SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Tecnológico Nacional de México
Coordinación Sectorial Académica
Dirección de Estudios de Posgrado e Investigación

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo Tecnológico

Subdirección Académica

Departamento de Ciencias Computacionales

TESIS DE MAESTRÍA EN CIENCIAS

Minería de Datos Orientada al Big Data en el Área de Salud

presentada por

Lic. Eduardo Pérez Luna

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis
Dr. Joaquín Pérez Ortega

Cuernavaca, Morelos, México. Febrero de 2016.

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Morelos a 29 de enero del 2016
OFICIO No. DCC/034/2016

Asunto: Aceptación de documento de tesis

C. DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **C. Eduardo Pérez Luna**, con número de control M13CE060, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado **"Minería de datos orientada a big-data en el área de salud"** 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. José Crispín Zavala Díaz
Doctor en Ciencias Computacionales
3406871

REVISOR 2

M.C. Humberto Hernández García
Maestro en Ciencias con Especialidad
en Sistemas Computacionales
7573641

REVISOR 3

Dra. Alicia Martínez Rebollar
Doctora en Informática
7399055

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

AMR/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 2 de febrero de 2016
OFICIO No. SAC/086/2016

Asunto: Autorización de impresión de tesis

C. EDUARDO PÉREZ LUNA
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 **"Minería de Datos Orientada a Big-Data en el Área de Salud"**, 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.

ATENTAMENTE
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO SUBDIRECCIÓN ACADEMICA

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

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

GVGR/mcr

Dedicatoria

A mi padre Efraín Pérez Ramírez† un hombre sabio que supo guiar a su familia por el camino del bien con sus acertados consejos y que donde quiera que se encuentre sigue, como en vida lo hizo, cuidando de los suyos.

A mi madre Eugenia Luna Ramírez por su apoyo incondicional, por sus valores, por la motivación constante que me ha permitido ser una persona de bien, pero sobre todo, por su amor.

A mis hermanos Edgardo, Teresa, Álvaro y Julio así como a mi tío Israel Pérez Ramírez por todo el apoyo mostrado durante esta etapa así como por la confianza depositada en mí.

A toda mi familia por el amor brindado, por su paciencia y sobre todo porque sin ustedes esto no hubiera sido posible.

Agradecimientos

Al Centro Nacional de Ciencia y Tecnología (CONACyT).

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por darme la oportunidad de continuar con mi formación profesional.

Mi más sincero respeto, agradecimiento y admiración al Doctor Joaquín Pérez Ortega por su valioso tiempo dedicado, su consejos, por aportar su vasta experiencia en la dirección de esta tesis y por ser además de un gran profesional una gran persona.

Al comité revisor Dra. Alicia Martínez Rebollar, Dr. José Crispín Zavala Díaz, M.C. Humberto Hernández García por su tiempo y disposición para contribuir y dar validez a este trabajo.

A mis amigos de generación, en especial a Sadher, Lupita, Yaír y Juan, con los cuales pase buenos momentos.

A todas las personas que han contribuido a lo largo de mi vida a mi formación personal y profesional.

Resumen

En el presente trabajo de investigación, se muestra que es factible el desarrollo de un prototipo de un sistema de Minería de Datos orientado a manejar grandes instancias como las que se presentan en el paradigma de Big Data en el dominio de la salud. En particular, el objetivo del prototipo es encontrar regiones del territorio mexicano y estadounidense con altas tasas de incidencia de mortalidad por diabetes, a partir de bases de datos poblacionales.

En esta investigación se propuso el uso del algoritmo N-Means para realizar la tarea de agrupamiento en el proceso de Minería de Datos. Para realizar las tareas de visualización se propuso un módulo cartográfico que hace uso de los mapas proporcionados por Google Maps, los cuales comprenden el territorio de México y de los Estados Unidos.

El prototipo se validó de manera sistemática con un conjunto de casos de prueba diseñado para tal fin. Como base para las pruebas se usaron los datos de mortalidad de los censos del año 2000 y 2010. Es destacable que el volumen de datos para la experimentación fue del orden de los tres gigabytes, con más de cuatro millones de registros.

Con base en las pruebas realizadas para la causa de mortalidad E11 (Diabetes mellitus no insulinodependiente) se observó que:

a) En varias de las delegaciones del Distrito Federal se encontraron los más altos índices de incidencia a nivel nacional para el año 2000 y 2010,
b) Contrastando los valores de los grupos con mayor incidencia en el año 2000 y 2010 se observó un incremento cercano al 100% en las incidencias de mortalidad.

Los resultados obtenidos para esta enfermedad hacen evidente la utilidad de las ciencias computacionales y en particular de la Minería de Datos en el área de salud, ya que proporcionan elementos de apoyo para la toma de decisiones de los funcionarios y autoridades encargadas de la salud de la población.

Abstract

In the present research is shown that the prototype development of a data mining system oriented to handle large instances of data such as those presented in the Big Data paradigm in the domain of health is feasible. Particularly, the prototype goal is to find regions of the Mexican and US territory with high incidence rates of mortality from diabetes, from population databases.

In this research, the use of N-Means algorithm it is proposed to perform the task of clustering in the data mining process. To perform the visualization tasks, a mapping module that uses maps provided by Google Maps, which comprise the territory of Mexico and the United States was proposed.

The prototype was validated in a systematic way with a set of test cases designed for this purpose. As a basis for testing, the mortality data from the 2000 and 2010 censuses were used. It is noteworthy that the volume of data for experimentation was around three gigabytes, with more than four million records.

Based on testing for the cause of death E11 (non-insulin dependent diabetes mellitus) it was observed that:

a) In several of the delegations of the Federal District the highest incidence rates nationwide for 2000 and 2010 were found,
b) Contrasting the values of the groups with the highest incidence in 2000 and 2010, an increase of nearly 100% was observed in the incidence of mortality.

The obtained results for this disease make clear the usefulness of computer science and in particular of data mining in the health area, since they provide elements of support for decision-making by officials and authorities responsible for the population health.

Contenido

Página

LISTA DE FIGURAS

Página

LISTA DE TABLAS

Página

Capítulo 1

Introducción

La Minería de Datos es un área interdisciplinaria, que tiene como objetivo descubrir relaciones ocultas entre los datos [HAND ET AL. 2001]. Estas relaciones se expresan como patrones de interés los cuales muestran el comportamiento de los datos en un determinado contexto. Con dichos patrones se genera un nuevo conocimiento que puede apoyar la toma de decisiones.

El presente trabajo forma parte de una línea de investigación en el área de Minería de Datos aplicada a bases de datos de epidemiología la cual se desarrolla en el Centro Nacional de Investigación y Desarrollo Tecnológico. Algunos trabajos dentro de esta línea de investigación se detallan en la parte de antecedentes (Sección 1.6.1).

En este capítulo se presenta el panorama general de la tesis. La motivación que impulsó este trabajo de investigación. Se presentan también los antecedentes así como los objetivos planteados, los alcances y las limitaciones de la investigación.

1.1. Contexto

Considerando el avance en las Tecnologías de Información y Comunicación (TIC's), en la actualidad se genera mayor información que antes. Esta información tiene dos fuentes de origen: a) los datos generados por humanos, como ejemplo están las redes sociales, correos electrónicos, documentos, fotos, entre otros, y b) los generados por dispositivos electrónicos como sensores, cámaras de vigilancia, dispositivos móviles, entre otros.

Este incremento en la generación de datos no es ajeno al sector salud. Por ejemplo, existen sistemas encargados de almacenar información correspondiente a historiales médicos. Este tipo de información se utiliza para obtener conocimiento que pueda ser útil en la toma de decisiones en organizaciones de este sector.

De acuerdo con datos de la Organización Mundial de la Salud (OMS), en el ámbito de la salud se cuenta con dos principales fuentes de información: a) los concernientes a la población, tales como el censo, las estadísticas vitales y las encuestas de hogares, y b) las que se vinculan a los servicios de salud y registros administrativos, como los sistemas de vigilancia, los registros de centros de salud y administrativos.

Actualmente, la Minería de Datos ha demostrado ser una actividad de interés en diferentes dominios, particularmente en el área de la salud, que es dominio al que va dirigida la investigación en este trabajo de tesis, esto porque permite la exploración de grandes volúmenes de datos a fin de extraer información previamente desconocida de manera implícita y que es potencialmente útil [WITTEN ET AL. 2011].

1.2. Justificación e impacto social

1.2.1. Justificación

Al trabajar con bases de datos reales, se aporta en la solución de problemas concretos. Esto contribuye con los expertos del tema en la prevención y estudio sobre el diagnóstico de enfermedades.

La infraestructura tecnológica actual permite generar información como nunca antes, parte de esta información ha servido para el desarrollo de la ciencia y la tecnología y por ello es de gran importancia desarrollar soluciones a nivel computacional que permitan continuar explotando en tiempo razonable toda esta información generada. Estas soluciones se pueden desarrollar en forma de métodos y herramientas computacionales que aporten soluciones innovadoras con lo que se permita contribuir a la solución de problemas dentro del dominio de la salud. En este sentido es que la presente tesis se espera que contribuya desde un enfoque computacional a resolver problemas reales del sector salud.

1.2.2. Impacto social

Según la Federación Internacional de la Diabetes (IDF por sus siglas en inglés) se calcula que en el mundo hay más de 415 millones de personas con este padecimiento y para el año 2040 esta cifra habrá aumentado hasta alcanzar los 642 millones de casos. En México existen 11.5 millones de personas que padecen esta enfermedad y se calcula que para 2040 habrá 20.6 millones de casos mientras que en Estados Unidos existen 29.3 millones de personas con este padecimiento [IDF 2015].

Este padecimiento es muy costoso, tanto así que en el Foro Económico Mundial 2015 se identificó a las enfermedades crónicas como uno de los factores que propician riesgos económicos [WEFORUM 2015]. El costo de este padecimiento fue de 673 mil millones de pesos en 2015 [IDF 2015].

Es de particular interés realizar Minería de Datos para diabetes ya que es una de las principales causas de muerte en México y Estados Unidos.

Los estudios epidemiológicos son de gran importancia para conocer el comportamiento de las enfermedades que aquejan a la población, por esto es importante tener infraestructura informática que permita continuar con estos estudios a pesar de la exponencial generación de información en este dominio.

1.3. Descripción del problema

Tanto las organizaciones públicas como las privadas del sector salud dentro del país generan muchos datos con respecto a sus pacientes. Por ejemplo diagnósticos, tratamientos, resultados de exámenes, datos de epidemiología poblacional, entre otros. Comprender, controlar y aprovechar toda esta información genera un nuevo reto para estas organizaciones. Los datos ahora son considerados como un activo empresarial y por ende se debe poner mayor énfasis en su tratamiento. Estas organizaciones parten del conocimiento generado por los datos que poseen dentro de ellas para la toma de decisiones, sin embargo, la generación de todo este cúmulo de datos hace casi imposible poderlos capturar, almacenar, procesar, compartir y visualizar en un tiempo razonable con los métodos tradicionales. Por todo esto, se requiere implementar nueva infraestructura tecnológica que permita realizar las actividades mencionadas.

En esta tesis se aborda el problema de procesar grandes instancias de datos poblacionales del área de epidemiología para lo cual se aplicarán técnicas de

Minería de Datos con el fin de encontrar patrones de interés en la toma de decisiones del sector salud.

1.4. Objetivo

Este trabajo tiene como propósito el desarrollo de un prototipo de un sistema de Minería de Datos orientado al sector salud que incorpore algunos principios del paradigma de Big Data.

1.5. Alcances y limitaciones

A continuación se mencionan los alcances y las limitaciones existentes para la realización de este trabajo.

1.5.1. Alcances

En el presente trabajo se plantearon los siguientes alcances:

a) La investigación estuvo orientada a la aplicación de Minería de Datos en bases de datos poblacionales obtenidas de forma gratuita.
b) Se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining).
c) El dominio fue el área de la salud, en particular el dominio de la epidemiología.

1.5.2. Limitaciones

Las limitaciones planteadas en esta tesis fueron las siguientes:

a) Se utilizó el algoritmo de agrupamiento K-Means en su versión mejorada desarrollada en una tesis previa del CENIDET.
b) El tipo de análisis realizado fue descriptivo.
c) Se utilizó software y hardware disponible en CENIDET.

1.6. Estado del arte

Los trabajos que se presentan a continuación se seleccionaron por su relación con el trabajo desarrollado en esta tesis.

1.6.1. Antecedentes

A continuación se describen de manera general los trabajos de investigación elaborados en CENIDET y que preceden a este trabajo de investigación.

- Tesis de maestría **"Metodología de Preparación de Datos Orientada a aplicaciones de Epidemiología Basada en el Modelo CRISP-DM"** [ITURBIDE, 2013] desarrollada por el M.C. Gregorio Emmanuel Iturbide Domínguez en el Centro Nacional de Investigación y Desarrollo Tecnológico, en el área de Ingeniería de Software del Departamento de Ciencias Computacionales, concluida en Febrero de 2013. En esta tesis se propone una metodología para la fase de preparación de datos, con un nivel de detalle mayor al propuesto en la metodología CRIPS-DM, la cual es factible de ser aplicada directamente a proyectos de Minería de Datos del dominio epidemiológico. La metodología propuesta fue validada mediante una aplicación en el área de epidemiologia con resultados satisfactorios, los cuales muestran que es factible para el dominio de la epidemiología desarrollar metodologías con un mayor nivel de detalle, las cuales puedan ser usadas en varias aplicaciones de dicho dominio. Estos resultados pueden ser utilizados en procesos de toma de decisiones en las

instituciones de salud de México en programas para el control y prevención de enfermedades.

- Tesis de maestría **"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"** [MEXICANO 2007] desarrollada por la M. C. Adriana Mexicano Santoyo en el Centro Nacional de Investigación y Desarrollo Tecnológico, en el área de Ingeniería de Software del Departamento de Ciencias Computacionales, concluida en Noviembre de 2007. En esta tesis se propone un nuevo enfoque en la selección de atributos que consiste en la integración de esquemas de bases de datos para conformar un meta-esquema al cual se le aplican técnicas de selección de atributos. El enfoque propuesto puede ser de utilidad en la aplicación de técnicas de Minería de Datos. Los resultados de los experimentos fueron satisfactorios ya que se encontraron patrones de interés en los datos estudiados.

- Tesis de maestría **"Incremento de la Eficiencia del Algoritmo K-Means Mediante la Mejora de la Heurística Early Classification"** [LÓPEZ 2015] desarrollada por el M. C. Vitervo López Caballero en el Centro Nacional de Investigación y Desarrollo Tecnológico, en el área de Ingeniería de Software del Departamento de Ciencias Computacionales, concluida en Marzo de 2015. En esta tesis se propone una nueva heurística denominada N-Means, la cual permite reducir de manera importante la complejidad del algoritmo K-Means en su etapa de clasificación. Para lograr el desarrollo de esta heurística, se propuso la integración de dos heurísticas: Early Classification y Grupos Estables obteniendo resultados alentadores.

1.6.2. Trabajos relacionados

En esta sección se presentan algunos trabajos relacionados con el presente trabajo de investigación. Es importante mencionar que estos trabajos fueron estudiados únicamente con el propósito de observar cómo se ha abordado el tema de la Minería de Datos en el sector de la salud implementando principios de Big Data, por lo anterior, el estudio detallado de las técnicas utilizadas no está dentro del alcance de la tesis.

- En el trabajo **"Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer"** [SALAZAR ET AL. 2011] se realiza un sistema clasificador de expresiones genéticas que puede predecir la recaída en la enfermedad en pacientes con cáncer colorrectal en sus etapas tempranas. Con esto se puede mejorar considerablemente la precisión en el diagnóstico de factores patológicos de pacientes con cáncer colorrectal en etapa II y III y facilita la identificación de pacientes en etapa II de la enfermedad quienes podrían ser tratados sin necesidad de una quimioterapia. Este trabajo fue realizado con grandes bases de datos obtenidas de diferentes fuentes como son: a) Netherlands Cancer Insitute, b) Leiden University Medical Center, c) Slotervaart General Hospital y del d) Instituto Catalán de Oncología.

- El segundo trabajo de investigación **"The WU-Minn Human Connectome Project: An Overview"** [VAN ESSEN ET AL. 2013] se realiza una revisión general a un sistema que estudia métodos de neuroimagen no invasiva, así también realiza actividades de análisis y visualización de la estructura del cerebro humano, la función y la conectividad en un detalle sin precedentes. Estos avances hacen que sea factible la exploración sistemática del conectoma humano, es decir, generar mapas de conectividad cerebral que sean integrales a la resolución espacial de los métodos de imagen disponibles. Los datos utilizados y generados no tienen un formato en

específico puesto que aparte de utilizar datos en formato de texto utilizan imágenes. La cantidad de datos utilizados asciende al rango de terabytes.

- En el siguiente trabajo de investigación **"Twitter Improves Seasonal Influenza Prediction"** [ACHREKAR ET AL. 2012] se describe un enfoque para lograr la detección y predicción casi en tiempo real de la aparición y propagación de la epidemia de la influenza a través de un seguimiento continuo de los tweets relacionados con la gripe, originados en Estados Unidos. Se demuestra que al aplicar un algoritmo de clasificación a los tweets relacionados con la gripe se aumenta significativamente la correlación entre el flujo de datos de twitter y el registro de visitas por enfermedades relacionadas con la influenza a los centros hospitalarios.

- En esta investigación **"Geographical Variability of the Incidence of Type 1 Diabetes in Subjects Younger than 30 Years in Catalonia, Spain"** [ABELLANA ET AL. 2009] se evalúa la variabilidad geográfica en la incidencia de diabetes de tipo 1 en personas menores de 30 años de edad, en Cataluña (España). Los datos se obtuvieron del registro prospectivo catalán de diabetes mellitus. Se utilizaron modelos mixtos lineales generalizados a fin de determinarlos efectos de los factores de riesgo y de conocer la distribución geográfica.

- Por último, en la siguiente investigación **"Application of Data Mining: Diabetes Health Care in Young and Old Patients"** [ALJUMAH ET AL. 2013] se realiza un análisis predictivo del tratamiento de diabetes usando técnicas de Minería de Datos basadas en Regresión. El *dataset* se analizó y estudió para identificar la efectividad de los diferentes tipos de tratamientos para los diferentes grupos de edad (jóvenes y ancianos). Se concluye que el tratamiento farmacológico para pacientes en el grupo de jóvenes se puede retrasar para evitar efectos secundarios, mientras que en el grupo de ancianos, el tratamiento debe ser inmediato.

1.7. Organización del documento

El presente documento está dividido en siete capítulos que se organizan de la siguiente manera:

Seguido del presente capítulo introductorio, en el segundo capítulo se presenta el marco conceptual mediante el cual se describen conceptos y términos básicos referentes a Minería de Datos, Big Data y Salud.

El tercer capítulo muestra la metodología seguida para la concepción de la investigación, se describen las actividades que conlleva el proceso de Minería de Datos.

El cuarto capítulo describe las actividades realizadas para la obtención de los datos así como las relacionadas con la preparación de estos.

El quinto capítulo describe el diseño y desarrollo del prototipo que muestra que es factible el uso de la metodología propuesta.

En el sexto capítulo se muestran los resultados experimentales del uso del prototipo con datos de mortalidad por cáncer y diabetes para los años 2000 y 2010.

En el séptimo capítulo se presentan las conclusiones a las que se llegó después de realizado el presente trabajo así como algunas aportaciones y los trabajos a los que da pauta esta investigación.

Finalmente en los anexos A y B se presentan los resultados del cáncer de pulmón y de estómago.

Capítulo 2

Marco conceptual

En este capítulo se presenta un marco de términos que han sido utilizados en el trabajo de investigación. Se describen conceptos relacionados con la Minería de Datos, así como conceptos relacionados con las grandes cantidades de datos, también se abordan conceptos sobre epidemiología. Todo esto con el fin de tener un panorama más claro de la terminología utilizada en el presente documento.

2.1. Base de Datos

Una base de datos es un conjunto de datos persistentes que es utilizado por los sistemas de aplicación de alguna empresa dada [DATE 2001].

El término empresa es un término genérico conveniente para identificar a cualquier organización independiente ya sea de tipo comercial, técnico, científico u otro. Una empresa puede ser un solo individuo, una corporación o un consorcio familiar.

Dado que la información es tan importante en las organizaciones, los científicos informáticos han desarrollado un amplio conjunto de conceptos y técnicas para la gestión de los datos.

2.2. Almacén de Datos

Un almacén de datos [HAN 2011] es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.
Se caracteriza por ser:

- **Integrado:** Los datos almacenados deben integrarse en una estructura consistente. Por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.
- **Temático:** Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes

pueden ser consolidados en una única tabla del almacén de datos. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

- **Histórico:** El tiempo es parte implícita de la información contenida en un almacén de datos. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el almacén de datos sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el almacén se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.
- **No volátil:** La información de un almacén de datos existe para ser leído mas no modificado. La información es por tanto permanente, la actualización del almacén de datos significa la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

2.3. Descubrimiento de Conocimiento en Bases de Datos

El Descubrimiento de Conocimiento en Bases de Datos (por sus siglas en ingles, Knowledge Discovery in Databases, KDD) es el proceso organizado de identificación válida, novedosa, útil y comprensible de patrones a partir de conjuntos de datos grandes y complejos [FAYYAD ET AL. 1996].

La Minería de Datos es el núcleo del proceso KDD, lo que implica la inferencia de algoritmos de exploración de datos, el desarrollo de un modelo y el descubrimiento de patrones previamente desconocidos.

La creciente accesibilidad y la abundancia de los datos hacen del Descubrimiento de Conocimiento en Bases de Datos y de la Minería de Datos una cuestión de considerable importancia y necesidad.

2.4. Minería de Datos

La Minería de Datos [FRAWLEY ET AL. 1991] algunas veces conocida como Descubrimiento de Conocimiento en Bases de Datos, es un proceso de extracción no trivial de información implícita, previamente desconocida y potencialmente útil de datos en bases de datos. En la literatura especializada existen otros términos para hacer referencia a la Minería de Datos, entre ellos destacan: Extracción de Conocimiento, Arqueología de Datos y Análisis de Datos, entre otros.

Según [FAYYAD ET AL. 1996] la Minería de Datos es la aplicación de algoritmos específicos para extraer patrones de datos. Produce información que puede ser clasificada en: asociaciones, secuencias, clasificaciones, agrupamientos y pronósticos.

Es difícil optar por una definición que ilustre por completo este concepto, es por eso que en [GORUNESCU 2011] se especifican una serie de definiciones que ayudan a tener un panorama más completo:

a) Búsqueda automática de patrones en grandes bases de datos, usando técnicas computacionales de estadística, máquinas de aprendizaje y reconocimiento de patrones;
b) Ciencia de extraer información útil de grandes bases de datos;
c) Exploración y análisis automático y semiautomático de grandes cantidades de datos, con el fin de descubrir patrones significativos;
d) Proceso automático de descubrimiento de información. La identificación de patrones y relaciones escondidas en los datos.

Esta extracción de conocimiento tiene diferentes aplicaciones como pueden ser el tratamiento de información, toma de decisiones, procesos de control, entre otras. Debido a esto, muchos investigadores han mostrado un gran interés en ello.

2.5. Minería de grandes datos

La generación de información por parte de las grandes empresas tanto comerciales como de servicio, al día de hoy producen cantidades exponencialmente más grandes de datos que nunca, de manera que dichas organizaciones se replanteen cómo digerir esos datos. Y existen para ello algoritmos y técnicas analíticas avanzadas que pueden ser aprovechadas para descubrir patrones ocultos, y utilizar el conocimiento obtenido para lograr una ventaja competitiva sobre los competidores [KUAN-CHING ET AL. 2015].

La minería de grandes datos [WEI Y BIFET 2012] es la capacidad de extraer información de utilidad de grandes conjuntos de datos que debido a su volumen, variedad y velocidad no había sido posible realizarlo anteriormente.

Esta tendencia ha demostrado ser de mucho interés en la comunidad dedicada a la Minería de Datos, tanto así que KDD (la mayor comunidad de Minería de Datos) dedicó su conferencia anual en 2012 a este tema denominándola “Mining de Big Data”. Incluso existe un taller específico “BigMine 2015” perteneciente también a esta comunidad.

Por su parte existen retos que se deben afrontar en la generación de concomimiento en este tipo de paradigma, dentro de estos destacan:

a) Manejar diferentes tipos de datos: Esto debido a que las bases de datos de la actualidad contienen datos complejos como hipertexto, multimedia, datos espaciales, entre otros.
b) Eficiencia y escalabilidad de los algoritmos de Minería de Datos: Muchos algoritmos ven mermado su funcionamiento ante grandes cantidades de datos.
c) Protección de la privacidad y seguridad de los datos: Las fuentes de información son diversas y es posible que se esté invadiendo la privacidad de las personas.

2.6. Big Data

Big Data [WEI Y BIFET 2012] es el término que se emplea hoy en día para describir el conjunto de datos que debido a su gran tamaño y complejidad, no se puede manipular con las metodologías o sistemas de Minería de Datos tradicionales.

Según [GARTNET 2015], Big Data es la información que se caracteriza por ser de gran volumen, generarse a gran velocidad y/o ser de gran variedad, lo cual demanda formas innovadoras y rentables de procesamiento de la información que permite una visión mejorada, la toma de decisiones y la automatización de procesos.

Doug Laney [LANEY 2001] especifica tres V para definir Big Data:

Volumen: Muchos factores contribuyen para incrementar el volumen de datos. Ejemplo de éstos son: el almacenamiento de los datos que genera una transacción bancaria, los datos no estructurados que entran desde los medios de comunicación social y los provenientes de sensores.

Velocidad: Los datos son transmitidos a una velocidad sin precedentes y deben ser procesados de manera oportuna. Reaccionar con la rapidez suficiente para hacer frente a la velocidad de los datos es un reto para la mayoría de las organizaciones.

Variedad: Actualmente, los datos son generados en diversos formatos. Estructurados: datos numéricos en bases de datos tradicionales. No estructurados: documentos, correo electrónico, video, audio, transacciones financieras y demás. La gestión y fusión de los diferentes tipos de datos es algo con lo que aún lidian las organizaciones.

2.7. Mortalidad

Los datos de mortalidad [OMS 2015] indican el número de defunciones por lugar, intervalo de tiempo y causa. Los datos de mortalidad de la OMS reflejan las defunciones recogidas en los sistemas nacionales de registro civil, con las causas básicas de defunción codificadas por las autoridades nacionales.

La causa básica de defunción se define como "la enfermedad o lesión que desencadenó la sucesión de eventos patológicos que condujeron directamente a la muerte, o las circunstancias del accidente o acto de violencia que produjeron la lesión mortal", según lo expuesto en la Clasificación Internacional de Enfermedades [CEMECE].

Las estadísticas de mortalidad proporcionan información acerca de defunciones generales y fetales que permiten caracterizar el fenómeno de la mortalidad en el país. Se obtiene de los registros civiles y, en el caso de las defunciones accidentales y violentas, de los registros de las agencias del Ministerio Público [INEGI 2015].

2.8. Epidemiología

La epidemiología [LOPEZ 2000] es la rama de la salud pública que tiene como propósito describir y explicar la dinámica de la salud poblacional, identificar los elementos que la componen y comprender las fuerzas que la gobiernan, a fin de intervenir en el curso de su desarrollo natural. Actualmente, se acepta que para cumplir con su cometido la epidemiología investiga la distribución, frecuencia y determinantes de las condiciones de salud en las poblaciones humanas así como las modalidades y el impacto de las respuestas sociales instauradas para atenderlas.

Estudia la frecuencia y distribución de los problemas de salud y sus determinantes en las poblaciones humanas con el fin de controlarlos [HERNANDEZ 2005].

Capítulo 3

Metodología

El presente capítulo tiene como objetivo explicar la metodología que se siguió para la realización de este trabajo de investigación mediante la explicación detallada de cada uno de los pasos seguidos.

La calidad del conocimiento que va a ser descubierto, depende de la preparación que se haga de los datos, si los datos son de calidad es posible la generación de patrones y reglas de calidad. Sobre la preparación de los datos, algunos autores refieren que el proceso que toma dicha actividad, oscila entre el 50 y el 70% [CHAPMAN ET AL. 2000].

Existen diversos métodos para llevar a cabo un proyecto de Minería de Datos. Dentro de los ambientes académico e industrial el que más se utiliza [KDNUGGETS] es el estándar CRISP-DM (Cross-Industry Standard Process for Data Mining) [CHAPMAN ET AL. 2000], el cual se utilizará para la realización de este proyecto.

Las actividades llevadas a cabo, se ilustran en la Figura 3.1.

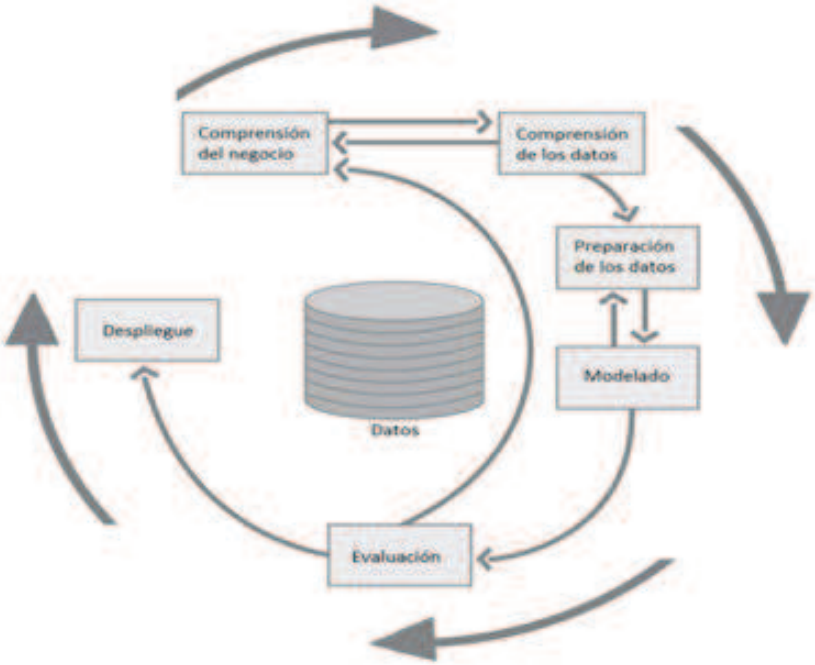

Figura 3.1 Estándar CRISP-DM

A continuación se describen las fases de la metodología a utilizar.

3.1. Comprensión del negocio

Esta fase inicial se enfoca en el entendimiento de los objetivos y requerimientos del proyecto desde una perspectiva del negocio. Después, esta información se convierte en un problema de Minería de Datos y se diseña un plan para lograr los objetivos. Figura 3.2.

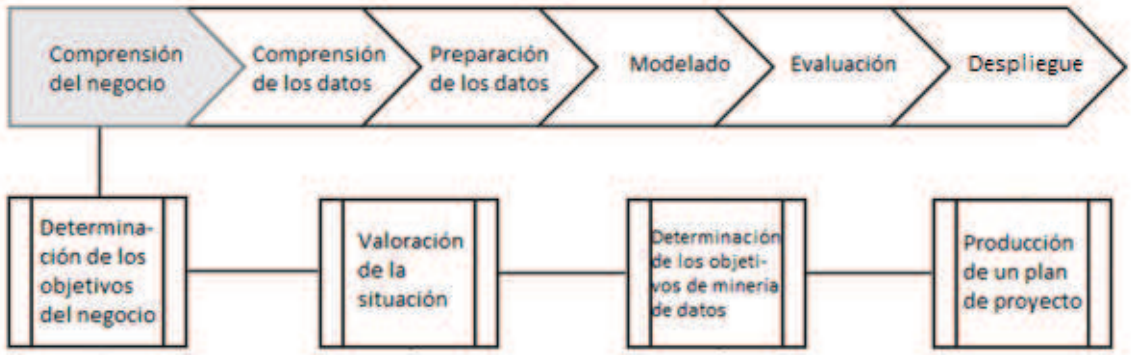

Figura 3.2 Fase 1 del estándar CRISP-DM Comprensión del negocio

3.2. Comprensión de los datos

Esta fase comienza con una colección de datos y continúa con actividades que permiten familiarizarse con los datos, identificar problemas en la calidad de los datos y/o detectar subconjuntos para formar hipótesis sobre la información oculta en estos datos. Figura 3.3.

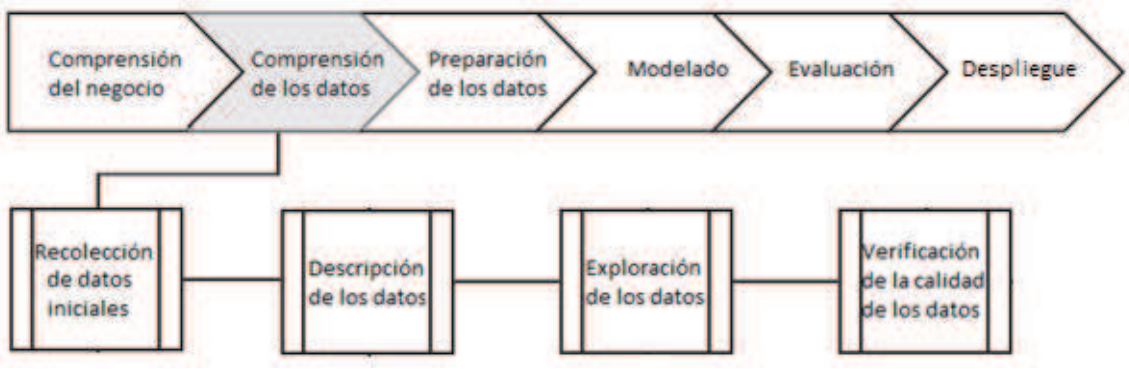

Figura 3.3 Fase 2 del estándar CRISP-DM Comprensión de los datos

3.3. Preparación de los datos

Esta fase cubre las actividades necesarias para la construcción del 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 sin algún orden prescrito. Las tareas incluyen la selección de tablas, columnas, registros y atributos, así como la transformación y limpieza de los datos mediante las herramientas de modelado. Figura 3.4.

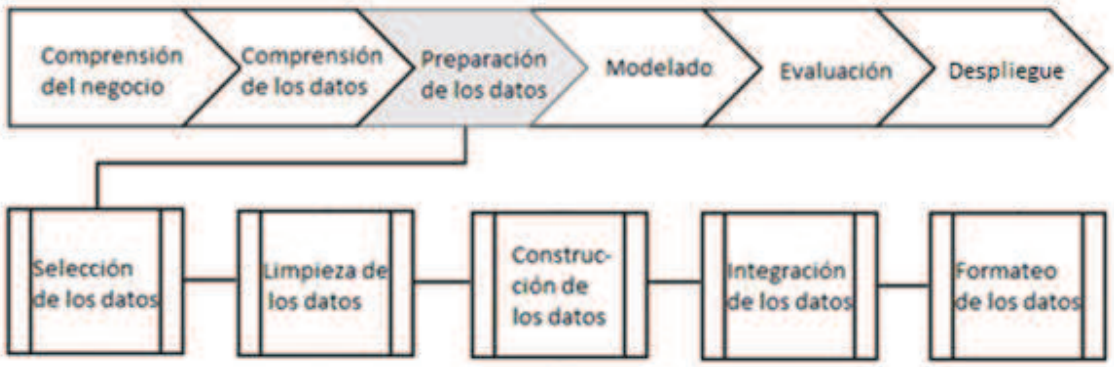

Figura 3.4 Fase 3 del estándar CRISP-DM Preparación de los datos

3.4. Modelado

En esta fase, se seleccionan y aplican varias técnicas de modelado. Normalmente existen diferentes técnicas para el mismo tipo de problema de Minería de Datos, algunas de éstas necesitan requerimientos específicos en la forma de los datos, por tanto, en ocasiones se necesita regresar a la preparación de los datos. Figura 3.5.

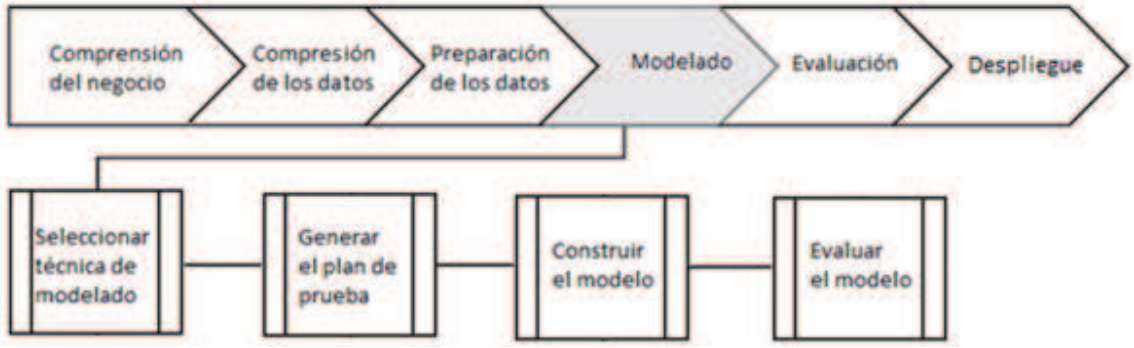

Figura 3.5 Fase 4 del estándar CRISP-DM Modelado

3.5. Evaluación

Antes de proceder al despliegue final del modelo, es importante evaluar a fondo y revisar los pasos ejecutados en su creación, para asegurar éste cumple con los objetivos del negocio. Un objetivo clave es determinar si hay un aspecto de negocio importante que no se consideró con suficiencia. Al final de esta fase se deberá tomar una decisión sobre el uso de los resultados. Figura 3.6.

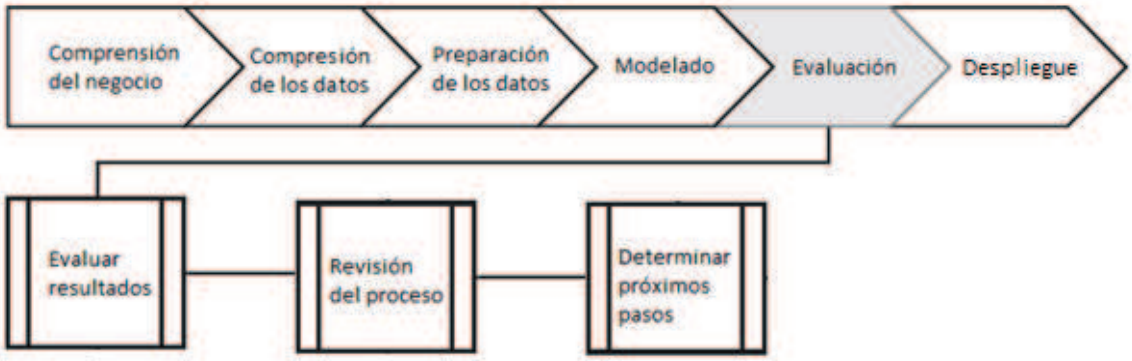

Figura 3.6 Fase 5 del estándar CRISP-DM Evaluación

3.6. Despliegue

La creación del modelo no es la parte final del proyecto. Aunque el propósito del modelo es incrementar el conocimiento a partir de los datos, la información obtenida se tendrá que organizar y presentar de una forma que el cliente pueda utilizar este conocimiento. Figura 3.7.

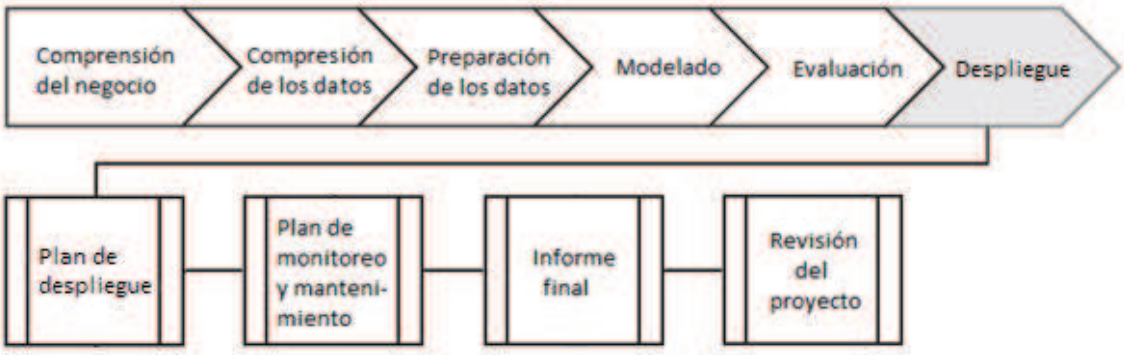

Figura 3.7 Fase 6 del estándar CRISP-DM Implantación

Capítulo 4

Obtención y preparación de los datos

En este capítulo se detalla la diversidad de fuentes de la cuales se obtuvieron los datos de estudio así también se describen las actividades realizadas durante el proceso de preparación de datos.

4.1. Obtención

La obtención de los datos para el presente estudio se llevó a cabo por medio de diferentes fuentes las cuales se detallan en el siguiente punto, posteriormente se describen los datos obtenidos.

4.1.1. Fuentes oficiales

Se estableció como objetivo de la Minería de Datos encontrar altas tasas de mortalidad tanto para diabetes como para cáncer de pulmón y estómago en México y en Estados Unidos de América, por lo que se optó por consultar instituciones que pudieran otorgar estos datos obteniéndolos de las siguientes fuentes:

	**INEGI:** Instituto Nacional de Estadística y Geografía. Se obtuvieron registros de la población total por municipio en México para los años 2000 y 2010 [INEGI].
	**SIMBAD:** Sistema Municipal de Bases de Datos. Se obtuvieron registros de la ubicación geográfica de los municipios de México [SIMBAD].
	**SINAIS:** Sistema Nacional de Información de Salud. Se obtuvieron registros de las defunciones por diferentes causas de muerte ocurridas en México en los años 2000 y 2010 [SINAIS].
	**United States Census Bureau.** Se obtuvieron registros de la población total y de la ubicación geográfica de los condados de Estados Unidos de América [CENSUS].

	**CEMECE:** Centro Mexicano para la Clasificación de Enfermedades y Centro Colaborador para la Familia de Clasificaciones Internacionales de la OMS en México. Se obtuvieron registros de la Clasificación de las enfermedades, causas externas de daños y circunstancias sociales de mortalidad [CEMECE].

4.1.2. Descripción de los datos

Comprender los datos obtenidos es de vital importancia puesto que de ello depende el buen curso de la preparación de los datos y por ende de todo el proceso de Minería de Datos. Como actividad inicial en la comprensión de los datos se deben estudiar las características de las bases de datos obtenidas.

La Tabla 4.1 muestra estas características, las cuales están orientadas al tamaño y volumen de los datos, el año al que corresponden en dado caso de que esto aplique y el formato en que fueron conseguidos.

Tabla 4.1 Características de las bases de datos utilizadas

Fuente	Tipo de datos		Ámbito	N° de Registros	N° de Atributos	Formato de archivo
INEGI	Población	2000	México	2,475	3	XLS
		2010				
CENSUS	Población	2000	EEUU	3,219	4	TXT
		2010			12	
SIMBAD	Geográfica		México	2,475	11	XLS
CENSUS	Geográfica		EEUU	3,219	12	TXT
SINAIS	Mortalidad	2000	México	437,667	38	DBF
		2010		592,018	40	
		2010		1,201,039		

CEMECE	Catálogo	Intl.	14,259	24	XLS

Una vez realizada la recolección de las bases de datos y teniendo comprendida su descripción, se procedió a realizar una inspección minuciosa de los datos con la finalidad de comprender su estructura, significado, rango, el tipo de dato que maneja (numérico, booleano, entre otros), etc. Con esto se determina qué actividades se llevarán a cabo en la fase de preparación de datos.

Los archivos de datos usualmente vienen con un archivo que los describe. Estos archivos detallan los atributos de los datos, así como sus valores y rangos. La Tabla 4.2 ilustra como son descritos los atributos de los datos de mortalidad.

Tabla 4.2 Ilustración de la descripción de los atributos de la bases de datos de mortalidad

Variable	**Tipo**	**Etiqueta**	**Longitud**	**Rango**	**Valores**
Entres	Numérico	Entidad de residencia	2	1 a 32	Aguascalientes a Zacatecas
				33	Estados Unidos de Norteamérica
				34	Otros países Latinoamericanos
				35	Otros países
Mpores	Numérico	Municipio de residencia	3	1 a 570	Municipios según entidad
Entdef	Numérico	Entidad de defunción	2	1 a 32	Aguascalientes a Zacatecas
Mpodef	Numérico	Municipio de defunción	3	1 a 570	Municipios según entidad
Causa	Cadena	Causa de defunción	4	De acuerdo a la clasificación internacional de enfermedades (CIE-10)	

4.2. Preparación de los datos

Gracias al entendimiento de los datos, su preparación se hace menos compleja puesto que ahora se tiene el conocimiento necesario para identificar y corregir posibles anomalías. Este conocimiento también es útil para determinar los atributos que representen mayor interés y sean necesarios para lograr el objetivo de la Minería de Datos.

En esta etapa se incluyen tareas de limpieza y selección de los datos, así como la generación de nuevas variables y la integración de los diferentes orígenes de los datos.

La preparación de los datos se realizó en dos partes, la primera parte fue un proceso manual que se describe a continuación y la segunda un proceso automático que se aborda más adelante.

Para la preparación de datos manual se utilizó Microsoft Office Excel para abrir y manipular la información contenida.

Para los datos de mortalidad y población, la preparación de datos se tuvo que realizar dos veces puesto que para cada año (2000 y 2010) los datos estaban contenidos en archivos diferentes. Cabe mencionar también que estos datos varían en el número de defunciones y de habitantes para cada año. En el caso de los datos geográficos la preparación se realizó una sola vez puesto que la ubicación y el número de municipios no varían en ambos años. Así también para el catálogo de enfermedades.

A continuación se describen las actividades realizadas para la preparación de los datos:

4.2.1. Limpieza

En esta etapa, las actividades consisten en: a) detectar, b) corregir o en su defecto c) eliminar aquellos registros y/o atributos que representen valores fuera contexto o que no coincidan con los valores de los demás registros. Cabe recordar que en esta etapa ya se tiene información necesaria para generar un criterio de qué tipo de datos se consideran anómalos.

En la revisión de los archivos se detectaron diversos errores, algunos de ellos representaban un obstáculo para lograr el objetivo que se planteó y fueron corregidos o eliminados. Existieron también anomalías que no representaban problema alguno y no hubo necesidad de corregir.

A continuación se describen las anomalías presentadas en los archivos:

a) Para la base de datos de mortalidad en México para el año 2000 se encontró que la ausencia de valor en la columna CVE_JUR (clave de la jurisdicción) en algunos registros produjo un error en los atributos localizados a la derecha de este. Estos registros presentaron un corrimiento a la izquierda por lo cual se tuvieron que desplazar los valores de los atributos para los registros que presentaron este problema. Puesto que la jurisdicción a la que pertenece cada registro no es de importancia, se optó por dejar en blanco los registros faltantes.
b) Para la base de datos de mortalidad en México para el año 2010 se encontró un valor no especificado para el atributo SEXO, este atributo no es de importancia y se optó por no corregirlo o eliminarlo.
c) En la base de datos geográfica para México se encontró una ausencia de valores para los estados de Quintana Roo, Campeche y Yucatán. Para corregir esta anomalía se tuvo que requerir esta información de los Anuarios Estadísticos de los Estados (AEE) del INEGI para obtener la información correspondiente. Una vez corregidos estos registros no se realizó eliminación alguna.

d) En cuanto a la información poblacional de México tanto para 2000 como para 2010 se encontró un encabezado que no representaba información útil por tanto fue eliminada. Además para la base de datos de 2010 se encontró una nota al pie de la página que al igual fue eliminada. En la información correspondiente al Catalogo Internacional de Enfermedades se localizó de igual manera un encabezado que al igual fue eliminado.
e) Las bases de datos, poblacional y geográfica correspondientes a Estados Unidos estuvieron libres de anomalías tanto en 2000 como en 2010.

4.2.2. Selección

Una vez que la información está libre de anomalías, se procede a realizar la selección de los atributos que representen información de utilidad. Esta actividad se compone de dos tareas: a) selección horizontal y b) selección vertical de datos.

La selección horizontal consiste en elegir los registros de interés, para esto se realizaron las siguientes acciones:

a) Para las bases de datos poblacionales de México se eliminaron los registros que representaban a las entidades federativas, esto debido a que únicamente son de interés los registros con información a nivel municipal. Se eliminaron también aquellos registros cuya población fuera menor a 100,000 habitantes quedando, de los 2475 registros, únicamente 168 para 2000 y 204 para 2010.
b) En el caso de Estados Unidos, se eliminaron de igual manera los registros cuya población fuera menor a 100,000 habitantes, así como los registros que representan información de Puerto Rico, Alaska y Hawái, que debido a su ubicación geográfica distante, significan valores atípicos en los datos. Como resultado, para el año 2000 resultaron 520 registros y para el año 2010, 574 registros.

c) Un escenario probable es que un municipio tenga en el año 2000 más de 100,000 habitantes y en el 2010 el número de habitantes sea menor. En una revisión a los datos de población de México se pudo establecer que esto no sucedía por lo que la selección de los registros en el archivo geográfico se realizó tomando como referencia los 204 municipios con más de 100,000 habitantes identificados para el año 2010. Este escenario aconteció con los datos de Estados Unidos, Portsmouth en Virginia y Cape May en Nueva Jersey presentaron este comportamiento, entonces, para los datos geográficos en Estados Unidos se tomaron como referencia los 574 registros obtenidos en los datos de población de 2010 y se agregaron los dos registros mencionados para un total de 576 registros.
d) Para el caso de las bases de datos de mortalidad en México, se eliminaron registros cuyo valor en el atributo ANODEF (año de defunción) no correspondiera a 2000 o 2010, esto significa que las defunciones no ocurrieron en 2000 ni 2010. Otro aspecto que se observó fue la existencia de registros cuyo valor en el atributo ENTRES (entidad de residencia) no correspondía a algún código para alguno de los estados de la república, es decir, que las defunciones no habían ocurrido en territorio mexicano. Dado este escenario, los registros se eliminaban.
e) En el archivo de clasificación de enfermedades, los registros cuya CAUSA excede los tres dígitos fueron excluidos, esto debido a que el cuarto dígito representa la ubicación específica de la enfermedad y para esta investigación no representa interés alguno.

Una vez determinados los registros que representaban información útil, se inició con la selección vertical, esta selección consiste en elegir los atributos que representen información de interés. Las actividades realizadas son las siguientes:

a) Para las bases de datos poblacionales de México en 2000 y 2010 no se realizó modificación alguna puesto que sus tres atributos (código, nombre y población total del municipio) son de interés. Para los datos de Estados Unidos en el 2000 se mantuvieron los cuatro atributos (código, nombre, año

y total de la población), mientras que para el 2010 se seleccionaron tres atributos (código, nombre y total de la población), el año no está presente para esta base de datos.

b) En la base de datos geográfica de México se consideraron como no útiles los atributos MUNICIPIO, CABECERA Y ALTITUD. Para Estados Unidos se consideraron útiles únicamente los atributos GEOID (identificador geográfico), INTPTLAT (latitud) e INTPTLONG (longitud).
c) De las bases de datos de mortalidad, se eliminaron los atributos que contenían información constante en todos los registros, por ejemplo: NACION. También se eliminó el atributo CONTROL el cual servía como clave primaria y por ende no resultaba útil. Otros atributos eliminados que no proporcionaban información de interés son: DIA_DEF, MES_DEF, DIA_NAC, MES_NAC Y TLOC (día y mes de defunción, día y mes de nacimiento y tamaño de la localidad respectivamente).
d) Del conjunto de datos resultante para 2000 y 2010 se eligieron los atributos de interés: ENT_RES (entidad de residencia), MPO_RES (municipio de residencia), GÉNERO, CAUSA, E_CIVIL, ENT_DEF (entidad de defunción), MUN_DEF (municipio de defunción), ESCO (escolaridad), OCUPA (ocupación), y EDAD.
e) Para los datos del CIE-10, únicamente se identificaron como atributos de interés CAUSA y NOMBRE por lo que los demás fueron desechados.

4.2.3. Formateo

En esta etapa se realizaron actividades enfocadas a estandarizar el formato de los archivos. A continuación se describen las actividades realizadas:

a) Algunos datos estaban almacenados en un formato distinto al que corresponde por lo que se tuvo que realizar la estandarización correspondiente. Por ejemplo: los datos de población en México con

información de EDAD, GÉNERO, EDO_CIVIL, presentaban información numérica pero el formato en que se almacenaron corresponde a una cadena de caracteres por lo que existió la necesidad de estandarizarlo al formato que se requiere para el estudio.

b) En el apartado 4.2.2 se menciona que en el catálogo de enfermedades fueron eliminados los registros que tuvieran más de tres caracteres en su atributo CAUSA, puesto que no representan información de interés. Esto conduce a realizar una modificación sobre el atributo CAUSA en los datos de mortalidad. La modificación consiste en reducir la longitud de los caracteres a tres, quedando así, por ejemplo, la causa E147 y E148 como E14.

c) Los valores de longitud en la base de datos geográfica de Estados Unidos estaban almacenados como números negativos ya que corresponden a valores de longitud oeste, para ser utilizados por el prototipo es necesario transformarlos a valores positivos, para esto únicamente se multiplicaron los valores de la longitud por -1.

d) Otra modificación a los datos geográficos consistió en separar los atributos: STATE ABBREVIATION (abreviación del nombre del estado), FIPS STATE CODE (código del estado), FIPS COUNTY CODE (código del condado) y NAME (nombre del condado); esto debido a que en el archivo de datos estaba toda esta información contenida en una sola columna, el resultado se muestra en la Figura 4.1.

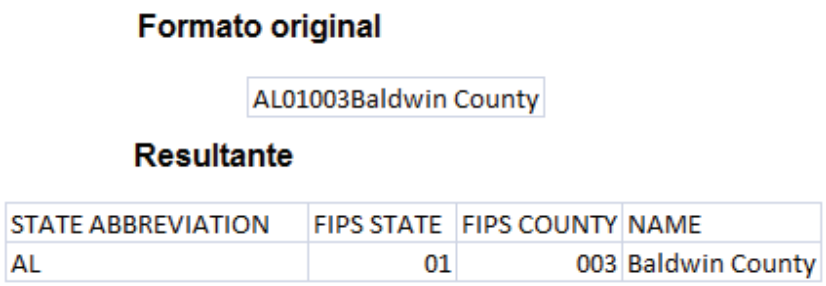

Figura 4.1 Formateo de los valores en registros geográficos

4.2.4. Construcción

Existen atributos que por sí solos no representan información de utilidad, es por ello que se realiza una nueva revisión a los datos con el fin de determinar qué información pueden servir para crear nuevos atributos y que representen información útil.

Esta etapa se realizó en dos fases: 1) construcción de datos manual y 2) construcción de datos automática, estas fases se describen a continuación.

1) Construcción de datos manual:

A continuación se detallan las actividades realizadas en esta fase:

a) En la base de datos poblacional de México se agregó el atributo AÑO para diferenciar los datos según su fecha de origen, mientras que para Estados Unidos se creó únicamente este atributo para los datos de 2010 ya que para 2000 ya se contaba con este atributo.
b) Para los datos de mortalidad tanto para 2000 como 2010 se realizó una concatenación de ENT_RES (entidad de residencia) y MPO_RES (municipio de residencia) para generar el atributo CLAVE, de igual forma se generó LUG_DEF (lugar de defunción) a partir de ENT_DEF y MPO_DEF (entidad y municipio de residencia).
c) También se creó el atributo CLAVE para los datos geográficos de Estados Unidos. Para los datos geográficos de México se realizó la misma actividad.
d) Otra actividad fue la conversión de grados sexagesimales a decimales tanto para valores de longitud como de latitud como se muestra en la siguiente expresión:

$$\text{Grados decimales} = \text{Grados} + (\text{Minutos}/60) + (\text{Segundos}/3600) \qquad (1)$$

La Expresión 1 muestra las operaciones realizadas para obtener los grados decimales, este nuevo formato es necesario para poder visualizar las coordenadas como puntos en un plano, lo que será importante en fase de visualización. Para almacenar estos resultados se crearon dos nuevos atributos denominados LAT_DECIMAL (latitud decimal) y LONG_DECIMAL (longitud decimal) dentro de la base de datos geográfica.

e) Para evitar que una variable domine a otra, los valores de ambos atributos son normalizados mediante:

$$VN = \frac{va - vm}{vM - vm} * 10 \qquad (2)$$

La Expresión 2 muestra las operaciones necesarias para la normalización de los registros. Donde: 'VN' se refiere al valor normalizado resultante, 'va' indica el valor actual que se va a normalizar, 'vm' y 'vM' muestran el valor mínimo y máximo respectivamente. El proceso de normalización fue realizado para los datos geográficos tanto de México como de Estados Unidos.

2) Construcción de datos automática:

La construcción de datos automática se compone de tres actividades: el cálculo de la incidencia, de la tasa de mortalidad y de la tasa de mortalidad normalizada. En esta fase se utilizó el prototipo de preparación de datos desarrollado por [ITURBIDE 2013]. Las actividades realizadas en esta fase se detallan a continuación:

a) En la base de datos de mortalidad se realizó el cálculo de la incidencia como se muestra a continuación:

Incidencia = Sumatoria de las defunciones por municipio en un año (3)

Como se muestra en la Expresión 3, la incidencia se crea a partir de la sumatoria de las defunciones en un municipio por año para una determinada causa. Para estos cálculos se utilizan los atributos CLAVE y CAUSA de los datos de mortalidad.

b) Una vez obtenidos los registros de incidencia, se calcula la tasa de mortalidad de una población para una determinada causa. Para calcular la tasa de mortalidad se llevaron a cabo las siguientes operaciones:

Tasa de Mortalidad = (Incidencia/Población)*100,000 (4)

La Expresión 4 muestra que para calcular la tasa de mortalidad de cada municipio es necesaria la incidencia calculada anteriormente, así como el número de habitantes de un municipio. Esto representa el número de defunciones en un lugar y tiempo determinados por cada 100,000 habitantes.

c) Para la normalización de la tasa de mortalidad se realizaron las operaciones descritas en la Expresión 2.

4.2.5. Integración

Como último paso en la preparación de los datos se realizó la integración. Esta actividad se realizó utilizando el prototipo de Minería de Datos de [ITURBIDE 2013]. Este prototipo utiliza el atributo CLAVE para establecer una conexión entre los datos geográficos y poblacionales, el atributo CAUSA para establecer una conexión entre los datos del catálogo de enfermedades y los datos de mortalidad. También se utilizan los atributos AÑO, INCIDENCIA, TASAMORTALIDAD,

TASAMORTALIDADNORMALIZADA. Todos estos registros se almacenan en HECHOS.

En la Figura 4.2 se muestra la estructura del almacén desarrollado.

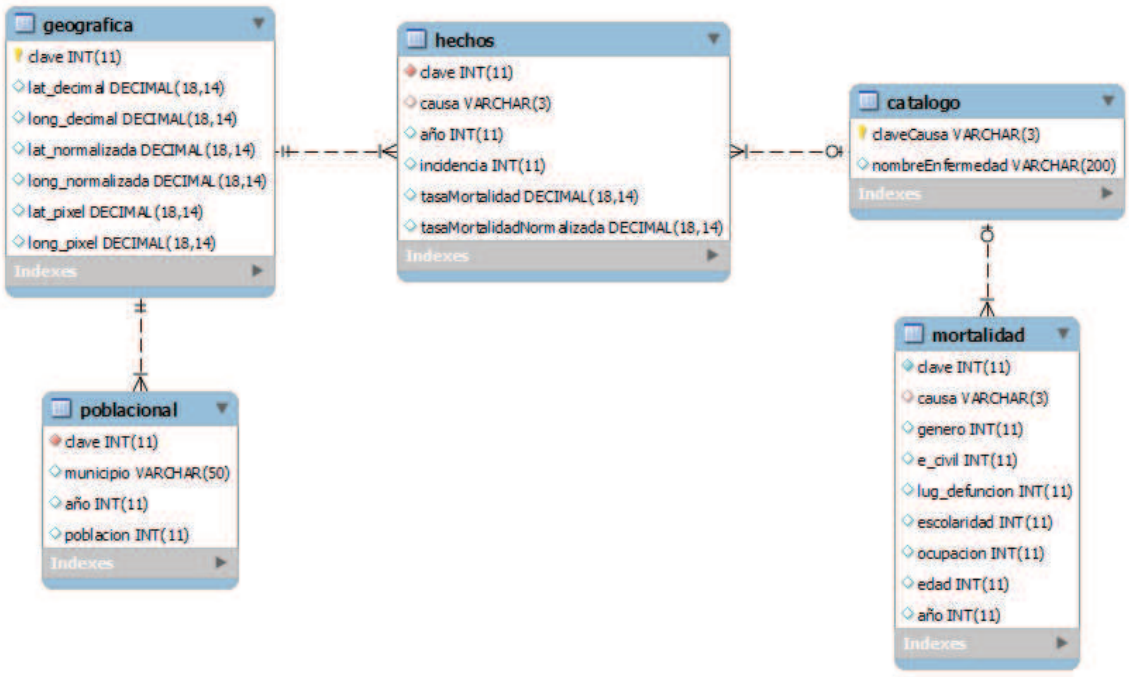

Figura 4.2 Esquema del almacén de datos

Teniendo el esquema del almacén de datos, se crearon las tablas correspondientes y se poblaron con los datos resultantes de la preparación. En la tabla hechos se almacenan registros generados durante la ejecución del prototipo de [ITURBIDE 2013]. Como resultado de la ejecución de este prototipo, se obtiene un conjunto de datos compuesto por los siguientes atributos:

Tabla 4.3 Descripción del conjunto de datos final

Atributo	**Descripción**
CAUSA	Causa de muerte
LATITUD_NORM	Latitud del municipio
LONGITUD_NORM	Longitud del municipio
TASAM_NORM	Tasa de mortalidad del municipio

La Tabla 4.3 muestra la estructura de este conjunto de datos, el cual está integrado por un total de 688 registros para el año 2000 y de 778 registros para el año 2010. Este conjunto de datos es utilizado por el prototipo de Minería de Datos desarrollado en esta investigación.

En la Figura 4.3 se muestra de manera resumida el proceso de preparación de los datos, como puede observarse, este proceso estuvo dividido en dos partes generales: a) la preparación de datos manual, y b) la preparación de datos automática. El prototipo de [ITURBIDE 2013] se utiliza para generar la incidencia y la tasa de mortalidad que se almacenan en la tabla HECHOS, así también para generar el conjunto de datos final que será utilizado en el prototipo desarrollado en la presente investigación.

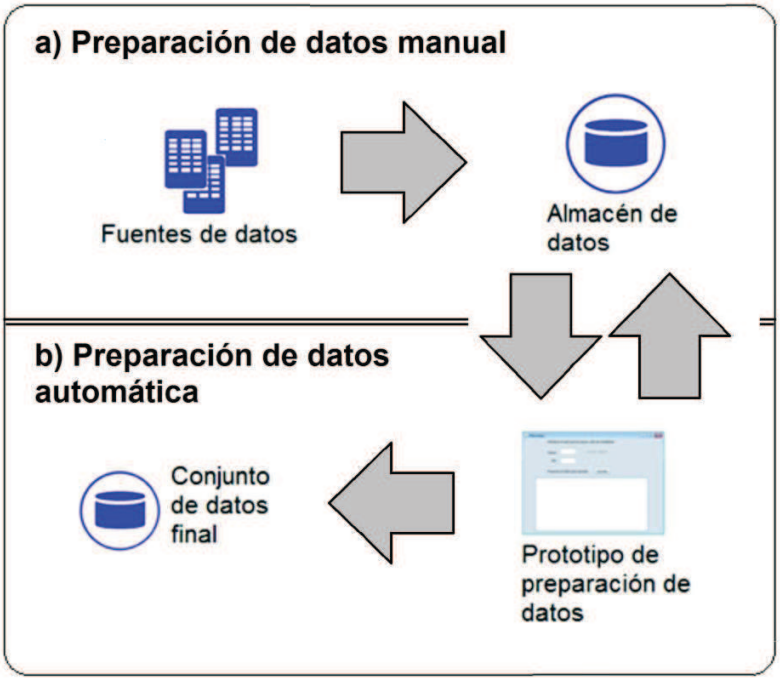

Figura 4.3 Proceso de preparación de datos

Capítulo 5

Diseño y desarrollo del prototipo

Este capítulo presenta el diseño y se detalla el proceso de desarrollo del prototipo realizado en esta investigación así como las características generales de su implementación.

5.1. Representación general del prototipo

En la Figura 5.1 se muestra de forma general el funcionamiento del prototipo, éste inicia con la selección de las características necesarias (causa, año y número de grupos) para ejecutar el algoritmo de agrupamiento. A partir de la ejecución del algoritmo se obtiene un modelo el cual es representado mediante un mapa y mostrado al usuario.

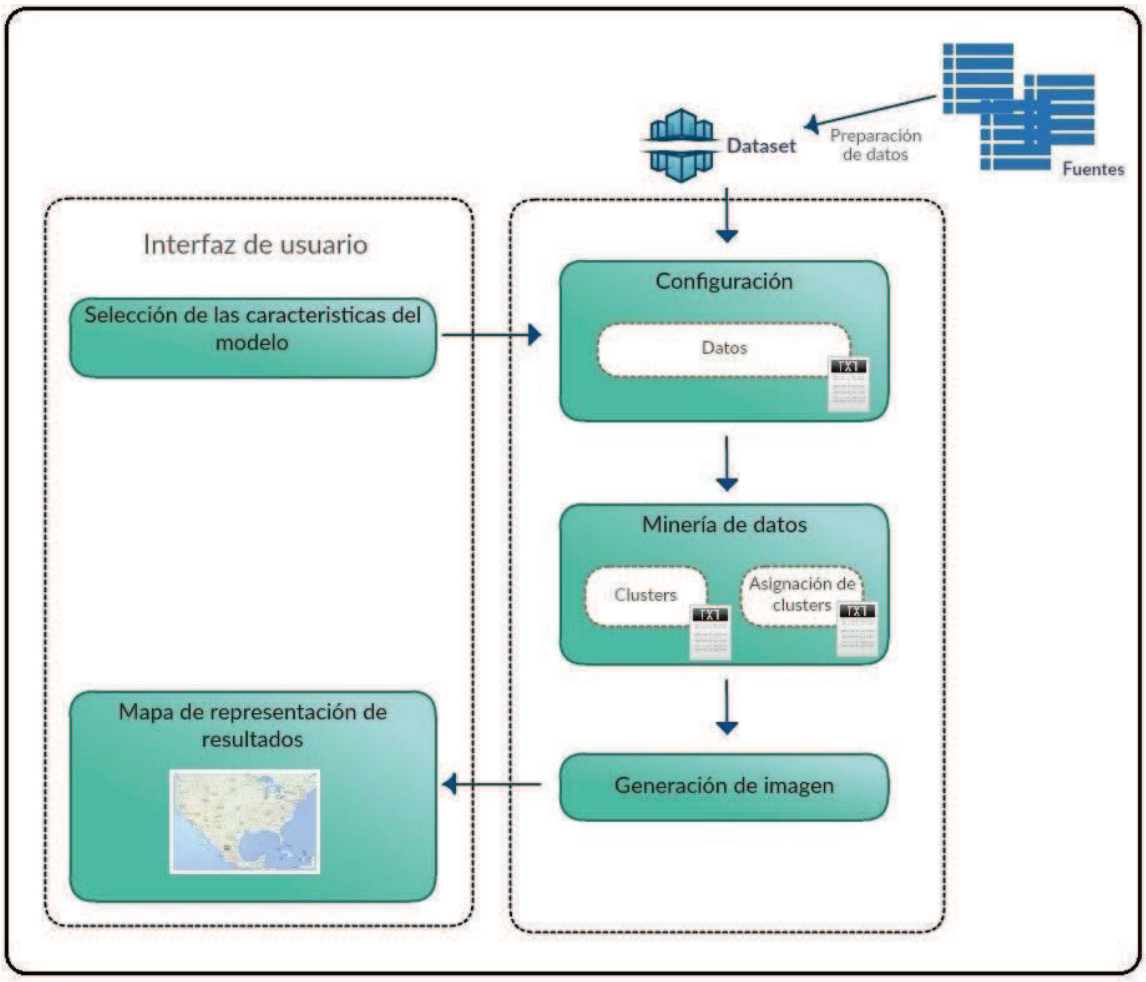

Figura 5.1 Representación general del prototipo

El esquema anterior consta de dos partes principales que son la Minería de Datos y la generación de la imagen.

Minería de Datos: Para la ejecución del algoritmo de Minería de Datos se requiere que en la parte de configuración se genere el *dataset* con la información a procesar, así como que el usuario haya ingresado las características para el modelado. Teniendo esto, el algoritmo arroja como resultado dos archivos de texto, uno con la descripción de los *clusters* y otro con la asignación de los *clusters*. Como resultado se genera lo siguiente:

- ✓ *Clusters*: Archivo de texto que se compone únicamente de los centroides finales que se obtuvieron con el algoritmo.
- ✓ Asignación de *clusters*: Archivo de texto que se compone de los objetos que conforman el *dataset* y el centroide al cual fueron asignados por el algoritmo.

Generación de imagen: A partir de la ejecución del algoritmo se genera una imagen que muestra el agrupamiento obtenido y la tasa de cada grupo.

Todos los archivos de texto generados en este proceso son almacenados en una carpeta cuyo nombre es definido por el usuario en la parte de selección de características del modelado.

5.2. Diseño del prototipo

La implementación del prototipo sigue el diseño presentado en la Figura 5.2 el cual consta de cuatro secciones, de las cuales se destacan: la sección de Minería de Datos y la visualización como base del prototipo.

La sección de Minería de Datos se encarga de recibir la información de la clase de configuración y realiza el agrupamiento utilizando el algoritmo N-Means.

La sección de visualización se encarga de mostrar los resultados obtenidos en la sección anterior de manera gráfica.

La implementación del prototipo se describe en las siguientes secciones del presente capítulo.

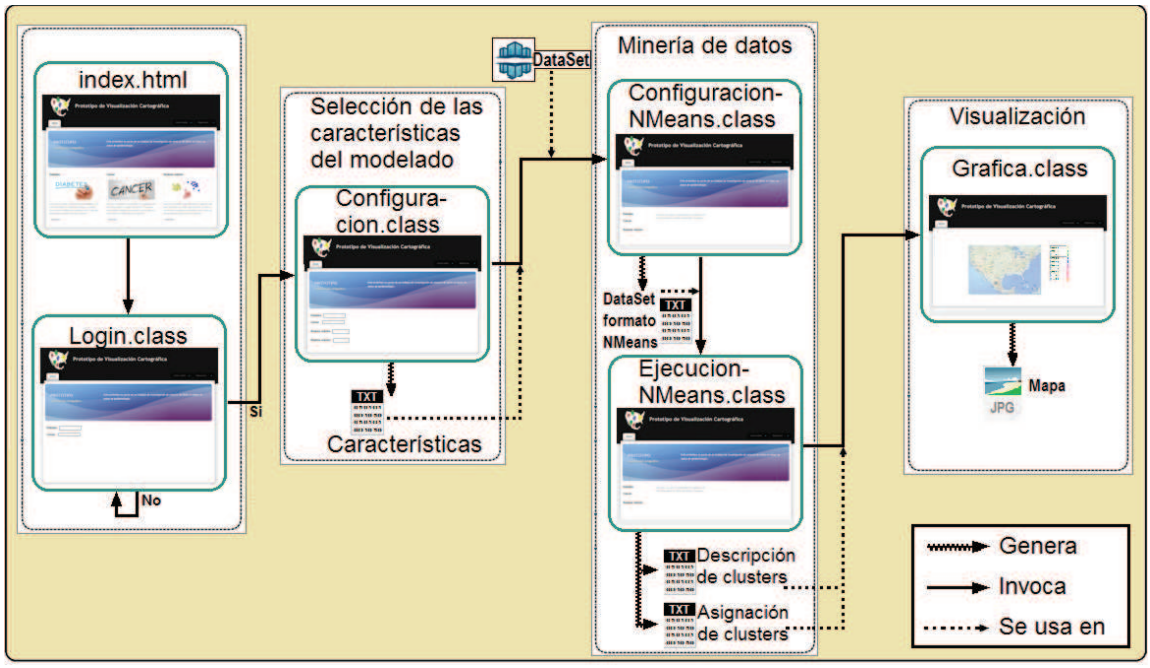

Figura 5.2 Diseño del prototipo

5.3. Características generales de la implementación

El prototipo fue desarrollado en una plataforma web utilizando: tecnología Java (JavaServlets) para la creación de clases que lleven a cabo la funcionalidad, HTML para la generación de interfaces y como servidor se utilizó el contenedor web Tomcat. El prototipo está constituido por una página de inicio y cinco clases contenidas en cuatro secciones.

Como parte inicial se cuenta con una página de bienvenida (index.html) en la cual se muestra información introductoria a las enfermedades estudiadas, también cuenta con un enlace a la clase de gestión de usuarios y un enlace a la clase de

configuración, en esta clase se especifican algunas características del estudio a realizar (causa, año, número de grupos, etc.).

5.4. Módulo de Minería de Datos

La sección de Minería de Datos es la encargada de recibir el conjunto de datos a procesar y devolver los resultados de la ejecución del algoritmo NMeans. Esta sección consta de dos clases: ConfiguraciónNMeans.class y EjecuciónNMeans.class.

La clase ConfiguraciónNMeans recibe de la clase Configuración.class el conjunto de datos seleccionado con el siguiente formato: “causa latitud longitud tasa de mortalidad” y devuelve el *dataset* con el formato que recibe N-Means (latitud longitud tasa) ya que el algoritmo trabaja únicamente con datos numéricos.

La clase de EjecuciónNMeans recibe el *dataset* de la clase ConfiguraciónNMeans y realiza la agrupación. Una vez realizado el agrupamiento, los resultados se guardan en dos archivos de texto creados.

El primer archivo creado se denomina centroides.txt, el cual contiene la descripción de los centroides de los grupos.

El segundo archivo creado es asigClusters.txt, el cual está conformado por cada uno de los ejemplares de entrada, así como la clase a la que pertenece cada ejemplar.

5.5. Módulo de visualización de resultados

Este módulo tiene como finalidad facilitar la interpretación de los resultados generados por el algoritmo de agrupamiento, esto se hace al mostrar el

agrupamiento de forma geográfica usando un mapa de México y Estados Unidos, el cual se usa como una referencia para ubicar cada uno de los municipios y condados así como los centroides de los grupos formados.

La visualización se realiza de la siguiente manera:

a) Los valores de las tasas de mortalidad se codifican en una escala de colores como se muestra en la Figura 5.3, para cada color se asigna un intervalo fijo entre 0 y 10.

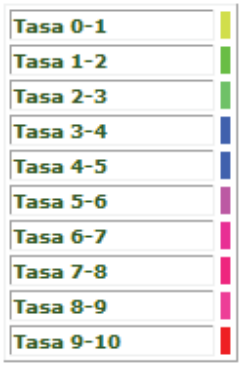

Figura 5.3 Codificación del rango de tasas y colores

b) A partir del archivo centroides.txt se muestra una tabla con información de los centroides y el color asignado al grupo al que pertenece cada centroide como se muestra en la Figura 5.4.

Cluster 1	Longitude	Latitude	Rate	Color
Centroids	4.7881767045	6.3145585585	1.8772922084009251	

Figura 5.4 Visualización de la información de cada centroide

c) A partir de los archivos centroides.txt y asigClusters.txt se genera un objeto de tipo imagen en el que se muestra la ubicación (latitud y longitud) de cada municipio y condado además de la ubicación de cada uno de los centroides de grupo. Estos elementos son coloreados dependiendo del valor de la tasa de mortalidad resultante para cada centroide según la codificación mostrada en la Figura 5.3. Una ejemplificación se muestra en la Figura 5.5.

Figura 5.5 Visualización del agrupamiento resultante

Cabe mencionar que el prototipo permite mostrar uno o varios de los grupos generados, esto con el fin de otorgarle al usuario la flexibilidad de visualizar algún grupo en específico.

Capítulo 6

Resultados experimentales

En este capítulo se describe el plan de pruebas desarrollado, se detallan los casos de prueba llevados a cabo y los resultados obtenidos, todo esto con el objetivo de probar el correcto funcionamiento del prototipo desarrollado en esta tesis.

6.1. Plan de pruebas

A continuación se presenta el plan de pruebas seguido para comprobar el funcionamiento del prototipo de sistema resultante de esta tesis, se define el ambiente en el que se trabajó, el alcance del plan, la descripción de los datos de entrada y salida así como los resultados obtenidos.

Objetivo.

Este apartado tiene la finalidad describir los pasos a seguir para la aplicación de las pruebas necesarias en el presente prototipo. Esto con el fin de verificar la funcionalidad del sistema.

Ambiente de pruebas.

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

Hardware:

- HP Pavilion G4-1420la-NoteBook PC
- Procesador AMD Dual Core E2-1800
- Velocidad de procesamiento 1.70GHz
- Memoria RAM de 6GB

Software:

- Sistema Operativo Microsoft Windows 7
- MySQL versión 5.5
- Tomcat 5.0
- Google Chrome versión 47.0.2526.80 m

Descripción de los casos de prueba.

Para validar la funcionalidad del prototipo desarrollado, se utilizó el conjunto de datos resultante en la fase de preparación de datos para los años 2000 y 2010 con las causas de muerte E11 (Diabetes mellitus no insulinodependiente) y E14 (Diabetes mellitus no especificada). Cabe recordar que, para el territorio estadounidense, los datos de mortalidad fueron generados de forma sintética, es decir, son datos ficticios.

A continuación se describen los casos que componen el plan de pruebas.

6.1.1. Diabetes para el año 2000 y 2010

En esta sección se muestran los resultados obtenidos de aplicar Minería de Datos a la causa E11 del año 2000.

a) Diabetes mellitus no insulinodependiente (E11), año 2000

Los resultados obtenidos para esta causa y año fueron los siguientes:

Debido a su alta tasa de mortalidad, se localizaron dos grupos de interés, los cuales se detallan a continuación.

Tabla 6.1 Grupo de interés C24 para la causa E11 año 2000

Municipio	Incidencia	Tasa de Mortalidad
Venustiano Carranza	247	57.3115
Orizaba	66	55.6525
Iztacalco	177	46.0546
Cuauhtémoc	236	45.7138
Promedio		51.1831

La tabla 6.1 muestra el grupo de interés C24 el cual está conformado por el municipio de Orizaba y tres delegaciones del Distrito Federal. Es destacable que

es el grupo con mayores tasas de incidencia, nótese que tres de los elementos pertenecen a la zona metropolitana de la Ciudad de México.

Tabla 6.2 Grupo de interés C08 para la causa E11 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Azcapotzalco	164	39.5456
Piedras Negras	49	38.2424
Miguel Hidalgo	143	37.999
Gustavo A. Madero	449	37.8656
Netzahualcóyotl	417	37.5484
Benito Juárez	128	35.5084
El Mante	35	31.0829
Tampico	91	30.8013
Matamoros	127	30.3725
Promedio		35.4406

La tabla 6.2 muestra el grupo de interés número C08, este grupo se conforma por cuatro municipios y cinco delegaciones del Distrito Federal.

El grupo con la tasa más alta de mortalidad se puede observar en la Figura 6.1.

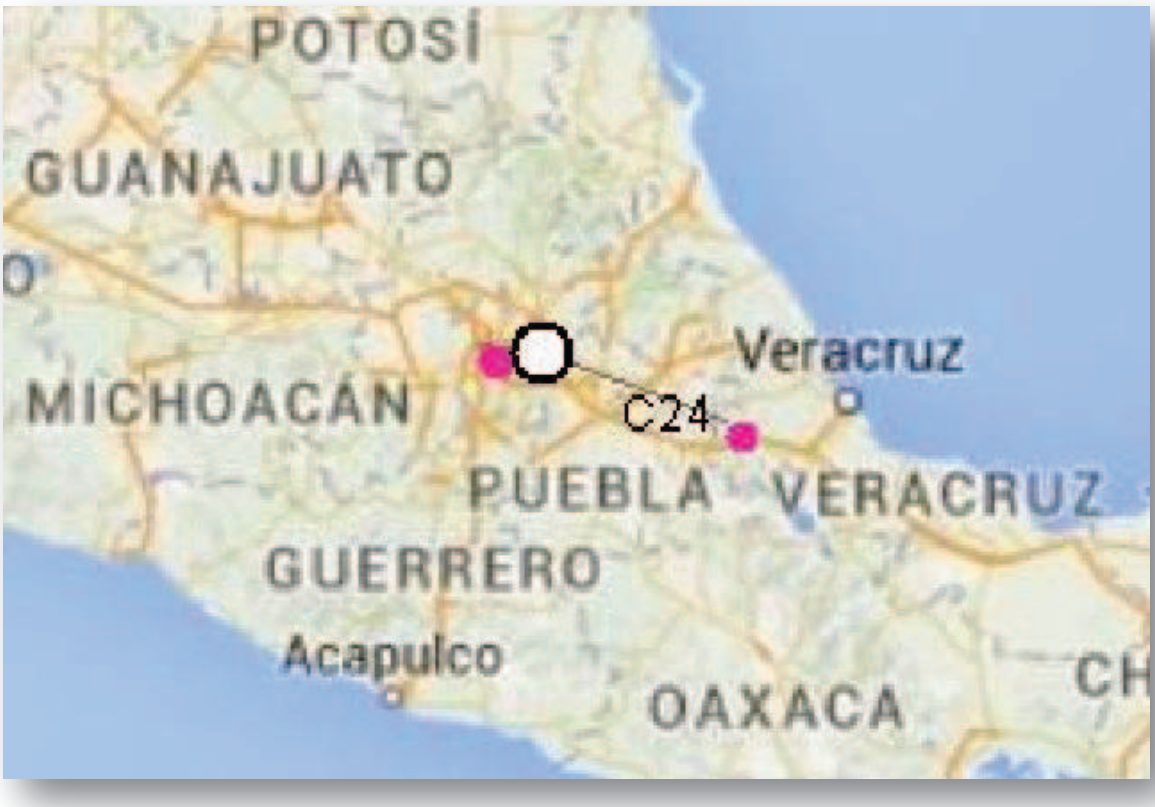

Figura 6.1 Grupo con la mayor tasa de mortalidad para E11 año 2000

En la siguiente sección se muestran los resultados obtenidos de aplicar Minería de Datos a la causa E11 del año 2010.

b) Diabetes mellitus no insulinodependiente (E11), año 2010

Los resultados obtenidos para esta causa y año fueron los siguientes:

Debido a su alta tasa de mortalidad, se localizaron dos grupos de interés, los cuales se detallan a continuación.

Tabla 6.3 Grupo de interés C63 para la causa E11 año 2010

Municipio	Incidencia	Tasa de Mortalidad
Iztacalco	350	91.0685
Cuauhtémoc	465	90.0717
Poza Rica de Hidalgo	136	88.9831
Orizaba	105	88.5381

Gustavo A. Madero	1028	86.6945
Venustiano Carranza	373	86.5473
Apatzingán	102	86.478
San Martín Texmelucan	104	85.9
Azcapotzalco	355	85.6017
Macuspana	114	85.0841
Promedio		87.4967

La Tabla 6.3 muestra el grupo de interés C63 el cual está conformado por cinco municipios y cinco delegaciones del Distrito Federal. Es destacable que es el grupo con mayores tasas de incidencia, nótese que la mitad de los elementos pertenecen a la zona metropolitana de la Ciudad de México.

Tabla 6.4 Grupo de interés C51 para la causa E11 año 2010

Municipio	**Incidencia**	**Tasa de Mortalidad**
Tecámac	133	76.9618
San Francisco del Rincón	77	76.8164
Zamora	124	76.5819
Chalco	164	75.239
Toluca	500	75.0079
Atlixco	87	74.2884
Martínez de la Torre	75	73.9951
Tláhuac	224	73.9786
Cuernavaca	245	72.3341
San Pedro Cholula	87	72.2237
Coyoacán	446	71.8872
Netzahualcóyotl	798	71.8553
Promedio		74.2641

La tabla 6.4 muestra el primer grupo de interés el cual se conforma por nueve municipios y tres delegaciones del Distrito Federal, nótese que tres municipios corresponden al Estado de México.

El grupo con la mayor tasa de mortalidad se puede observar en la Figura 6.2.

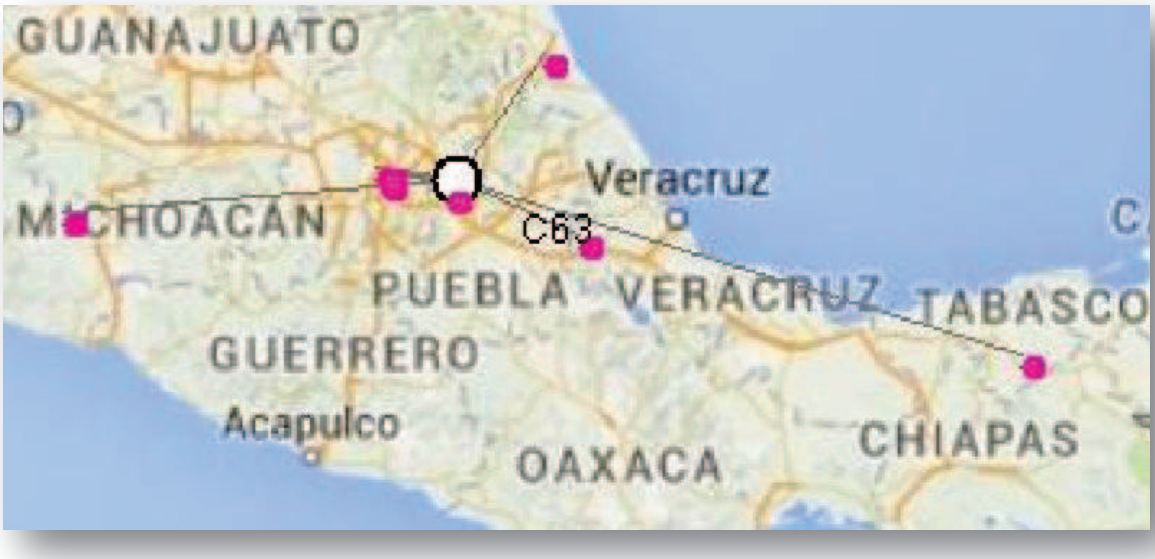

Figura 6.2 Grupo con la mayor tasa de mortalidad para E11 año 2010

En el siguiente inciso se muestra el resultado obtenido de aplicar Minería de Datos a la causa E14 del año 2000.

c) Diabetes mellitus no especificada (E14), año 2000

Los resultados obtenidos para esta causa y año se describen a continuación:

De los grupos generados, se identificaron tres grupos de interés debido a su alta tasa de mortalidad.

Los grupos localizados se describen en las siguientes tablas.

Tabla 6.5 Grupo de interés C69 para la causa E14 año 2000

Municipio	Incidencia	Tasa de Mortalidad
Venustiano Carranza	228	52.9029
Azcapotzalco	216	52.0844

Atlixco	60	51.2334
Iztacalco	187	48.6566
Guadalajara	717	47.9538
Benito Juárez	171	47.437
Martínez de la Torre	48	47.3568
Promedio		49.6607

La Tabla 6.5 muestra el grupo de interés C69 el cual está conformado por tres municipios y cuatro delegaciones del Distrito Federal. Se destaca que es el grupo con la mayor tasa de incidencia promedio.

Tabla 6.6 Grupo de interés C20 para la causa E14 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Gustavo A. Madero	549	46.2989
Acámbaro	49	44.9417
Apatzingán	53	44.9346
Taxco de Alarcón	43	42.8949
Cuauhtémoc	220	42.6145
Promedio		44.3369

La Tabla 6.6 muestra el grupo de interés C20 el cual está conformado por tres municipios y dos delegaciones del Distrito Federal.

Tabla 6.7 Grupo de interés C10 para la causa E14 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
El Mante	48	42.628
San Martín Texmelucan	49	40.4721
Túxpam	51	40.2792
La Magdalena Contreras	87	39.1803
Córdoba	67	37.7916

Irapuato	166	37.7157
Álvaro Obregón	243	35.3701
Promedio		39.0624

La Tabla 6.7 muestra el grupo de interés C10, el cual está conformado por cinco municipios y dos delegaciones del distrito federal.

El grupo con la mayor tasa de mortalidad se puede observar en la Figura 6.3.

Figura 6.3 Grupo con la mayor tasa de mortalidad para E14 año 2000

En el siguiente inciso se muestra el resultado obtenido de aplicar Minería de Datos a la causa E14 del año 2010.

d) Diabetes mellitus no especificada (E14), año 2010

Los resultados obtenidos para esta causa y año se describen a continuación:

De los grupos generados, se identificó como grupo de interés el grupo con la mayor tasa de mortalidad. El grupo localizado se describe en la siguiente tabla.

Tabla 6.8 Grupo de interés C16 para la causa E14 año 2010

Municipio	**Incidencia**	**Tasa de Mortalidad**
Atlixco	85	72.5807
Papantla	101	63.6826
Acámbaro	68	62.3681
Poza Rica de Hidalgo	90	58.8858
Tuxpan	74	58.4444
Tlajomulco de Zúñiga	70	56.6255
Macuspana	73	54.4837
Córdoba	94	53.021
Orizaba	60	50.5932
Promedio		58.9650

La tabla 6.8 muestra el grupo identificado como grupo de interés, el cual se compone de nueve municipios y tiene como tasa de mortalidad promedio de 58.9650.

El grupo con la mayor tasa de mortalidad se puede observar en la Figura 6.4.

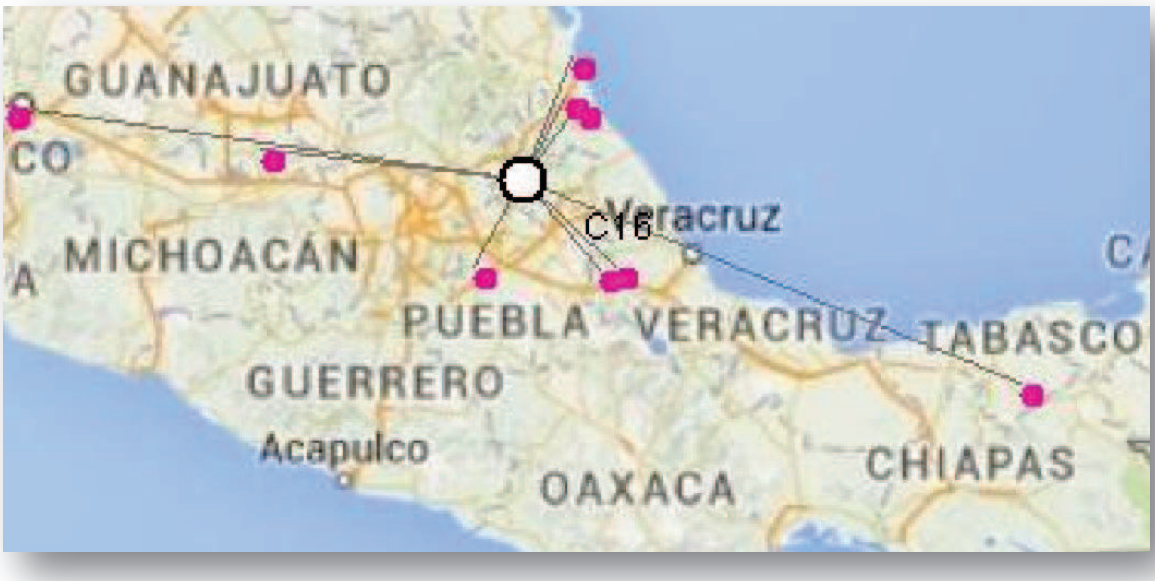

Figura 6.4 Grupo con la mayor tasa de mortalidad para E14 año 2010

El siguiente caso de prueba tiene únicamente la finalidad de demostrar que fue posible realizar la escalabilidad en el prototipo desarrollado.

e) Prueba de escalabilidad abarcando los territorios de México y Estados Unidos

Con la finalidad de validar la escalabilidad del prototipo se diseñó una prueba que incluye las siguientes causas de mortalidad:

a) Diabetes mellitus insulinodependiente (E10),
b) Diabetes mellitus no insulinodependiente (E11),
c) Diabetes mellitus asociada con desnutrición (E12),
d) Otras diabetes mellitus especificadas (E13),
e) Diabetes mellitus no especificada (E14).

En el caso de los Estados Unidos se usaron datos sintéticos en las incidencias de mortalidad debido a que no se tenían datos oficiales de dicho país. Desde el punto de vista computacional de prueba de escalabilidad del prototipo los resultados fueron exitosos.

De manera particular los resultados se muestran en la Tabla 6.9 y en la Figura 6.5.

Tabla 6.9 Grupo de interés C64, agrupación de los diferentes tipos de diabetes año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Orizaba	142	119.7372
Venustiano Carranza	495	106.9562
Iztacalco	383	93.1146
Cuauhtémoc	475	92.0087
Azcapotzalco	401	90.928
Gustavo A. Madero	1040	84.1735
Benito Juárez	303	84.055
Acámbaro	92	83.0939
Miguel Hidalgo	287	81.3861
Promedio		92.8281

La Tabla 6.9 muestra el grupo identificado como grupo de interés, el cual se compone de nueve localidades y tiene como tasa de mortalidad promedio de 92.8281. Este grupo está conformado por siete delegaciones del Distrito Federal y únicamente por dos municipios (Orizaba y Acámbaro) localizados en los estados de Veracruz y Guanajuato respectivamente.

Los grupos con mayores tasas de mortalidad se muestran de manera visual en la Figura 6.5. En dicha figura se destacan los puntos del centro de México y en el caso de los Estaos Unidos los de la Florida y del noreste.

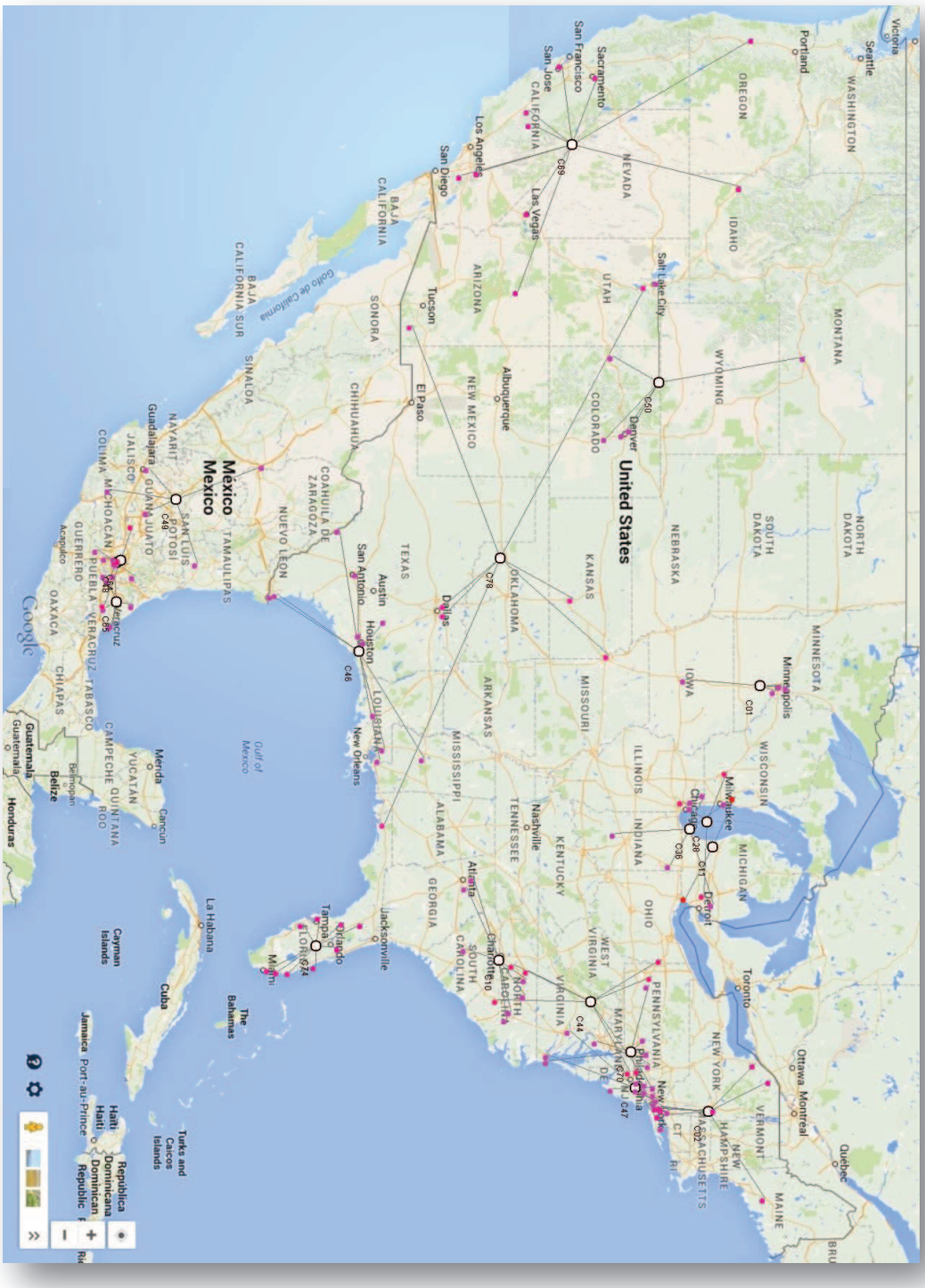

Figura 6.5 Grupos con las mayores tasas de mortalidad para la agrupación de los diferentes tipos de diabetes del año 2010

Capítulo 7

Conclusiones y trabajos futuros

En este capítulo se presentan las conclusiones a las que se llegó con la investigación así como los trabajos que se derivan a partir de los resultados obtenidos y que continúan con la línea de investigación.

7.1. Conclusiones

En el presente trabajo de investigación se muestra que es factible el desarrollo de un prototipo de un sistema de Minería de Datos orientado a manejar grandes instancias como las que se presentan en el paradigma de Big Data en el dominio de la salud. En particular el objetivo del prototipo es encontrar patrones de interés de regiones de México y de Estados Unidos con altas tasas de incidencia de mortalidad por diabetes, a partir de bases de datos poblacionales.

En esta investigación se propuso el uso del algoritmo N-Means para realizar la tarea de agrupamiento en el proceso de Minería de Datos. Para realizar las tareas de visualización se propuso un módulo cartográfico que hace uso de los mapas proporcionados por Google Maps, los cuales comprenden el territorio de México y de los Estados Unidos.

El prototipo se validó de manera sistemática con un conjunto de casos de prueba diseñado para tal fin. Como base para las pruebas se usaron los datos de mortalidad de los censos del año 2000 y 2010, además de otras bases de datos poblacionales. Se generó un almacén de datos que integra datos de mortalidad de los años 2000 y 2010 correspondientes a 2049 causas de muerte para localidades mayores a 100,000 habitantes en México. Es destacable que el volumen de datos para la experimentación fue del orden de los tres gigabytes, con más de cuatro millones de registros.

Con base en las pruebas realizadas para la causa E11 se observó lo siguiente:

a) Para el año 2000 el grupo con una mayor tasa de incidencia promedio fue el identificado como C24 el cual constaba del municipio de Orizaba y de las delegaciones Venustiano Carranza, Iztacalco y Cuauhtémoc del Distrito Federal, dicho grupo tuvo una tasa de 51.1831.
b) Para el año 2010 el grupo con mayor tasa de incidencia promedio identificado como C63, el cual constaba de los municipios Poza Rica de Hidalgo y Orizaba en el estado de Veracruz, Apatzingán en el estado de

Michoacán, San Martín Texmelucan en Puebla, Macuspana en Tabasco y las delegaciones Iztacalco, Gustavo A. Madero, Cuauhtémoc, Venustiano Carranza y Azcapotzalco en el Distrito Federal, dicho grupo tuvo una tasa de 87.4967.

El análisis de los dos patrones encontrados permitió encontrar el siguiente conocimiento:

c) En varias de las delegaciones del Distrito Federal se encontraron los más altos índices de incidencia a nivel nacional para el año 2000 y 2010,
d) Contrastando los valores de los grupos con mayor incidencia en el año 2000 y 2010 se observó un incremento cercano al 100% en las incidencias de mortalidad.

Los dos patrones encontrados se pudieron corroborar, en lo general, con lo reportado en el portal del SIVANE [SINAVE]. Es importante destacar que la información del SINAVE es más general ya que su granularidad es a nivel de estados y del Distrito Federal, en contraste, en esta investigación el nivel de detalle es de municipio y delegación en el caso del Distrito Federal.

Como es fácil darse cuenta el problema de la diabetes se ha incrementado notablemente en un periodo de diez años en el área metropolitana de la Ciudad de México y en otros municipios del país.

Los resultados obtenidos para esta enfermedad hacen evidente la utilidad de las ciencias computacionales y en particular de la Minería de Datos en el área de salud, ya que proporcionan elementos de apoyo para la toma de decisiones de los funcionarios y autoridades encargadas de la salud de la población.

De manera complementaria se realizaron otras pruebas con resultados satisfactorios para el cáncer de pulmón (C34) y de estómago (C16) los cuales son descritos en el anexo A y B.

Con relación al escalamiento del sistema es destacable mencionar que tiene la infraestructura para manejar datos poblacionales de México y de Estados Unidos

así como datos cartográficos de los 2473 municipios y 3141 condados, lo cual permite la visualización de los patrones de interés sobre mapas que comprenden el territorio mexicano y de Estados Unidos. Este escalamiento se probó con datos reales de México y datos sintéticos de mortalidad de los Estados Unidos ya que durante el desarrollo de la presente investigación no fue posible acceder a los datos reales de mortalidad de Estados Unidos, aún cuando se realizaron varios esfuerzos para conseguirlos. Las pruebas realizadas para validar el escalamiento fueron exitosas.

Finalmente, se identificó y resaltó la importancia de la estandarización y normalización de los datos como parte indispensable y relevante para el éxito del proceso de Minería de Datos.

7.2. Trabajos futuros

Debido a los resultados alentadores obtenidos en este trabajo de tesis, se proponen las siguientes investigaciones para dar continuidad a la misma:

- ✓ Incorporar en el estudio otras causas de mortalidad.
- ✓ Explorar el uso de otras mejoras del algoritmo K-Means.
- ✓ Incorporar en el estudio datos reales de mortalidad por diabetes en los Estados Unidos.
- ✓ Explorar posibles mejoras a la interfaz con el usuario para hacerla más amigable.

REFERENCIAS

[ABELLANA ET AL. 2009] Abellana, R., Ascaso, C., Carrasco, J. L., Castell, C., & Tresserras, R. (2009). Geographical variability of the incidence of Type 1 diabetes in subjects younger than 30 years in Catalonia, Spain. *Medicina clínica*, *132*(12), 454-458.

[ACHREKAR ET AL. 2012] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2012). Twitter Improves Seasonal Influenza Prediction. In HEALTHINF (pp. 61-70).

[ALJUMAH ET AL. 2013] Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), 127-136.

[CEMECE] Centro Colaborador para la Familia de Clasificadores Internacionales de la OMS en México. ”Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud, Décima Revisión (CIE-10)”. Fecha de consulta: Enero 2015. Disponible en: http://www.dgis.salud.gob.mx/contenidos/cemece/documentos.html

[CENSUS] United States Census Bureau. Fecha de consulta: Enero 2014. Disponible en: http://www.census.gov/data.html

[CHAPMAN ET AL. 2000] Chapman, P., Clinton, J., Kerber, R., and et al. CRISP-DM 1.0 Step-by-step data mining guide. USA: CRISP-DM Consortium, 2000.

[DATE 2001] Christopher J. Date. (2001). Introducción a los Sistemas de Bases de Datos. México: Pearson Educación.

[FAYYAD ET AL. 1996] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[FRAWLEY ET AL. 1991] Frawley, W. J., & Matheus, C. J. (1991). Knowledge discovery in databases(pp. 1-27). G. Piatetsky-Shapiro (Ed.). Menlo Park, CA: AAAI Press.

[GARTNET 2015] Gartner.Big Data. Consultado el 05 de Octubre 2015 de http://www.gartner.com/it-glossary/big-data/.

[GORUNESCU 2011] Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Springer Science & Business Media.

[HAN 2011] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques.USA: Elsevier.

[HAND ET AL. 2001] David Hand, Heikki Mannila, Padhraic Smyth, “Principles of Data Mining”, Massachusetts Institute of Technology, pp. 1-4, 9-13, 18-21, 31-33, 167-168, 298-299; Estados Unidos, 2001.

[HERNADEZ 2005] Hernandez I., Gil de Miguel A., Delgado M., Bolumar M. F.: Concepto y aplicaciones de la epidemiología, primera edición, Médica Panamericana, Madrid España, ISBN 84-7903-955-8 (2005).

[IDF 2015] International Diabetes Federation. IDF Diabetes Atlas, séptima edición, Bruselas Bélgica, 2015

[INEGI 2015] Instituto Nacional de Estadística y Geografía (INEGI). “Mortalidad”. Fecha de consulta: Enero de 2015. Disponible en: http://www.inegi.org.mx/est/contenidos/Proyectos/registros/vitales/mortalidad/

[INEGI] Instituto Nacional de Estadística y Geografía (INEGI). “Censo de Población y Vivienda”. Fecha de consulta: Enero 2014. Disponible en: http://www.inegi.org.mx/sistemas/olap/Proyectos/bd/censos/cpv2010/PT.asp?s=est&c=27770&proy=cpv10_pt.

[ITURBIDE 2013] Gregorio Emmanuel Iturbide Domínguez, Tesis de maestría: “Metodología de Preparación de Datos Orientada a Aplicaciones de Epidemiología Basada en el Modelo CRIPS-DM”, Centro Nacional de Investigación y Desarrollo

Tecnológico (CENIDET) Ingeniería de Software, Departamento de Ciencias Computacionales, Febrero del 2013.

[KDNUGGETS] KDNuggets: Data Mining, Analytics, Big Data and Data Science. “Metodologías más utilizadas en proyectos de Minería de Datos”. Fecha de consulta: Enero 2014. Disponible en: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

[KUAN-CHING ET AL. 2015] Kuan-Ching Li, Hai Jiang, Laurence T. Yang, and Alfredo Cuzzocrea. Big Data: Algorithms, Analytics, and Applications (1st ed.). Chapman & Hall/CRC. 2015.

[LANEY 2001] Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety(). META Group Research Note.

[LOPEZ 2000] López Moreno, Sergio; Garrido Latorre, Francisco y Hernández Ávila, Mauricio. Desarrollo histórico de la epidemiología: su formación como disciplina científica. Salud pública Méx [online]. 2000, vol.42, n.2, pp. 133-143. issn 0036-3634.

[LÓPEZ 2015] Vitervo López Caballero, Tesis de maestría: “Incremento de la Eficiencia del Algoritmo K-Means Mediante la Mejora de la Heurística Early Classification”, Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) Ingeniería de Software, Departamento de Ciencias Computacionales, Noviembre del 2015.

[MEXICANO 2007] Adriana Mexicano Santoyo, Tesis de maestría: “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”, Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) Ingeniería de Software, Departamento de Ciencias Computacionales, Noviembre del 2007.

[OMS 2015] Organización Mundial de la Salud Mortalidad. Consultado el 25 de Agosto de 2015 de http://www.who.int/topics/mortality/es/.

[SALAZAR ET AL. 2011] Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., ... & Bruin, S. (2010). Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. Journal of clinical oncology, JCO-2010.

[SIMBAD] Sistema Estatal y Municipal de Bases de Datos (SIMBAD). "Área geográfica". Fecha de consulta: Enero 2014. Disponible en: http://sc.inegi.org.mx/sistemas/cobdem/contenido-arbol.jsp.

[SINAIS] Sistema Nacional de Información en Salud (SINAIS). "Bases de datos sobre defunciones". Fecha de consulta: Enero 2014. Disponible en: http://www.sinais.salud.gob.mx/basesdedatos/estandar.html.

[SINAVE] Sistema Nacional de Vigilancia Epidemiológica. SINAVE/DGE/SALUD/. "Panorama epidemiológico y estadístico de la mortalidad en México 2009".

[VAN ESSEN ET AL. 2013] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn human connectome project: an overview. Neuroimage, 80, 62-79.

[WEFORUM 2015] World Economic Forum, global risks 2015, décima edición, Ginebra Suiza, 2015.

[WEI Y BIFET 2012] Wei Fan, Albert Bifet, Mining big data: current status, and forecast to the future, ACM SIGKDD Explorations Newsletter, v.14 n.2, December 2012.

[WITTEN ET AL. 2011] Witten, I., Frank, E., and Hall M.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011.

Anexo A. Cáncer de estómago (C16) para el año 2000 y 2010

Este anexo está dividido en dos partes, la primera se enfoca en el estudio realizado para la causa C16 en el año 2000 y la segunda parte para la causa C16 en el año 2010.

a) Cáncer de estómago (C16) para el año 2000

Los resultados obtenidos para esta causa y año se mencionan a continuación.

De los grupos generados, se identificaron tres grupos de interés debido a sus altas tasas de mortalidad. Estos se describen en las siguientes tablas.

Tabla A.1 Grupo de interés C55 para la causa C16 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Poza Rica de Hidalgo	12	11.4057
Comitán de Domínguez	12	11.4057
Atlixco	12	10.24
Ciudad Madero	18	9.8724
Promedio		10.73095

La Tabla A.1 muestra el grupo de interés C55 el cual está conformado por los municipios de Poza Rica de Hidalgo, Comitán de Domínguez, Atlixco y Ciudad Madero. Es destacable que es el grupo con mayores tasas de incidencia con una media de 10.73095.

Tabla A.2 Grupo de interés C14 para la causa C16 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Minatitlán	14	9.1502
Cuauhtémoc	45	8.7166
Iztacalco	33	8.5864

Comalcalco	14	8.5035
Venustiano Carranza	36	8.353
Pénjamo	12	8.3087
Miguel Hidalgo	29	8.2236
La Magdalena Contreras	18	8.1062
Iguala de la Independencia	10	8.0671
Gustavo A. Madero	93	7.8429
Benito Juárez	28	7.7674
Tapachula	21	7.7229
Promedio		8.2790

La Tabla A.2 muestra el grupo de interés C14 el cual está conformado por doce localidades, de las cuales siete con delegaciones del Distrito Federal.

Tabla A.3 Grupo de interés C73 para la causa C16 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Córdoba	13	7.3327
Tlajomulco de Zúñiga	9	7.2804
Tampico	21	7.1079
Fresnillo	13	7.0946
San Cristóbal de las Casas	9	6.7965
Orizaba	8	6.7457
Allende	9	6.6725
Oaxaca de Juárez	17	6.6372
San Felipe del Progreso	8	6.59
Tulancingo de Bravo	8	6.5426
Zitácuaro	9	6.5193
Dolores Hidalgo	8	6.2018
Promedio		6.7934

La Tabla A.3 muestra el grupo de interés C73 conformado por doce municipios, ubicados en Veracruz, Jalisco, Tamaulipas, Zacatecas, Guanajuato, Oaxaca, Estado de México, Hidalgo, Michoacán y Chiapas.

El grupo con la mayor tasa de mortalidad se puede observar en la Figura A.1.

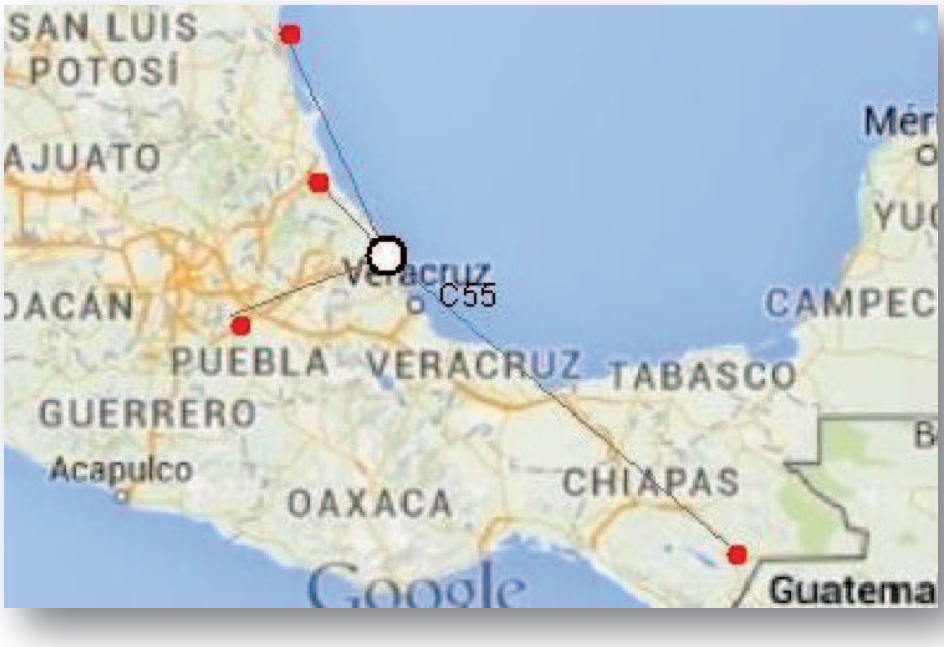

Figura A.1 Grupo con la mayor tasa de mortalidad para C16 año 2000

b) Cáncer de estómago (C16) para el año 2010

Los resultados obtenidos para esta causa y año se mencionan a continuación.

De los grupos generados, se identificaron tres grupos de interés debido a sus altas tasas de mortalidad. Estos se describen en las siguientes tablas.

Tabla A.4 Grupo de interés C14 para la causa C16 año 2010

Municipio	Incidencia	Tasa de Mortalidad
Atlixco	17	14.5161
Acámbaro	13	11.9233

Promedio		13.2197

La Tabla A.4 muestra el grupo de interés C14 el cual está conformado por dos municipios: Atlixco y Acámbaro. Es destacable que es el grupo con mayores tasas de incidencia con una media de 13.2197.

Tabla A.5 Grupo de interés C73 para la causa C16 año 2010

Municipio	**Incidencia**	**Tasa de Mortalidad**
Cunduacán	12	11.4986
Tepatitlán de Morelos	12	10.0673
Papantla	16	10.0883
Temapache	10	9.7138
Tlajomulco de Zúñiga	12	9.7072
Ciudad Valles	14	9.5495
Promedio		10.1041

La Tabla A.5 muestra el grupo de interés C73 el cual está conformado por seis municipios y tiene como tasa de mortalidad media 10.1041.

Tabla A.6 Grupo de interés C45 para la causa C16 año 2010

Municipio	**Incidencia**	**Tasa de Mortalidad**
Hidalgo del Parral	11	10.9104
Decilias	11	9.448
Chihuahua	58	8.6336
Promedio		9.6640

La tabla A.6 muestra el grupo de interés C45, este grupo se conforma por tres municipios todos estos localizados en el estado de Chihuahua, tiene como tasa de mortalidad media 9.6640.

El grupo con la mayor tasa de mortalidad se puede observar en la Figura A.2.

Figura A.2 Grupo con la mayor tasa de mortalidad para C16 año 2010

Anexo B. Cáncer de pulmón (C34) para el año 2000 y 2010

Este anexo está dividido en dos partes, la primera se enfoca en el estudio realizado con la causa C34 en el año 2000 y la segunda parte para la causa C34 en el año 2010.

a) Cáncer de pulmón (C34) para el año 2000

Los resultados obtenidos para esta causa y año se mencionan a continuación.

De los grupos generados, se identificó un grupo de interés debido a sus altas tasas de mortalidad. Este se describe en la siguiente tabla.

Tabla B.1 Grupo de interés C59 para la causa C34 año 2000

Municipio	**Incidencia**	**Tasa de Mortalidad**
Culiacán	108	14.4862
Navojoa	20	14.2196
Ahome	51	14.2003
Guasave	39	14.059
Hidalgo del Parral	14	13.8859
Delicias	16	13.7426
La Paz	27	13.712
Guaymas	17	13.0439
Mazatlán	48	12.6146
Promedio		13.7737

La Tabla B.1 muestra el grupo C59 el cual está compuesto por nueve municipios localizados en la parte noroeste de México. Se destaca que es el grupo con mayores tasas de incidencia con una media de 13.7737. Este grupo se puede observar en la Figura B.1.

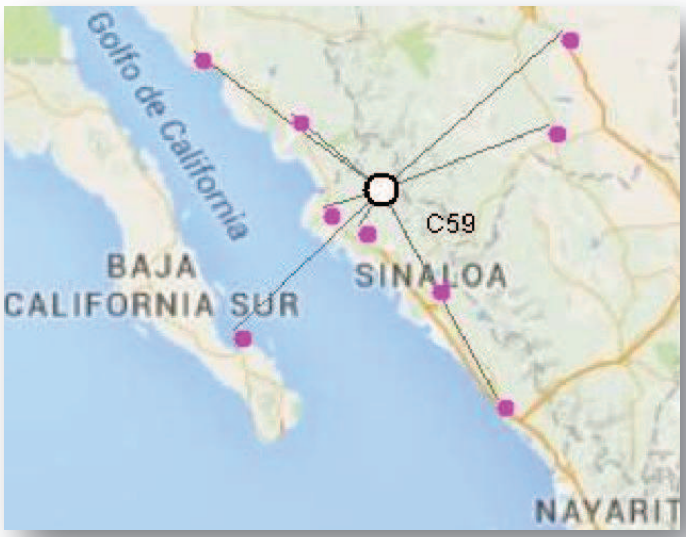

Figura B.1 Grupo con la mayor tasa de mortalidad promedio para C34 año 2000

b) Cáncer de pulmón (C34) para el año 2010

Los resultados obtenidos para esta causa y año se mencionan a continuación.

De los grupos generados, se identificó un grupo de interés debido a sus altas tasas de mortalidad. Este se describe en la siguiente tabla.

Tabla B.2 Grupo de interés C62 para la causa C34 año 2010

Municipio	**Incidencia**	**Tasa de Mortalidad**
Tepatitlán de Morelos	17	14.262103
Cuernavaca	40	11.809652
Cuauhtémoc	60	11.622163
Benito Juárez	41	11.373787
Miguel Hidalgo	40	11.343012
Zamora	18	11.116738
Altamira	14	10.966286
Veracruz	50	10.9319
Promedio		11.6782

La Tabla B.2 muestra el grupo C62 el cual está compuesto por cinco municipios localizados en los Estados de Jalisco, Morelos, Michoacán, Tamaulipas y Veracruz respectivamente y por tres delegaciones del Distrito Federal. Se destaca que es el grupo con mayores tasas de incidencia con una media de 11.6782. Este grupo se puede observar en la Figura B.2.

Figura B.2 Grupo con la mayor tasa de mortalidad promedio para C34 año 2010