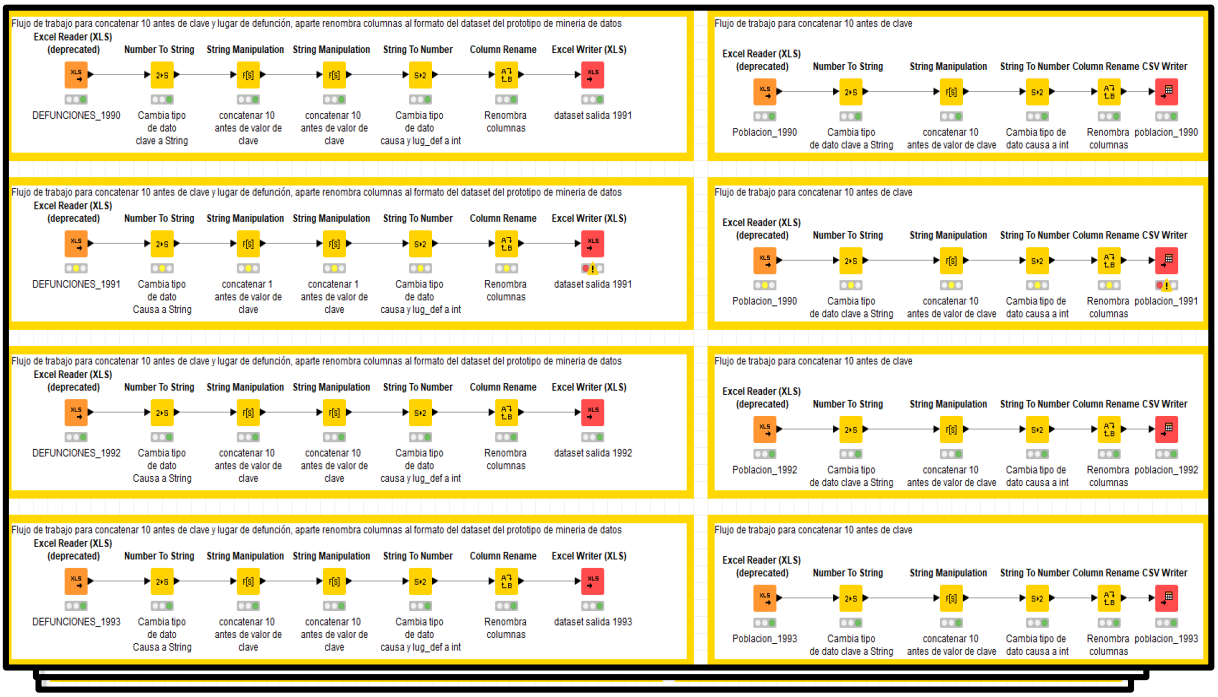

Figura 13. Flujo de trabajo de concatenación de identificador 10

Por otra parte, en la Figura 14, se muestra el flujo de datos para procesar la información poblacional de los años 1990, 1995, 2000, 2005 y 2010; misma que fue obtenida por [19]. El objetivo de estos flujos de trabajo es identificar los municipios con una población mayor a 100,000 habitantes. Los archivos de salidas de estos flujos de trabajo son archivos csv.

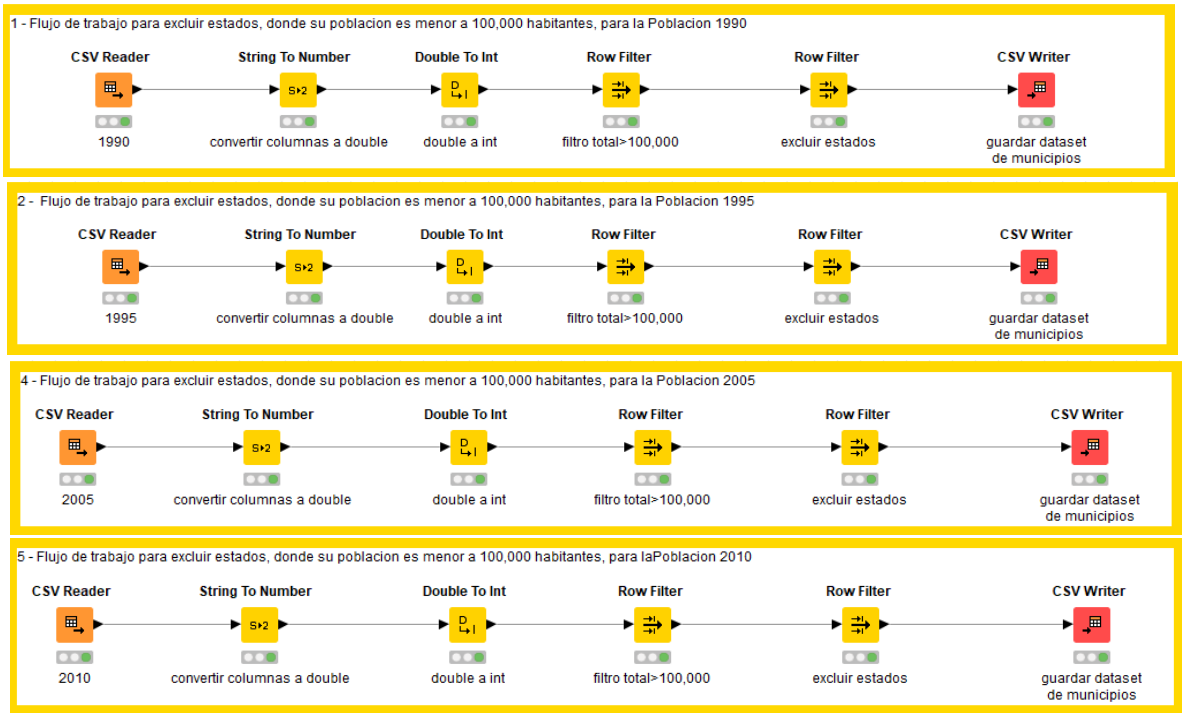

Figura 14. Flujo de trabajo para excluir municipios con una población menor a 100,000 habitantes.

Derivado de la investigación previa [19], se maneja información de la población por cada cinco años, se creó un nuevo flujo de trabajo para obtener los datos de los años faltantes correspondientes a la población (Figura 15). Para realizar el cálculo se aplicó la siguiente operación;

(Valor_poblacion_Año_Mayor – Valor_poblacion_Año_Menor) / 5

Ejemplo: La población del año 1995 menos la población del año 1990; el resultado de esa diferencia, se dividió entre cinco.

Clave	Municipio	Población 1995	Operación	Población 1990	Resultado resta	Operación	Resultado división	Población 1991
101001	Aguascalientes	582113	-	504387	77726	77726 ÷ 5	15545	519932

Una vez realizada esta ope ración, se procedió a sumar el resultado calculado al año menor de la población. En el ejemplo seria el año 1990. De esta forma se obtuvo el dato poblacional del año 1991. A este año (1991) se le sumó nuevamente el resultado para obtener el año 1992, este proceso se repitió para obtener cada año faltante.

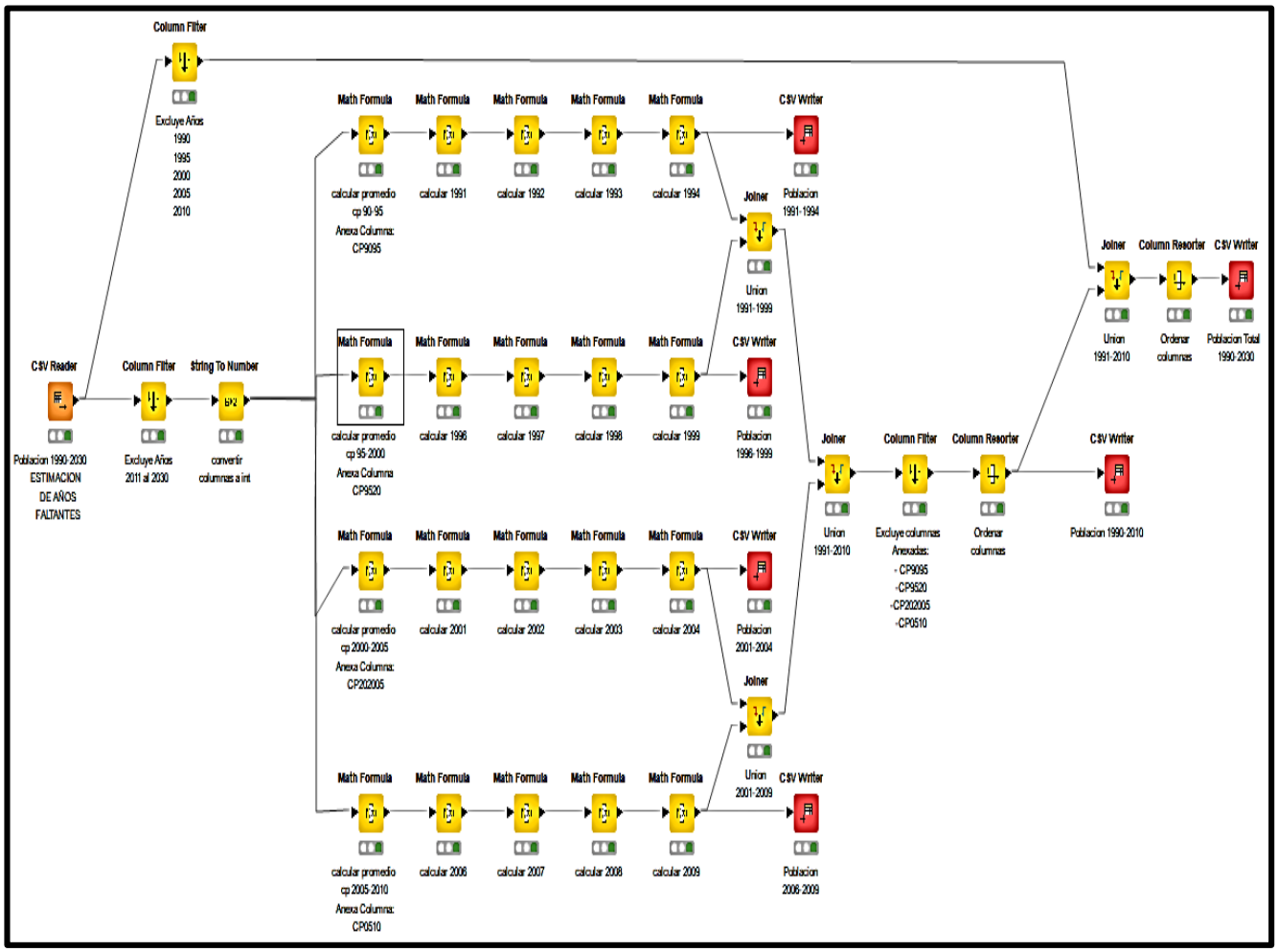

Figura 15. Flujo de trabajo para excluir municipios con población menor a 100,000 habitantes.

3.1.8.6. Construcción de datos

Para la construcción de la base de datos de mortalidad, el atributo 'clave' hace referencia al lugar del suceso, aquí fue necesario realizar una concatenación del valor 10 + ID_Estado + ID_Municipio.

3.1.8.7. Construcción de datos automáticos

Los atributos calculados a partir de la información original, son los siguientes: cálculo de incidencia, tasa de mortalidad y tasa de mortalidad normalizada. Para llevar a cabo esta actividad se utilizó únicamente el trabajo desarrollado por [10], el cual genera patrones de interés basados en el agrupamiento de distritos con altas tasas de mortalidad por varias enfermedades.

El trabajo antes mencionado calcula la incidencia de defunciones por causa de diabetes con los datos de mortalidad. Para ello se utilizó la siguiente Expresión (1):

Incidencia = Sumatoria de las defunciones por municipio por año (1)

Esta Expresión, indica que la incidencia se obtuvo a partir de la sumatoria de las defunciones de un municipio, en un año determinado por una causa de defunción. Para realizar este cálculo se utilizan los atributos clave, causa y año del dato de mortalidad.

Para calcular la tasa de mortalidad, se toma en cuenta el valor calculado de la incidencia y el total de habitantes de una población. La Expresión 2 indica el cálculo que se realizó.

$$\text{Tasa de mortalidad} = \frac{incidencia}{población} * 100\,000 \quad (2)$$

Una vez obtenido el valor de la tasa de mortalidad, fue necesario convertir tal valor en un dato normalizado. Es de suma importancia que todos los valores de los atributos que se utilicen en la técnica de minería de datos se encuentren en un mismo rango de valores.

La normalización es una técnica ampliamente utilizada en estadística para estandarizar una escala genérica entre cero y el valor de escalabilidad, la más utilizada es la normalización lineal uniforme, para obtener esta normalización se aplica la siguiente formula:

$$v' = \frac{v - min}{\max - min} * 10 \quad (3)$$

Dónde:

- v'se refiere al valor normalizado
- v es el valor por normalizar
- Max y min son los valores máximo y mínimo del conjunto de valores dados para ese atributo a normalizar.

3.1.8.8. Integración

En esta investigación se utilizaron datos provenientes de diversas fuentes y diferentes años, estos datos se agregaron al almacén de datos desarrollado en [10].

La Figura 16, muestra la estructura del *Data werehouse* utilizado en el trabajo de [10] y su respectiva descripción de cada tipo de dato y de cada atributo. Tal estructura, cuenta con una arquitectura elaborada con un modelo multidimensional, el cual es posible observar en Figura 17. Las tablas geográfica y población, hacen referencia a una dimensión de tipo lugar, las tablas catálogo y mortalidad indican la dimensión de tipo causa. Por último, la tabla año, hace referencia a la dimensión tipo tiempo en su atributo año. Es preciso mencionar que se utilizó la estructura del esquema del *Data werehouse* del trabajo de [10] para almacenar la nueva información minada de los años 1990 al 2015.

La implementación final de la información se hizo sobre un esquema tipo estrella, para lo que se utilizó el lenguaje de *MySQL*. Este esquema contiene las siguientes tablas:

1. Mortalidad: Tiene datos de mortalidad
2. Catálogo: Tiene datos de la CIE-10
3. Geográfica**:** Contiene datos geográficos
4. Poblacional: Tiene datos de población
5. Hechos: Almacena datos de los atributos derivados de incidencia y tasa de mortalidad.

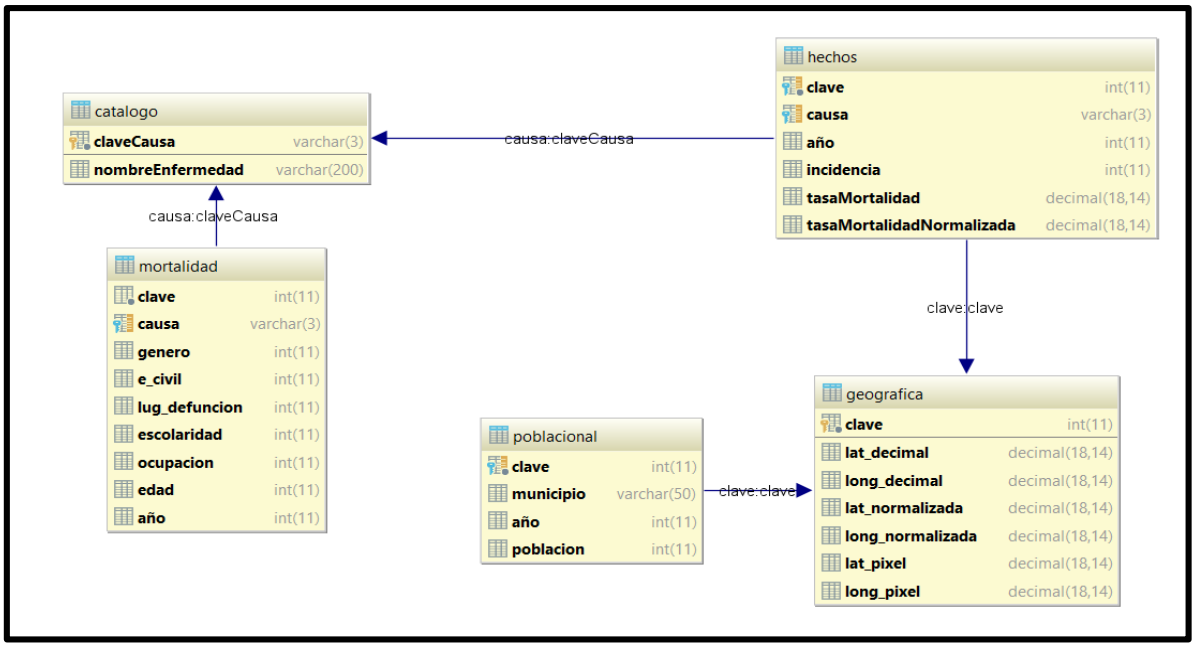

Figura 16. Estructura Data werehouse

3.1.8.9. Arquitectura del almacén de datos

El almacén de datos está basado en el modelo multidimensional a nivel conceptual. Los datos están organizados con relación a los hechos, los cuales contienen los atributos o medidas que pueden verse en mayor o menor detalle según ciertas dimensiones.

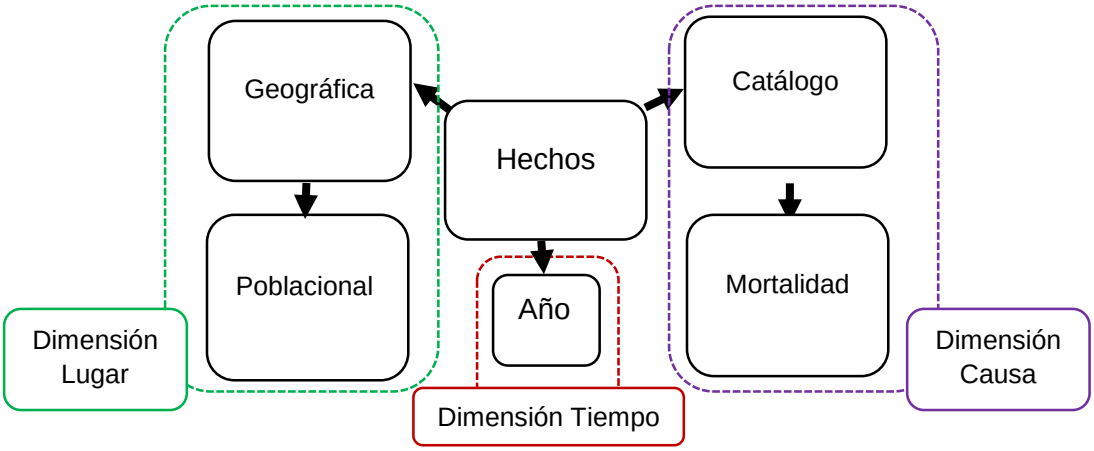

Figura 17. Representación del modelo multidimensional

Las dimensiones del almacén de datos presentado, muestra la relación entre *lugar* y *tiempo* en relación con un hecho, como lo son las defunciones ocurridas por una determinada *causa* de muerte.

3.1.8.10. Población del conjunto de datos final

En la tabla *hechos* se almacenan los registros que contienen la incidencia y mortalidad de cada municipio para determinada causa. Esta actividad emplea el trabajo desarrollado por [31], en el que se obtuvo como resultado el conjunto de datos que incluye los registros con las siguientes características:

	Atributo	**Descripción**
1.	Causa:	causa de muerte
2.	Latitud_norm:	latitud del municipio
3.	Longitud_norm:	longitud del municipio
4.	Tasam_norm:	tasa de mortalidad del municipio

Cabe mencionar que con este prototipo se genera un archivo de entrada para la fase de modelado.

3.2. Modelado

En esta fase, se desarrolló el prototipo de predicción el cual fue programado con el lenguaje de R. Dicho prototipo es capaz de realizar proyecciones de grado dos o grado tres de las tasas de mortalidad de municipios o alcaldías de México.

En la Figura 18, se muestra el diagrama de flujo correspondiente al proceso del algoritmo de regresión polinomial, donde el valor de *x* corresponde a los datos de año y la variable *y* corresponde a los datos de tasa de mortalidad.

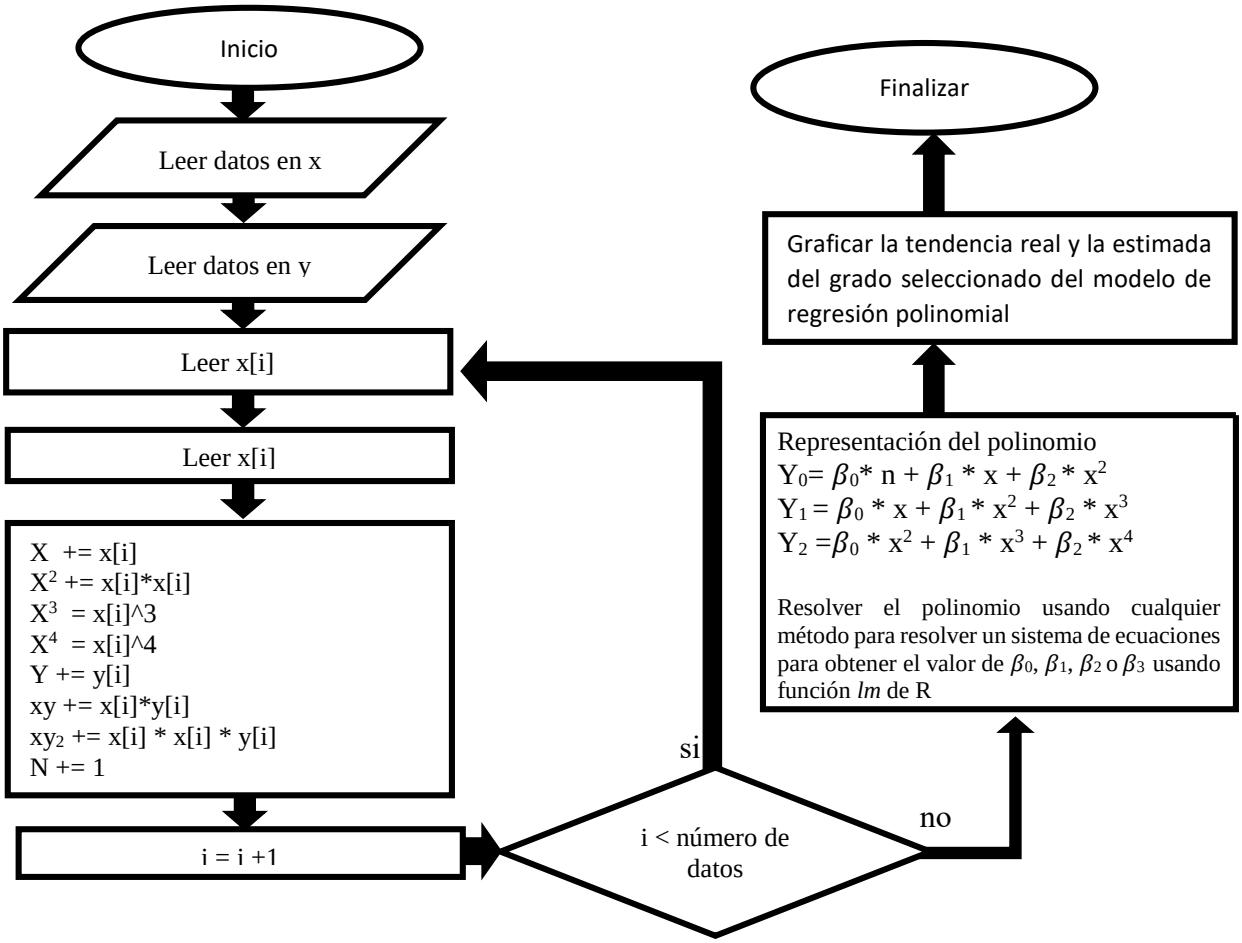

Figura 18. Diagrama de flujo de la regresión polinomial

3.2.1. Seleccionar técnica de modelado

La técnica seleccionada en la presente investigación es la técnica de regresión polinomial. Esta se utilizó para determinar la correlación de los datos de año y tasa de mortalidad del D*ata werehouse* mostrado en la Figura 16.

3.2.2. Generar el plan de prueba

Para validar el modelo seleccionado se realizaron un conjunto de pruebas. En particular se aplicaron las técnicas de regresión a los siguientes datos de entrada:

- Nombre del municipio
- Año (variable independiente)
- Tasa de mortalidad normalizada (variable dependiente)

Los datos anteriormente listados son obtenidos del almacén de datos descrito en la Figura 16. de la tabla de *hechos*. Una vez que se tiene estos datos se realizan predicciones tomando los valores históricos de las tasas de mortalidad desde el año 1998 hasta el 2015.

3.2.2.1. Objetivo del plan de pruebas

Validar el módulo del modelo de predicción, utilizando los datos correspondientes a los municipios con mayor tasa de mortalidad y menor tasa de moralidad de los grupos C08, C24 y C51 por diabetes Tipo E11. Tales pruebas tienen como objetivo: identificar la tendencia de las tasas de mortalidad para el año 2020.

3.2.2.2. Ambiente de pruebas

Las pruebas se llevaron a cabo en un equipo portátil con las siguientes características.

- Hardware:
    - Laptop Acer VX5
    - Procesador Intel Core (TM) i5-7300HQ
    - CPU 2.50 GHz
    - Memoria RAM de 8Gb.

- Software:
    - Sistema operativo Windows 10 Home Single Language
    - MySQL versión 6.4
    - Tomcat 5.0
    - Google Chrome 52.0001
    - R Versión 3.5.2
    - R Studio 1.2.1335
        - Librería ggplot2
        - Librería dplyr

3.2.3. Construir el modelo

Este modelo de regresión se programó en el ambiente de desarrollo de *R Studio*. Se utilizaron las librerías *dplyr* y *ggplot2*, la primera se ocupó para manipular y realizar operaciones con datos ordenados, además esta librería permite que cada variable tenga su propia columna y su propia fila. La segunda librería se utilizó para graficar y visualizar los datos de la regresión.

3.2.4. Evaluar el modelo

Para determinar qué grado se debe tomar en cuenta al momento de realizar la visualización de la predicción, es importante identificar el valor del coeficiente de β_0, β_1, β_2 en caso de tener una expresión de grados dos u observar a β_3 en un modelo de grado tres. Ambos coeficientes se compararán y el que este más aproximado a 0.05, es el coeficiente que se utiliza.

3.3. Despliegue

Con la finalidad de realizar pruebas y verificar que es posible obtener patrones de relación entre los municipios dada su localización y su tasa de mortalidad por diabetes y a su vez identificar los grupos con mayor tasa de mortalidad, se toma como referencia el prototipo de aplicación de minería de datos [11]. En este prototipo, se utiliza el algoritmo de agrupamiento *K-Means*, porque este resultó ser una opción viable para encontrar patrones de interés con altas tasa de mortalidad por cáncer de estómago.

Particularmente, en esta investigación se utilizó una versión *N-Means*, que cuenta con un módulo que recibe un conjunto de datos para agrupar. Asimismo, es necesario configurar los parámetros necesarios que están previamente establecidos por el usuario por medio de una interfaz.

Como resultado, este módulo genera dos archivos de texto con información importante. El primer archivo detalla la descripción de los grupos (información de los valores de los centroide al final de la ejecución); el segundo archivo especifica la asignación de grupos (información de la asignación de cada objeto del conjunto de datos a un grupo en específico generado por el algoritmo). Posteriormente, será necesario utilizar estos archivos en la interfaz de usuario para la creación de imágenes del mapa de la República Mexicana donde se representarán los grupos.

La misma imagen será desplegada en la interfaz del usuario, para observar los grupos y la tasa de mortalidad por medio de una paleta de colores que indica los valores más altos de las tasas de mortalidad.

Una vez realizado el proceso de agrupamiento y de visualización de los patrones en un mapa, se procedió a identificar los grupos con mayor tasa de mortalidad mediante la paleta de colores. Se seleccionaron los tres grupos con mayor tasa de mortalidad para un posterior análisis, de cada municipio y de cada grupo. Se inspecciona cómo ha sido su tasa de mortalidad a través del rango de tiempo de 1990 al 2015. La finalidad de esta actividad es identificar a los dos municipios con mayor tasa de mortalidad de cada grupo y el que presente menor tasa de mortalidad en el periodo de tiempo mencionado.

Una vez identificados los municipios, se toman los valores correspondientes al año y el respectivo valor de la tasa de mortalidad para que el prototipo desarrollado pueda calcular las predicciones. El prototipo muestra una gráfica con dos tendencias, una que es de color rojo, indica la tendencia real de la tasa de mortalidad, y la segunda línea azul muestra la proyección de la tasa de mortalidad normalizada.

En la Figura 18, se muestra el esquema del prototipo de minería de datos para el pronóstico de diabetes mellitus. En él se resalta el proceso general que se debe realizar antes de ejecutar el prototipo que realiza las predicciones de las tasas de mortalidad por causa Tipo E11 de municipios y alcaldías con altos índices de mortalidad por la causa antes mencionada.

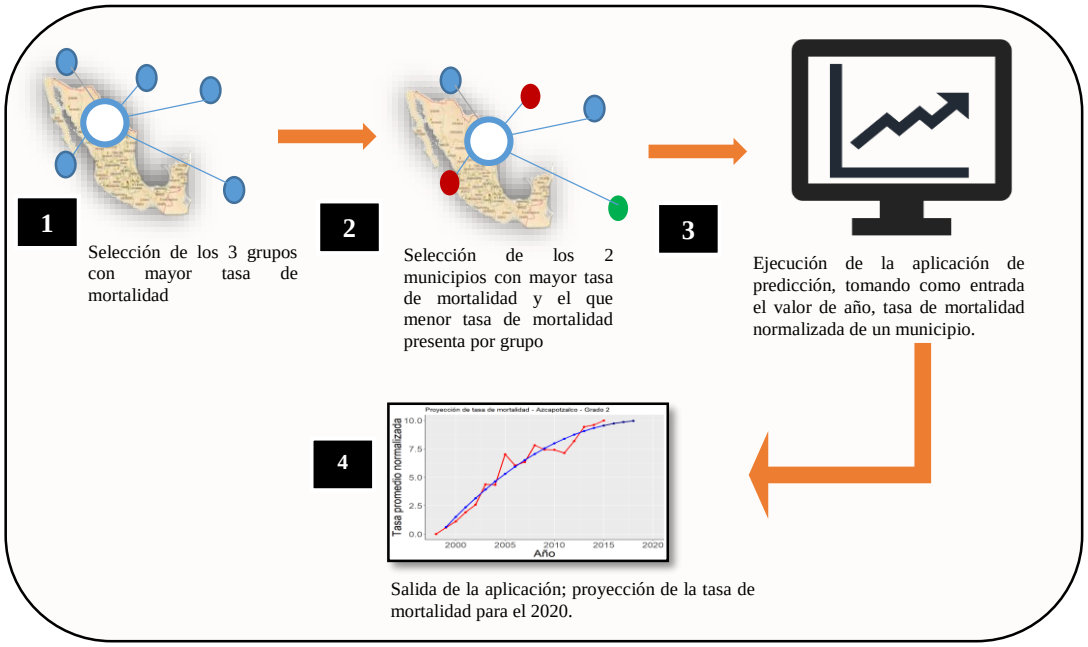

Figura 19. Esquema del prototipo de minería de datos para el pronóstico de *diabetes mellitus*

3.3.1. Evaluar resultados

En esta parte de la metodología se realizaron múltiples pruebas de grado dos y grado tres de los municipios y alcaldías pertenecientes a los grupos C08, C24 y C51, se evaluaron los valores de los coeficientes de β_2 y β_3 de los modelos que se utilizaron.

En el capítulo 4, se muestra a detalle los patrones obtenidos una vez que se ejecutó el prototipo basado en el algoritmo de regresión polinomial, aplicado a las regiones de los municipios con mayor y menor tasa de mortalidad de los grupos antes mencionados.

Capítulo 4

4. Patrones obtenidos

En el presente capítulo se reportan los resultados obtenidos una vez que se ejecutó el prototipo basado en el algoritmo de regresión polinomial, aplicado a las regiones de los municipios con mayor y menor tasa de mortalidad de los grupos C08, C24 y C51.

4.2.1. Grupo C08

Los municipios y alcaldías que conforman el grupo C08 son 9: los cuales son; Piedras Negras, Azcapotzalco, Gustavo A. Madero, Benito Juárez, Miguel Hidalgo, Nezahualcóyotl, El Mante, Matamoros y Tampico. De este conjunto de municipios y alcaldías, los que mayor tasa de mortalidad presentaron durante el periodo de 1998 al 2015, fueron Azcapotzalco y Miguel Hidalgo en la Ciudad de México (CDMx), mientras que el que menor tasa presentó fue Matamoros del estado de Tamaulipas. Estos fueron los municipios de mayor interés para la aplicación del algoritmo predictivo.

En la Tabla 4 se muestra la tasa de mortalidad de la alcaldía de Azcapotzalco, misma que posteriormente se normalizó para proceder a la aplicación del modelo predictivo. Cabe mencionar que este municipio es el que mayor tasa de mortalidad presentó en el grupo C08.

Tabla 4. Tasa de mortalidad de la alcaldía de Azcapotzalco.

Primer lugar en tasa de mortalidad del grupo C08		
Alcaldía	**Año**	**Tasa mortalidad**
Azcapotzalco	1998	28.89494880176260
	1999	33.24990491893630
	2000	37.41428726916520
	2001	43.60037013931250
	2002	48.69346992475920
	2003	62.39143535751000
	2004	62.03651564751810
	2005	82.63467283415250
	2006	75.03981951713090
	2007	77.34441125823310
	2008	88.71357977623650
	2009	85.70390161486780
	2010	85.60178051703480
	2011	83.45920360197420
	2012	91.49640186768850
	2013	101.03592253101100
	2014	102.31606362198900
	2015	105.30200516668700
Promedio:		71.94048302033160

En la Gráfica 1, se muestra en color azul, la proyección de la tasa de mortalidad de Grado 2 del modelo de regresión polinomial aplicado a la alcaldía de Azcapotzalco de CDMx y de color rojo se muestra la tendencia real de la tasa de mortalidad. En tal proyección se observa que para futuros años se espera un incremento en su tasa de mortalidad. Desde el año 1998, se presenta un aumento constante en la tendencia de la tasa de mortalidad por diabetes Tipo E11.

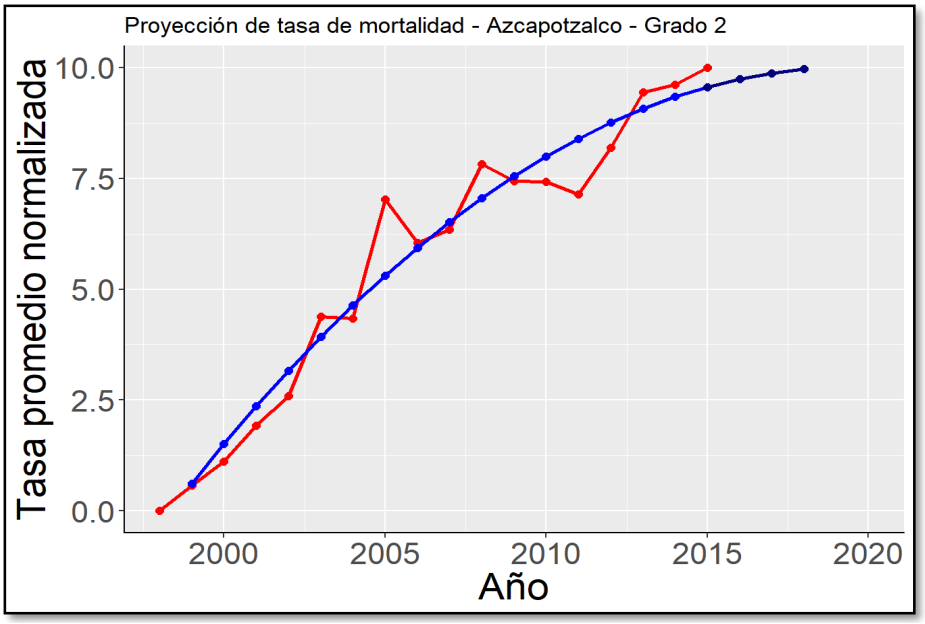

Figura 20. Tendencia real y proyección de la tasa de mortalidad Grado 2 - Azcapotzalco

En la Tabla 5, se muestra la tasa de mortalidad del municipio de Miguel Hidalgo, el cual arrojó datos interesantes en cuanto a su comportamiento, ya que éste es la alcaldía que mayor decremento presenta en su tasa de mortalidad por diabetes Tipo E11, para el año 2020.

Tabla 5. Tasa de mortalidad de la alcaldía de Miguel Hidalgo.

Segundo lugar con mayor mortalidad del grupo C08		
Alcaldía	Año	Tasa mortalidad
Miguel Hidalgo	1998	30.43334812337170
	1999	35.37674798525110
	2000	37.99909255898370
	2001	38.98989780542570
	2002	49.93440569193130
	2003	64.62680255171630
	2004	68.14688505691920
	2005	69.60563866199800
	2006	136.39912297756900
	2007	71.55677379484500
	2008	76.29950332239030
	2009	79.20151003170950
	2010	66.23955118010990
	2011	69.17966645895160
	2012	73.41873414523750
	2013	73.82924163443750
	2014	68.30224267786820
	2015	75.73366992742190
Promedio:		**65.84849081034100**

En la Gráfica 2, se muestra en color rojo la tendencia real de la tasa de mortalidad y en color azul, la proyección de la tasa de mortalidad en un Grado 3 del modelo de regresión polinomial correspondiente a la alcaldía de Miguel Hidalgo. En tal proyección se observa que, para futuros años, existirá un decremento constante en su tasa de mortalidad.

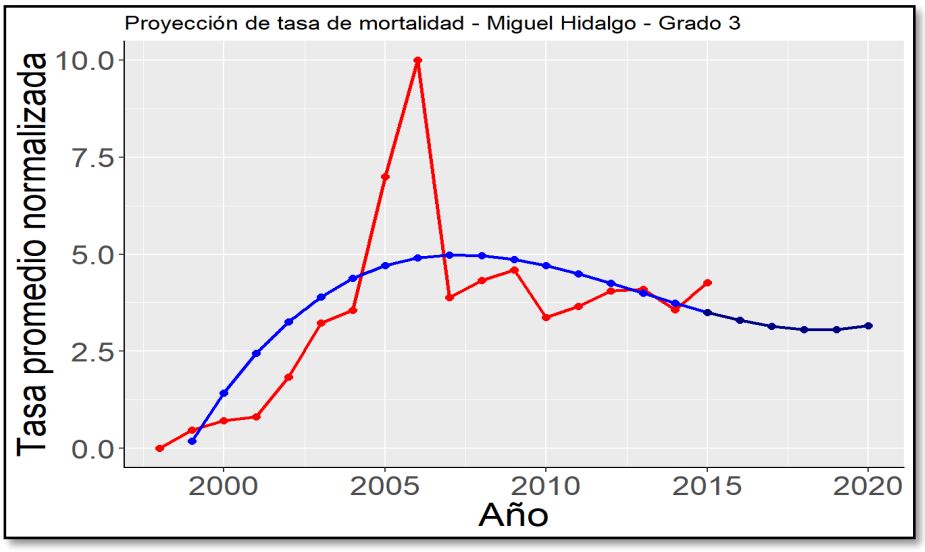

Figura 21. Tendencia real y proyección de la tasa de mortalidad Grado 3 - Miguel Hidalgo

Tabla 6. Tasa de mortalidad del municipio de Matamoros.

Lugar con menor tasa de mortalidad de grupo C08		
Municipio	**Año**	**Tasa mortalidad**
Matamoros	1998	18.79331669024300
	1999	24.01774837323090
	2000	31.56829873176750
	2001	29.43669093611050
	2002	30.57140523447800
	2003	40.38855626974970
	2004	33.85782872001610
	2005	41.75077894072940
	2006	37.91057924297170
	2007	46.06570743301550
	2008	34.79716243016550
	2009	40.04425944464930
	2010	38.83947644385750
	2011	40.53216346383100
	2012	45.63955375532870
	2013	50.03338803910000
	2014	49.49675093311970
	2015	60.76757640236900
Promedio:		**38.58395786026290**

En la Gráfica 3, se muestra de color rojo la tendencia real de la tasa de mortalidad del municipio de Matamoros, y de color azul, la proyección de la tasa de mortalidad en un Grado 2 del modelo de regresión polinomial. En tal proyección se observa que para los próximos años la tasa de mortalidad ira creciendo constantemente, esto indicaría que, dentro de pocos años, este municipio dejará de ser el que menor tasa de mortalidad presente dentro del grupo C08.

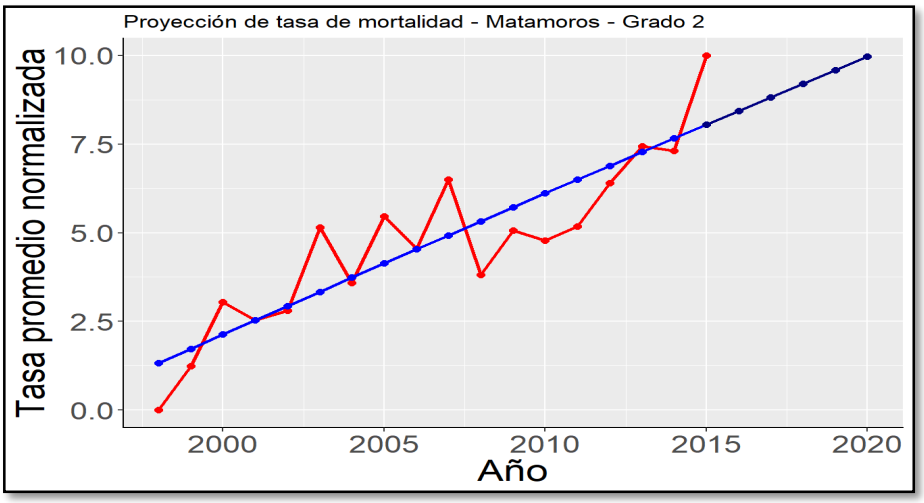

Figura 22. Tendencia real y proyección de la tasa de mortalidad Grado 2 - Matamoros

4.2.2. Grupo C24.

El grupo C24, está conformado por las siguientes alcaldías y municipios; Cuauhtémoc, Iztacalco, Venustiano Carranza y Orizaba. De estos municipios y alcaldías, los dos que mayor tasa de mortalidad presentaron durante el periodo de 1998 al 2015, fueron la alcaldía de Venustiano Carranza de Ciudad de México y el municipio de Orizaba Veracruz, siendo el primero de estos el que mayor proyección tiene en su tasa de mortalidad respecto a las alcaldías que corresponden a la zona metropolitana del país.

De igual manera el municipio de Orizaba es el que mayor proyección de tasa de mortalidad presenta por parte de los municipios de provincia del país. Por otro lado, la alcaldía de Cuauhtémoc es el municipio que menor tasa de mortalidad presentó en el grupo C24.

En la Tabla 7, se muestra la tasa de mortalidad de la alcaldía de Venustiano Carranza. Estas tasas se utilizaron para aplicar el modelo predictivo.

Tabla 7. Tasa de mortalidad de la alcaldía de Venustiano Carranza.

Primer lugar - mortalidad en grupo C24		
Alcaldía	**Año**	**Tasa mortalidad**
Venustiano Carranza	1998	33.52136613445270
	1999	52.94183744911100
	2000	53.37009459687210
	2001	55.38484714889880
	2002	58.12544323455750
	2003	63.21213127204180
	2004	73.73441784372910
	2005	75.90106238133160
	2006	49.05608119368230
	2007	82.56722741609400
	2008	94.36847808303950
	2009	86.27161258802920
	2010	86.54734116358610
	2011	99.23487131926730
	2012	101.89463253663900
	2013	102.12677839430300
	2014	116.95629527913300
	2015	104.98825131473400
	Promedio:	**77.23348718608340**

En la Gráfica 4. se muestra en color azul, la proyección de la tasa de mortalidad en un Grado 2 del modelo de regresión polinomial correspondiente al municipio de Venustiano Carranza. Se observa que para futuros años se espera un incremento constante en su tasa de mortalidad.

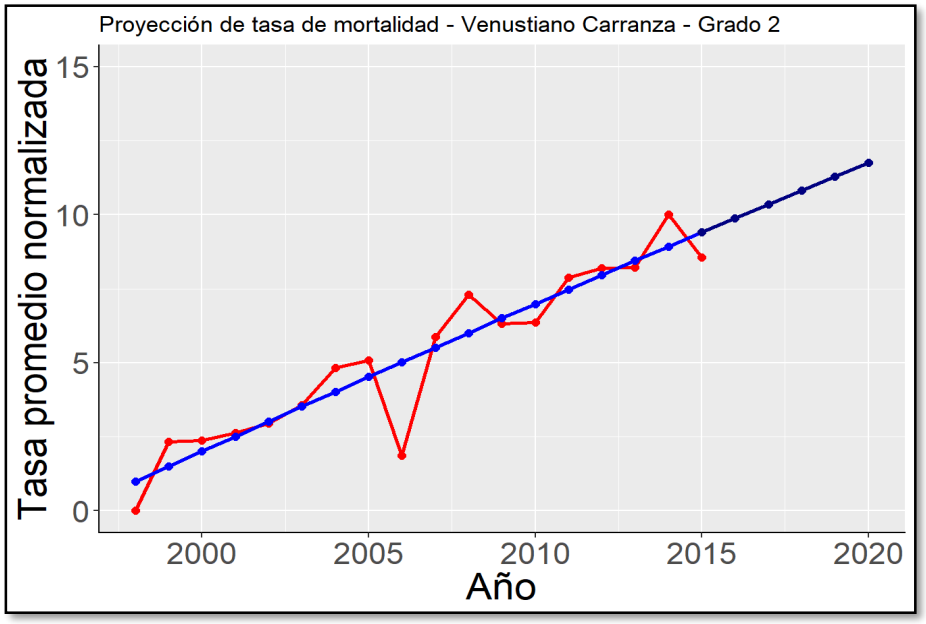

Figura 23. Tendencia real y proyección de la tasa de mortalidad Grado 2 - Venustiano Carranza

Para aplicar el modelo predictivo al municipio de Orizaba, se utilizó la tasa de mortalidad que se muestra en la Tabla 8.

Tabla 8. Tasa de mortalidad - Orizaba.

Segundo lugar - mortalidad grupo C24		
Municipio	Año	Tasa mortalidad
Orizaba	1998	20.58848760401480
	1999	24.71555801764180
	2000	56.49574595465160
	2001	53.52136606915300
	2002	62.22351025835540
	2003	61.57635467980300
	2004	61.78295305353660
	2005	61.12995712293150
	2006	105.40026392911600
	2007	66.52452025586350
	2008	77.24826403626420
	2009	96.32120587390370
	2010	86.78044547295340
	2011	79.92040579993800
	2012	73.08932327407680
	2013	79.24250632727160
	2014	99.02424886484400
	2015	88.97866916768870
Promedio:		69.69798809788930

En la Gráfica 5, se observa el crecimiento constante de la tasa de mortalidad del municipio de Orizaba, con el cual se prevé que para el año 2020 sería el municipio de provincia con mayor tasa de mortalidad por causa de diabetes tipo E11.

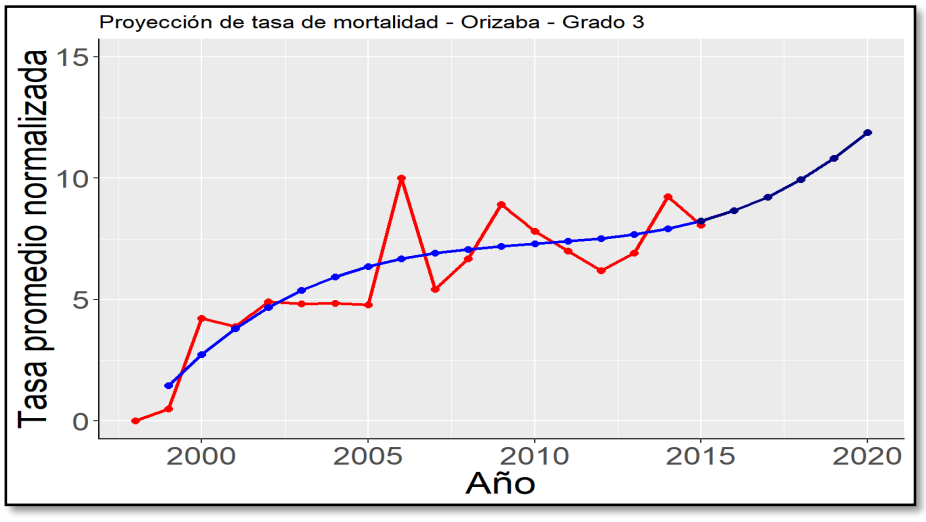

Figura 24. Tendencia real y proyección de la tasa de mortalidad Grado 3 - Orizaba

4.2.3. Grupo C51

Respecto al grupo C51, los municipios que lo conforman son diez: Coyoacán, Tláhuac, San Francisco del Rincón, Chalco, Netzahualcóyotl, Tecámac, Toluca, Zamora, Cuernavaca, Atlixco, San Pedro Cholula, y Martínez de la Torre. El municipio y alcaldía que resaltaron por su alto índice de mortalidad fueron Netzahualcóyotl de Ciudad de México y Cuernavaca Morelos; mientras que el municipio que menor tasa de mortalidad presentó fue San Pedro Cholula Puebla. Este es el municipio que proyecta el mayor decremento en su tasa de mortalidad para el 2020 respecto a los municipios de provincia utilizados para este trabajo de investigación.

Tabla 9. Tasa de mortalidad del municipio de San Pedro Cholula

Menor lugar - mortalidad grupo C51		
Municipio	Año	Tasa mortalidad
San Pedro Cholula	2002	30.31666242885360
	2003	48.76183191509700
	2004	32.73291809288670
	2005	49.42158441938790
	2006	51.18672378027420
	2007	49.36095196121640
	2008	57.12159109595560
	2009	64.60607297085930
	2010	72.22374417851720
	2011	51.79795559943990
	2012	68.83469268511330
	2013	75.22607414915350
	2014	85.42252803661410
	2015	81.47113594040970
Promedio:		58.46317623241280

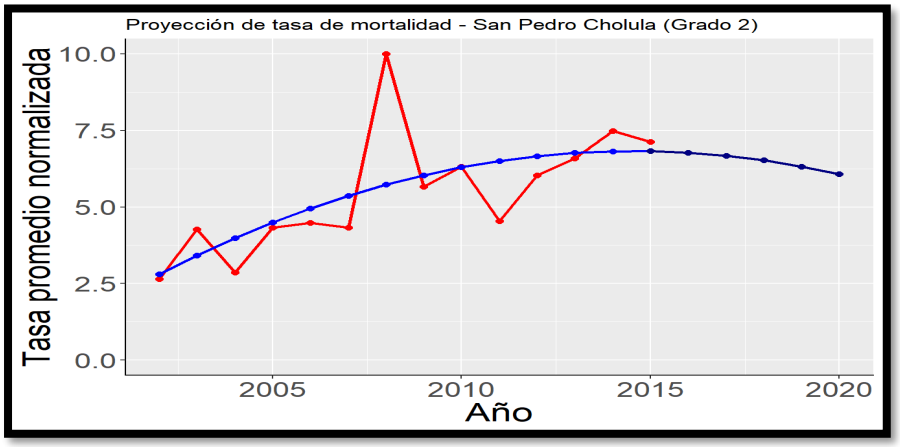

Figura 25. Tendencia real y proyección de la tasa de mortalidad Grado 2 – San Pedro Cholula

Capítulo 5

5. Conclusiones y trabajos futuros

En este capítulo se exponen las conclusiones del proceso de investigación y desarrollo de la presente tesis. Del mismo modo, se sugieren algunos temas para futuras investigaciones con el fin de dar continuidad a esta línea de investigación.

5.1. Conclusiones

Con lo expuesto en esta investigación, se muestra que es factible predecir la evolución de la *diabetes mellitus* en México, mediante el desarrollo de un prototipo de minería de datos. Dicho prototipo utiliza técnicas predictivas y el lenguaje de programación R para su desarrollo, entre otros recursos computacionales.

El minado de los datos se realizó a partir de bases de datos oficiales de defunciones, poblacionales y geográficas. Como resultado de aplicar el proceso de minería de datos, se identificaron 60 grupos. Para esta investigación fue de particular interés los tres grupos con mayor promedio en sus tasas de mortalidad, los cuales se denominaron en la presente investigación como: C08, C24 y C51.

A continuación, se describen los patrones para estos tres grupos, mencionando el pronóstico para el año 2020 del primer y segundo municipio con mayor promedio en sus tasas de mortalidad. Asimismo, se menciona el pronóstico del municipio con menor de mortalidad de cada grupo:

Grupo C08. La alcaldía de Azcapotzalco presentó la mayor tasa de mortalidad y su pronóstico para el año 2020 se estima que crecerá un 4.5%. La alcaldía de Miguel Hidalgo quedo en segundo lugar de acuerdo a su tasa de moralidad, se pronostica que tendrá un decremento del 10.7% para el año 2020. El municipio de Matamoros tuvo el menor índice de tasa de mortalidad en este grupo, pero se pronostica que seguirá creciendo un 19.2 % más en el año 2020.

Grupo C24. La alcaldía de Venustiano Carranza tuvo la mayor tasa de mortalidad y se pronostica un aumento del 20% para el año 2020. El municipio con segunda tasa de mortalidad fue el de Orizaba, para el cual se pronostica un crecimiento del 30% en su tasa de mortalidad para el 2020 y se predice que la alcaldía de Cuauhtémoc disminuirá su tasa de mortalidad en un 40.7% para el año en mención.

Grupo C51. El municipio con mayor tasa de mortalidad es la alcaldía de Netzahualcóyotl para la cual se pronostica el crecimiento del 17.5 % en su tasa de mortalidad para el año 2020. El segundo lugar en tasa de mortalidad es el municipio de Cuernavaca, se le pronostica un incremento del 7.8 % para el año 2020. El municipio con menos tasa de mortalidad es San Pedro Cholula y se pronostica un decremento en su tasa de mortalidad del 23% para el año 2020.

Es destacable que los municipios y alcaldías del grupo C24 muestran mayores porcentajes de crecimiento en los próximos 5 años, por ejemplo, el municipio de Orizaba y la alcaldía de Venustiano Carranza se predice que crezca el 30% y 20% respectivamente para el año 2020. Por otra parte, los municipios del grupo C51, aun cuando presentan incremento en sus tasas de mortalidad, sus incrementos están por debajo del 18% para el 2020.

De manera individual se observó, que para la alcaldía Cuauhtémoc el pronóstico del valor de la tasa de mortalidad para el año 2020 fue el menor, mientras que el pronóstico para la alcaldía de Venustiano Carranza muestra un incremento mayor en su tasa de mortalidad.

Se considera que los resultados obtenidos en esta investigación son de suma importancia, porque pueden ser útiles para los funcionarios o personas responsables del sector de salud pública del país para la toma de decisiones, en cuanto a la promoción y prevención de la salud por la enfermedad de *diabetes mellitus*, puesto que les permite identificar regiones con altas tasas de mortalidad en México, de tal manera que se les podría facilitar centrar su atención en aquellos municipios donde se pronostica un crecimiento en su tasa de mortalidad por la causa E11.

Finalmente es destacable desde el punto de vista computacional, los resultados obtenidos con la metodología usada para el desarrollo del prototipo y la preparación de los datos, así como el uso y selección de las técnicas para llevar acabo la predicción. Si bien el prototipo se centra en una enfermedad es factible con pocas modificaciones realizar minería de datos para obtener patrones y predecir el comportamiento de otras enfermedades, desde el punto de vista epidemiológico.

5.2. Trabajos futuros

La presente investigación da pauta al desarrollo de trabajos posteriores que den continuidad al trabajo realizado, por lo que se propone lo siguiente:

- Aplicar un algoritmo predictivo diferente a los datos minados para la causa E11.
- Comparativa de precisión de los algoritmos predictivos aplicados a bases de datos de mortalidad por alguna causa de acuerdo a la clasificación CIE-10.
- Utilizar otro lenguaje de programación para el desarrollo del algoritmo predictivo y aplicarlo a alguna otra enfermedad con altos índices de mortalidad en México.
- Identificar los municipios de la frontera norte y sur del país México con altos índices de mortalidad y aplicar proyecciones en sus tasas de mortalidad.

Referencias

[1] Sistema Nacional de Información en Salud. Fecha de consulta: septiembre 2018. Disponible en: http://www.dgis.salud.gob.mx/contenidos/sinais/estadisticas.html

[2] Instituto Nacional de Estadística y Geografía. Fecha de consulta: septiembre 2018. Disponible en: http://www.inegi.org.mx/default.aspx

[3] Federación Mexicana de Diabetes. Fecha de consulta: julio 2018. Disponible en: http://fmdiabetes.org/diabetes-en-mexico/.

[4] World Health Organization / *Diabetes programme*. Fecha de consulta: mayo 2018. Disponible en: http://www.who.int/diabetes/en/

[5] Secretaria de Salud. *Programa de Acción Especifico Prevención y Control de la Diabetes Mellitus 2013 – 2018*. Programa Sectorial de Salud. Edición Electrónica 2013, Disponible en: https://www.gob.mx/cms/uploads/attachment/file/37607/PAE_PrevencionControl1DiabetesMellitus2013_2018.pdf

[6] Organización Panamericana de la Salud. *OPS*, julio 2018. Disponible en: https://www.paho.org/hq/index.php?option=com_content&view=article&id=11889: diabetes-in-the-americas&Itemid=1926&lang=es

[7] I. Blanco, "Aplicación de minería de datos en el área de Salud pública." Maestría, Ciencias computacionales, CENIDET, Cuernavaca, México, 2017.

[8] G. Iturbide, "Metodología de Preparación de Datos Orientada a aplicaciones de Epidemiología Basada en el Modelo CRISP-DM." Maestría, Ciencias computacionales, CENIDET, Cuernavaca, México, 2013.

[9] L. Sánchez, "Desarrollo de una Aplicación de Ciencia de Datos." Maestría, Ciencias computacionales, CENIDET, Cuernavaca, México, 2018.

[10] J. Pérez-Ortega, *et al.*, "An Epidemiological Data Mining Application Based on Census Databases" en conf. DBKDA '13, Sevilla, España., pp. 217-253, 2013.

[11] J. Pérez-Ortega, *et al.*, "A Data Mining System for the Generation of Geographical C16 Cancer Patterns" en conf. *Fifth International Conference on Software Engineering Advances* '10, Recife, Brasil, 2010.

[12] J. Pérez-Ortega, *et al.,* "A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases", *J Med Syst* '15, pp. 1173—1182, 2015

[13] CRISP-DM 1.0 *Step by step data mining guide*. SPSS Inc. P. Chapman, *et al*., Estados Unidos de América, 2000.

[14] J. Pérez-Ortega, *et al.,* "Improving the Efficiency and Efficacy of the K-Means Clustering Algorithm through a new convergence condition", *Springer-Verlag Berlin Heidelberg.'*07, pp. 674–682, 2007

[15] J. Pradeep, et al., "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes". *Procedia Comput. Sci* '16, vol. 82, pp. 115–121, 2016.

[16] Nahla H. Barakat, *et al,* "Intelligible Support Vector Machines for Diagnosis" en conf. *IEEE Transactions on Information Technology in Biomedicine*, '10, vol. 14, No. 4, pp. 1114—1120, 2010.

[17] Gagandeep Singh, *et al.," ,"* Diabetes Classification using k means" en conf. *Appejay Journal Of Computer Science And Applications*

[18] J. Pradeep Kandhasamy, et al, "Performance Analysis of Classifier Models to Predict *Diabetes Mellitus*", *Procedia Computer Science,* vol. 47, pp. 45–51, 2015.

[19] Consejo Nacional de Población. Fecha de consulta: septiembre 2018. Disponible en: *https://www.gob.mx/conapo*

[20] Centro Mexicano para la Clasificación de Enfermedades. Fecha de consulta: septiembre 2018. Disponible en: *http://www.dgis.salud.gob.mx/contenidos/cemece/cindex_gobmx.html*

[21] K. Mehmed. *Data Mining: Concepts, models, methods and algorithms.* EUA: Institute of Electrical and Electronics Engineers. 2003

[22] D. Hand, *et al.,* "Principles of data Mining", *Cambridge, Massachusetts*: The MIT press, 2001.

[23] Daniel T. Larose, *Data Mining and Predictive Analytics*, Wiley", ,2nd. ed, John Wiley, 2015

[24] World Health Organization / *Diabetes programme*. Fecha de consulta: junio 2018. Disponible en: http://www.who.int/topics/epidemiology/es/

[25] R. Beaglehole, *et al.*, *Epidemiologia Básica*. Segunda Edición, Washington, D.C: Organización Mundial de la Salud. 2008

[26] M. Ramírez, *et al*., *Introducción a la Minería de datos*. España: Prentice Hall, pp. 29 – 93. 2015

[27] Secretaria de Salud. *Salud: México 2002. Información para la rendición de cuentas. México*. DF: Secretaria de Salud, 2003.

[28] Ronald E. Walpole, *et al*., *Probabilidad y estadística para ingeniería y ciencias*, 9 ed, Pearson, pp. 389-505, México, 2012

[29] E. Pérez., "Minería de datos Orientada al Big Dara en el Área de Salud." Maestría, Ciencias Computacionales, CENIDET, Cuernavaca, México, 2017

[30] Knime. Fecha de consulta: agosto 2018. Disponible en: https://www.knime.com

[31] E. Iturbe., "Metodología de Preparación de Datos Orientada a Aplicaciones de Epidemiologia Basa en el Modelo CRIPS-DM" Maestría, Ciencias computacionales, CENIDET, Cuernavaca, México, 2013

[32] E. Pérez, "Minería de Datos Orientada al Big Data en el Área de Salud." Maestría, Ciencias computacionales, CENIDET, Cuernavaca, México, 2017.

[33] Microsoft, "Conceptos de minería de datos | Microsoft Docs.", 13-Jun-2018, Disponible en: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-analysis-services-2017

[34] Secretaria de Salud, "Cubos dinámicos". Fecha de consulta: junio 2018, Disponible en: http://www.dgis.salud.gob.mx/contenidos/basesdedatos/BD_Cubos_gobmx.html

[35] E. R. Nathan Landman, Hannah Pang, "K-Means Clustering | Brilliant Math Science Wiki.", Fecha de consulta: junio 2018, Disponible en: https://brilliant.org/wiki/K-Means-clustering/

[36] Rodolfo, J. Pérez-ortega, F. Miranda-henriques, and G. Reyes-salgado, "Spatial Data Mining of a Population-Based Data Warehouse of Cancer in Mexico," Int. J. Comb. Optim. Probl. Informatics, vol. 1, no. 1, pp. 61–67, 2010.

[37] R. Sanakal and S. T. Jayakumari, “Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machines,” Int. J. Comput. Trends Technol., vol. 11, no. 2, pp. 94–98, 2014.

[38] G. Singh and G. Singh, “Diabetes Classification Using K-Means,” Apeejay J. Comput. Sci. Appl., vol. ISSN, pp. 974–5742.

[39] Silvia Acid Carrillo, Nicolás Marín Ruiz, Juan Miguel Medina Rodríguez, Olga Pons Capote, Amparo Vila Miranda. ’05. Introducción a las bases de datos. El modelo relacional. Madrid, España: Paraninfo. 2005.

[40] Chapra, Steven y Canale, Raymond. “Métodos numéricos para ingenieros,” McGraw-Hill Interamericana de México, S.A. De C.V., vol. 4, 2003

[41] B. D. Crockett, R. Johnson, and B. Eliason, “What is Data Mining in Healthcare?,” Heal. Catal., pp. 1–13, 2014.

[42] Departamento de Sociología de la Universidad Complutense de Madrid, “Análisis de regresión lineal: El procedimiento Regresión lineal,” Guía para el análisis datos, pp. 67-95, 2013.

[43] A. Mexicano, “Desarrollo de una Metodología para la Selección de Atributos y Generación de Indicadores para la Aplicación de minería de datos a una Base de Datos Real de Registros de Cáncer de Base Poblacional,” pp. 111-123, 2007.

[44] J. Molina and J. García, “Técnicas de minería de datos basadas en Aprendizaje Automático,” Técnicas de Análisis de Datos, pp. 96–266, 2008.

[45] J. P. Ortega, M. Del Rocío Boone Rojas, M. J. S. García, and M. V. M. Hernández, “Data warehouse development to identify regions with high rates of cancer incidence in México through a spatial data mining clustering task,” CEUR Workshop Proc., vol. 686, pp. 37–47, 2010.

[46] E. Pérez, “Minería de datos Orientada al Big Data en el Área de Salud,” Centro Nacional de Investigación y Desarrollo Tecnológico”, pp. 104, 2016.

[47] L. Rincón, “Una introducción a la probabilidad y a la estadística”, pp. 3–132, 2006.

[48] J. Salinas, “Adecuación de una Metodología de minería de datos para su Aplicación a una Base de Datos real de registros de cáncer de base poblacional,” pp. 99-106, 2007.

[49] C. A. Vega, G. Rosano, J. M. López, J. L. Cendejas, and H. Ferreira, “Data Mining Aplicado a la Predicción y Tratamiento de Enfermedades,” CISCI, Conferencia Iberoam. en Sist. Cibernética e Informática, 2012.

[50] E. Vilches-González and I. A. Escobar-Broitman, “Minería de datos,” pp. 2–8, 2007.

[51] Microsoft, “Algoritmo de regresión lineal de Microsoft | Microsoft Docs.” Fecha de consulta: junio 2018, Disponible en: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/microsoft-linear-regression-algorithm?view=sql-analysis-services-2017.