SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Desarrollo de una aplicación de Ciencia de Datos

presentada por

Ing. Lorenzo Sánchez Méndez

como requisito para la obtención del grado de

Maestro en Ciencias en Ciencias Computacionales

Director de tesis

Dr. Joaquín Pérez Ortega

Cuernavaca, Morelos, México. Septiembre de 2018.

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Morelos a 24 de agosto del 2018
OFICIO No. DCC/222/2018

Asunto: Aceptación de documento de tesis

DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **Ing. Lorenzo Sánchez Méndez**, con número de control M16CE089, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "**Desarrollo de una aplicación de Ciencia de Datos"** 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 María Rodríguez Lelis
Doctor en Ciencias en Ingeniería Mecánica
450002

REVISOR 2

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

REVISOR 3

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

C.p. M.T.I. María Elena Gómez Torres - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 19 de septiembre de 2018
OFICIO No. SAC/398/2018

Asunto: Autorización de impresión de tesis

ING. LORENZO SÁNCHEZ MÉNDEZ
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 **"Desarrollo de una Aplicación de Ciencia de Datos"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

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

A T E N T A M E N T E
EXCELENCIA EN EDUCACIÓN TECNOLÓGICA®
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

SEP TecNM
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

C.p. M.T.I. María Elena Gómez Torres .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Dedicatoria

Dedico este logro a la gran y hermosa familia que Dios me ha dado. A ellos quienes confiaron en mí desde el primer momento en que les comuniqué la decisión de venir a estudiar la maestría a Cuernavaca. Recuerdo muy bien que les dije que este esfuerzo lo haría por ustedes también.

A mi familia Méndez, quienes siempre han orado al Padre por mí y muestran a cada momento el cariño que me tienen.

A mi familia Sánchez, ellos me han alentado a esforzarme cada día para conseguir mis metas y siempre me reciben con los brazos abiertos.

Especialmente a mis padres Lorenzo y Teresa, a ellos quienes me trajeron a este mundo, quienes han trabajado arduamente para proveernos de estudio y comodidad a mí y mis hermanas. Por enseñarme a través de su vida la humildad, honestidad, responsabilidad, respeto, trabajo duro, pero por sobre todas las cosas, por enseñarme el camino del Señor. Quienes me han apoyado en cada una de las decisiones que he tomado.

A ustedes dedico estos dos años de esfuerzo.

Agradecimientos

Porque Jehová da la sabiduría y de su boca viene el conocimiento y la inteligencia.

Proverbios 2:6

Primeramente quiero agradecer a Dios por darme la vida, la fortaleza, protección, dirección e inteligencia en todo este tiempo, no habría obtenido este logro sin su presencia en mi vida.

Al CONACYT por el apoyo económico recibido en estos dos años, que sin ello no hubiese podido estudiar la maestría.

Agradezco el profesionalismo y dirección del Dr. Joaquín Pérez Ortega, su tiempo dedicado a esta tesis fue muy valioso. Mis agradecimientos a mis revisores, los Dres. Noé Alejandro Castro Sánchez, Moisés Gonzáles García y José María Rodríguez Lelis por su valiosa contribución a esta tesis. A la Dra. Leticia Sánchez Lima por apoyarme en la redacción de este documento.

Quiero agradecer al Sr. Manuel Cortina y la Sra. Lupita por haberme recibido tan calurosamente cuando llegué a Cuernavaca.

A Marianita y Anahí, que juntos hemos compartido muchas experiencias, muchas gracias por su sincera amistad chicas. A mis compañeros de generación: Luis, Alber, Omar, Sócrates y Bryan; mis compañeros de línea: Celia, Daniel, Gudiel, Andrea, con quienes compartí gratos momentos.

Quiero agradecer a la INP. Bethel Centro por haberme recibido con cariño y confianza, estos dos años fueron de mucha bendición para mí. Al Pbro. Israel

Campos y su familia por enseñarme a servir con esfuerzo al Señor. A la familia Coctecón Bagazo por tantas muestras de afecto, realmente fueron mi familia adoptiva. Al equipo de liderazgo del Min. Juvenil: Jean, Iseni, Itzel, Noel, Miriam y Eva por todo su apoyo en las actividades realizadas.

Agradezco también a mis hermanas Nury del Rocío y Rosa del Carmen por ser un gran apoyo para mí, siempre estuvieron alentándome para dar lo mejor de mí.

Finalmente, quiero agradecer a una persona muy especial para mí, una mujer que es de gran bendición en mi vida, mujer por la que agradezco a Dios su vida y de la que estoy profundamente enamorado. Muchas gracias Analy Hernández Torres -mi amada novia- por apoyarme en todo momento y confiar en mí, a Dios gracias por convertirte en la mujer que eres: una Mujer Virtuosa.

RESUMEN

En esta investigación se muestra que es factible la asimilación de conceptos de Ciencia de Datos y la creación de una infraestructura de conocimiento que apoye el desarrollo de aplicaciones de Ciencia de Datos.

Se validaron los conceptos de Ciencia de Datos por medio del desarrollo de un caso práctico. Se utilizó la metodología “Foundational Methodology for Data Science” -propuesta por la empresa IBM- para desarrollar un caso práctico. Además, se utilizó el lenguaje de programación estadística R como apoyo a las actividades realizadas.

El estudio desarrollado tuvo como objetivo la proyección de las tasas de mortalidad por diabetes mellitus tipos E11-E14 en regiones de municipios de México para el periodo 2016-2020. Las regiones de análisis fueron clasificadas como C24, C08 y C51. Representan a 25 municipios del país con las mayores tasas de mortalidad por diabetes mellitus. En el análisis se utilizaron datos poblacionales obtenidos de instituciones oficiales como, SINAIS, INEGI, CONAPO y CEMECE.

Una de las mayores preocupaciones del gobierno mexicano en materia de Salud Pública ha sido el incremento en las tasas de mortalidad por diabetes mellitus a nivel nacional. En particular, se tiene interés en conocer si seguirá creciendo en los próximos años, se mantendrá o disminuirá y en qué proporción.

Entre los hallazgos obtenidos al aplicar la proyección de las tasas de mortalidad por diabetes mellitus se destaca que en la región C24, de continuar con la tendencia actual, se prevé un descenso para el año 2020 en un rango del 6.2 al 11.1% con respecto al año 2003 –punto máximo en la tasa de mortalidad. En la región C08, para el año 2020 se prevé un descenso en un rango del 18.8 al 21.9% con respecto al año 2002. Sin embargo, los modelos de predicción aplicados a la región C51, prevén que la tasa de mortalidad oscilará entre un – 1.3 y +16.6% para el año 2020 con respecto al 2015.

ABSTRACT

This research shows that the assimilation of Data Science concepts and the creation of a knowledge infrastructure that supports the development of Data Science applications is feasible.

The concepts of Data Science were validated through the development of a practical case. The methodology "Foundational Methodology for Data Science", proposed by the IBM Company, was used to develop a practical case. In addition, the statistical programming language R was used to support the activities carried out.

The objective of the study was to project the mortality rates for diabetes mellitus types E11-E14 in regions of municipalities of Mexico for the period 2016-2020. The regions of analysis were classified as C24, C08 and C51. They represent the 25 municipalities of the country with the highest mortality rates for diabetes mellitus. In the analysis, population data obtained from official institutions such as SINAIS, INEGI, CONAPO and CEMECE were used.

One of the major concerns of the Mexican government in matters of Public Health has been the increase nationwide in death rates from diabetes mellitus. In particular, it is interesting to know if the death rates will continue to grow in the coming years, it will be maintained or decreased and in what proportion.

Among the findings obtained when applying the projection of mortality rates for diabetes mellitus, it is highlighted that in the C24 region, if this current trend continues, a decrease is expected in 2020 in a range of 6.2 to 11.1% with respect to the year 2003 - maximum point in the mortality rate. In the C08 region, a decrease in the range from 18.8 to 21.9% is foreseen in 2020 with respect to the year 2002. However, the prediction models applied to the C51 region, foresee that the mortality rate will vary between a - 1.3 and + 16.6% for the year 2020 with respect to 2015.

TABLA DE CONTENIDO

Página

Resumen……I
Abstract II
Lista de tablas..... V
Lista de figuras VI
Organización del documento VIII
1. INTRODUCCIÓN 1
1.1. Contexto de la investigación 2
1.2. Planteamiento de problema 3
1.3. Enfoque general de solución 4
1.4. Objetivos 5
1.5. Alcances y limitaciones 5
1.5.1. Alcances 5
1.5.2. Limitaciones 5
2. CONCEPTOS BÁSICOS DE CIENCIA DE DATOS 6
2.1. Introducción a la Ciencia de Datos 7
2.2. Definiciones y conceptos 9
2.3. Desarrollo de la Ciencia de Datos..... 11
3. ESTADO DEL ARTE 14
3.1. Trabajos relacionados 15
3.2. Casos prácticos de Ciencia de Datos 19
3.3. Casos prácticos en el dominio de la Salud Pública 22
4. SELECCIÓN DE UNA METODOLOGÍA DE CIENCIA DE DATOS 33
4.1. Revisión de metodologías de Ciencia de Datos 34
4.2. Selección de una metodología de Ciencia de Datos 41
5. APLICACIÓN DE LA METODOLOGÍA FMDS EN UN CASO EN EL ÁREA DE SALUD 43
5.1. Fase de entendimiento del negocio 44
5.2. Fase de enfoque analítico..... 45
5.3. Fase de requerimiento de datos…….. 47
5.4. Fase de recopilación de datos 50
5.5. Fase de entendimiento de los datos 51
5.6. Fase de preparación de los datos 54
5.6.1. Selección, limpieza y transformación de las bases de datos..... 54

5.6.2. Creación de datos 56
5.6.3. Integración de datos 56
5.7. Fase de modelado 57
5.8. Fase de evaluación 67
5.9. Fase de despliegue 74
5.10. Fase de retroalimentación 78
6. CONCLUSIONES Y TRABAJOS FUTUROS 79
6.1. Conclusiones 80
6.2. Trabajos futuros 81
REFERENCIAS 82
Anexo A Reseña de fuentes de información de Ciencia de Datos 88
Anexo B Principios de Ciencia de Datos 92
Anexo C Disciplinas que integran la Ciencia de Datos 95
Anexo D Técnicas de Ciencia de Datos 97
Anexo E Herramientas de Ciencia de Datos 102

Lista de tablas

		Página
2.1	Eventos relevantes en el desarrollo de la Ciencia de Datos--------	8
2.2	Programas académicos en Ciencia de Datos ----------	12
3.1	Tasas de clasificación incorrecta para la bases de datos Avisos de granja----------	15
3.2	Tasas de clasificación incorrecta para la bases de datos Trombina----------	16
3.3	Comparación de estudios relacionados con aplicaciones de Ciencia de Datos y diabetes ----------	31
4.1	Comparativa de las metodologías usadas en Ciencia de Datos---	42
5.1	Casos prácticos de predicción de tasas de mortalidad----------	46
5.2	Atributos para bases de datos de mortalidad----------	47
5.3	Atributos para base de datos del catálogo de enfermedades------	48
5.4	Atributos para bases de datos de población y proyección de población----------	48
5.5	Atributos para conjunto de datos de incidencias y tasa de mortalidad----------	48
5.6	Grupo C24----------	49
5.7	Grupo C08----------	49
5.8	Grupo C51----------	50
5.9	Estructura del conjunto de datos para implementar la regresión polinomial----------	50
5.10	Reporte de recopilación de datos----------	50
5.11	Características de bases de datos de mortalidad----------	51
5.12	Características de la base de datos del catálogo de enfermedades----------	52
5.13	Bases de datos poblacional (municipios con más de 100,000 habitantes) ----------	52
5.14	Bases de datos de proyección de población 2010-2030----------	53
5.15	Coeficientes de regresión del modelo F2-C24----------	62
5.16	Coeficientes de regresión del modelo F2-C08----------	63

5.17 Coeficientes de regresión del modelo F2-C51---------- 63
5.18 Coeficientes de regresión del modelo F3-C24---------- 64
5.19 Coeficientes de regresión del modelo F3-C08---------- 65
5.20 Coeficientes de regresión del modelo F3-C51---------- 66
5.21 Evaluación del modelo de regresión de grado 2 del grupo C24--- 68
5.22 Evaluación del modelo de regresión de grado 2 del grupo C08--- 69
5.23 Evaluación del modelo de regresión de grado 2 del grupo C51--- 70
5.24 Evaluación del modelo de regresión de grado 3 del grupo C24--- 71
5.25 Evaluación del modelo de regresión de grado 3 del grupo C08--- 72
5.26 Evaluación del modelo de regresión de grado 3 del grupo C51--- 73
A.1 Fuentes de información de Ciencia de Datos---------- 90

Lista de figuras

Página

1.1 Enfoque general de solución ---------- 4
2.1 Relación de la Ciencia de Datos con los procesos de una organización ---------- 10
2.2 Ciencia de Datos en el contexto de otras disciplinas---------- 11
4.1 Metodología Fundacional para la Ciencia de Datos---------- 34
4.2 Diagrama de flujo de la metodología FMDS---------- 35
4.3 Metodología DSP---------- 39
4.4 Fases del método científico---------- 40
5.1 Diagrama Entidad-Relación de base de datos de integración----- 56
5.2 Tasa promedio de mortalidad normalizada del Grupo C24-------- 57
5.3 Tasa promedio de mortalidad normalizada del Grupo C08-------- 58
5.4 Tasa promedio de mortalidad normalizada del Grupo C51-------- 59
5.5 Tasas promedio de mortalidad normalizada de los grupos C24, C08 y C51---------- 59
5.6 Modelo de la función de grado 2 para el grupo C24---------- 61
5.7 Modelo de la función de grado 2 para el grupo C08---------- 62
5.8 Modelo de la función de grado 2 para el grupo C51---------- 63

5.9	Modelo de la función de grado 3 para el grupo C24----------	64
5.10	Modelo de la función de grado 3 para el grupo C08----------	65
5.11	Modelo de la función de grado 3 para el grupo C51----------	66
5.12	Proyección de tasas de mortalidad para el grupo C24, modelo F2-C24----------	74
5.13	Proyección de tasas de mortalidad para el grupo C08, modelo F2-C08----------	75
5.14	Proyección de tasas de mortalidad para el grupo C51, modelo F2-C51----------	75
5.15	Proyección de tasas de mortalidad para el grupo C24, modelo F3-C24----------	76
5.16	Proyección de tasas de mortalidad para el grupo C08, modelo F3-C08----------	77
5.17	Proyección de tasas de mortalidad para el grupo C51, modelo F3-C51----------	77
C.1	Disciplinas de la Ciencia de Datos----------	95
D.1	Técnicas de clasificación----------	98
E.1	Lenguaje R----------	103
E.2	Lenguaje Phyton----------	103
E.3	Lenguaje Scala----------	104
E.4	Lenguaje SQL----------	105
E.5	Software Excel----------	105
E.6	Software SAS----------	106
E.7	Software SPSS----------	107
E.8	Uso de herramientas de Ciencia de Datos----------	107

Organización del documento

En esta sección se describe la organización de este documento con el objeto de facilitar la lectura del mismo. El presente documento posee seis capítulos y una sección de anexos. En el capítulo 1 se describe el problema a resolver, el contexto en el que desarrolla la investigación, se presenta el enfoque de solución, los objetivos, alcances y limitaciones establecidas. En el capítulo 2 se elabora una introducción teórica a la Ciencia de Datos, donde se incluyen definiciones y conceptos de esta área. El capítulo 3 de este documento integra algunas investigaciones desarrolladas aplicando la Ciencia de Datos, se incluyen casos prácticos de Ciencia de Datos desarrollados en coordinación con instituciones gubernamentales. Asimismo, aplicaciones enfocadas al análisis de la diabetes mellitus. En el capítulo 4 se presentan las metodologías propuestas para desarrollar aplicaciones de Ciencia de Datos, se realiza una comparación entre ellas y se selecciona una para desarrollar el caso práctico definido en el capítulo 1. El capítulo 5 describe el desarrollo del caso práctico utilizando la metodología de Ciencia de Datos seleccionada. Por último, en el capítulo 6 se exponen las conclusiones obtenidas de la investigación y se incluyen propuestas de temas para investigaciones posteriores.

Uno de los principales objetivos de esta investigación es crear una infraestructura de conocimiento que facilite la asimilación de los conceptos de la Ciencia de Datos. Esta infraestructura de conocimiento está descrita en los anexos de este documento. En el anexo A se presenta la reseña de algunas fuentes de información sobre Ciencia de Datos. En el anexo B se describen los principios de los que se apoya la Ciencia de Datos. El Anexo C muestra las disciplinas que integran la Ciencia de Datos. En el anexo D se explican algunas técnicas de Ciencia de Datos que apoyan el análisis de grandes cantidades de datos. Finalmente, en el Anexo E se describen las herramientas de Ciencia de Datos más utilizadas en el desarrollo de proyectos en esta área.

CAPÍTULO 1

INTRODUCCIÓN

La Ciencia de Datos es un área de la Computación enfocada en obtener conocimiento de grandes volúmenes de datos. Se sustenta en principios de diferentes disciplinas con el principal objetivo de apoyar la toma de decisiones basadas en los datos.

Esta investigación es la primera en la línea de investigación en Ciencia de Datos del Departamento de Ciencias Computacionales del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), atiende a la necesidad de resolver diversos problemas nacionales con aplicación del conocimiento tecnológico.

En este capítulo se presenta el contexto en el cual se desarrolla esta investigación. Se expone el problema a resolver y se presenta la propuesta general de solución. Asimismo, se incluyen los objetivos, alcances y limitaciones establecidos para esta investigación.

1.1. Contexto de la investigación

La Ciencia de Datos es un campo reciente. Se define como un área enfocada en adquirir conocimiento de datos de gran volumen, variedad y velocidad. Utiliza herramientas y aplica principios de diferentes disciplinas para proveer soluciones a problemas del mundo real [1, 2, 3]. Uno de sus principales objetivos es facilitar la toma de decisiones basadas en los datos. Ésta es empleada en muchos dominios tales como: energía, transporte, economía, finanzas, seguridad, salud, marketing, entre otros.

En el CENIDET se han realizado investigaciones en Minería de Datos y se han desarrollado tesis que son el antecedente de esta investigación. Algunas de las tesis relacionadas con Minería de Datos son:

1. **Minería de Datos Orientada al Big Data en el Área de Salud** [4]: El objetivo de este trabajo fue desarrollar un prototipo de Minería de Datos enfocado en el Sector Salud en el área de Epidemiología. La metodología de Minería de Datos utilizada fue *Cross Industry Standard Process for Data Mining* (CRISP-DM, por sus siglas en inglés). La técnica utilizada para realizar el análisis de datos fue el agrupamiento. En la investigación se encontraron grupos de interés que representan a regiones de municipios de México con altas tasas de mortalidad por diabetes mellitus para los años 2000 y 2010.
2. **Aplicación de Minería de Datos en el área de Salud Pública** [5]: En este trabajo se encontraron patrones de comportamiento en la tasa de mortalidad por diabetes mellitus tipos E11 y E14 en las regiones y municipios de México y condados de Estados Unidos para los años 2000 y 2010. Se utilizó la metodología CRISP-DM para guiar el proceso de Minería de Datos. Al igual que en [4], en esta investigación se utiliza la técnica de agrupamiento

Esta investigación tiene diferencias significativas con los trabajos mencionados anteriormente. En este trabajo se desarrolló una aplicación de Ciencia de Datos en el dominio de Salud Pública; se analizaron datos de mortalidad por diabetes mellitus en regiones de municipios de México en el periodo 1990-2015. Se utilizó la Metodología Fundacional para Ciencia de Datos (*FMDS*, por sus siglas en inglés) para guiar el proceso de desarrollo de la aplicación. El enfoque de la aplicación fue

predictivo ya que se proyectaron las tasas de mortalidad de regiones de municipios de México para el periodo 2016-2020.

1.2. Planteamiento de problema

El hombre ha utilizado los avances en la ciencia y tecnología para resolver problemas de muchos dominios. La ciencia y tecnología han avanzado a un ritmo acelerado en los últimos años, lo que ha permitido resolver problemas en menor tiempo y mayor calidad. Este avance desafía al hombre a actualizarse para aprovechar los beneficios que éstas ofrecen.

Una de las áreas de la Computación creadas recientemente es la Ciencia de Datos. Ésta posibilita la generación de soluciones efectivas a problemas específicos. Se apoya de varias disciplinas del conocimiento y el análisis riguroso de grandes cantidades de datos para lograrlo.

La Ciencia de Datos es un área nueva de conocimiento. Por ello, aún no existe un acuerdo sobre sus conceptos fundamentales, principios y disciplinas que la integran. Asimismo, la información de que se dispone es limitada. Como en toda área emergente, las metodologías que apoyan el desarrollo de aplicaciones de Ciencia de Datos aún no están completamente maduras a causa del proceso natural de madurez de la misma. Sin embargo, existen problemas que deben ser resueltos en esta área. Se requiere conocer los conceptos fundamentales y metodologías que apoyan a la Ciencia de Datos.

En esta investigación se plantea generar una infraestructura de conocimiento que facilite la asimilación de los conceptos de la Ciencia de Datos (ver Anexos A, B, C, D y E). En este sentido, se investigará sobre sus conceptos, principios, disciplinas que la integran, las metodologías y herramientas en las que se apoya; de acuerdo con la información descrita en la literatura especializada. Se realizará la validación del conocimiento asimilado mediante el desarrollo de una aplicación utilizando una metodología específica de Ciencia de Datos.

1.3. Enfoque general de solución

En la Figura 1.1 se muestra el diagrama que representa el enfoque general de solución propuesto para desarrollar esta investigación.

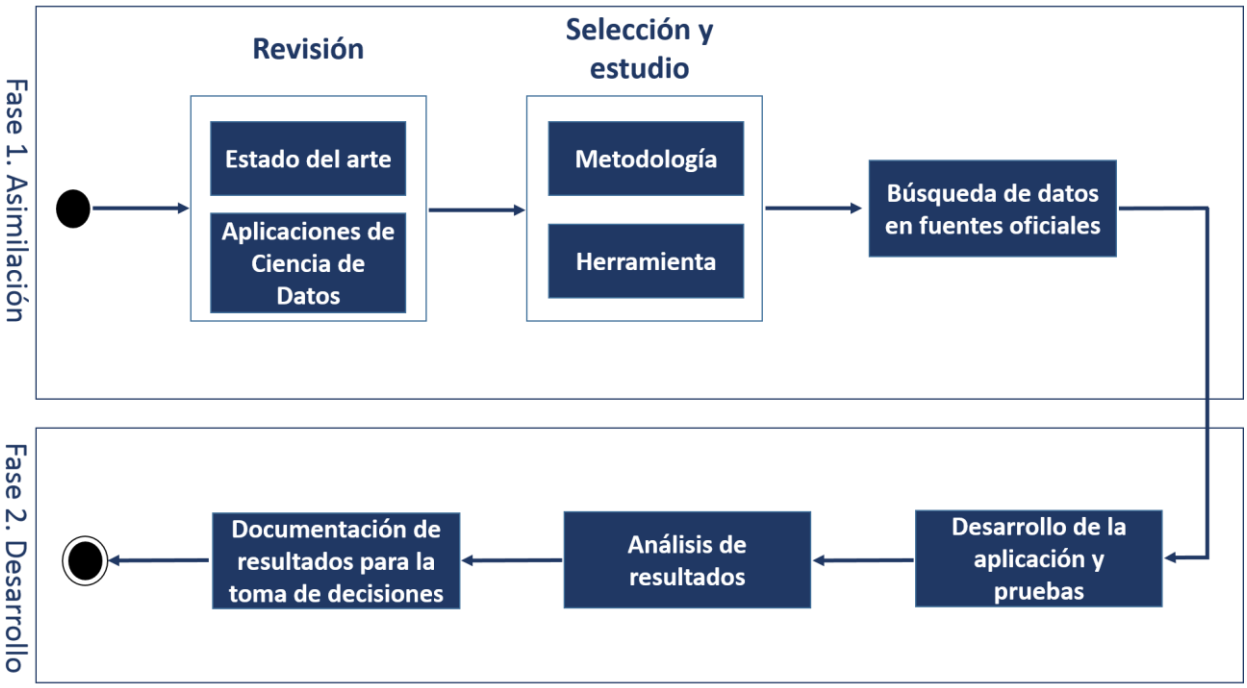

Figura 1.1 Enfoque general de solución

La investigación se dividió en 2 fases. La primera se enfocó en la asimilación de los conceptos fundamentales de la Ciencia de Datos. Con ese fin se hizo una revisión del estado del arte en el tema de Ciencia de Datos y de su aplicación en diversos dominios. Se seleccionó y estudió la Metodología Fundacional para Ciencia de Datos (*FMDS*, por sus siglas en inglés) y la herramienta R –lenguaje de programación estadística- como apoyo al desarrollo de la aplicación de Ciencia de Datos. La última actividad de esta fase fue la recopilación de los datos que se utilizaron en la aplicación.

En la segunda fase se desarrolló una aplicación de Ciencia de Datos dentro del campo de la Salud Pública, tomando como caso la enfermedad conocida como diabetes mellitus. Para realizar el análisis, se aplicó la metodología FMDS. Se probó la aplicación y se analizaron los resultados obtenidos.

1.4. Objetivos

El objetivo general de esta investigación es mostrar que es factible asimilar los conceptos de Ciencia de Datos y desarrollar una aplicación en el área de salud.

Los objetivos específicos planteados para esta investigación fueron los siguientes:

1. Asimilar los conceptos generales relacionados con la Ciencia de Datos,
2. Aplicar los conceptos de la Ciencia de Datos y utilizar una de sus metodologías para desarrollar una aplicación con el objetivo de realizar una proyección de la tasa de mortalidad por diabetes mellitus tipos E11-E14 en regiones de municipios de México con altas tasas de incidencia para el periodo 2016-2020.

1.5. Alcances y limitaciones

1.5.1. Alcances

Los alcances planteados para esta investigación son los siguientes:

a) Se seleccionará y empleará una metodología y herramienta para desarrollar la aplicación de Ciencia de Datos,

b) Se implementará la aplicación computacionalmente,

c) El caso de estudio se enfocará en las defunciones a nivel municipal por diabetes mellitus,

d) Se utilizarán bases de datos de fuentes oficiales del periodo 1990-2015.

1.5.2. Limitaciones

Las limitaciones establecidas para esta investigación son:

a) La aplicación será desarrollada sólo para mostrar las fases de la Ciencia de Datos,

b) Se realizarán pruebas de la aplicación únicamente en equipo disponible en el CENIDET.

c) La validación de los resultados será de manera experimental.

CAPÍTULO 2

CONCEPTOS BÁSICOS DE CIENCIA DE DATOS

La Ciencia de Datos es un área emergente de conocimiento, el interés en ella se ha incrementado y cada vez se genera más información que permite entender mejor sus alcances, límites, objetivos, entre otros. En este capítulo se elabora una introducción teórica a la Ciencia de Datos. Asimismo se incluyen definiciones y conceptos acerca de esta temática y de otras áreas relacionadas con el análisis de datos.

2.1. Introducción a la Ciencia de Datos

Ciencia de Datos es un área que cada vez se utiliza con mayor frecuencia en distintos campos del conocimiento, tales como salud pública, finanzas, transporte, comunicaciones, educación, entre otros. En 1974, Naur, en su libro *Concise Survey of Computer Methods* utilizó el concepto de Ciencia de Datos como sustituto *de* las Ciencias Computacionales. En 1996, se utilizó por primera vez el término *Ciencia de Datos* en una conferencia de la Federación Internacional de Sociedades de Clasificación (*IFCS*, por sus siglas en inglés) titulada *Ciencia de Datos, clasificación y métodos relacionados*. Wu, en 1997 afirmó que la estadística debía renombrarse como *Ciencia de Datos* y los estadísticos como científicos de datos.

Fue hasta 2001 cuando Cleveland publicó el artículo *Data Science: an action plan for expanding the technical áreas of the field of Statistics*, en el cual introdujo a la Ciencia de Datos como una disciplina independiente. Con ello extendió el campo de la estadística al incluir a la computación para mejorar el análisis de los datos. En 2002 se creó el *Data Science Journal* con el objetivo de realizar descripciones a los sistemas de datos, su publicación en internet, aplicaciones y usos legales. A partir de ese momento, la Ciencia de Datos comenzó a generar interés tanto en el sector empresarial como en el académico. En la Tabla 2.1 se sintetiza la historia de la Ciencia de Datos.

Tabla 2.1 Eventos relevantes en el desarrollo de la Ciencia de Datos

Año	Acontecimiento
1962	John W. Tukey publicó el artículo *The future of data analysis*. Resalta el incremento de la importancia del análisis de datos.
1974	Peter Naur publicó *Concise survey of Computer Methods in Sweden and the United States*, en esta publicación usa por primera vez el término "**Data Science**", definiéndola como: La ciencia de trabajar con datos, una vez que se ha establecido la relación de los datos y lo que representan se delega a otros campos y ciencias.
1977	Se fundó *The International Association for Statistical Computing (IASC)*, cuya misión fue relacionar la estadística tradicional, la computación moderna y el conocimiento del dominio de los expertos con el objetivo de convertir los datos en información y conocimiento.
1989	Gregory Piatetsky-Shapiro organizó y presidió el primer taller *Knowledge Discovery in Databases (KDD)*, que en 1995 se convirtió en *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).*
1994	La revista BussinessWeek publicó un artículo titulado *Database Marketing*, en éste se menciona que las compañías están generando grandes cantidades de datos y los utilizan para realizar promociones especializadas.
1996	La revista *International Federation of Classification Societies (IFCS),* tituló a una conferencia "Data Science, classification and related methods".
1997	C. F. Jeff Wu propuso que la estadística sea nombrada Ciencia de Datos y a los estadísticos sean nombrados científicos de datos.
2001	William S. Cleveland publicó *Data Science: an action plan for expanding the technical áreas of the field of Statistics,* en la que propuso crear una nueva área al expandir la estadística y computación para mejorar el análisis de los datos, esta área se llamaría "Ciencia de Datos".
2002	Se creó el *Data Science Journal*, la primera revista del área.
2003	Se creó el *Journal of Data Science,* su objetivo es proveer una plataforma para intercambio de ideas entre quienes trabajan con datos.
2005	Thomas H. Davenport, Don Cohen y Al Jacobson publicaron *Competing in Analytics,* describieron el surgimiento de una nueva forma de competencia basada en el uso extensivo del análisis de datos y toma de decisiones basadas en ellos.
2007	Se estableció *The Research Center fot Dataology and Data Science*
2009	Troy Sadkoswky creó el grupo de científicos de datos de LinkedIn
2009	Kirk D. Bourne publicó el artículo *The Revolution in Astronomy Education: Data Science for the masses,* donde afirmó que los especialistas deben aprender a aplicar las nuevas técnicas de investigación de Ciencia de Datos.
2010	Kenneth Cukier escribió en *The Economist Special Report:* un nuevo tipo de profesional ha emergido, el científico de datos, quien combina las habilidades de un programador de sistemas, es estadístico y todo un artista en extraer el oro escondido bajo las montañas de datos.
2010	Mike Loukides publicó el libro *What is Data Science?,* donde se incluyen conceptos fundamentales de Ciencia de Datos.
2010	Drew Conway diseñó *The Data Science Venn Diagram,* en ella incluye a las disciplinas: computación, Matemática-Estadística y conocimiento del dominio.
2011	Harlan Harris publicó *Data Science, Moore´s Law and Moneyball.*
2011	D. J. Patil escribió el libro *Building Data Science Teams.*
2012	Tom Davenport y D. J. Patil publicaron el artículo *Data Scientist: the sexiest Job of the 21st Century.*
2013	Se creó el taller *Data Science for Social Good* de la Universidad de Chicago.
2014	Se creó el laboratorio *Imperial Bussines Analytics* del Imperial College London.
2016	La empresa O´Reilly Media realizó la encuenta *Data Science Salary Survey* para conocer el salario de los científicos de datos alrededor del mundo.
2018	Se celebró en Chile la conferencia "Women in Data Science" de la Universidad de Stanford.

2.2. Definiciones y conceptos

Con la gran cantidad de datos disponible actualmente, la mayoría de las industrias se han enfocado en explotarlos para tomar ventaja frente a sus competidores. El volumen, variedad y velocidad con que se generan los datos ha sobrepasado las capacidades de los sistemas tradicionales para analizar los datos. Al mismo tiempo, las computadoras se han vuelto más poderosas y cada vez se están desarrollando nuevos algoritmos y mejorando los existentes para realizar análisis de datos más profundos. Otro aspecto a tomar en cuenta es que las empresas e instituciones requieren tomar más y mejores decisiones. Las preguntas que se hacen las personas que toman decisiones cada vez son más complejas, por lo que se necesitan formular soluciones eficaces a partir de un análisis riguroso de los datos. La convergencia de estos fenómenos ha dado lugar a la Ciencia de Datos. [1]

La Ciencia de Datos se conceptualiza como: "Área enfocada a adquirir conocimiento de grandes cantidades de datos, al utilizar herramientas y al aplicar principios de diferentes disciplinas para proveer soluciones a problemas del mundo real" [1], [2].

Los objetivos de la Ciencia de Datos son [3]:

a) Facilitar la toma de decisiones basadas en los datos (DDD, por sus siglas en inglés),
b) Descubrir patrones,
c) Predecir el futuro,
d) Entender el pasado/presente de las personas y el mundo,
e) Crear nuevas industrias/productos basados en los datos.

La Ciencia de Datos ha sido confundida con disciplinas similares tales como la Minería de Datos, *Big Data*, *Bussiness Intelligence,* entre otros. Entre ellas existen similitudes y diferencias. La Figura 2.1 [1] muestra el papel de la Ciencia de Datos en el contexto de varios procesos relacionados dentro de una organización. La toma de decisiones basadas en los datos se refiere a la práctica de tomar decisiones basadas en el análisis de los datos más que sólo en la intuición [6].

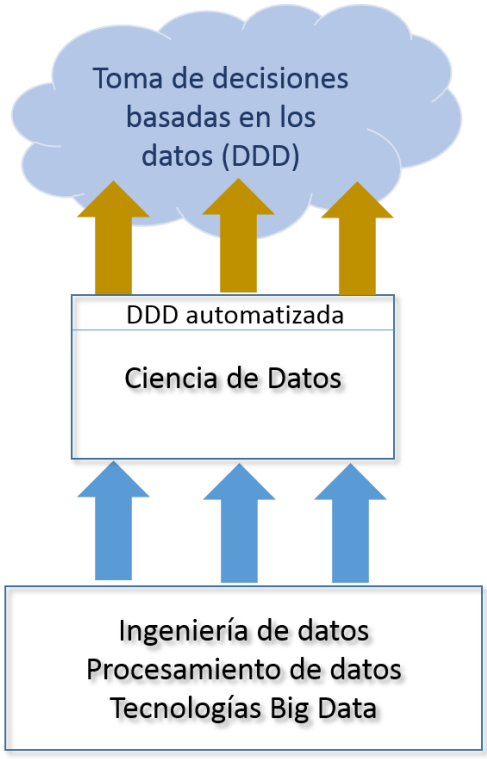

Figura 2.1 Relación de la Ciencia de Datos con los procesos de una organización

Los conocimientos de la ingeniería y el procesamiento de datos son fundamentales para comprender a la Ciencia de Datos. Actualmente, muchas técnicas, sistemas y tecnologías de procesamiento de datos a menudo son erróneamente llamadas Ciencia de Datos. Esta necesita acceder a los datos y se beneficia de tecnologías para su procesamiento. Las tecnologías de procesamiento de datos son importantes para muchas tareas de negocios orientados a los datos, tareas que no involucran extracción de conocimiento o toma de decisiones basadas en datos, tales como procesamiento eficiente de transacciones, procesamiento de sistemas web modernos y administración de campañas de publicidad *online*.

Las tecnologías *Big Data* (tales como *Hadoop, HBase* y *MongoDB*), recientemente han recibido atención considerable. Estas tecnologías se usan para otras tareas, incluyendo la ingeniería de datos. Sin embargo, son comunes para el procesamiento de datos en apoyo a las técnicas de minería de datos y otras actividades de la Ciencia de Datos como se muestra en la Figura 2.1.

La Figura 2.2 representa a la Ciencia de Datos en el contexto de otras disciplinas que están estrechamente relacionadas con ella. Utiliza el conocimiento de otras disciplinas para lograr su objetivo; mientras que, por ejemplo, la Minería de Datos encuentra patrones y relaciones en los datos, la Ciencia de Datos utiliza esos patrones para proponer soluciones que faciliten la toma de decisiones.

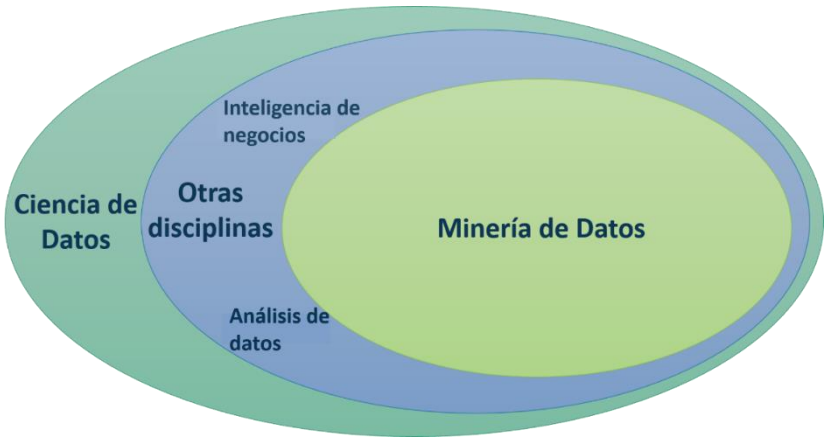

Figura 2.2 Ciencia de Datos en el contexto de otras disciplinas

2.3. Desarrollo de la Ciencia de Datos

Uno de los principales objetivos de la Ciencia de Datos es apoyar la toma de decisiones basadas en los datos (*Data-Driven Decision Making*), la cual se refiere a la práctica de tomar decisiones basadas en el análisis de los datos más que sólo en la intuición. En el estudio “Strenght in Numbers: How Does Data-Driven Decision Making Affect Firm Performance?” [7] se desarrolló un modelo para medir el efecto del uso de la Toma de Decisiones Basados en los Datos en los diferentes procesos de la empresa. Usando los datos financieros públicos y privados de inversión tecnológica de 179 empresas registradas en Estados Unidos, se examinaron las relaciones entre DDD y productividad, comportamiento financiero y valor de mercado. Los resultados mostraron que DDD está asociado a un incremento de entre un 5% y un 6% en la productividad de las empresas.

Los resultados de [7] reflejan la importancia de utilizar la Ciencia de Datos en los procesos de una organización, ya sea que genere pocas o grandes cantidades de datos. Esta situación ha propiciado que muchas organizaciones empresariales utilicen cada vez más la Ciencia de Datos para la toma de decisiones.

El sector académico también ha puesto mucho interés en la Ciencia de Datos. A partir de que en 2002 se creara el *Data Science Journal* y en 2003 el *Journal of Data Science,* se han promovido congresos dedicados a explorar los temas inherentes a la Ciencia de Datos.

Se han creado laboratorios e institutos especializados en Ciencia de Datos, como el *Data Science Lab* de la Universidad John Hopkins, formado por un grupo de investigadores, educadores, desarrolladores de software y expertos en creación de contenido para hacer accesible la Ciencia de Datos al mundo [8]. El *Data Science for Social Good* de la Universidad de Chicago se enfoca en resolver problemas sociales a través de la Ciencia de Datos uniendo al sector gubernamental con el académico [9]. El *Data Science Institute* de la Universidad de Columbia tiene la misión de avanzar en el estado del arte de la Ciencia de Datos, transformar todas las áreas, profesiones y sectores a través de la aplicación de la Ciencia de Datos y asegurarse del uso responsable de los datos para beneficio de la sociedad [10].

Las universidades –principalmente de los Estados Unidos- han creado programas especializados, tanto de licenciatura como de posgrado, para satisfacer la demanda de especialización en Ciencia de Datos. Algunos de los programas creados se muestran en la Tabla 2.2.

Tabla 2.2 Programas académicos en Ciencia de Datos

Universidad	Programa
Carnegie Mellon University	Master of Computational Data Science
University of California, Berkeley	Master of Information and Data Science
New York University	Master of Science in Data Science
Stanford University	Master of Science in Statistics: Data Science
Columbia University	Master of Science in Data Science

En México -aunque la Ciencia de Datos aún está en desarrollo- hay instituciones públicas y privadas donde se imparten maestrías y doctorados en Ciencia de Datos. Una de las primeras instituciones en ofrecer un posgrado en Ciencia de Datos fue el ITAM (Instituto Tecnológico Autónomo de México), el cual ofrece una Maestría en Ciencia de Datos. El INFOTEC (Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación) ofrece la maestría en Ciencia de

Datos en las modalidades presencial (Maestría en Ciencias en Ciencia de Datos, MCCD) y a distancia (Maestría en Ciencia de Datos e Información, MCDI), también ofrece un doctorado en Ciencia de Datos [11].

El futuro de la Ciencia de Datos es prometedor. En 2012, la Revista de Negocios de Harvard (*Harvard Business Review*) publicó un artículo donde nombró al científico de datos como el trabajo más rentable y demandado del siglo XXI. En éste se resaltan las cualidades y responsabilidades que tienen los científicos de datos, así como la importancia que tienen para las empresas de hoy en día [12]. Lo anterior ha derivado en el incremento de la demanda por parte de empresas para contratar científicos de datos y en el incremento del salario para los mismos.

CAPÍTULO 3

ESTADO DEL ARTE

En este capítulo se integran investigaciones desarrolladas aplicando Ciencia de Datos en países como México, Estados Unidos, China, Arabia Saudita, Australia, entre otros. Se incluyen casos prácticos de Ciencia de Datos en diferentes dominios y específicamente en el dominio de la salud. Estos últimos aplicados al análisis de la diabetes mellitus.

3.1. Trabajos relacionados

Con el presente trabajo, en el CENIDET se inicia una serie de tesis donde se realizan investigaciones en Ciencia de Datos. En el ITAM se han realizado investigaciones en diferentes dominios aplicando la Ciencia de Datos. Algunas de ellas se muestran a continuación:

1. **Estudio numérico de algoritmos a gran escala para aprendizaje automático**

En [13] se describieron los modelos de *Máquina de Soporte Vectorial* (MSV) y *Regresión Logística* (RL) utilizadas en el *Aprendizaje Automático* para resolver problemas de clasificación. Se proponen algunos cambios para optimizar (reducir el tiempo de ejecución) el método de *Puntos Interiores* (del modelo MSV) mediante su implementación en paralelo y del método *Descenso por Coordenadas* (del modelo RL). Para la implementación de ambos métodos se utilizó la librería *glmnet* del lenguaje R. Se utilizaron las bases de datos *Avisos de granja* (avisos de texto encontrados en doce sitios de internet que involucraban temas relacionados con animales de granja en formato de bolsa de palabras) y *Trombina* (proporcionada por los laboratorios de farmacéutica *DuPont*). Ambas bases de datos se separaron en conjuntos de entrenamiento y de prueba. Los resultados obtenidos (tasa de clasificación incorrecta) en ambos métodos para la base de datos *Avisos de granja* se muestran en la Tabla 3.1.

Tabla 3.1 Tasas de clasificación incorrecta para la bases de datos *Avisos de granja*

Conjunto	MSV	RL
Entrenamiento	1.072e-3	2.681e-3
Prueba	7.729e-2	5.797e-2

Los resultados obtenidos (tasa de clasificación incorrecta) en ambos métodos para la base de datos *Trombina* se muestran en la Tabla 3.2.

Tabla 3.2 Tasas de clasificación incorrecta para la bases de datos *Trombina*

Conjunto	MSV	RL
Entrenamiento	1.047e-3	2.2e-3
Prueba	2.3659e-1	2.3659e-1

De los resultados obtenidos se concluyó que ambos métodos presentan bajas tasas de clasificación incorrecta, por lo que se sugiere su uso para problemas de clasificación en bases de datos de gran escala.

2. **Métodos de escalamiento en optimización en Máquinas de Soporte Vectorial (MSV)**

En esta investigación [14] se estudiaron dos métodos de escalamiento usados en el modelo de *Máquina de Soporte Vectorial* (MSV), los cuales son el método de *Puntos Interiores* y *Optimización Mínima Sucesiva* (OMS). Se presentan conceptos, métricas y técnicas de calendarización en cómputo distribuido y se comenta sobre el modelo de cómputo de *MapReduce*. Se propone paralelizar el método de *Puntos Interiores* como una estrategia para el entrenamiento de MSV. Según su implementación, esta estrategia presenta una gran oportunidad de optimización puesto que se consigue una reducción en términos de almacenamiento y también en cuanto a complejidad logarítmica. También se explica la forma de implementar el método OMS en paralelo.

3. **Importancia de factores socioeconómicos en la clasificación de desastres hidrometereológicos**

En este estudio [15] se desarrolló un proyecto de Ciencia de Datos para recopilar todas las variables involucradas en el proceso de declaratoria de un desastre para determinar qué variables son las que influyen para que se active o no el Fondo Nacional para la Atención de Desastres Naturales (FONDEN). Se utilizaron datos socioeconómicos, meteorológicos y políticos de México para el periodo de estudio 2003-2013. Para determinar la importancia de las variables envueltas en el proceso de atención a desastres, se usaron diversas técnicas de aprendizaje de máquina tales como *Árboles, Bosques Aleatorios, Bosques*

Adaptativos y *Regresión Logística*. Para realizar el análisis de los datos se utilizó la herramienta R. Los resultados indicaron que las variables con mayor peso en la declaración de un desastre son : los daños netos, el nivel de lluvia, los cultivos afectados, la tasa de mora de las pequeñas y medianas empresas (PYMES) así como el partido político que gobierna a nivel estatal.

4. Sistema de recomendación de hoteles similares

En otro estudio [16] se desarrolló un sistema de recomendación de hoteles similares "inteligente, flexible y empático" para la empresa *Best Day Travel*. Se analizaron los sistemas de recomendación de hoteles de *Best Day* y de sus competidores (*Price Travel, Booking, Despegar, Expedia,* y *Trip Advisor*). También se explicaron los criterios que se utilizaron para desarrollar el sistema y se presentaron sus modelos matemáticos. La información utilizada fue tomada de la base de datos de *Best Day*. Para acceder a los datos se utilizó el lenguaje *Structured Query Language* (SQL, por sus siglas en inglés) y para analizarlos se empleó el lenguaje R. Se describieron las propiedades teóricas del modelo, así como los resultados y el desempeño del sistema de recomendación una vez implementado.

5. Cómo detectar corrupción, colusión y fraude en contratos que otorga el Banco Mundial

En esta tesis [17] se presenta el desarrollo de un sistema para detectar corrupción, colusión y fraude en los contratos que otorga el Banco Mundial a los distintos países del mundo. Presenta información del proceso en el que el Banco Mundial otorga los préstamos. Los datos utilizados en este proyecto provienen de fuentes privadas y públicas; la base de datos privada (la más valuable) fue la de investigaciones de la Vicepresidencia de Integridad del Banco Mundial, aunque con un acceso remoto a los datos a través del servicio *Amazon Elastic Compute Cloud* (Amazon EC2) perteneciente a *Amazon Web Sevice´s* (AWS, por sus siglas en inglés); las bases de datos públicas fueron: *World Bank´s Development Indicators* (WDI, por sus siglas en inglés) y *Historic & Major Awards*, ambas pertenecientes al Banco Mundial. Los lenguajes utilizados para realizar las tareas de preparación de datos y modelado fueron Phyton y R. Los modelos que se tomaron en cuenta para

clasificar los contratos (clasificar si pertenecían a corrupción, colusión, fraude o ninguno de los anteriores) fueron los de *Máquina de Soporte Vectorial* (MSV), *Bosque Aleatorio* y *Regresión Logística*. También se mostraron las interfaces de la aplicación desarrollada, la cual se ejecuta en los servidores del Banco Mundial.

6. Propuesta para la creación del área de Analítica Avanzada en Grupo Bolsa Mexicana de Valores

En otra investigación [18] se promovió la creación del área de analítica avanzada para la Bolsa Mexicana de Valores (BMV). La propuesta realizada implica un importante esfuerzo de inversión y replanteamientos de estrategia por parte de los siguientes ejes:

- Negocios: Se requerirá contar con un portafolio de casos de negocio con metas y objetivos bien definidos,
- Tecnológico: Consiste en mejorar los recursos tecnológicos actuales en BMV o migrar a otra tecnología que cumpla con las restricciones impuestas, se analizaron las tecnologías: *Teradata, Oracle, IBM, EMC* y *Cloudera,*
- Humano: Se presentó un organigrama con los puestos necesarios para el funcionamiento del área de analítica avanzada y se propone la capacitación del personal en lenguajes como *SQL y Phyton*.

Se desarrolló un prototipo que consistió en un sitio Web que expone algunas de las métricas más representativas del mercado de capitales y derivados en BMV y se presentan sus interfaces. Las fuentes de datos provienen principalmente del datawarehouse en producción de BMV (a través de consultas SQL) y de fuentes de noticias bursátiles como *Yahoo*. Para la manipulación de los datos se utilizaron los lenguajes R y Phyton.

7. Etiquetador automático de contenido editorial en español con BM25

El objetivo de esta tesis [19] fue desarrollar un prototipo de etiquetador para clasificar el contenido editorial en español generado por el Grupo Expansión. Se mencionan y describen los métodos para la *Recuperación de Información* (IR, por sus siglas en inglés), los cuales son: algebraicos, probabilísticos y de Aprendizaje de Máquina. La colección de documentos con la que se trabajó fue de 3, 895

etiquetas generadas por los editores de Grupo Expansión y se utilizó el modelo probabilístico *Okapi Best Match 25* (BM25) para realizar la clasificación. Para ejecutar las actividades del proyecto (obtención, gestión y limpieza de datos, generación de Score con *BM2* y *TF/IDF* y podado de etiquetas sugeridas) se utilizó el lenguaje R, se utilizó *ShinyDashboard* –paquete del lenguaje R- para mostrar los resultados obtenidos. También se explica el procedimiento para realizar 205 encuestas enviadas a cinco personas para seleccionar la mejor recomendación de contenido entre el modelo *BM25, TF/IDF* o "ninguna". Los resultados mostraron que el 60% de las personas seleccionaron el modelo BM25, mientras que el modelo TF/IDF fue seleccionado el 25%, la opción "ninguna" obtuvo el 15%.

Las investigaciones desarrolladas en el ITAM analizadas anteriormente, aunque son de diferentes dominios, tienen una similitud con el presente trabajo. Todas las investigaciones presentadas –a excepción de Métodos de escalamiento en optimización en Máquinas de Soporte Vectorial (MSV) – utilizan como herramienta para la preparación y modelado de datos el lenguaje R, al igual que se emplea en este trabajo.

3.2. Casos prácticos de Ciencia de Datos

Tanto empresas como instituciones académicas y gubernamentales se han percatado de la importancia de analizar los datos y tomar decisiones con base a ellos. Cada vez se está utilizando más la Ciencia de Datos en diferentes dominios de aplicación; desde la salud, astronomía, transporte, finanzas, hasta la metereología, entre otros.

Desde hace dos décadas se empezó a aplicar la Ciencia de Datos en países como Estados Unidos, España, Francia e Inglaterra. En México, su uso ha ido en aumento. Se han creado programas de posgrado en instituciones tanto públicas como privadas; se han fundado empresas que aplican la Ciencia de Datos; y se han desarrollado proyectos de Ciencia de Datos en el sector social. Algunos proyectos con aplicación de Ciencia de Datos desarrollados en México por instituciones gubernamentales son los siguientes [9]:

1. **Mejora en la Distribución de Servicios Sociales**

La Secretaría de Desarrollo Social (SEDESOL) proporciona una variedad de servicios sociales a los ciudadanos necesitados. En la actualidad están combinando datos de hogares, beneficiarios y geográficos para construir un nuevo sistema que les ayude a prestar servicios a las personas que más los necesitan. En 2016, el proyecto Ciencia de Datos para el Bien Social (*DSSG*, por sus siglas en inglés) se unió a SEDESOL para desarrollar un sistema con el propósito de mejorar las condiciones de vida de las poblaciones pobres en México.

DSSG ayudó a SEDESOL a concentrarse en tres metas, todas basadas en una orientación más precisa de los programas de servicios sociales a individuos y familias elegibles. Se usaron datos para identificar individuos que califican para recibir apoyos de programas en particular – aunque no los hayan usado. También se utilizaron conjuntos de datos combinados para predecir con mayor exactitud las necesidades de los hogares e informar potencialmente sobre el diseño de nuevos programas de servicios sociales. Además, se detectó a las personas que tienen ingresos insuficientes, para recibir asistencia.

2. **Mejorar la respuesta del gobierno a las solicitudes de los ciudadanos en línea**

Desde la promulgación de la Constitución de 1917, todos los ciudadanos mexicanos tienen derecho a solicitar información a las instancias de gobierno y recibir una respuesta oficial satisfactoria. Estas solicitudes comprenden desde serios llamamientos por servicios públicos hasta solicitudes frívolas, como invitar al Presidente a una fiesta de cumpleaños. Sin embargo, todas las peticiones, independientemente de la importancia o los medios de presentación, se revisan y se responden manualmente, un proceso a menudo ineficiente y lento para hacer frente a más de un millar de solicitudes al mes.

DSSG trabajó con la Oficina de la Presidencia de México para mejorar este sistema como parte de su Estrategia Digital Nacional (EDN). Con ese fin, se han creado nuevos algoritmos para clasificar las peticiones en función de su contenido e importancia y encaminarlas a la agencia gubernamental correcta para automatizar

parcialmente las respuestas a las solicitudes comunes. El proyecto utilizó el *Aprendizaje Automático* para clasificar más de 27.000 peticiones digitales anónimas y el análisis histórico de la página web *gob.mx* para revelar patrones en los datos. Estos conocimientos ayudaron al Gobierno de México a identificar las necesidades de la comunidad y proporcionar servicios específicos, respuestas más rápidas y reducir la asimetría de información entre el gobierno y los ciudadanos.

3. Mejorar la solidez financiera a largo plazo identificando las causas del abandono del hogar

INFONAVIT (Instituto del Fondo Nacional de la Vivienda para los Trabajadores) es el mayor proveedor de hipotecas en México, atiende a familias de bajos ingresos que no pueden obtener financiamiento de una institución privada para adquisición y otras soluciones de vivienda. El principal objetivo del INFONAVIT es aumentar la calidad de vida y el valor patrimonial de los trabajadores mexicanos y sus familias a través de dos mandatos: proveer financiamiento para la vivienda y administrar los ahorros de los trabajadores.

Para avanzar en esta misión, la organización quiere entender la relación entre la política, las influencias sociales y el abandono de la vivienda. La investigación encontró que los trabajadores abandonan su hogar por varias razones, incluyendo la distancia a los trabajos y las escuelas, la carencia de servicios, de finanzas, y de seguridad. Causas de abandono del hogar también varían según la región, y de acuerdo con si el propietario compró la propiedad para la inversión o para satisfacer una necesidad de vivienda.

El proyecto se basó en los resultados preliminares, explorando datos del INFONAVIT, censos y encuestas domiciliarias, préstamos y en una investigación social para identificar los factores que elevan el riesgo de abandono de viviendas. DSSG incorporó esos hallazgos en herramientas y recomendaciones basadas en evidencia, ayudando a INFONAVIT a ofrecer servicios y apoyar políticas locales para mitigar el abandono y mejorar los ingresos económicos para los ciudadanos de México. Los resultados también ayudarán a mejorar el valor de la vivienda y el origen del crédito para los trabajadores, así como mejorar la gestión

del riesgo de la cartera y el desempeño de la recaudación para INFONAVIT, para mejor su misión social.

4. Reducción de las tasas de mortalidad materna

Las muertes maternas en México por complicaciones de embarazo, parto o posparto han disminuido de 89 muertes por 100.000 nacidos vivos en 1990 a 43 en 2011. A pesar de esta mejora, la tasa de disminución se ha desacelerado significativamente y México no estaba en camino de alcanzar su Objetivo de Desarrollo del Milenio, cuyo objetivo es reducir la mortalidad materna en un 75% para 2015.

El objetivo del proyecto fue identificar los factores que contribuyen a la mortalidad materna y determinar qué se podría hacer para reducirla. Si bien el proyecto inicial se centró en los municipios y localidades, la intención en 2014 fue desarrollar modelos de riesgo a nivel individual utilizando todos los datos disponibles.

Los proyectos anteriores son relevantes porque fueron aplicados por instituciones de gobierno en México y cuyos objetivos fueron el beneficiar a la sociedad. Con ello se facilita la identificación de los problemas relacionados con ese sector.

Cabe destacar que estos proyectos fueron financiados por la Universidad de Chicago e instituciones públicas, por lo que la información que existe de ellos no incluye las bases de datos utilizadas, las técnicas de modelado y las herramientas empleadas para desarrollar dichos proyectos.

3.3. Casos prácticos en el dominio de la Salud Pública

En esta sección se presentan las investigaciones que se han realizado con enfoque en el análisis de casos de diabetes mellitus en países como Estados Unidos, China, Canadá, entre otros; aplicando la Ciencia de Datos o áreas relacionadas.

1. **Data Mining Technologies for Blood Glucose and Diabetes Management [20]**

Se describen los métodos de Minería de Datos usados en el análisis de datos para mejorar la administración de la glucosa y la diabetes mellitus. Las áreas de análisis son: a) monitoreo de glucosa en casa de pacientes con diabetes y b) monitoreo de glucosa de pacientes en unidades de hospitalización.

Para la interpretación de datos de glucosa en pacientes con diabetes se sugiere la aplicación de la técnica de Inteligencia Artificial llamada Abstracciones Temporales (TA`s, por sus siglas en inglés), la cual provee una representación basada en intervalos de datos monitoreados. Para el monitoreo de glucosa, se aplican tres tipos de abstracciones temporales, las cuales son las TA`s estado, TA`s complejos y TA`s de tendencia. Para la interpretación de los datos de glucosa en pacientes en unidades de hospitalización, se aconseja dividirlos en tres secciones:

1. Pacientes en riesgo,
2. Rol de la última medida de glucosa,
3. Calidad del proceso regulatorio.

En este estudio se concluyó que el enfoque de Minería de Datos ayudó a mejorar el análisis de glucosa. Asimismo, existen diferencias cuando los datos se originan del monitoreo en casa del paciente y cuando son registrados en hospitales. Es factible aplicar los métodos presentados en el análisis de varios tipos de datos coleccionados por organizaciones de salud.

2. **Type 2 Diabetes Mellitus Trajectories and Associated Risks [21]**

Se presenta un método para observar las trayectorias de la diabetes mellitus tipo 2, usando datos de un Registro Electrónico de Salud (*EHR*, por sus siglas en inglés). Los datos se tomaron del *Rochester Epidemiology Project*, específicamente del hospital *Mayo Clinic* en un periodo de 15 años (1999-2013). Para este estudio se recolectaron datos de 69, 747 pacientes.

Para desarrollar el estudio, los datos se dividieron en dos periodos:

- Periodo base (1999-2004),

- Periodo seguido (2005-2013).

Por otra parte, se excluyeron los registros de los pacientes que presentaban los siguientes casos:

- Pacientes con diabetes = 389,
- Pacientes con glucosa desconocida = 14,559,
- Pacientes con lípidos desconocidos = 1,023,
- Pacientes con presión sanguínea desconocida = 498,
- Pacientes que no sobrevivieron a 5 años= 10,089.

Del total de pacientes iniciales, al final quedaron 43,509 casos, con los cuales se desarrolló el estudio.

También se definieron trayectorias para determinar el orden de los factores de riesgo (Hiperlipidemia (HLD, por sus siglas en inglés), hipertensión (HTN, por sus siglas en inglés) / glucosa (IFG, por sus siglas en inglés), diabetes (DM, por sus siglas en inglés) de la siguiente manera:

- Trayectoria típica: HLD / HTN / IFG / DM.
- Trayectoria atípica: orden cambiante de factores de riesgo.

Para encontrar la relación en las trayectorias de los factores de riesgo se utilizó el **Modelo de regresión logística multivariada** usando datos demográficos, nivel de glucosa, comorbilidades escalonadas y tres trayectorias.

De esta investigación se concluyó que se pueden inferir la progresión de enfermedades usando datos almacenados por un EHR.

3. Application of Data Mining Methods and Techniques for Diabetes Diagnosis [22]

Se exploran los métodos y técnicas de Minería de Datos para identificar aquellos que provean una clasificación eficiente –baja tasa de clasificación incorrecta- de datos de diabetes. Los datos que se utilizaron para hacer el estudio se obtuvieron del National Institute of Diabetes and Digestive and Kidney Diseases, la base de

datos utilizada fue *Pilma Indians Diabetes Database* que contiene 768 registros y 8 atributos.

En este estudio se propuso un proceso para explorar los métodos:

1. Análisis de características relevantes: se utilizaron técnicas de reducción de atributos tales como: *Fisher, runs, relief* y *step disc*.
2. Comparación de técnicas de clasificación: se aplicaron diferentes técnicas de clasificación a los datos y se tabularon los resultados.
3. Selección de técnica de clasificación: se seleccionó la técnica C4.5, la cual es una técnica de inducción de aprendizaje por árboles de decisión, debido a que es usada en la mayoría de las aplicaciones médicas.
4. Obtención de reglas de clasificación: al aplicar la técnica C4.5, se obtuvieron las reglas de clasificación.
5. Evaluación: los resultados de la aplicación de la técnica C4.5 son los patrones de datos que se usan para clasificar si una persona es afectada o no por la diabetes.

En el estudio descrito se determinó que la técnica C4.5 clasificó correctamente el 91% de los casos.

4. **Application of data mining: Diabetes health care in young and old patients [23]**

Se identificó la efectividad de diferentes tipos de tratamientos para la diabetes, aplicados a diferentes tipos de edades. Se utilizó el método *Support Vector Machine* (SVM, por sus siglas en inglés). La información se obtuvo de la base de datos *Non-Communicable Diseases (NCD*, por sus siglas en inglés*),* a cargo de la Organización Mundial de la Salud. Esa información corresponde a Arabia Saudita en el año 2005.

Para llevar a cabo el estudio, se siguieron los siguientes pasos:

- Recolección y descripción de datos: se obtuvieron los datos y se realizó una descripción de ellos.
- Selección de herramientas y técnicas: se utilizó la técnica regresión descriptiva, aplicando el método *Support Vector Machine* (SVM, por sus

siglas en inglés). La herramienta que se utilizó para realizar la minería de datos fue *Oracle Data Miner* (ODM), versión 10.2.0.3.0.1.

- Experimento: se subdividieron los grupos de edad en dos: jóvenes y adultos.

En esta investigación se concluyó que los tratamientos para pacientes del grupo de edad joven pueden ser retardados para evitar efectos secundarios. En contraste, los tratamientos para los pacientes del grupo de edad adulta deben ser prescritos inmediatamente.

5. Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining [24]

Se analizó la asociación entre la diabetes tipo 2 y las enfermedades de comorbilidad usando la técnica de Reglas de Asociación (ARM, por sus siglas en inglés). Los datos para el estudio se obtuvieron del Keimyung University Dongsal Medical Center, con datos de 411,414 pacientes, correspondientes al periodo 1996-2007.

Para llevar a cabo el estudio, se siguió el siguiente proceso:

1. Estudio de la población: se definió el caso de estudio y los datos a utilizar.
2. Colección de datos: se recolectó la información requerida.
3. Selección de herramientas y técnicas: se utilizó la herramienta *DX Analyze Tool* y se seleccionó el algoritmo *A priori*, el cual es un algoritmo de la técnica de Reglas de Asociación (ARM, por sus siglas en inglés).

De esta investigación se concluyó que la hipertensión juega un rol importante en el desarrollo de la diabetes mellitus. Del total de pacientes que padecían diabetes, el 34.68% de ellos también padecían hipertensión, el 15.61% gastritis y duodenitis, el 15.43% catarata senil, 13.64% lipidemias y otros desórdenes de la lipoproteína y 12.78% enfermedad de la retina. Se sugiere el uso de la herramienta *Dx Analyze* en estudios de comorbilidad que tienen una gran cantidad de información clínica.

6. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors [25]

Se comparó el comportamiento de los modelos de Regresión Logística, Redes Neuronales Artificiales (ANNs, por sus siglas en inglés) y Árboles de Decisión para

la predicción de diabetes usando factores de riesgo comunes. La herramienta utilizada en la comparación fue *SPSS* (versión 13.0). Los datos fueron obtenidos a través de un cuestionario que se aplicó a voluntarios de 2 comunidades de Guangzho (China). El número total de voluntarios fue de 1487, donde 735 tenían diagnóstico de diabetes o prediabetes y 752 no tenían ese diagnóstico.

Para comparar el comportamiento de cada modelo se establecieron las métricas exactitud, sensibilidad y especificidad. La exactitud de clasificación mide la proporción de casos clasificados correctamente. La sensibilidad mide la fracción de casos positivos que son clasificados como positivos. La especificidad mide la fracción de casos negativos que son clasificados como negativos.

Los resultados obtenidos por cada modelo fueron los siguientes:

a) **Regresión logística**

   Exactitud=75.95%,

   Sensibilidad=79.68%,

   Especificidad=72.40%.

b) **Redes neuronales artificiales**

   Exactitud=73.52%,

   Sensibilidad=83.47%,

   Especificidad=64.08%.

c) **Arboles de decisión**

   Exactitud=78.27%,

   Sensibilidad=81.87%,

   Especificidad=74.86%.

Con los resultados se concluyó que el modelo generado por la técnica Árboles de Decisión, tuvo una mayor exactitud y especificidad (78.27% y 74.86% respectivamente) en comparación con los modelos generados por las técnicas Regresión Logística y Redes Neuronales Artificiales. Para la métrica de sensibilidad, la técnica Redes Neuronales Artificiales obtuvo el mejor resultado con un valor de 83.47%.

7. Data-Mining Technologies for Diabetes: A Systematic Review [26]

Se revisó sistemáticamente la aplicación de técnicas de Minería de Datos en la investigación de la diabetes. Se buscaron artículos de *MEDLINE* (Librería Nacional de Medicina que contiene citas de congresos y artículos de la literatura biomédica de todo el mundo) utilizando los términos: *"diabetes mellitus", "data mining", "data", "mining"*. En ella se encontraron 31 artículos. Se realizó un filtrado de los artículos usando los criterios de inclusión y exclusión. Se incluyeron aquellos artículos que tenían información sobre la aplicación de métodos de minería de datos y se excluyeron aquellos que no lo tenían. Al final, quedaron 16 artículos, los que se dividieron por temas, quedando de la siguiente manera:

a) Predicción e interpretación del nivel de glucosa: 4 artículos.
b) Selección de características: 4 artículos.
c) Análisis de datos genómicos: 2 artículos.
d) Otros estudios: 6 artículos que hablan sobre análisis del flujo de la salud, análisis del efecto de las medicinas, detección de fraude, enriquecimiento de las guías clínicas y predicción de mortalidad temprana.

Del estudio se concluyó que, el uso de la minería de datos para procesar grandes cantidades de datos clínicos de pacientes que se generan en las investigaciones, es un elemento valioso que ayuda a los investigadores y clínicas a proveer mejores soluciones para pacientes con diabetes.

8. Machine Learning and Data Mining Methods in Diabetes Research [27]

Se hizo una revisión sistemática de las aplicaciones de Aprendizaje Automático, técnicas de Minería de Datos y herramientas en la investigación de la diabetes. Se revisaron artículos de dos fuentes diferentes: *PubMed* (base de datos usada en las Ciencias Biomédicas) y *DBPL* (base de datos de publicaciones de Ciencias de la Computación). Para buscar los artículos se utilizaron los términos: *"machine learning and diabetes" (PubMed), "minería de datos y diabetes" (PubMed) y "diabetes" (DBLP).* El número de artículos encontrados fue de 1,287. Realizaron una inspección manual de los artículos para analizar si reportaban el uso de métodos de Aprendizaje Automático o Minería de Datos en los estudios, quedando 103 artículos. Los artículos finales se dividieron en 5 categorías:

1. Predicción y diagnóstico de biomarcadores: 48 artículos,
2. Complicaciones de la diabetes: 31 artículos,
3. Medicinas y terapias: 13 artículos,
4. Fondo genético y ambiental: 6 artículos,
5. Salud y administración de la diabetes: 5 artículos.

De este trabajo se concluyó que, a la fecha existen múltiples investigaciones en casi todos los aspectos de la diabetes mellitus, especialmente en la identificación de biomarcadores y predicción.

9. **Performance Analysis of Data Mining Classification Techniques to Predict Diabetes [28]**

Se analizaron las técnicas *Adaboost* y *Bagging Ensemble* usando el Árbol de Decisión J48 para clasificar pacientes con diabetes mellitus usando factores de riesgo de diabetes. Para ejecutarlo se utilizó la herramienta *WEKA*. La información se obtuvo de la base de datos *Canadian Primary Sentinel Surveillance Network (CPCSSN),* contando con 667,907 registros de pacientes, pertenecientes al periodo comprendido entre 2003 y 2013.

Para realizar el estudio, se seleccionaron las características más relevantes de la base de datos. Los datos se dividieron en tres grupos de edades:

1. D18-35, para pacientes de 18-35 años,
2. D36-55, para pacientes de 36-55 años,
3. D>55, para pacientes mayores de 55 años.

Para la técnica *Adaboost,* el área bajo la curva (*AROC*, por sus siglas en inglés) para un conjunto grande de datos fue de 0.98%, mostrando mayor sensibilidad y especificidad que las técnicas *Bagging Ensemble* y Árbol de Decisión.

10. **Can Data Science Inform Environmental Justice and Community Risk Screening for Type 2 Diabetes? [29]**

Se desarrolló un enfoque para evaluar niveles de riesgo comunitario al identificar regiones con potencial de presentar susceptibilidades genéticas para desarrollar diabetes mellitus tipo 2. Para el estudio, primero se consultó la base de datos

*Database of Single Nucleotide Polymorphisms (*dbNSP, por sus siglas en inglés*)* con el fin de encontrar los genotipos relacionados con diabetes tipo 2. Después se calcularon las variables de riesgo atribuibles a la población, los cuales son:

$$CC = p * p * n \quad // \textit{número de portadores del genotipo}$$

Donde:

p es la frecuencia de los casos reportados

n son las poblaciones del control y caso de estudio

$$PAR = 100 * (E * (OR - 1))/(1 + (E * (OR - 1))) \quad // \textit{riesgo atribuible a la población}$$

Donde:

E es la frecuencia del genotipo CC

OR es la relación reportada para desarrollar diabetes mellitus tipo 2 en las poblaciones de estudio.

Se consultó información demográfica del estado de California en los Estados Unidos. La base de datos utilizada fue *American Communities Survey*, para poblaciones caucásicas, asiáticas y mexicanas, entre 2007 y 2011. Se unieron los datos demográficos con los cálculos realizados anteriormente y se graficaron a través de la herramienta *ArcGIS* (versión 10.1).

De este estudio se concluyó que la frecuencia de riesgo de padecer diabetes mellitus tipo 2 para la población mexicana fue de 81% (CC/CT), para la población caucásica 73.6% y asiática 55.6%.

En la siguiente tabla (3.3) se establece una comparación entre los estudios analizados en este apartado.

Tabla 3.3 Comparación de estudios relacionados con aplicaciones de Ciencia de Datos y diabetes

N.	Tìtulo	Campo	Enfoque	Técnica	Algoritmo	Datos	País
1	Data Mining Technologies for Blood Glucose and Diabetes Management	Minería de datos	Analítico-descriptivo	Abstracciones temporales (TA´s)	-	-	-
2	Type 2 Diabetes Mellitus Trajectories and Associated Risks	Big data Analtics-Minería de datos	Analítico	Regresión logística multivariada	-	Rochester Epidemilogy Project, Mayo Clinic (1999-2013)	USA
3	Application of Data Mining Methods and Techniques for Diabetes Diagnosis	Minería de datos	Analítico	Clasificación	C4.5	National Institute of Diabetes and Digestive and Kidney Diseases, Pima Indians Diabetes Database (768 registros)	USA
4	Application of data mining: Diabetes health care in young and old patients	Minería de datos	Aplicativo	Regresión descriptiva	Support Vector Machine (SVM)	Non Communicable Diseases (NCD), OMS (2005)	Arabia Saudita
5	Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining	Minería de datos	Analítico	Reglas de asociación (ARM)	A priori	Keimyung University Dongsal Medical Center (411,414 pacientes, 1996-2007)	Corea del Sur
6	Comparison of three data mining models for predicting diabetes or prediabetes by risk factors	Minería de datos	Comparativo	Regresión logística, redes neuronales artificiales, árboles de decisión	-	Cuestionario a 1487 personas	China
7	Data-Mining Technologies for Diabetes: A Systematic Review	Minería de datos	Analítico	-	-	MEDLINE (31 artículos)	-
8	Machine Learning and Data Mining Methods in Diabetes Research	Machine learning-Minería de datos	Analítico	-	-	PubMed, DBPL (103 artículos)	-
9	Performance Analysis of Data Mining Classification Techniques to Predict Diabetes	Minería de datos	Analítico	Adaboost, bagging ensemble, árbol de decision J48	-	Canadian Primary Care Sentinel Surveillance Network (CPCSSN), (667,907 registros, 2003-2013)	Canadá
10	Can Data Science Inform Environmental Justice and Community Risk Screening for Type 2 Diabetes?	Ciencia de Datos	Predictivo	-	-	Single Nucleotide Polymorphism (dbSNP), American Communities Survey: caucasian, Asian and Mexican population (2007-2011)	USA
11	Desarrollo de una aplicación de Ciencia de Datos	Ciencia de Datos	Predictivo	Regresión	Regresión polinomial	Defunciones generales (1990-2015) Población nacional (1990-2015) Clasificación de enfermedades (2013)	México

La presente investigación, tiene una similitud respecto a las reportadas en la Tabla 3.3, las cuales todas analizan casos de diabetes mellitus. Sin embargo, tienen diferencias significativas; la mayoría de ellas utiliza la Minería de Datos como medio de extracción de conocimiento; sus enfoques son analíticos, descriptivos o comparativos; la técnica que utilizan es la clasificación con sus diferentes algoritmos, utilizan datos de pacientes de hospitales particulares o de encuestas; y ninguna de las investigaciones se ha realizado en México. La investigación que tiene más similitudes con el presente trabajo (N. 11) es [29] ya que ambas emplean la Ciencia de Datos, su enfoque es predictivo; utilizan datos poblacionales, aunque los empleados en [29] pertenecen a Estados Unidos y los de la presente investigación corresponden a México.

En este capítulo se mostraron las investigaciones relevantes desarrolladas por instituciones educativas aplicando la Ciencia de Datos a problemas específicos, tales como la identificación de variables que influyen en la declaración de desastres naturales, etiquetar contenido editorial, entre otros. Se realizaron aportaciones de conocimiento en áreas como el Aprendizaje Automático, Minería de Datos y Ciencia de Datos. En México, las instituciones gubernamentales han empleado la Ciencia de Datos para mejorar los servicios a sus usuarios. A nivel internacional, en años recientes se han iniciado investigaciones utilizando la Ciencia de Datos en el área de Salud. En particular, en esta investigación fue de interés el análisis de casos de diabetes mellitus. Sin embargo, aún no se han desarrollado proyecciones de mortalidad por diabetes para regiones de México en el periodo 2016-2020. Esto representa una oportunidad para aplicar los conocimientos de Ciencia de Datos a un problema de salud pública en México que ha provocado altos índices de mortalidad en las últimas tres décadas.

CAPÍTULO 4

SELECCIÓN DE UNA METODOLOGÍA DE CIENCIA DE DATOS

En este capítulo se presentan diversas metodologías para apoyar el análisis de los casos de diabetes mellitus con Ciencia de Datos. Para esto, se realizó una comparación entre las distintas fases de las metodologías y sus respectivas actividades. Por último, se seleccionó una metodología para apoyar el desarrollo del caso práctico mencionado en el capítulo uno.

4.1. Revisión de metodologías de Ciencia de Datos

Una metodología es una estrategia general que guía los procesos y actividades dentro de un dominio dado [30, 2]. Actualmente se proponen diversas metodologías de Ciencia de Datos que guían sus procesos aplicándolos en distintos campos del conocimiento. Las metodologías de Ciencia de Datos propuestas hasta la actualidad son las siguientes:

I. Metodología Fundacional para Ciencia de Datos (FMDS)

Esta metodología tiene similitudes con las metodologías reconocidas para la Minería de Datos, pero incluye nuevas actividades enfocadas en la Ciencia de Datos como el uso de volúmenes de datos muy grandes, la incorporación de la analítica de texto en el modelado predictivo y la automatización de algunos procesos.

Esta metodología fue desarrollada en 2015 por John B. Rollins en los laboratorios de IBM. La metodología FMDS consta de 10 etapas, cada etapa desempeña un papel vital en el contexto de la metodología general. [30]. Como referencia se utiliza la metodología representada en la Figura 4.1.

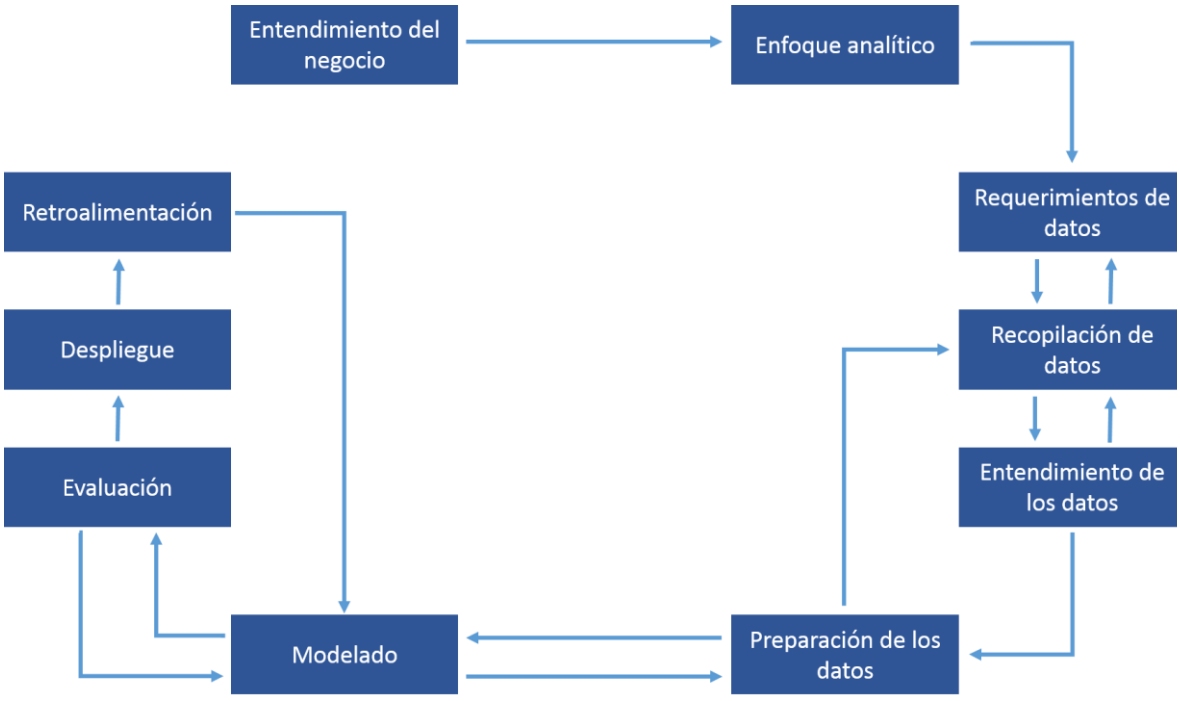

Figura 4.1 Metodología fundacional para la Ciencia de Datos

Sin embargo, en la Figura 4.1 no se especifican las condiciones necesarias para continuar el desarrollo de las actividades en cada una de las etapas. Por tanto, se ha diseñado un diagrama de flujo en la que se detallan dichas condiciones (Figura 4.2).

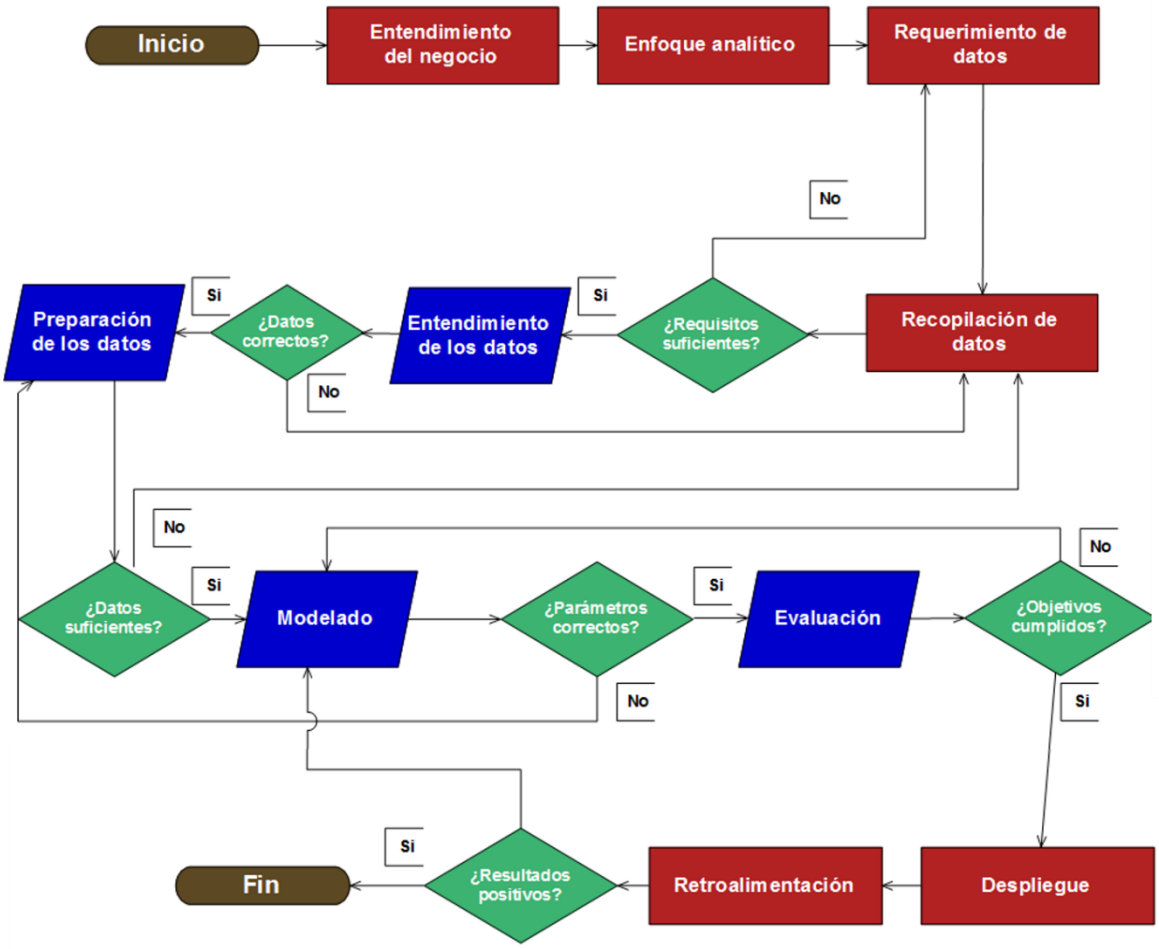

Figura 4.2 Diagrama de flujo de la metodología FMDS

A continuación se describen las actividades involucradas en cada una de las etapas de esta metodología y se detallan las condiciones que deben existir para avanzar de una etapa a otra.

1. Entendimiento del negocio

Cada proyecto comienza con la comprensión del negocio. Los clientes que necesitan la solución analítica desempeñan el papel más crítico en esta etapa al definir el problema, los objetivos del proyecto y los requisitos de la solución desde

una perspectiva empresarial. Una vez que se ha definido el problema a resolver, los objetivos y requisitos, se continúa con la etapa de “Enfoque analítico”.

2. Enfoque analítico

Esta etapa implica expresar el problema en el contexto de las técnicas estadísticas y de aprendizaje automático, para que la organización pueda identificar las más adecuadas para el resultado deseado. Cuando se haya definido el enfoque analítico a seguir, la siguiente etapa a desarrollar es “Requerimiento de datos”.

3. Requerimientos de datos

El enfoque analítico elegido determina los requisitos de datos. Específicamente, en esta etapa se requiere definir el contenido de los datos, su formato y representación, todo ello guiado por el conocimiento del dominio. Cuando se hayan definido estos requerimientos se prosigue con la etapa “Recopilación de datos”.

4. Recopilación de datos

En la etapa inicial de recopilación de datos, los científicos de datos identifican y recopilan los recursos de datos disponibles -estructurados, semiestructurados y no estructurados- relevantes para el dominio del problema. Si el científico de datos considera que los requisitos definidos en la etapa anterior nos son suficientes para efectuar el análisis seleccionado, debe regresarse a la etapa “Requerimiento de datos” para realizar las modificaciones pertinentes. Si los requisitos fuesen los correctos, se continúa con la etapa “Entendimiento de los datos”.

5. Entendimiento de los datos

Después de la recopilación de datos original, los científicos usan típicamente estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar su calidad y descubrir conocimientos iniciales acerca de ellos. Si los datos analizados son correctos de acuerdo al enfoque analítico adoptado, se prosigue con la etapa “Preparación de datos”, de lo contrario, se requiere regresar a la etapa “Recopilación de datos” para continuar la búsqueda.

6. Preparación de datos

Esta etapa abarca todas las actividades para construir el conjunto de datos que se utilizará en la fase de modelado. Las actividades de preparación de datos incluyen la limpieza de datos (tratar valores perdidos o no válidos, eliminar duplicados, formatear adecuadamente), combinar datos de múltiples fuentes (archivos, tablas, plataformas) y transformar datos en variables más útiles. Una vez terminada la preparación de los datos el científico de datos debe preguntarse si la cantidad de datos es suficiente para generar los modelos. Si se cumple esta condición, se continúa con la etapa “Modelado”, de lo contrario, es necesario regresarse a la etapa “Recopilación de datos”.

7. Modelado

A partir de la primera versión del conjunto de datos preparado, esta etapa se centra en el desarrollo de modelos predictivos o descriptivos de acuerdo con el enfoque analítico previamente definido. El proceso de modelado suele ser altamente iterativo a medida que las organizaciones obtienen ideas intermedias, lo que lleva a refinamientos en la preparación de los datos y la especificación del modelo. Para una técnica dada, los científicos de datos pueden probar múltiples algoritmos con sus respectivos parámetros para encontrar el mejor modelo para las variables disponibles. La siguiente etapa de esta metodología es “Evaluación”

8. Evaluación

Durante el desarrollo del modelo y antes del despliegue, el científico de datos evalúa el modelo para entender su calidad y asegurarse de que aborda adecuadamente y completamente el problema del negocio. La evaluación del modelo implica el cálculo de diversas medidas de diagnóstico y otros resultados tales como tablas y gráficos, lo que permite al científico de datos interpretar la calidad del modelo y su eficacia para resolver el problema. Si el modelo generado cumple con los objetivos definidos en la etapa de “Entendimiento del problema”, se prosigue a la etapa de “Despliegue”. En caso de no ser así, se realizan modificaciones al modelo.

9. Despliegue

Una vez que se ha desarrollado un modelo satisfactorio y es aprobado por el cliente, se implementa en el entorno de producción o un entorno de prueba comparable. Por lo general, se despliega de forma limitada hasta que se ha evaluado completamente su rendimiento. El despliegue de un modelo en un proceso empresarial operacional suele implicar grupos, habilidades y tecnologías adicionales dentro de la empresa.

10. Retroalimentación

Mediante la recopilación de los resultados del modelo implementado, la organización obtiene retroalimentación sobre el rendimiento del modelo y su impacto en el entorno en el que se implementó. Se realiza el monitoreo del modelo implementado por un tiempo definido y se evalúa si los resultados del mismo han proporcionado cambios positivos en la organización. Si no ha habido cambios o las medidas de análisis se han modificado, es necesario regresarse a la etapa de "Modelado". El proyecto de Ciencia de Datos termina cuando los cambios en los modelos satisfacen las necesidades y objetivos de la organización en la implementación real.

II. Proceso de Ciencia de Datos

De acuerdo con [4], los cinco pasos de la metodología "Proceso de Ciencia de Datos" son:

1. Definir preguntas de investigación: Al inicio de todo proyecto de Ciencia de Datos, la organización define el problema a resolver, establece objetivos concretos y formulan preguntas que guíen el proceso de análisis.
2. Obtener los datos: Una vez obtenidas las preguntas, el científico de datos recopila en diferentes fuentes los datos que necesita para realizar el análisis.
3. Explorar los datos: En este paso se utilizan diferentes técnicas estadísticas para obtener conocimiento preliminar de los datos.
4. Modelar los datos: en este paso del proceso, se seleccionan y utilizan técnicas especiales para generar modelos cuyos resultados cumplan con los

objetivos planteados y respondan las preguntas establecidas al inicio del proyecto.

5. Comunicar y visualizar los resultados: este último paso implica realizar reportes y presentaciones orales de los hallazgos y conclusiones del proyecto.

III. Fases de la Ciencia de Datos

Para desarrollar proyectos de Ciencia de Datos también se utiliza la metodología "Fases de la Ciencia de Datos" (DSP, por sus siglas en inglés) [31]. Su proceso se ilustra en la Figura 4.3.

Figura 4.3 Metodología DSP

De acuerdo a esta metodología un proyecto de Ciencia de Datos inicia cuando se identifican los diferentes roles que participarán en la elaboración del mismo. Posteriormente se analiza el problema a resolver y establecen los objetivos que se desean alcanzar. En la fase de investigación de recolectan los datos; se realizan modificaciones a su contenido, estructura y presentación de acuerdo a los requerimientos de la técnica seleccionada; se generan y refinan modelos para que cumplan con los objetivos establecidos en la fase "Preguntar". Se elaboran reportes del trabajo realizado y se desarrolla una presentación oral de los hallazgos y conclusiones obtenidas. Por último se aplica el conocimiento adquirido al problema planteado.

IV. Principios del Método Científico

Es importante incluir los principios que rigen el método científico, ya que todas las metodologías revisadas tienen sus fundamentos en él. El método científico rige el proceso de descubrimiento del conocimiento. Sin embargo, éste se debe adecuar dependiendo del dominio en el que se trabaja, es decir, se especifican las actividades inherentes al dominio de estudio. En la figura 4.4 se representan las fases del método científico [32].

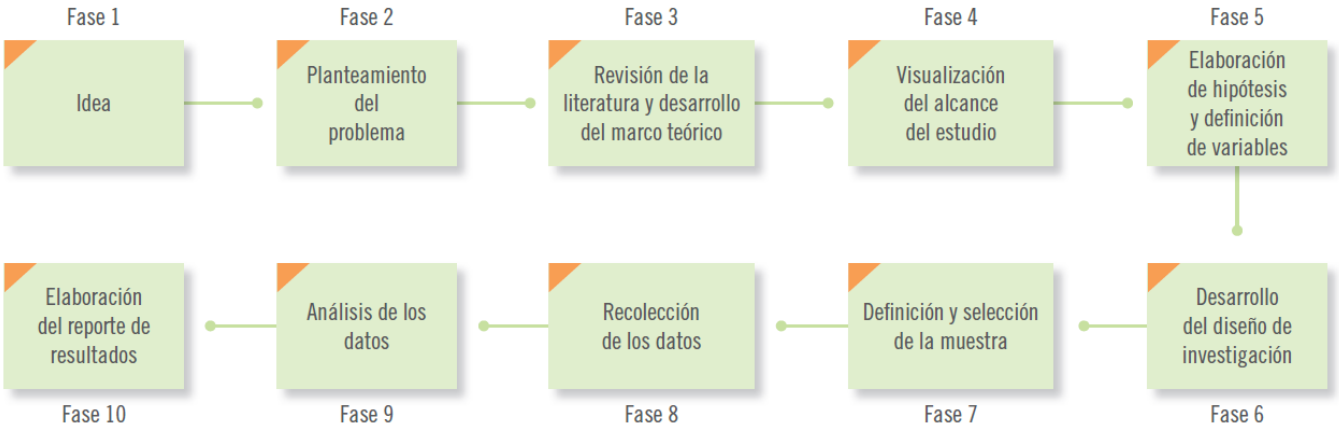

Figura 4.4 Fases del método científico

Como se observa en la Figura 4.4, el método científico es secuencial y probatorio. Cada fase precede a la siguiente y no se pueden eludir pasos. El método científico parte de una idea general acerca del fenómeno que se desea estudiar. En la siguiente fase se define el problema a resolver, se derivan objetivos y preguntas de investigación. En la fase 3 se revisa la literatura y se construye un marco o una perspectiva teórica. Se determina el alcance del estudio. De las preguntas se establecen hipótesis y determinan variables. La fase 6 corresponde al diseño de un plan para probar las hipótesis. Se definen los datos necesarios para realizar el análisis. En la fase 8 del proceso se recolectan los datos definidos en la etapa anterior. Se analizan las mediciones obtenidas (con frecuencia utilizando métodos estadísticos). En la última fase se establece una serie de conclusiones respecto de la(s) hipótesis.

4.2. Selección de una metodología de Ciencia de Datos

Para seleccionar la metodología más adecuada se contrastaron las existentes y se realizó un análisis comparativo de las mismas (ver Tabla 4.1).

Como se puede observar en la Tabla 4.1, la “Metodología Fundacional para Ciencia de Datos” tiene mayor nivel de especificidad. Se compone por 10 etapas con actividades claramente definidas. La metodología “Proceso de Ciencia de Datos” presenta cinco etapas, sus actividades son más generales y no presentan suficiente información que aclare su funcionamiento. Por último, la metodología “Fases de la Ciencia de Datos” se constituye por seis etapas. Sin embargo, no describe con detalle las actividades a realizar en cada una de las etapas. Por tanto, se empleará la metodología FMDS para desarrollar el caso práctico propuesto en el capítulo uno.

Es conveniente destacar que el método científico (ver Figura 4.4) es de propósito general para todas las áreas de la ciencia. En cambio, la metodología FMDS está orientada a casos específicos relacionados a la Ciencia de Datos. Si bien la metodología FMDS está inspirada en el método científico y algunas de sus etapas presentan similitudes, debido a los requerimientos específicos de proyectos de Ciencia de Datos, algunas etapas de la metodología FMDS difieren con las del método científico. Algunas de las etapas entre la metodología FMDS y el método científico que presentan mayor similitud son: Planteamiento del problema – Entendimiento del negocio, Elaboración de hipótesis y definición de variables – Enfoque analítico, Definición y selección de la muestra – Requerimiento de datos, Recolección de datos – Recopilación de datos, Análisis de los datos – Modelado y Evaluación.

Tabla 4.1 Comparativa de las metodologías usadas en Ciencia de Datos

Metodología Fundacional para Ciencia de Datos	**Fase**		**Entendimi-ento del negocio**	**Enfoque analítico de solución**	**Requerimientos de datos**	**Recopilación de datos**	**Entendimiento de los datos**	**Preparación de los datos**	**Modelado**	**Evaluación**	**Despliegue**	**Retroalimentación**
	Actividades		Definir el problema, los objetivos y requisitos de la solución.	Escoger las técnicas a usar de acuerdo al problema definido.	Determinar los requisitos de los datos en cuanto a contenido, formato y representación.	Identificar y recopilar los recursos de datos.	Descubrir conocimientos iniciales de los datos.	Limpieza, integración, transformación y construcción de datos.	Desarrollo de modelos de acuerdo al enfoque analítico seleccionado.	Evaluar el modelo en cuanto a su eficacia y eficiencia para resolver el problema planteado.	Desplegar el modelo en un entorno de producción.	Recopilar resultados del modelo implementado, corregir anomalías.
Proceso de Ciencia de Datos	**Fase**		**Definir preguntas de investigación**	----------	----------	**Obtención de datos**	**Exploración de datos**	----------	**Modelado**	----------	**Visualización y comunicación de resultados**	----------
	Actividades		Se define el problema a resolver	----------	----------	Buscar los datos que se utilizarán en el análisis	Uso de herramientas de visualización para entender, manipular y transformar los datos.	----------	Se escoge, ajusta y valida la efectividad de los modelos.	----------	Presentar los resultados al cliente de una forma clara, precisa y veraz.	----------
Fases de la Ciencia de Datos	**Fase**	**Identificar**		**Preguntar**	----------	----------	**Investigación**	----------	**Presentación**	**Hallazgos**	**Aprendizaje**	----------
	Actividades	Identificar los roles principales en el proyecto de Ciencia de Datos.		Establecer los objetivos del proyecto	----------	----------	Recolección y preparación datos, generación de modelos.	----------	Crear reportes de resultados.	Elaborar presentación de hallazgos a la organización.	Utilizar los hallazgos para generar conocimiento útil para la organización	----------

CAPÍTULO 5

APLICACIÓN DE LA METODOLOGÍA FMDS EN UN CASO EN EL ÁREA DE SALUD

Se desarrolló el caso práctico presentado en el capítulo uno utilizando la Metodología Fundacional para Ciencia de Datos (*FMDS*, por sus siglas en inglés). En este capítulo se describen las actividades realizadas en el desarrollo del caso práctico en cada una de las etapas de esta metodología.

En el CENIDET se han desarrollado investigaciones en el área de salud pública para analizar casos de diabetes mellitus. Estos análisis se han realizado aplicando la metodología de Minería de Datos. La presente investigación es una continuación de un trabajo previo [4], en la cual se identificaron regiones de municipios de México con altas tasas de mortalidad por diabetes mellitus. Sin embargo, no se realizaron proyecciones de tasas de mortalidad para esas regiones.

En este capítulo se describe el desarrollo de un caso para validar los conceptos asimilados y la metodología de Ciencia de Datos antes seleccionada. Este caso se enfoca en proyectar las tasas de mortalidad por diabetes mellitus en regiones de municipios de México para el periodo 2016-2020. Para ello se seleccionaron tres grupos representativos de estas regiones reportados en [4]. Los grupos fueron identificados como C24, C08 y C51. Los datos utilizados en el análisis fueron obtenidos de fuentes oficiales tales como INEGI, CEMECE y CONAPO. Los datos pertenecen al periodo 1990-2015.

5.1. Fase de entendimiento del negocio

Una de las enfermedades que ha sido motivo de mayor preocupación para los gobiernos e instituciones a nivel mundial en los últimos años es la diabetes, también llamada diabetes mellitus. La diabetes es un problema importante de salud pública y constituye una de las cuatro enfermedades no transmisibles (ENT) seleccionadas por los dirigentes mundiales para intervención con carácter prioritario. En las últimas décadas, el número de casos y la prevalencia de la enfermedad ha aumentado sin pausa [33].

De acuerdo con la OMS (Organización Mundial de la Salud), la diabetes es una enfermedad crónica que aparece cuando el páncreas no produce insulina suficiente o cuando el organismo no utiliza eficazmente la insulina que produce, afectando físicamente al paciente hasta llevarlo a la muerte. Según sus estimaciones, 422 millones de adultos en todo el mundo sufrían de diabetes en 2014, frente a 108 millones que en 1980 sufrían este padecimiento. La prevalencia mundial (normalizada por edades) de la diabetes casi se ha duplicado desde ese año. En la

población adulta ha pasado del 4.7% al 8.5%. Tan solo en 2012, la diabetes provocó 1.5 millones de muertes [33].

En México, según datos de la Federación Mexicana de Diabetes, ésta se encuentra entre las primeras causas de muerte. Además esta enfermedad representa un gasto de 3,450 millones de dólares al año en su atención y complicaciones. De acuerdo a la Federación Internacional de Diabetes (IDF, por sus siglas en inglés), México se encuentra en el sexto lugar a nivel mundial en mortalidad por diabetes y en el octavo en prevalencia de diabetes [34]. Ante esta situación, en 2016, la Secretaría de Salud en coordinación con el Comité Nacional de Seguridad en Salud, a través del Subcomité de Enfermedades Emergentes, emitió la declaratoria de emergencia epidemiológica EE-4-2016 para todo el territorio nacional.

La preguntas que guiaron la presente investigación y que fueron respondidas aplicando la metodología *FMDS* son las siguientes: ¿Cómo han evolucionado las tasas de mortalidad por diabetes mellitus en municipios de México?, ¿Cómo se comportarán las tasas de mortalidad por diabetes mellitus en municipios de México con las mayores tasas para el periodo 2016-2020?

5.2. Fase de enfoque analítico

Para realizar la proyección de las tasas de mortalidad fue necesario seleccionar una técnica de predicción numérica. En la actualidad, existen diferentes técnicas para realizar predicciones. Cada una de ellas se utiliza en contextos diferentes. En la Tabla 5.1 se realiza una comparación de las técnicas utilizadas para la proyección de tasas de mortalidad en diferentes dominios.

Tabla 5.1 Casos prácticos de predicción de tasas de mortalidad

Estudio	Objetivo	Técnica utilizada
Estimation of Mortality Rates in Fish Populations [35].	Estimar la tasa de mortalidad natural de pescados en el Lago Opeongo, Canadá.	Regresión lineal simple.
Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis [36].	Estimar la mortalidad global para el año 2004 si hubiese una pandemia similar a la de 1918-1920.	Regresión lineal simple.
Bayesian mortality forecasting with overdispersion [37].	Proponer dos métodos para mejorar el modelo Bayesiano Poisson Lee-Carter en la proyección de mortalidad teniendo en cuenta la sobredispersión.	Modelo de Poisson Lee-Carter, Modelos de sobredispersión: Modelo Poisson Gamma LC, Modelo Poisson log-normal LC.
Estimación de las proyecciones de tasas de incidencia, prevalencia y mortalidad por melanoma en España [38].	Realizar estimaciones de tasas de incidencia y de mortalidad por melanoma cutáneo (CMC) en España.	Regresión de Poisson.
Modeling and forecasting mortality rates [40].	Proponer un modelo que exprese los cambios en la tasa de mortalidad como una transformación lineal dependiente de un grupo de edad.	Modelo Gaussiano Inverso Normal.
Mortality modeling and forecasting: a review of methods [41].	Realizar una revisión de los métodos existentes para el modelado y proyección de la mortalidad.	Modelo Lee – Carter. Modelo de Poisson Lee – Carter.
Predicting Human Mortality: Quantitative Evaluation of Four Stochastic Models [42].	Comparar cuantitativamente la exactitud y precisión en la predicción de tasas de mortalidad de cuatro modelos de mortalidad diferentes.	Modelo Lee – Carter. Modelo de Will and Sherris. Proceso Feller. Proceso Ornstein – Uhlenbeck. Modelo de Cairns-Blake-Dowd.
A Bayesian forecasting model: predicting U.S. male mortality [43].	Presentar un enfoque Bayesiano para la predicción de tasas de mortalidad.	Modelo Lee- Carter.
Forecasting of literacy rate using statistical and Data mining methods. [44].	Realizar una predicción de la tasa de alfabetismo usando técnicas de minería de datos.	Regresión lineal múltiple, Curva Logística.

Como se observa en la Tabla 5.1, se han desarrollado investigaciones enfocadas en la proyección de tasas de mortalidad y tasas de alfabetismo. Para realizar la proyección se utilizaron diferentes técnicas de predicción de valores numéricos tales como el Modelo Lee-Carter, Modelo de Poisson, regresión lineal simple y regresión lineal múltiple.

La familia de los modelos *Lee- Carter, Poisson, Will and Sherris, Feller, Ornstein – Uhlenbeck* y *Cairns – Blake – Dowd* –cada una con sus variantes-, predicen valores de tasas de mortalidad tomando en cuenta grupos de edades. La regresión lineal es una de las técnicas más utilizadas para predicción de valores numéricos. Existen varios tipos de regresión, las cuales se utilizan según el tipo de datos, su

distribución y número de variables predictoras. En la regresión lineal simple sólo existe una variable predictora –o independiente. Mientras que, en la regresión lineal múltiple puede haber dos o más variables predictoras. La regresión polinomial permite construir un modelo ajustándose a la distribución de los datos.

Para esta aplicación se ha seleccionado la técnica regresión polinomial, aunque en la Tabla 5.1 no se haya reportado alguna investigación que utilice esta técnica. Las razones para emplear esta técnica son las siguientes: para realizar la proyección sólo se utilizará una variable predictora (ANIO_OCUR); la implementación de esta técnica es sencilla y sus resultados son fácilmente interpretables.

5.3. Fase de requerimiento de datos

En esta etapa del proceso metodológico se determinaron los datos requeridos para desarrollar la aplicación. Las fuentes para la selección fueron las siguientes:

a) Bases de datos de mortalidad a nivel nacional del periodo 1990 – 2015,
b) Catálogo de enfermedades CIE_CAT_2013_DGIS,
c) Base de datos de población de México a nivel municipal del periodo 1990 – 2015,
d) Proyección de población de México a nivel municipal hasta el año 2020.

Los atributos para las bases de datos de mortalidad se encuentran especificados en la Tabla 5.2.

Tabla 5.2 Atributos para bases de datos de mortalidad

Atributo	Clave_mun	Causa	Año	Genero	Edo_civil	Edad
Descripción	Clave del municipio	Causa de defunción	Año de defunción	Género del fallecido	Estado civil del fallecido	Edad del fallecido
Tipo dato	Cualitativo	Cualitativo	Cualitativo	Cualitativo	Cualitativo	Cuantitativo

Un dato *cuantitativo* es aquel que se describe usando números y con el cual se pueden realizar operaciones matemáticas. Un dato *cualitativo* es aquel que se describe usando datos categóricos y se utiliza para incluir a cierto registro en alguna categoría. Sin embargo, en estos casos, se puede describir un dato cualitativo con

un valor numérico. Aunque esté descrito por un número, no se podrían realizar procedimientos matemáticos con éste. En la sección 5.6 se explican las actividades realizadas para preparar los datos utilizados en esta investigación.

Se espera que hayan trabajos futuros que continúen con esta investigación, por lo que los atributos *Genero, Edo_civil* y *Edad* se incluyen en las bases de datos de mortalidad (ver Tabla 5.2), aun cuando no se usan en el análisis realizado, los atributos utilizados fueron *Clave_mun, Causa* y *Año*. Los atributos para la bases de datos del catálogo de enfermedades se muestran en la Tabla 5.3.

Tabla 5.3 atributos para base de datos del catálogo de enfermedades

Atributo	Causa	Nombre
Descripción	Clave de la enfermedad	Nombre de la enfermedad
Tipo dato	Cualitativo	Cualitativo

Los atributos para las bases de datos de población y proyección de población en México a nivel municipal se encuentran especificados en la Tabla 5.4.

Tabla 5.4 Atributos para bases de datos de población y proyección de población

Atributo	Clave_mun	Año	Nombre	Población
Descripción	Clave del municipio	Año de análisis	Nombre del municipio	Cantidad de habitantes
Tipo dato	Cualitativo	Cualitativo	Cualitativo	Cuantitativo

Antes de implementar la técnica *regresión polinomial* se necesita calcular las incidencias y tasas de mortalidad históricas para cada municipio que en 2010 contaba con más de 100,000 habitantes. La estructura requerida del conjunto de datos de incidencias y tasas de mortalidad por cada municipio es la siguiente (Tabla 5.5):

Tabla 5.5 Atributos para conjunto de datos de incidencias y tasa de mortalidad

Atributo	Clave_mun	Causa	Año	Incidencia	Tasa mortalidad	Tasa normalizada
Descripción	Clave del municipio	Causa de defunción	Año de defunción	Número de defunciones	Tasa de mortalidad	Tasa de mortalidad normalizada
Tipo dato	Cualitativo	Cualitativo	Cualitativo	Cuantitativo	Cuantitativo	Cuantitativo

En [4] se identificaron tres regiones de municipios de México los cuales presentan las mayores tasas de mortalidad por diabetes mellitus a nivel nacional. Los grupos se identifican como C24, C08 y C51. Los municipios que integran dichos grupos se presentan en las Tablas 5.6, 5.7 y 5.8 respectivamente.

Tabla 5.6 Grupo C24

Estado	Municipio
Ciudad de México	Venustiano Carranza
Veracruz	Orizaba
Ciudad de México	Iztacalco
Ciudad de México	Cuauhtémoc

Tabla 5.7 Grupo C08

Estado	Municipio
Ciudad de México	Azcapotzalco
Veracruz	Piedras Negras
Ciudad de México	Miguel Hidalgo
Ciudad de México	Gustavo A. Madero
Estado de México	Nezahuatcóyotl
Ciudad de México	Benito Juárez
Tamaulipas	El Mante
Tamaulipas	Tampico
Tamaulipas	Matamoros

Tabla 5.8 Grupo C51

Estado	Municipio
Estado de México	Tecámac
Guanajuato	San Francisco del Rincón
Michoacán	Zamora
Estado de México	Chalco
Estado de México	Toluca
Puebla	Atlixco
Veracruz	Martínez de la Torre
Ciudad de México	Tláhuac
Morelos	Cuernavaca
Puebla	San Pedro Cholula
Ciudad de México	Coyoacán
Estado de México	Netzahualcóyotl

La implementación computacional de la técnica regresión polinomial requirió de una estructura de datos específica, la cual se presenta en la Tabla 5.9

Tabla 5.9 Estructura del conjunto de datos para implementar la regresión polinomial

Atributo	Grupo	Año	Tasa_Normalizada
Descripción	Grupo de análisis	Año de análisis	Tasa normalizada
Tipo dato	Cualitativo	Cualitativo	Cuantitativo

5.4. Fase de recopilación de datos

En esta etapa se obtuvieron los datos listados en la fase previa. Los cuales fueron obtenidos de las siguientes instituciones (Tabla 5.10):

Tabla 5.10 Reporte de recopilación de datos

Fuente	Conjunto de datos
CEMECE- Centro Mexicano para la Clasificación de Enfermedades	Clasificación de enfermedades
INEGI– Instituto Nacional de Estadística y Geografía	Población de México por municipios
SINAIS- Sistema Nacional de Información en Salud	Defunciones generales (1990-2015)
CONAPO- Consejo Nacional de Población	Proyección de población por municipios de 2010-2030

5.5. Fase de entendimiento de los datos

Esta etapa se enfoca en analizar las características de los datos ya recopilados. Usualmente se utilizan técnicas descriptivas y de visualización para comprender el contenido de los datos, evaluar su calidad e identificar conocimientos iniciales. En la Tabla 5.11 se muestra la distribución de los registros de mortalidad, de acuerdo con el SINAIS, así como el tamaño que ocupa en memoria.

Tabla 5.11 Características de bases de datos de mortalidad

Año	Fuente	Fecha de consulta	Número de registros	Tamaño en bytes
1990	SINAIS	24/07/2017	422,804	59,534,000
1991	SINAIS	24/07/2017	411,132	57,970,000
1992	SINAIS	24/07/2017	409,815	57,943,000
1993	SINAIS	24/07/2017	416,336	58,656,000
1994	SINAIS	24/07/2017	419,075	57,282,000
1995	SINAIS	24/07/2017	430,279	58,819,000
1996	SINAIS	24/07/2017	436,322	61,501,000
1997	SINAIS	24/07/2017	440,438	61,999,000
1998	SINAIS	24/07/2017	444, 665	74, 097,000
1999	SINAIS	24/07/2017	443,950	39,454,000
2000	SINAIS	24/07/2017	437, 667	35,904,000
2001	SINAIS	24/07/2017	443,128	38,083,000
2002	SINAIS	24/07/2017	459,687	46,689,000
2003	SINAIS	24/07/2017	472, 140	47,954,000
2004	SINAIS	24/07/2017	473,417	44,847,000
2005	SINAIS	24/07/2017	495, 240	50,300,000
2006	SINAIS	24/07/2017	494, 471	46,842,000
2007	SINAIS	24/07/2017	514,420	48,731,000
2008	SINAIS	24/07/2017	539, 530	51,110,000
2009	SINAIS	24/07/2017	564, 573	53,492,000
2010	SINAIS	24/07/2017	592,018	56,082,000
2011	SINAIS	24/07/2017	590, 693	55,956,000
2012	SINAIS	24/07/2017	602, 354	71,767,000
2013	SINAIS	24/07/2017	623,599	79,779,000
2014	SINAIS	24/07/2017	633, 642	81,064,000
2015	SINAIS	24/07/2017	655, 688	83,884,000

En la Tabla 5.12 se presenta el número de registros del CEMECE, así como el tamaño en memoria que ocupa la base de datos.

Tabla 5.12 Características de la base de datos del catálogo de enfermedades

Año	Fuente	Fecha de consulta	Número de registros	Tamaño en bytes
2013	CEMECE	24/07/2017	14,423	2,892,000

La distribución por periodos de los registros de población, de acuerdo con el INEGI, así como el tamaño en memoria que ocupa cada base de datos, se muestra en la Tabla 5.13.

5.13 Bases de datos poblacional (municipios con más de 100,000 habitantes)

Año	Fuente	Fecha de consulta	Número de registros	Tamaño en bytes
1990	INEGI	23/08/2017	142	7,000
1995	INEGI	23/08/2017	158	8,000
2000	INEGI	23/08/2017	167	8,000
2005	INEGI	23/08/2017	178	9,000
2010	INEGI	23/08/2017	202	10,000

En la Tabla 5.14 se presentan las bases de datos que representan la proyección de la población, de acuerdo con el CONAPO, en los estados de la República Mexicana.

5.14 Bases de datos de proyección de población 2010-2030

Estado	Fuente	Fecha de consulta	Número de registros	Tamaño en bytes
Aguascalientes	CONAPO	27/10/2017	207	567,000
Baja California	CONAPO	27/10/2017	105	543,000
Baja California Sur	CONAPO	27/10/2017	105	543,000
Campeche	CONAPO	27/10/2017	207	566,000
Chiapas	CONAPO	27/10/2017	2026	983,000
Chihuahua	CONAPO	27/10/2017	1159	781,000
Coahuila	CONAPO	27/10/2017	666	669,000
Colima	CONAPO	27/10/2017	190	562,000
Distrito Federal	CONAPO	27/10/2017	292	586,000
Durango	CONAPO	27/10/2017	683	674,000
Guanajuato	CONAPO	27/10/2017	802	701,000
Guerrero	CONAPO	27/10/2017	1397	836,000
Hidalgo	CONAPO	27/10/2017	1448	846,000
Jalisco	CONAPO	27/10/2017	2145	1,006,000
Estado de México	CONAPO	27/10/2017	2145	1,010,000
Michoacán	CONAPO	27/10/2017	1941	956,000
Morelos	CONAPO	27/10/2017	581	650,000
Nayarit	CONAPO	27/10/2017	360	616,000
Nuevo León	CONAPO	27/10/2017	887	720,000
Oaxaca	CONAPO	27/10/2017	9710	2,717,000
Puebla	CONAPO	27/10/2017	3709	1,363,000
Querétaro	CONAPO	27/10/2017	326	594,000
Quintana Roo	CONAPO	27/10/2017	190	578,000
San Luis Potosí	CONAPO	27/10/2017	1006	745,000
Sinaloa	CONAPO	27/10/2017	326	593,000
Sonora	CONAPO	27/10/2017	1244	798,000
Tabasco	CONAPO	27/10/2017	309	589,000
Tamaulipas	CONAPO	27/10/2017	751	688,000
Tlaxcala	CONAPO	27/10/2017	1040	770,000
Veracruz	CONAPO	27/10/2017	3624	1,343,000
Yucatán	CONAPO	27/10/2017	1822	933,000
Zacatecas	CONAPO	27/10/2017	1006	744,000

5.6. Fase de preparación de los datos

Esta etapa comprende todas las actividades de construcción del conjunto de datos que se utilizaron en la etapa de modelado. Las actividades de preparación de datos incluyen la limpieza de datos (tratar valores perdidos o no válidos, eliminar duplicados, formatear adecuadamente), combinar datos de múltiples fuentes (archivos, tablas, plataformas) y transformar datos en variables más útiles [30]. Las herramientas utilizadas para realizar las actividades de preparación de datos fueron Knime (herramienta de software libre) y Excel. El lenguaje R fue utilizado para generar los modelos de datos. En esta sección se presentan las actividades realizadas en la preparación de los datos.

5.6.1. Selección, limpieza y transformación de las bases de datos

En la siguiente sección se describen las actividades realizadas en esta fase para cada una de las bases de datos obtenidas.

5.6.1.1. Datos de mortalidad

Los archivos de las bases de datos de mortalidad de los años 1990-2015 se convirtieron del formato *.dbf* (dBase) al formato *.xlsx* (Microsoft Excel) para facilitar su manipulación.

Se realizó la selección vertical (columnas o atributos) de estas bases de datos. Las columnas seleccionadas fueron: ENT_RESID, MUN_RESID, CAUSA, ANIO_OCUR, SEXO, EDAD y EDO_CIVIL. Estas últimas no fueron incluidas en el análisis. También se realizó la selección horizontal (filas o registros). En las bases de datos correspondientes a los años 1990 - 1997 se incluyó una codificación de claves en los atributos ENT_RESID y MUN_RESID mayor a las claves oficiales, por lo que se excluyeron aquellas claves que no coincidían con las oficiales. Para estas mismas bases de datos (1990 - 1997), se utilizó la codificación de la Clasificación de Enfermedades, la cual difiere con la codificación de la CEMECE (CIE_CAT_2013_DGIS) utilizada en años posteriores, por lo que en las de los años 1990 - 1997 se incluyeron únicamente los registros que tuvieran la causa 181 (correspondiente a la enfermedad diabetes mellitus). Para todas las bases de datos

de mortalidad se realizó una selección horizontal de los registros cuyos municipios tuvieran más de 100,000 habitantes (de acuerdo con datos de población del 2010).

En las bases de datos de mortalidad se modificaron los datos para que cumplieran con la estructura de la Tabla 5.2. Para los correspondientes a los años 1990 - 1997 se modificaron los valores del atributo EDAD. Originalmente la codificación era 4 000 (el 4 correspondía a la clave 'años' y los últimos 3 dígitos, a los años de la persona al momento de fallecer). Se realizó una resta de 4,000 al valor del atributo EDAD de tal manera que sólo quedó la edad de la persona. También se unificaron las columnas ENT_RESID y MUN_RESID, quedando únicamente la columna *clave_mun*.

5.6.1.2. Datos de clasificación de enfermedades

Para procesar esta base de datos no se realizó la selección vertical ni horizontal. Solo se transformaron los valores del atributo *clave,* ya que éste tenía cuatro dígitos. El último dígito representaba la especificación de la enfermedad, lo cual no interesaba, por lo que se eliminó el último dígito.

5.6.1.3. Datos de población

Para todas las bases de datos de población se hizo una selección vertical. Los atributos seccionados fueron: *Clave, Municipio y Total (población).* En la selección horizontal de los datos, se incluyeron a aquellos municipios que tuvieran más de 100,000 habitantes (de acuerdo con los datos de población de 2010).

Para la base de datos de población correspondiente al año 2000, se cambiaron columnas a filas, debido a que las filas en esta base de datos representan a las columnas en las demás bases de datos y las columnas representan a las filas.

La transformación de los datos consistió en darle el formato a los valores del atributo *Clave*, el cual debería quedar con el mismo formato que el atributo *clave_mun* de las bases de datos de mortalidad.

5.6.1.4. Datos de proyección de población

En este proceso se realizó la selección vertical de los datos, resultando seleccionados los atributos *Clave_municipio, Municipio, 2015-2020 (años de*

población estimada). Los registros resultantes de la selección horizontal fueron aquellos municipios que tuvieron más de 100,000 habitantes (de acuerdo a datos de población del 2010).

5.6.2. Creación de datos

Se creó un conjunto de datos con la población histórica de los municipios que, para el año 2010, contaban con más de 100,000 habitantes, éste se incluyó en la base de datos diseñada para realizar los cálculos de incidencias y tasas de mortalidad. La estructura del conjunto de datos se representa en la Tabla 5.5.

5.6.3. Integración de datos

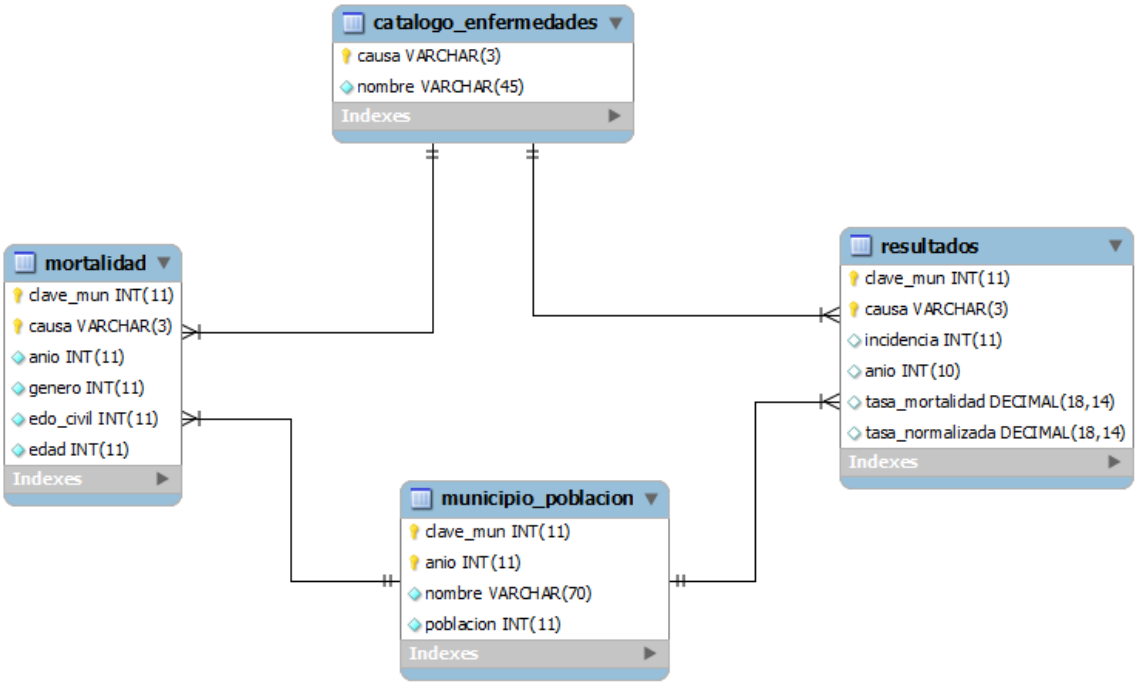

Figura 5.1 Diagrama Entidad-Relación de base de datos de integración

En la Figura 5.1 se muestra el diagrama Entidad-Relación de la base de datos, creada para almacenar los datos utilizados en el análisis. El Sistema Manejador de Base de Datos (DBMS, por sus siglas en inglés) utilizado fue Mysql. En la tabla *catálogo_enfermedades* se almacenan los datos de clasificación de enfermedades, en la tabla *mortalidad*, los registros de defunciones por diabetes mellitus durante el periodo 1990-2015, en la tabla *municipio_población* se almacenan los datos de población y proyección de población de municipios para el periodo 1990-2030. En

la tabla resultados se guardan los cálculos de incidencias y tasas de mortalidad para la diabetes mellitus (E11 - E14) en cada municipio seleccionado.

5.7. Fase de modelado

Una vez obtenidos los valores de las tasas de mortalidad para cada municipio, es necesario normalizarlos para obtener una escala uniforme entre dichos valores. La normalización es el proceso de ajustar los valores de los datos a una escala definida, se calcula mediante la fórmula [45]:

$$V' = \frac{V_{actual} - V_{min}}{V_{max} - V_{min}} * 10$$

Donde V' es el valor normalizado resultante, V_{actual} es el valor a normalizar, V_{min} y V_{max} los valores mínimos y máximos respectivamente y 10 es el factor de normalización.

Al obtener los cálculos de las tasas de mortalidad y normalizarlos, se creó un conjunto de datos donde se calcularon los promedios de los grupos C24, C08 y C51; los cuales están compuestos por los municipios definidos en la Tablas 5.6, 5.7 y 5.8 respectivamente. En la Figura 5.2, se muestra la tendencia real de la tasa promedio de mortalidad normalizada para el grupo C24.

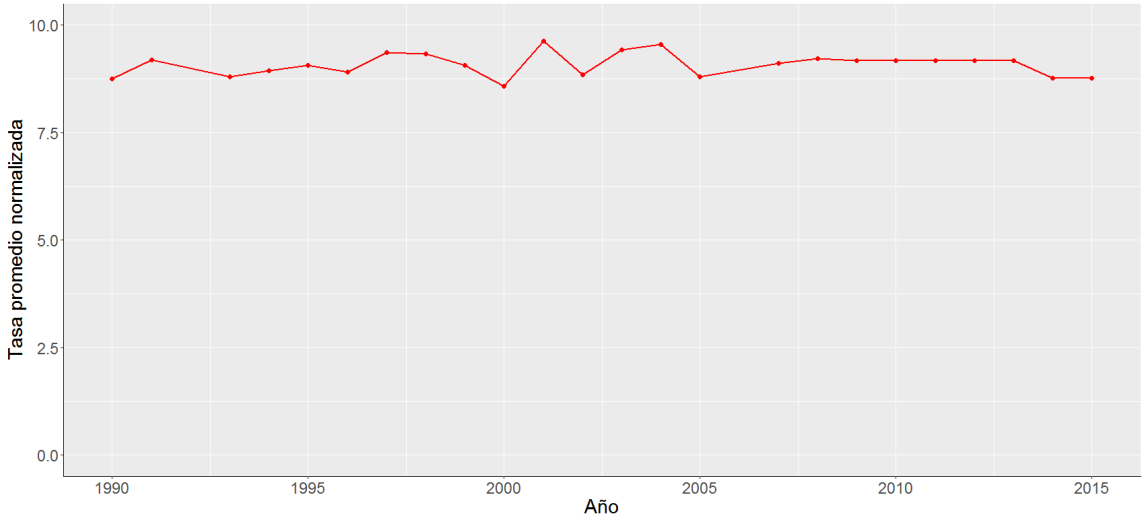

Figura 5.2 Tasa promedio de mortalidad normalizada del Grupo C24

El grupo C24 lo integran los municipios con las mayores tasas de mortalidad a nivel nacional. La tasa promedio de mortalidad normalizada para 1990 fue de 8.75 y para 2015 su valor fue de 8.76. Sin embargo, a partir de 1995 y hasta 2005 se observa una variación significativa en la tasa promedio de mortalidad normalizada (teniendo el año 2000 el valor más bajo con 8.58 y el 2001 el año con el valor más alto con 9.63). Para el periodo 2007-2013 su valor se mantiene constante con 9.1, pero para los años 2014 y 2015, el valor de la tasa promedio de mortalidad normalizada disminuye a 8.76. La tendencia real de la tasa promedio de mortalidad normalizada para el grupo C08 se ilustra en la Figura 5.3.

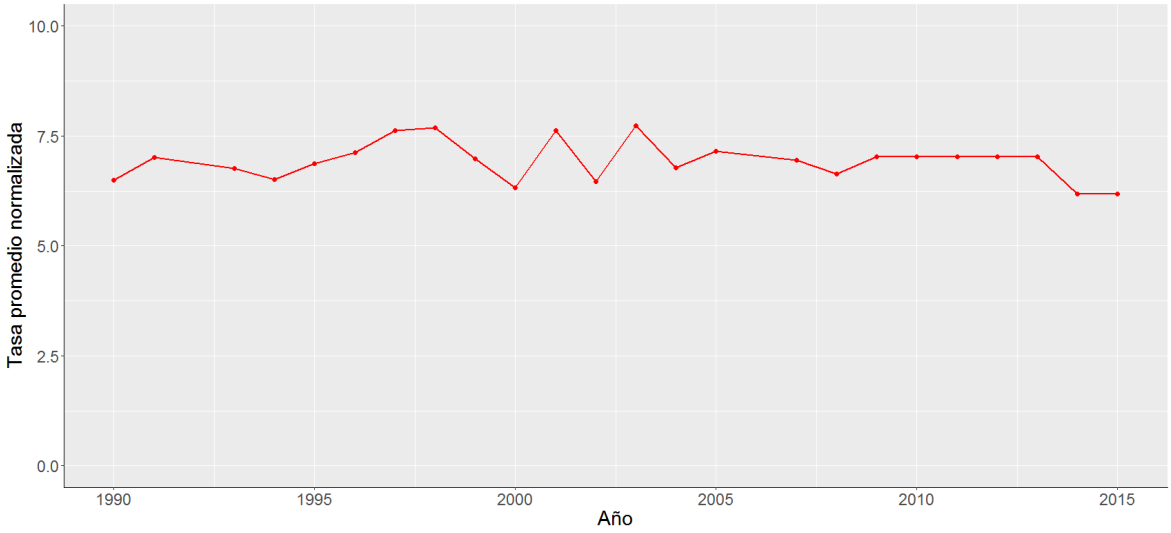

Figura 5.3 Tasa promedio de mortalidad normalizada del Grupo C08

De acuerdo a la Figura 5.3, la tasa promedio de mortalidad normalizada para el grupo C08 ha ido variando a través del tiempo. Para este grupo, el periodo 1990-2009 tuvo cambios significativos en la tasa promedio de mortalidad normalizada, para ese periodo el año con el menor valor fue 2000 con 6.32 y el mayor fue el 2003 con un valor de 7.73. Al igual que en el grupo C24, para los años 2014 y 2015 se observa una disminución en la tasa promedio de mortalidad normalizada al pasar el 2013 de 7.02 a 6.18 para los años mencionados.

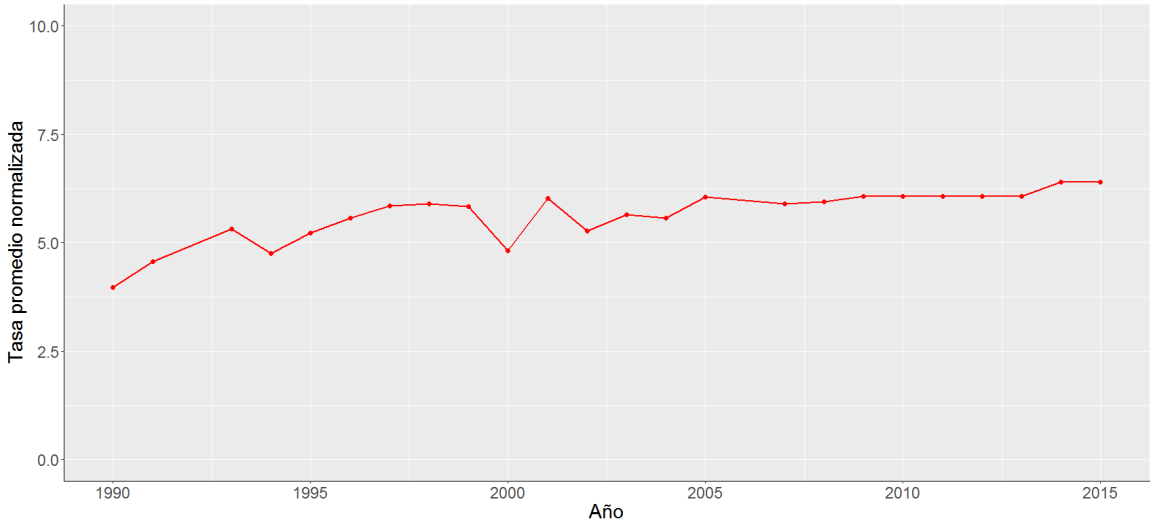

Figura 5.4 Tasa promedio de mortalidad normalizada del Grupo C51

En la Figura 5.4, se muestra la tendencia real de la tasa promedio de mortalidad normalizada para el grupo C51. Su valor ha ido incrementándose gradualmente. El cambio más significativo de un año a otro lo presentan los años 2000 y 2001, con valores de 4.82 y 6.01 respectivamente. Para 1990 su valor era de 3.97, mientras que para 2015 tenía un valor de 6.40, esto significa un aumento del 62% en 25 años para ese grupo.

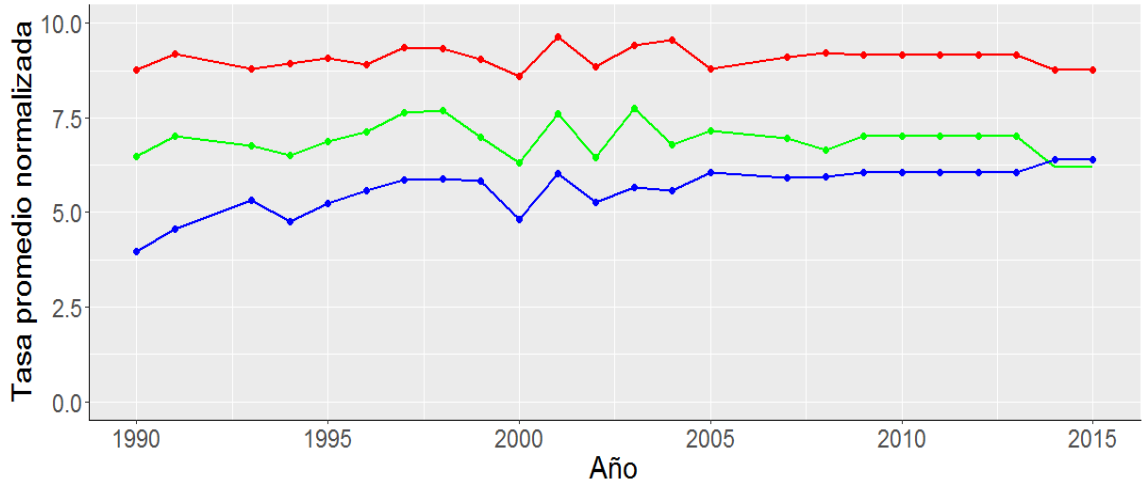

Figura 5.5 Tasas promedio de mortalidad normalizada de los grupos C24, C08 y C51

En la Figura 5.5 se observan las tendencias reales de las tasas promedio de mortalidad normalizada para los grupos C24, C08 y C51. El grupo C24 (línea roja)

posee los valores más altos en todo el periodo 1990-2015. Le continúa el grupo C08 (línea verde), el grupo C51 (línea azul) tiene los valores más bajos en las tasas promedio de mortalidad normalizada. Al comparar en el tiempo los tres grupos se observa que, para el periodo 1990-2005, sus comportamientos son similares, haciéndose más evidente para el periodo 2000-2005. A partir del año 2007 y hasta el 2013 la tasa promedio se estabiliza en los tres grupos. Sólo en los grupos C24 y C08, para los años 2014 y 2015 su valor disminuye.

De acuerdo con la distribución de los datos en los tres grupos, se observa que no es conveniente utilizar la regresión lineal para realizar la proyección de las tasas de mortalidad. Se confirma que la regresión polinomial es adecuada para generar los modelos de proyección.

Para generar los modelos por grupo se dividió el conjunto de datos de las tasas promedio de mortalidad normalizada. Por cada grupo se generaron dos subconjuntos de datos (método Holdout [46]), los cuales son:

a) Subconjunto de entrenamiento: corresponde a los datos utilizados para generar el modelo. En cada grupo se seleccionaron de 1990 al 2015.
b) Subconjunto de prueba: corresponde a los datos que se utilizaron para evaluar la precisión del modelo construido. Se seleccionaron los datos del periodo 1990-2015 por cada grupo.

El objetivo de la regresión polinomial es encontrar una función que se ajuste lo más posible al subconjunto de entrenamiento, para ello es necesario determinar el orden del polinomio. A continuación se presentan las fórmulas generales de las funciones polinómicas utilizadas para generar los modelos:

1. $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

2. $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

La función uno es de grado 2 (denominada función cuadrática). La función dos es de grado 3 (llamada función cúbica).

Para generar los modelos se utilizó la herramienta R. Ésta consiste en un lenguaje de programación especializado en el manejo de grandes volúmenes de datos.

Posee paquetes que facilitan la construcción de modelos de datos. Es utilizado con frecuencia en los sectores empresarial, gubernamental y académico, además es de libre acceso.

Para identificar los modelos generados se utiliza la nomenclatura: “F2/F3” significa que pertenece a la función polinómica de grado dos o tres -presentadas anteriormente. La nomenclatura “C24/C08/C51” representa al grupo de análisis.

Para cada función polinómica y para cada grupo se construyó un modelo. La Figura 5.6 muestra el modelo construido utilizando la función polinómica de grado 2 para el grupo C24. La línea roja representa los datos reales observados y la línea azul el ajuste del modelo.

1. **Modelo F2-C24**

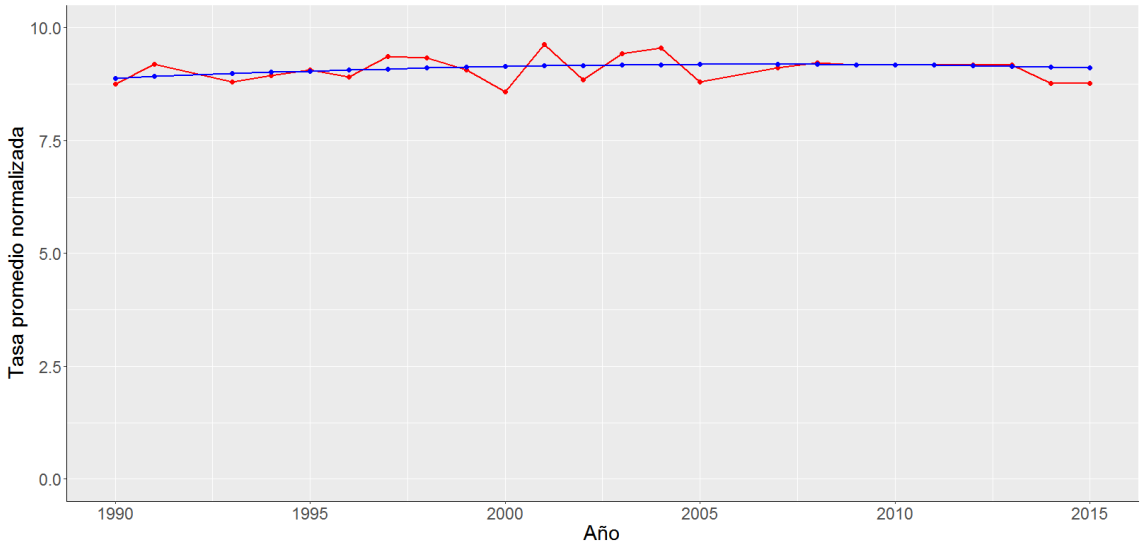

Figura 5.6 Modelo de la función de grado 2 para el grupo C24

En la Figura 5.6 se observa el contraste de la tendencia real y el ajuste del modelo para el grupo C24. 2000 fue el año en el que existe una mayor diferencia entre el valor real y predicho con un valor de 6.84% (ver Tabla 5.21 para mayor detalle de los valores del error porcentual absoluto medio para este modelo y grupo). El año con la menor diferencia fue 1995, con un valor de 0.11%. En promedio, el error porcentual absoluto medio para este grupo y este modelo fue de 2.12%

Cada modelo de regresión posee coeficientes que definen la relación entre sus variables, éstos son llamados *coeficientes de regresión*. Un modelo construido con una función polinómica de grado 2 tiene tres coeficientes: β0, β1 y β2. Los coeficientes de regresión del modelo F2-C24 se muestran en la Tabla 5.15.

Tabla 5.15 Coeficientes de regresión del modelo F2-C24

Coeficiente	Valor
β0	9.08224
β1	0.09058
β2	-0.50406

2. Modelo F2-C08

El modelo construido utilizando la función polinómica de grado 2 para el grupo C08 se muestra en la Figura 5.7. La línea roja representa los datos reales observados y la línea azul el ajuste del modelo.

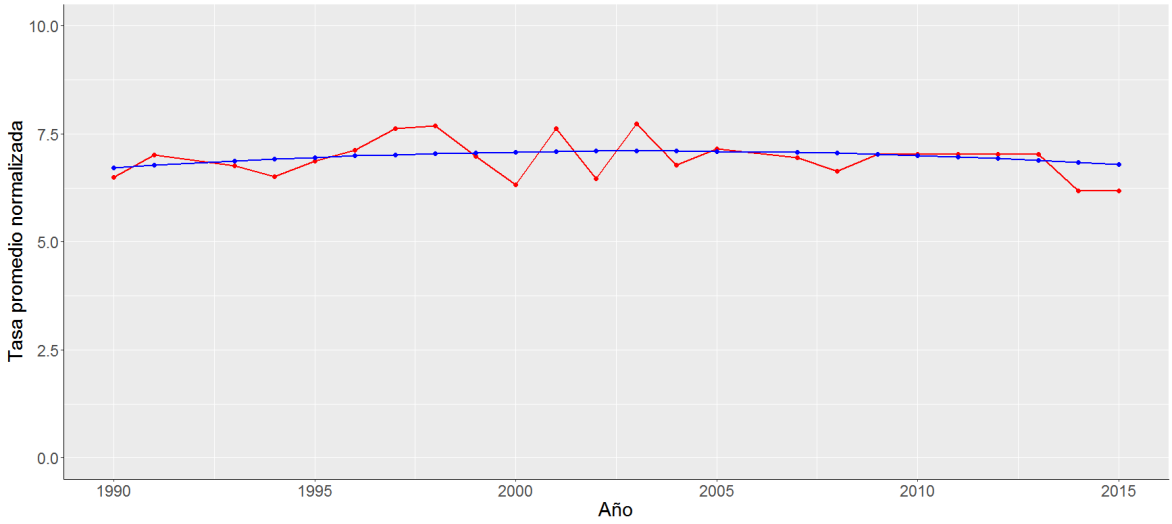

Figura 5.7 Modelo de la función de grado 2 para el grupo C08

Para este grupo, los valores del error porcentual absoluto varían significativamente en el periodo 1996-2005. En el año 2000 el error absoluto entre el valor real y predicho fue de 13.02%, siendo éste el mayor de todos. Mientras que el año 2005 registró el menor error absoluto con 0.63%. El error porcentual absoluto medio para este modelo y grupo fue de 4.55% (ver Tabla 5.22 para mayor detalle de los valores del error porcentual absoluto medio para este modelo y grupo). Los coeficientes de regresión del modelo F2-C08 se muestran en la Tabla 5.16.

Tabla 5.16 Coeficientes de regresión del modelo F2-C08

Coeficiente	Valor
β0	6.92737
β1	0.33758
β2	-0.97122

3. Modelo F2-C51

El modelo construido utilizando la función polinómica de grado 2 para el grupo C51 se muestra en la Figura 5.8. La línea roja representa los datos reales observados y la línea azul el ajuste del modelo.

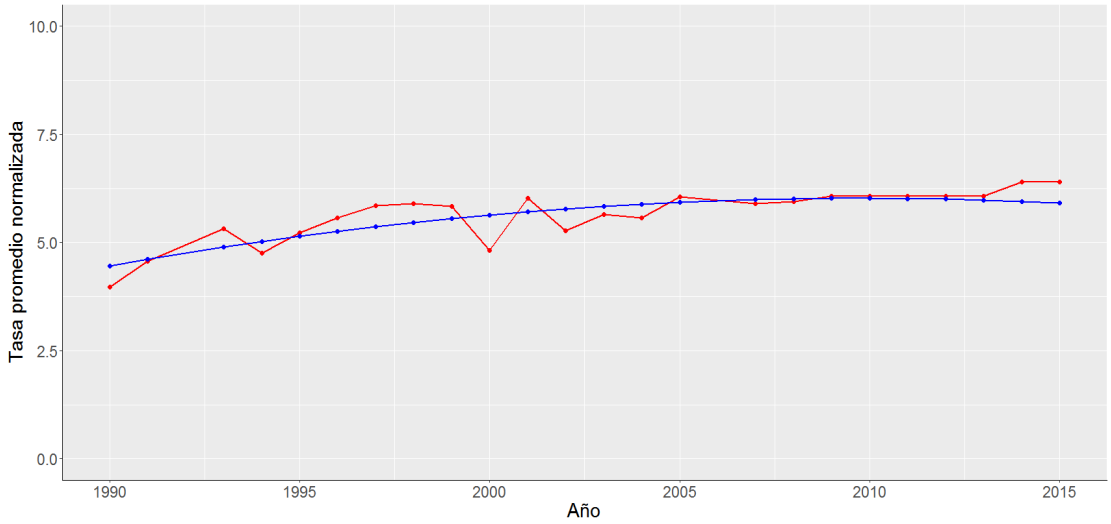

Figura 5.8 Modelo de la función de grado 2 para el grupo C51

Para el grupo C51, nuevamente el año 2000 registró el mayor error absoluto con 15.90%. Mientras que su valor más bajo lo obtuvo el año 2009 con 0.30%. El error porcentual absoluto medio para este modelo y grupo fue de 5.14% (ver Tabla 5.23 para mayor detalle de los valores del error porcentual absoluto medio para este modelo y grupo).

Los coeficientes de regresión del modelo F2-C51 se muestran en la Tabla 5.17.

Tabla 5.17 Coeficientes de regresión del modelo F2-C51

Coeficiente	Valor
β0	5.64231
β1	2.35848
β2	-0.66854

La Figura 5.9 muestra el modelo generado utilizando la función polinómica de grado 3 para el grupo C24.

4. Modelo F3-C24

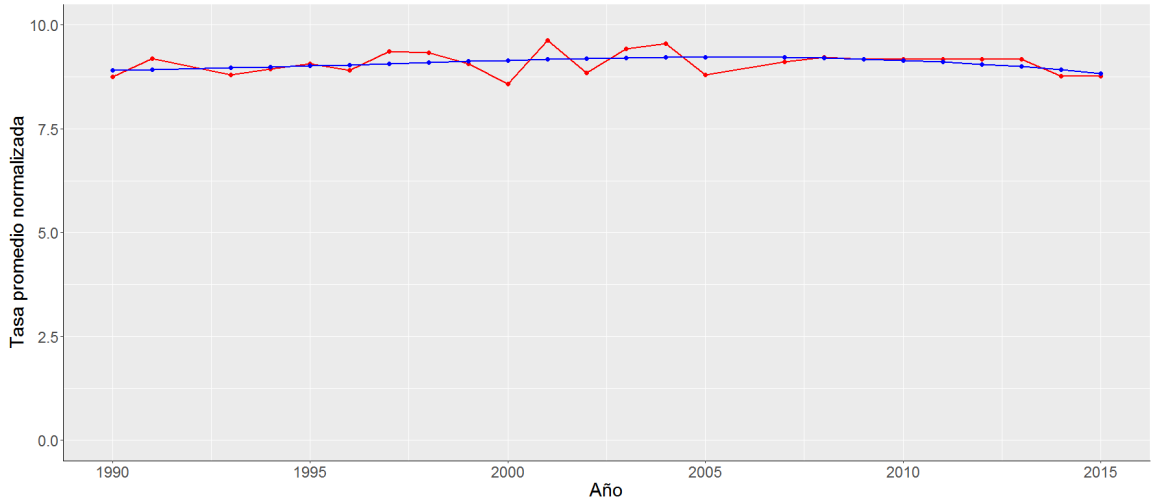

Figura 5.9 Modelo de la función de grado 3 para el grupo C24

Al generar el modelo de datos con la función de grado 3, el valor del error porcentual absoluto medio para este grupo disminuye respecto al generado por la función de grado 2. Sin embargo, 2000 fue el año con el mayor error absoluto con 6.53%. El menor porcentaje del error absoluto para este modelo y grupo lo obtuvo el año 2009 con un valor de 0.11%. El error porcentual absoluto medio para este modelo y grupo fue de 2.08% (ver Tabla 5.24 para mayor detalle de los valores del error porcentual absoluto medio para este modelo y grupo).

Un modelo generado con una función polinómica de grado 3 tiene cuatro coeficientes: β0, β1, β2 y β3. Los coeficientes de regresión del modelo F2-C24 se muestran en la Tabla 5.18.

Tabla 5.18 Coeficientes de regresión del modelo F3-C24

Coeficiente	Valor
β0	9.08224
β1	0.09058
β2	-0.50406
β3	-0.20139

En la Figura 5.10 se muestra el modelo generado utilizando la función polinómica de grado 3 para el grupo C08.

5. Modelo F3-C08

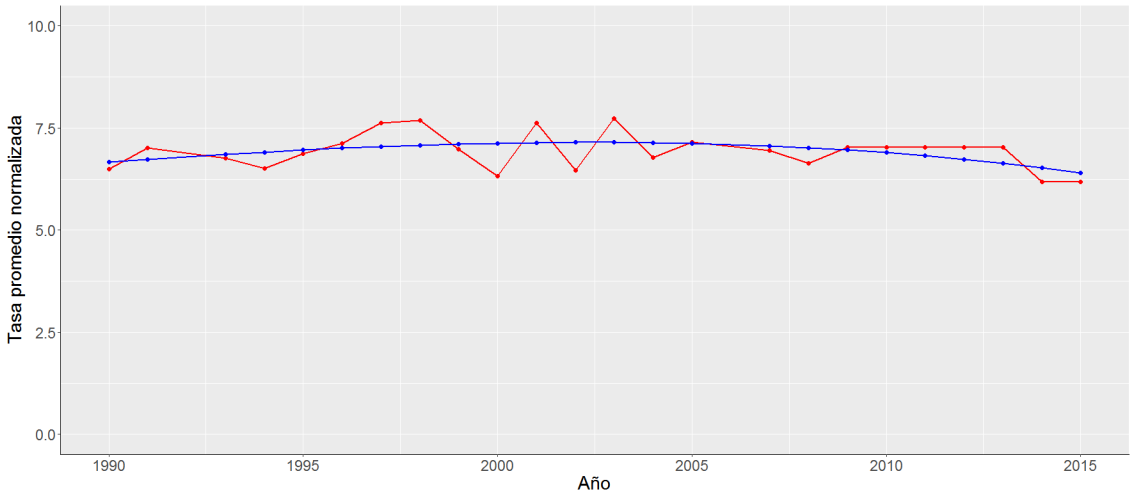

Figura 5.10 Modelo de la función de grado 3 para el grupo C08

Para este modelo y grupo, el año 2000 fue el que obtuvo el mayor error absoluto con 13.03%. Mientras que el año 2005 presentó el menor error absoluto con 0.83%. El error porcentual absoluto medio para este modelo y grupo fue de 2.08% (ver Tabla 5.25 para mayor detalle de los valores del error porcentual absoluto medio para este modelo y grupo). Los coeficientes de regresión del modelo F3-C08 se muestran en la Tabla 5.19.

Tabla 5.19 Coeficientes de regresión del modelo F3-C08

Coeficiente	Valor
β0	6.92737
β1	-0.33758
β2	-0.97122
β3	-0.10288

En la Figura 5.11 se presenta el modelo generado utilizando la función polinómica de grado 3 para el grupo C51.

6. Modelo F3-C51

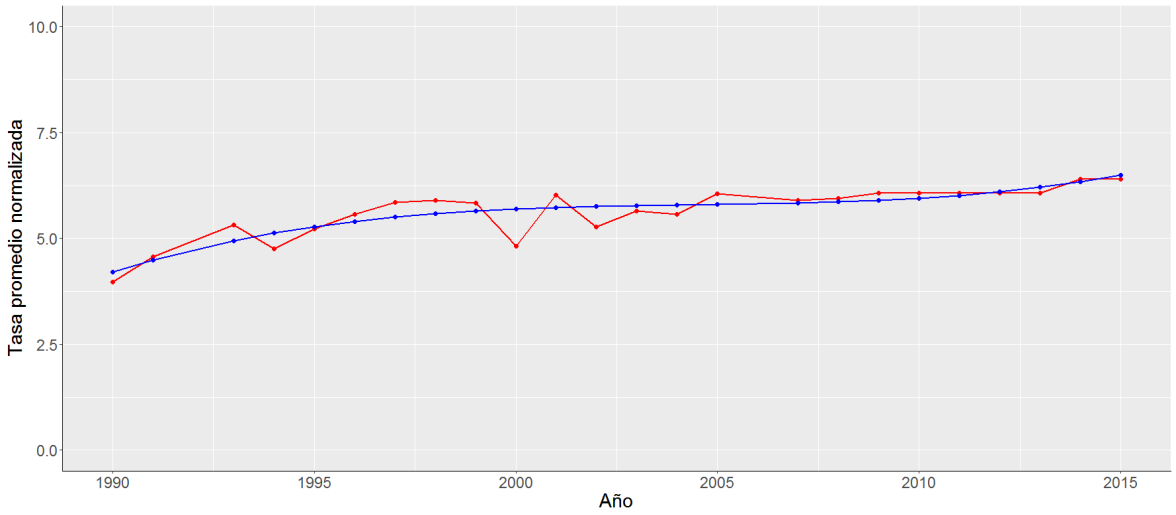

Figura 5.11 Modelo de la función de grado 3 para el grupo C51

De acuerdo a los valores del error absoluto para este modelo y grupo, el año 2000 registró el mayor con 18.69%. Mientras que su valor más bajo lo obtuvo el año 2015 con 0.00%, es decir, el valor real coincide con el predicho. El error porcentual absoluto medio para este modelo y grupo fue de 4.30% (ver Tabla 5.26 para mayor detalle de los valores del error porcentual absoluto medio para este modelo y grupo).Los coeficientes de regresión del modelo F3-C51 se muestran en la Tabla 5.20.

Tabla 5.20 Coeficientes de regresión del modelo F3-C51

Coeficiente	Valor
β0	5.64231
β1	2.35848
β2	-0.66854
β3	0.76724

5.8. Fase de evaluación

La evaluación de los modelos implica el cálculo de diversas medidas de diagnóstico y otros resultados tales como tablas y gráficos, lo que permite interpretar la calidad del modelo y su eficacia para resolver el problema [30].

Los datos utilizados para la construcción y evaluación de los modelos se escogieron de acuerdo al dominio del problema. Para construir los modelos se utilizaron los datos correspondientes al periodo 1990-2015. Con el objeto de calcular el error de predicción de los modelos construidos, se tomaron en cuenta los años correspondientes al mismo periodo.

Una de las medidas que se utilizan para evaluar la precisión de un modelo de regresión es el E*rror Porcentual Absoluto Medio (MAPE,* por sus siglas en inglés*)* [47]*.* Esta medida denota el promedio absoluto (en términos porcentuales) con el que difiere el valor predicho del valor real en cualquier punto. Es decir, entre menor sea su valor, mayor precisión tendrá el modelo. Generalmente, la medida MAPE ha sido utilizada en aplicaciones de predicción. Para esta investigación se utilizará esta medida para evaluar los modelos construidos. La fórmula para determinarla es la siguiente:

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|A_t - F_t|}{|A_t|}}{n}$$

Donde:

$n = número\ de\ datos$

$A_t = valor\ real\ del\ t\ ésimo\ dato$

$F_t = valor\ estimado\ del\ t\ ésimo\ dato$

En la Tabla 5.21 se muestran los valores de evaluación del periodo 1990-2015 para el modelo de grado 2 y grupo C24.

Tabla 5.21 Evaluación del modelo de regresión de grado 2 del grupo C24

Año	Real	Predicho	Error absoluto	% Error absoluto	MAPE
1990	8.757329299	8.83017839	0.072849091	0.83%	
1991	9.195706351	8.883041627	0.312664724	3.40%	
1992	8.994140807	8.929745349	0.064395458	0.72%	
1993	8.792575263	8.976449071	0.183873808	2.09%	
1994	8.946728427	9.016993277	0.07026485	0.79%	
1995	9.063628954	9.053431139	0.010197815	0.11%	
1996	8.903896646	9.085762657	0.181866011	2.04%	
1997	9.359420559	9.11398783	0.245432729	2.62%	
1998	9.32875631	9.13810666	0.19064965	2.04%	
1999	9.059911069	9.158119146	0.098208077	1.08%	
2000	8.58680178	9.174025287	0.587223507	6.84%	
2001	9.630792483	9.185825085	0.444967398	4.62%	
2002	8.842793162	9.193518539	0.350725377	3.97%	
2003	9.423313475	9.197105648	0.226207827	2.40%	
2004	9.549116709	9.196586414	0.352530295	3.69%	
2005	8.794548659	9.191960835	0.397412176	4.52%	
2006	8.954189617	9.181175741	0.226986124	2.53%	2.12 %
2007	9.113830574	9.170390646	0.056560072	0.62%	
2008	9.229057777	9.153446036	0.075611741	0.82%	
2009	9.171635664	9.132395081	0.039240583	0.43%	
2010	9.171635664	9.107237783	0.064397881	0.70%	
2011	9.171635664	9.07797414	0.093661524	1.02%	
2012	9.171635664	9.044604154	0.12703151	1.39%	
2013	9.171635664	9.007127823	0.164507841	1.79%	
2014	8.768641365	8.965545148	0.196903783	2.25%	
2015	8.768641365	8.919856129	0.151214764	1.72%	

En la Tabla 5.22 se muestran los valores de evaluación del periodo 1990-2015 para el modelo de grado 2 y grupo C08.

Tabla 5.22 Evaluación del modelo de regresión de grado 2 del grupo C08

Año	Real	Predicho	Error absoluto	% Error absoluto	MAPE
1990	6.49127	6.620794	0.129524	2.00%	
1991	7.017044	6.70865	0.308393	4.39%	
1992	6.356133	6.784638	0.428505	6.74%	
1993	6.757147	6.860626	0.103479	1.53%	
1994	6.513403	6.924746	0.411342	6.32%	
1995	6.876127	6.980953	0.104827	1.52%	
1996	7.11722	7.029249	0.087971	1.24%	
1997	7.629983	7.069633	0.560351	7.34%	
1998	7.686394	7.102104	0.58429	7.60%	
1999	6.989338	7.126664	0.137326	1.96%	
2000	6.32048	7.143311	0.822831	13.02%	
2001	7.622344	7.152047	0.470297	6.17%	
2002	6.462905	7.15287	0.689965	10.68%	
2003	7.738825	7.145781	0.593043	7.66%	
2004	6.78135	7.130781	0.34943	5.15%	
2005	7.152615	7.107868	0.044747	0.63%	
2006	7.152615	7.073087	0.079528	1.11%	
2007	6.949617	7.038306	0.088689	1.28%	
2008	6.638873	6.991657	0.352785	5.31%	
2009	7.028606	6.937096	0.09151	1.30%	
2010	7.028606	6.874623	0.153983	2.19%	
2011	7.028606	6.804238	0.224368	3.19%	
2012	7.028606	6.725941	0.302665	4.31%	
2013	7.028606	6.639732	0.388874	5.53%	
2014	6.184446	6.545611	0.361165	5.84%	
2015	6.184446	6.443577	0.259132	4.19%	
					4.55%

En la Tabla 5.23 se muestran los valores de evaluación del periodo 1990-2015 para el modelo de grado 2 y grupo C51.

Tabla 5.23 Evaluación del modelo de regresión de grado 2 del grupo C51

Año	Real	Predicho	Error absoluto	% Error absoluto	MAPE
1990	3.978758	4.525235	0.546477	13.73%	
1991	4.561757	4.656542	0.094784	2.08%	
1992	4.641884	4.779679	0.137794	2.97%	
1993	5.320447	4.902816	0.417632	7.85%	
1994	4.748884	5.017783	0.268899	5.66%	
1995	5.230027	5.127305	0.102722	1.96%	
1996	5.575259	5.23138	0.343879	6.17%	
1997	5.860263	5.330009	0.530255	9.05%	
1998	5.89734	5.423191	0.474148	8.04%	
1999	5.839987	5.510928	0.32906	5.63%	
2000	4.825998	5.593218	0.767219	15.90%	
2001	6.019495	5.670061	0.349434	5.81%	
2002	5.267724	5.741459	0.473735	8.99%	
2003	5.648205	5.80741	0.159206	2.82%	
2004	5.569935	5.867915	0.29798	5.35%	
2005	6.057804	5.922974	0.13483	2.23%	5.14%
2006	5.262864	5.969864	0.707	13.43%	
2007	5.905647	6.016753	0.111106	1.88%	
2008	5.951911	6.055473	0.103562	1.74%	
2009	6.070575	6.088747	0.018172	0.30%	
2010	6.070575	6.116574	0.046	0.76%	
2011	6.070575	6.138955	0.068381	1.13%	
2012	6.070575	6.15589	0.085316	1.41%	
2013	6.070575	6.167379	0.096804	1.59%	
2014	6.40156	6.173421	0.228139	3.56%	
2015	6.40156	6.174018	0.227543	3.55%	

En la Tabla 5.24 se muestran los valores de evaluación del periodo 1990-2015 para el modelo de grado 3 y grupo C24.

Tabla 5.24 Evaluación del modelo de regresión de grado 3 del grupo C24

Año	Real	Predicho	Error absoluto	% Error absoluto	MAPE
1990	8.7573293	8.91514691	0.15781761	1.80%	
1991	9.19570635	8.92779227	0.26791408	2.91%	
1992	8.99414081	8.94655545	0.02882219	0.32%	
1993	8.79257526	8.96531862	0.19615236	2.23%	
1994	8.94672843	8.98872763	0.0419992	0.47%	
1995	9.06362895	9.01425254	0.04937641	0.54%	
1996	8.90389665	9.04115738	0.13726074	1.54%	
1997	9.35942056	9.06870615	0.29071441	3.11%	
1998	9.32875631	9.09616287	0.20596476	2.21%	
1999	9.05991107	9.12279155	0.08794513	0.97%	
2000	8.58680178	9.1478562	0.56105442	6.53%	
2001	9.63079248	9.17062083	0.46017165	4.78%	
2002	8.84279316	9.19034946	0.3475563	3.93%	
2003	9.42331348	9.2063061	0.21700737	2.30%	
2004	9.54911671	9.21775476	0.32515725	3.41%	
2005	8.79454866	9.22395946	0.42627758	4.85%	2.08%
2006	8.95418962	9.22082624	0.26663662	2.98%	
2007	9.11383057	9.21769302	0.10386245	1.14%	
2008	9.22905778	9.2037499	0.02530788	0.27%	
2009	9.17163566	9.18161887	0.00998321	0.11%	
2010	9.17163566	9.15056394	0.06178655	0.67%	
2011	9.17163566	9.10984912	0.11289724	1.23%	
2012	9.17163566	9.05873842	0.11289724	1.23%	
2013	9.17163566	8.99649586	0.1751398	1.91%	
2014	8.76864137	8.92238546	0.15374409	1.75%	
2015	8.76864137	8.83567121	0.06702985	0.76%	

En la Tabla 5.25 se muestran los valores de evaluación del periodo 1990-2015 para el modelo de grado 3 y grupo C08.

Tabla 5.25 Evaluación del modelo de regresión de grado 3 del grupo C08

Año	Real	Predicho	Error absoluto	% Error absoluto	MAPE
1990	6.4912705	6.66420293	0.17293244	2.66%	4.83%
1991	7.01704357	6.73151239	0.28553118	4.07%	
1992	6.35613278	6.79322599	0.49880683	7.85%	
1993	6.7571468	6.8549396	0.15315856	2.27%	
1994	6.5134034	6.91030536	0.44753448	6.87%	
1995	6.87612674	6.96093788	0.13033443	1.90%	
1996	7.11722016	7.00646117	0.11075899	1.56%	
1997	7.62998323	7.04649922	0.583484	7.65%	
1998	7.68639435	7.08067604	0.57777872	7.52%	
1999	6.98933805	7.10861563	0.14060393	2.01%	
2000	6.32048024	7.12994198	0.82379886	13.03%	
2001	7.62234358	7.14427909	0.4710926	6.18%	
2002	6.46290521	7.15125098	0.68834577	10.65%	
2003	7.73882477	7.15048162	0.58834315	7.60%	
2004	6.78135043	7.14159503	0.34286478	5.06%	
2005	7.15261466	7.12421521	0.05927113	0.83%	
2006	7.15261466	7.09334353	0.0901428	1.26%	
2007	6.94961746	7.06247186	0.06773887	0.97%	
2008	6.63887251	7.01735632	0.37848382	5.70%	
2009	7.02860618	6.96224356	0.06636262	0.94%	
2010	7.02860618	6.89675756	0.20808386	2.96%	
2011	7.02860618	6.82052232	0.29544434	4.20%	
2012	7.02860618	6.73316184	0.39430605	5.61%	
2013	7.02860618	6.63430013	0.505045	7.19%	
2014	6.18444568	6.52356118	0.3391155	5.48%	
2015	6.18444568	6.40056899	0.21612331	3.49%	

En la Tabla 5.26 se muestran los valores de evaluación del periodo 1990-2015 para el modelo de grado 3 y grupo C51.

Tabla 5.26 Evaluación del modelo de regresión de grado 3 del grupo C51

Año	Real	Predicho	Error absoluto	% Error absoluto	MAPE
1990	3.97875826	4.2015245	0.22276624	5.60%	4.30%
1991	4.56175733	4.48605182	0.0757055	1.66%	
1992	4.64188418	4.71563595	0.3033359	6.53%	
1993	5.32044719	4.94522008	0.19497828	3.66%	
1994	4.748884	5.1254689	0.52768211	11.11%	
1995	5.23002697	5.27656611	0.17128867	3.28%	
1996	5.57525888	5.40131564	0.07273744	1.30%	
1997	5.86026315	5.50252144	0.2772757	4.73%	
1998	5.89733952	5.58298746	0.2518219	4.27%	
1999	5.83998747	5.64551763	0.14707158	2.52%	
2000	4.82599825	5.69291589	0.90198796	18.69%	
2001	6.01949515	5.7279862	0.26596265	4.42%	
2002	5.26772399	5.75353249	0.50463472	9.58%	
2003	5.64820471	5.77235871	0.13906409	2.46%	
2004	5.5699352	5.7872688	0.2311315	4.15%	
2005	6.05780421	5.8010667	0.239	3.95%	
2006	5.26286353	5.81880421	0.57367819	10.90%	
2007	5.90564674	5.83654172	0.04182002	0.71%	
2008	5.95191073	5.86382671	0.05069544	0.85%	
2009	6.07057457	5.90121529	0.11906317	1.96%	
2010	6.07057457	5.9515114	0.0530556	0.87%	
2011	6.07057457	6.01751897	0.03146739	0.52%	
2012	6.07057457	6.10204196	0.13730973	2.26%	
2013	6.07057457	6.2078843	0.26727538	4.40%	
2014	6.40156047	6.33784994	0.09318236	1.46%	
2015	6.40156047	6.49474282	6.40156047	0.00%	

Para esta investigación se utilizarán los modelos construidos por las dos funciones polinómicas descritas anteriormente. La estructura del dataset utilizado para implementar la regresión polinomial se encuentra en la Tabla 5.10.

5.9. Fase de despliegue

Una vez elegidos los modelos, se proyectaron las tasas promedio de mortalidad normalizada. El periodo de proyección está comprendido entre 2016 y 2020 para los tres grupos.

En la Figura 5.12 se muestra la proyección de las tasas promedio de mortalidad normalizada para el grupo C24, utilizando la función polinómica de grado 2. La línea roja representa los datos reales obtenidos entre 1990 y 2015; y la línea verde la proyección entre 1990 y 2020.

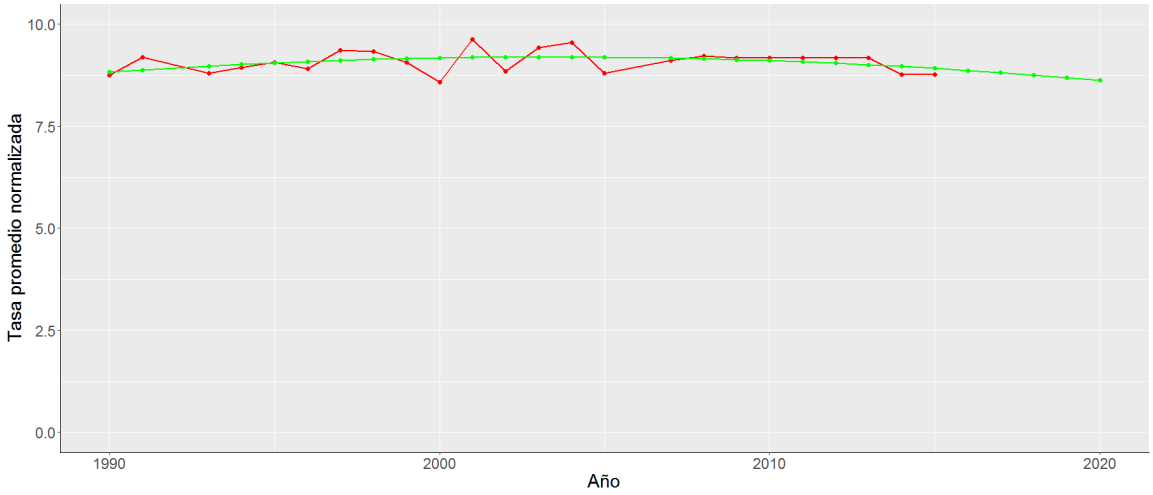

Figura 5.12 Proyección de tasas de mortalidad para el grupo C24, modelo F2-C24

El grupo C24 contiene a los municipios con las tasas promedio de mortalidad normalizada más altas. Según este modelo, el año con la mayor tasa promedio de mortalidad normalizada fue el 2003 con un valor de 9.197, a partir de allí se proyecta un descenso gradual. De continuar esta tendencia, para el año 2020 habrá disminuido la tasa promedio de mortalidad normalizada en un 6.2 % para este grupo con respecto al año 2003.

En la Figura 5.13 se muestra la proyección de las tasas promedio de mortalidad normalizada para el grupo C08, utilizando la función polinómica de grado 2.

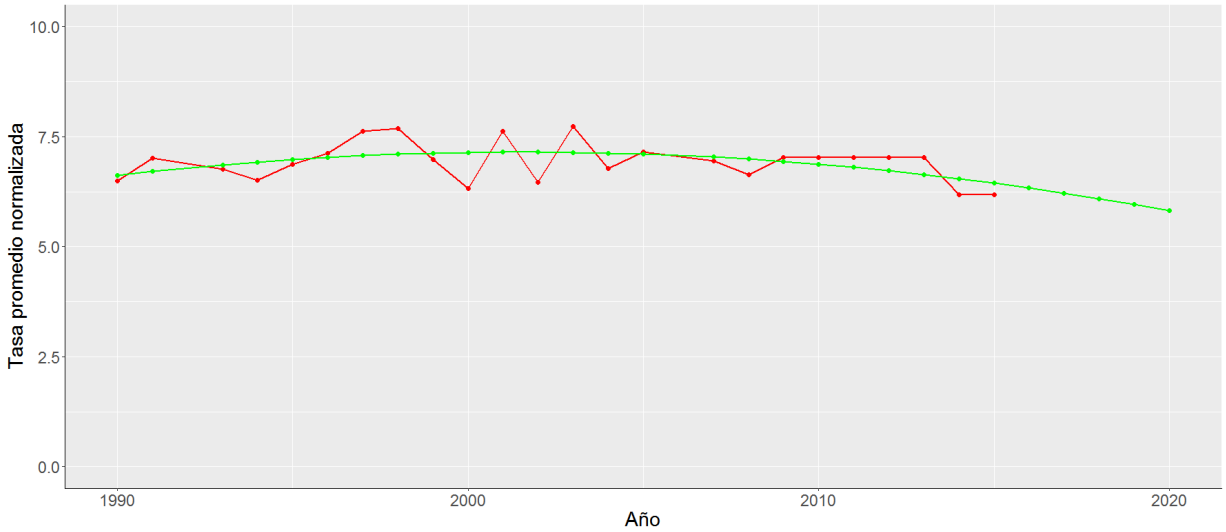

Figura 5.13 Proyección de tasas de mortalidad para el grupo C08, modelo F2-C08.

El grupo C08 tuvo un comportamiento irregular a través de los años. De acuerdo con este modelo, el año con la mayor tasa promedio de mortalidad normalizada fue 2002 con un valor de 7.152. A partir de allí, se proyecta un descenso gradual. De continuar la tendencia, para el año 2020 habrá disminuido la tasa promedio de mortalidad normalizada un 18.8 % con respecto al año 2002.

La proyección de las tasas promedio de mortalidad normalizada para el grupo C51 se presenta en la Figura 5.14, utilizando la función polinómica de grado 2.

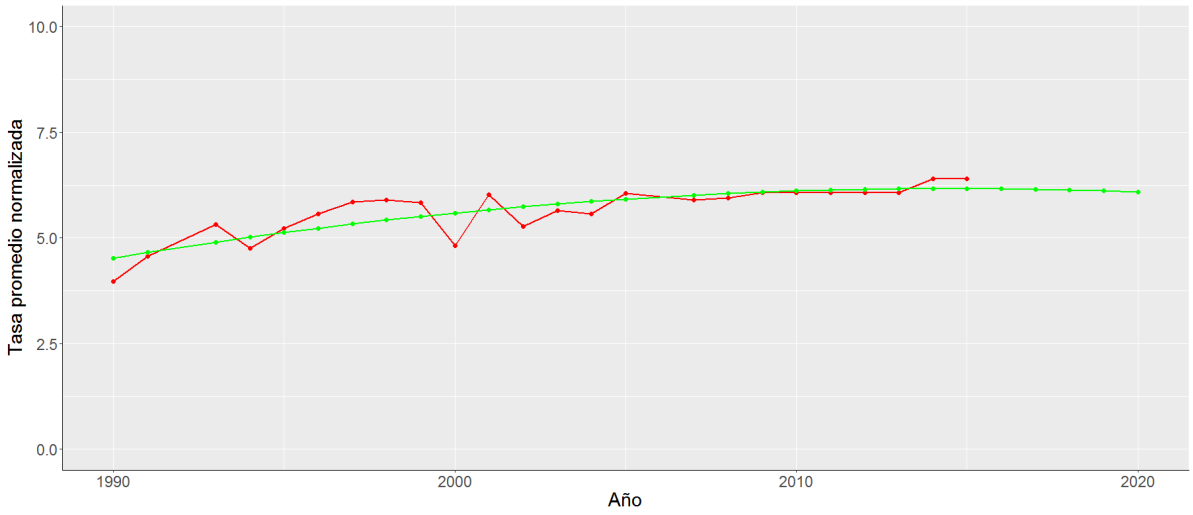

Figura 5.14 Proyección de tasas de mortalidad para el grupo C51, modelo F2-C51

De acuerdo con la proyección del modelo construido con la función polinomial de grado 2, para el grupo C51 la tasa promedio de mortalidad normalizada disminuirá para el año 2020 1.3% respecto al 2015.

En la Figura 5.15 se muestra la proyección de la tasa promedio de mortalidad normalizada para el grupo C24, utilizando la función polinómica de grado 3. La línea roja representa los datos reales de 1990-2015 y la línea azul la proyección de 1990-2020.

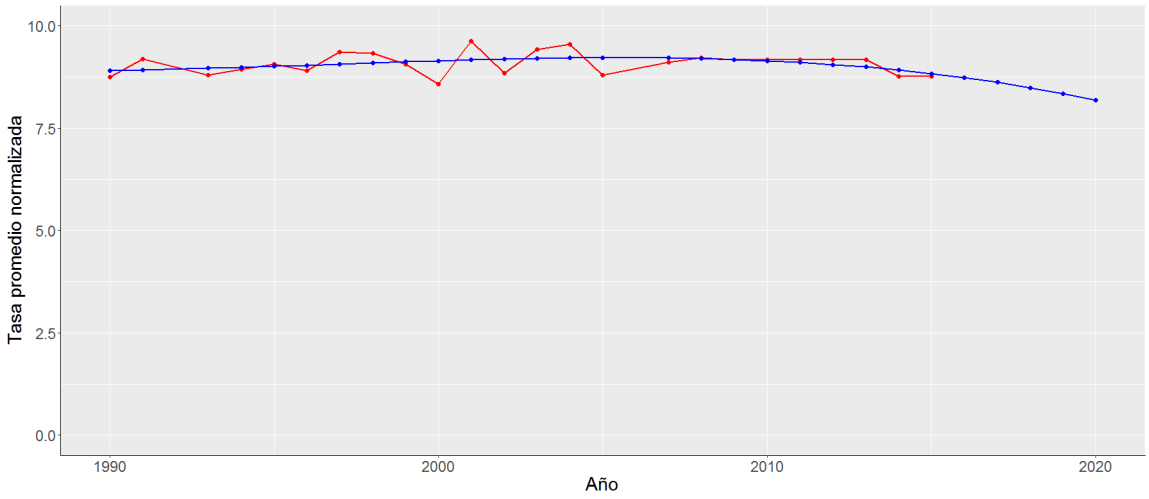

Figura 5.15 Proyección de tasas de mortalidad para el grupo C24, modelo F3-C24

Para este modelo, la tasa promedio de mortalidad normalizada para el año 2003 tuvo un valor de 9.206. De seguir esta tendencia en el grupo, para el año 2020 habrá disminuido en un 11.1 %, con respecto al año 2003.

En la Figura 5.16 se muestra la proyección de la tasa promedio de mortalidad normalizada para el grupo C08, utilizando la función polinómica de grado 3.

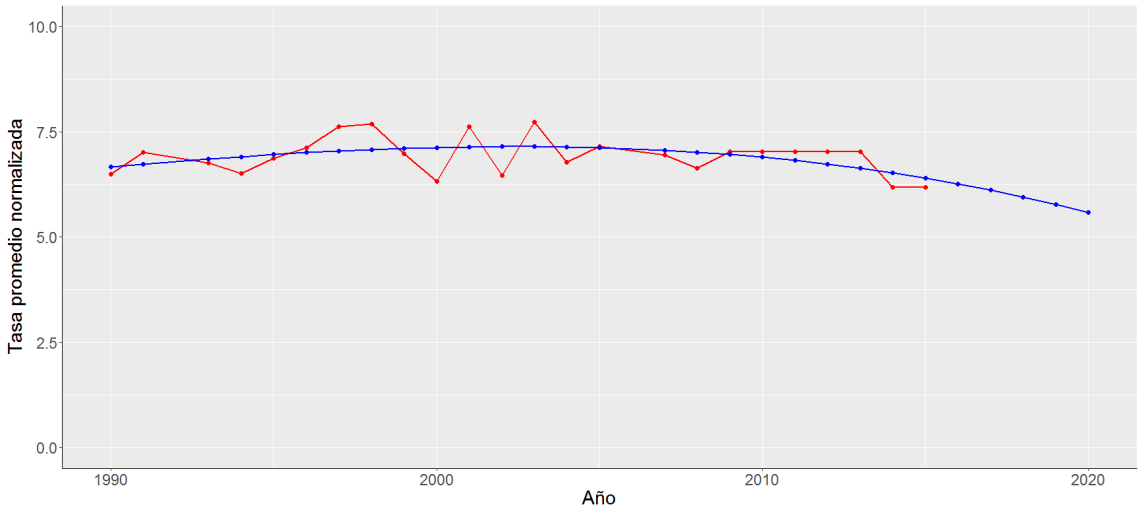

Figura 5.16 Proyección de tasas de mortalidad para el grupo C08, modelo F3-C08.

El año con la mayor tasa promedio de mortalidad normalizada de acuerdo con este modelo es el 2002 con un valor de 7.151. Para el año 2020 se tiene proyectado un descenso del 21.9 %, con respecto al año 2002. La proyección de la tasa promedio de mortalidad normalizada para el grupo C51 se presenta en la Figura 5.17, utilizando la función polinómica de grado 3.

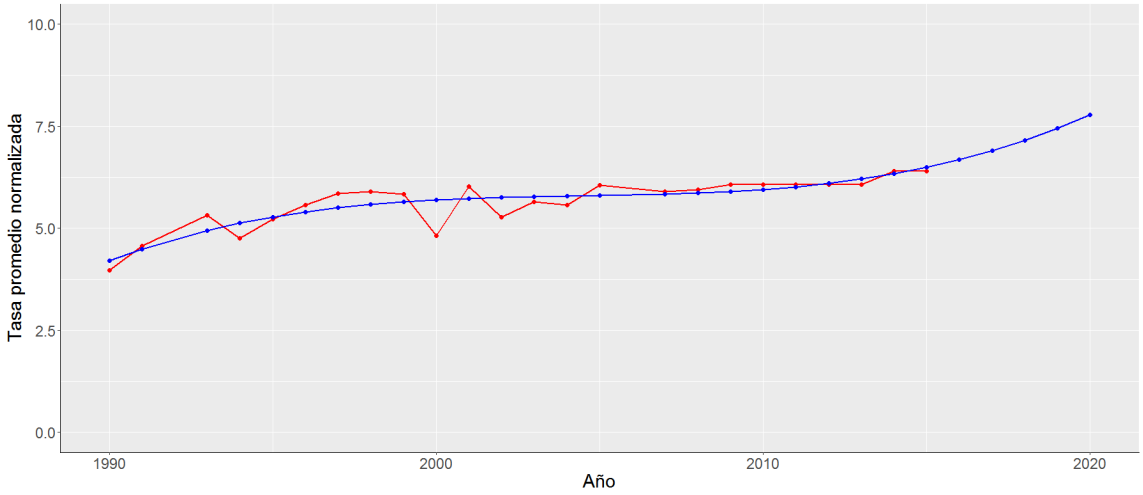

Figura 5.17 Proyección de tasas de mortalidad para el grupo C51, modelo F3-C51

Tal y como se observa en la tendencia real de este grupo, se proyecta que la tasa promedio de mortalidad normalizada se incremente. De continuar esta tendencia, se espera que este incremento sea del 16.6 % para el año 2020, con respecto al 2015.

5.10. Fase de retroalimentación

La etapa de retroalimentación es la última que comprende la metodología FMDS. Consiste en implementar la solución dentro del entorno organizacional. Para ello, se debe conocer bien el dominio en el que se aplica la solución propuesta. Lo anterior, implica relacionarse estrechamente con la organización y sus roles. Cada rol en la organización debe monitorear por un tiempo razonable (tres meses, seis o un año) el rendimiento del modelo desplegado y realizar registros. Estos registros servirán para evaluar el rendimiento del modelo y determinar si se están alcanzando los objetivos de la organización. Cada rol de la organización adquirirá el conocimiento de lo que se debe mejorar en el modelo, para que se realicen los cambios necesarios y se vuelva a desplegar [30].

En esta investigación se abordó un caso real con datos oficiales. Sin embargo, se trata de un proyecto académico que aún no está vinculado con una institución o empresa. Por lo tanto, las actividades propuestas en esta etapa solamente se describen, sin llegar a implantarse.

Las actividades propuestas son las siguientes:

1. Monitorear cada cinco años las estadísticas de las tasas promedio de mortalidad normalizada por diabetes mellitus para los municipios que conforman los grupos C24, C08 y C51,
2. Realizar registros anuales de las diferencias entre las tasas promedio de mortalidad normalizada reales y las predichas por el modelo,
3. Determinar y ejecutar los cambios al modelo de acuerdo a los registros obtenidos,
4. Desplegar la técnica computacional con el nuevo modelo generado.

CAPÍTULO 6

CONCLUSIONES Y TRABAJOS FUTUROS

En este capítulo se exponen las conclusiones obtenidas como resultado de la presente investigación. También se incluyen propuestas de temas para aplicaciones y estudios posteriores.

6.1. Conclusiones

Con los resultados de la presente investigación se muestra que es factible asimilar conceptos de Ciencia de Datos y crear una infraestructura de conocimiento que apoye el desarrollo de aplicaciones de Ciencia de Datos.

Se validaron los conceptos de Ciencia de Datos por medio del desarrollo de un caso práctico. Se utilizó la metodología “Foundational Methodology for Data Science” -propuesta por la empresa IBM- para desarrollar un caso práctico. Además, se utilizó el lenguaje de programación estadística R como apoyo a las actividades realizadas.

La aplicación de la metodología de Ciencia de Datos tuvo el objetivo de proyectar las tasas de mortalidad por diabetes mellitus tipos E11-E14 en regiones de municipios de México en el periodo 2016-2020. Las regiones de análisis fueron clasificadas como C24, C08 y C51. Representan a los 25 municipios del país con las mayores tasas de mortalidad por diabetes mellitus. En el análisis se utilizaron datos poblacionales obtenidos de instituciones oficiales como, SINAIS, INEGI, CONAPO y CEMECE.

Con los resultados obtenidos al aplicar la proyección de las tasas de mortalidad por diabetes mellitus, se destaca que en la región C24 de continuar con la tendencia actual, se prevé un descenso para el año 2020 en un rango del 6.2 al 11.1% con respecto al año 2003 –punto máximo en la tasa de mortalidad. En la región C08, para el año 2020 se prevé un descenso en un rango del 18.8 al 21.9% con respecto al año 2002. Sin embargo, los modelos de predicción aplicados a la región C51, prevén que la tasa de mortalidad oscilará entre un – 1.3 y +16.6% para el año 2020 con respecto al 2015.

Cabe destacar que las regiones C24 y C08 – cuyas proyecciones presentan un descenso en la tasa de mortalidad- tienen 1.45 y 4.64 millones de habitantes respectivamente, mientras que la región C51 cuenta con 4.94 millones de habitantes.

Esta investigación representó diversos retos computacionales y aborda un problema para la sociedad mexicana. Desde el enfoque computacional, es importante destacar que se realizó un estudio acerca de la Ciencia de Datos -un área emergente en ciencia y tecnología-, la cual ha generado mucho impacto en

los sectores empresarial, gubernamental y académico; se seleccionó una metodología de Ciencia de Datos que representara el proceso de desarrollo de casos prácticos y se construyeron modelos de proyección de datos mediante el lenguaje R –lenguaje de programación estadístico muy usado en Ciencia de Datos. Los beneficios tecnológicos aportados en este trabajo son: a) Iniciar con la investigación en Ciencia de Datos en el CENIDET, b) desarrollo de aplicación relacionada a una enfermedad de importancia nacional aplicando conocimientos de un área emergente en la computación.

Desde una perspectiva social, es necesario enfatizar que la diabetes mellitus es una enfermedad que ha causado un alto índice de mortalidad en México en los últimos años a pesar de que el gobierno ha destinado grandes cantidades de recursos económicos para detectarla, prevenirla y atenderla.

6.2. Trabajos futuros

Por ser la primera investigación en el área de Ciencia de Datos al interior del CENIDET, se abre la posibilidad de desarrollar nuevos trabajos relacionados con ella. Se sugieren algunos temas para profundizar en el estudio de la Ciencia de Datos aplicado al área de la salud, en el campo de la Epidemiología:

1. Realizar una adaptación de la Metodología Fundacional para Ciencia de Datos para aplicarlo específicamente al área de Epidemiología,
2. Realizar un agrupamiento de los municipios que para el año 2020 presentarían las mayores tasas de mortalidad por diabetes mellitus y graficarlos.

REFERENCIAS

[1] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Data Science and Big Data*, vol. 1, no. 1, pp. 51–59, March 2013.

[2] M. A. Waller and S. E. Fawcett, “Data Science , Predictive Analytics , and Big Data : A Revolution That Will Transform Supply Chain Design and Management,” *Journal of bussiness logistics,* vol. 34, no. 2, pp. 77–84, 2013.

[3] S. Ozdemir, *Principles of Data Science, 1st. ed.,* Birmingham, UK: *Pack Publishing Ltd*., 2016.

[4] E. P. Luna, “Data Mining Oriented to Big Data in the Health Area,” in Master Thesis, Department of Computational Sciences, National Center for Technological Research and Development, Cuernavaca, Morelos, México, 2016.

[5] I. I. B. Carrillo, “Aplicación de Minería de Datos en el Área de Salud Pública,” in Master Thesis, Department of Computational Sciences, National Center for Technological Research and Development, Cuernavaca, Morelos, México, 2017.

[6] F. Provost and T. Fawcett, “Data Science for Business,” 1st. ed., Gravenstein Highway North, Sebastopol, CA: O’Reilly Media, 2013.

[7] E. Brynjolfsson, L. M. Hitt and H. H. Kim, “Strength in Numbers: How does data-driven decision-making affect firm performance?” in *International Conference of Information Systems*, Shangai, China, p. 18, 2011.

[8] T. Johns. (2018, March 22). The Johns Hopkins Data Science Lab [Online]. Available: http://jhudatascience.org/.

[9] C. University. (2017, Aug 30). Data Science for Social Good [Online].

Available: https://dssg.uchicago.edu/.

[10] U. Columbia. (2018, Feb 22). Data Science Institute [Online]. Available: http://datascience.columbia.edu/columbia-data-science.

[11] INFOTEC. (2018, Feb 22). Master of Science in Data Science [Online].Available:https://www.infotec.mx/en_gb/infotec/maestria_en_ciencia_de_datos_mcd.

[12] T. H. Davenport and D. J. Patil, “Data scientist: the sexiest job of the 21st century.,” *Harvard Bussiness Review*, vol. 90, no. 10, pp. 70–77, 2012.

[13] E. P. Moreno, “Numerical study of large-scale algorithms for Machine Learning,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México, 2014.

[14] A. G. Iñigo, “Scaling methods in optimization in vector support machines,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México, 2014.

[15] A. G. Tapia, “Importance of socioeconomic factors in the classification of hydrometeorological disasters,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México, 2015.

[16] F. G. Valdés, “Recommendation system for similar hotels,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México 2016.

[17] C. E. P. Araiza, “How to detect corruption, collusion and fraud in contracts awarded by the World Bank,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México, 2016.

[18] A. F. D. del Castillo, “Proposal for the creation of the advanced analytical area in the Bolsa Mexicana de Valores Group,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México, 2016.

[19] L. M. Núñez, “Automatic labeling of editorial content in Spanish with BM25,” in Master Thesis, Autonomous Technological Institute of Mexico, Ciudad de México 2017.

[20] R. Bellazzi and A. Abu-Hanna, “Data mining technologies for blood glucose and diabetes management,” *Journal of diabetes Science and Technology*, vol. 3, no. 3, pp. 603–612, 2009.

[21] W. Oh *et al.*, “Type 2 Diabetes Mellitus Trajectories and Associated Risks,” *Big Data*, vol. 4, no. 1, pp. 25–30, 2016.

[22] K. Rajesh and V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis,” *International Journal of Engeneering and Innovative Technology*, vol. 2, no. 3, pp. 224–229, 2012.

[23] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, “Application of data mining: Diabetes health care in young and old patients,” *Journal of King Saud Univ. - Computer and Information Sciences.*, vol. 25, no. 2, pp. 127–136, 2013.

[24] H. S. Kim, A. M. Shin, M. K. Kim, and Y. N. Kim, “Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining,” *Korean Journal of Internal Medicine*, vol. 27, no. 2, pp. 197-202, 2012.

[25] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors,” *Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93–9, 2013.

[26] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, “Data-mining technologies for diabetes: a systematic review,” *Journal of Diabetes Science and Technology.*, vol. 5, no. 6, pp. 1549–1556, 2011.

[27] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” *Compututational Structural and Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[28] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,” *Procedia Computer Science*, vol. 82, no. 30, pp. 115–121, 2016.

[29] J. A. Davis and L. D. Burgoon, “Can data science inform environmental justice

and community risk screening for type 2 diabetes?," *PLoS One*, vol. 10, no. 4, pp. 1–14, 2015.

[30] J. B. Rollins. (30, Aug 2017). Foundational Methodology for Data Science [Online].Available:https://tdwi.org/~/media/64511A895D86457E964174EDC5C4C7B1.PDF

[31] D. Rose, *Data science create teams that aks the right questions and deliver real value*. 1st ed. Atlanta: Springer Science + Bussiness Media, 2012.

[32] R. Sampieri, C. Collado, Ma. Baptista *Investigation Methodology*. 5st ed. Mexico D.F: McGraw-Hill, 2010.

[33] O. M. de la Salud. (2017, abril 24). Informe Mundial de la diabetes [Online]. Available:http://apps.who.int/iris/bitstream/handle/10665/254649/9789243565255spa.pdf;jsessionid=A7DD4A67DDADA0259EB4DB70E7080E08?sequence=1

[34] F. M. de Diabetes. (2017,Octubre 07). Federación Mexicana de Diabetes – Diabetes en México [Online]. Available: http://fmdiabetes.org/diabetes-en-mexico/.

[35] J. E. Paloheimo, "Estimation of Motality Rates in Fish Populations," *Transactions of the American Fisheries Society*, vol. 124, no. 3, pp. 347–355, 2011.

[36] C. J. Murray, A. D. Lopez, B. Chin, D. Feehan, and K. H. Hill, "Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918-20 pandemic: a quantitative analysis," *Lancet*, vol. 368, no. 9554, pp. 2211–2218, 2006.

[37] J. S. T. Wong, J. J. Forster, and P. W. F. Smith, "Bayesian mortality forecasting with overdispersion," *Insurance: Mathematics and Economics*, no. 2007, 2017.

[39] M. B. Pérez, D. L. B. De Souza, F. J. G. Bernal, and G. J. G. Bernal, "Estimation of projections of incidence, prevalence and mortality rates due to melanoma in Spain," *Medicina Cutánea Ibero-Latino-Americana*, vol. 42, no.

1–3, pp. 23–29, 2014.

[40] D. Mitchell, P. Brockett, R. Mendoza-Arriaga, and K. Muthuraman, “Modeling and forecasting mortality rates,” *Insurance Mathematics and Economics*, vol. 52, no. 2, pp. 275–285, 2013.

[41] H. Booth and L. Tickle, “Mortality modeling and forecasting: A review of methods,” *A.A.S.,* vol. 43, no. 3, pp. 3–43, 2008.

[42] A. Novokreshchenova, “Predicting Human Mortality: Quantitative Evaluation of Four Stochastic Models,” *Risks*, vol. 4, no. 4, p. 45, 2016.

[43] C. Pedroza, “A Bayesian forecasting model: Predicting U.S. male mortality,” *Biostatistics*, vol. 7, no. 4, pp. 530–550, 2006.

[44] S. Jain and N. Mishra, “Forecasting of literacy rate using statistical and data mining methods,” *International Journal of Advanced Computational Engenieeering and Networking*, vol. 3, no. 8, pp. 26–31, 2015.

[45] J. Hernández, M. Ramirez and C. Ferri, Introducción a la Minería de Datos. 1st ed,. Spain: Prentice Hall, pp. 49-93, 2001.

[46] F. Gorunescu, *Data Mining: Concepts and Techniques*. 2nd. ed., San Francisco: Morgan Kaufmann Publishers, 2011.

[47]

[48] W. S. Cleveland, “Data science: An action plan for expanding the technical areas of the field of statistics,” *International Statistical Review*, vol. 7, no. 6, pp. 414–417, 2001.

[49] J. Stanton, *An Introduction to Data Science*. 1st. ed., New York, pp. 1–157, 2012.

[50] Z. Ingvaldsen, Jon Espenzgbek and J. A. Gulla, *Data Science from Scratch*. 1st. ed., California: O´Reilly Media, 2015.

[51] R. Schutt and C. O´neil, *Doing Data Science straight talk from the frontline*. 1st. ed., California: O´Reilly Media, 2014.

[52] B. Steele, J. Chandler, and S. Reddy, *Algorithms for Data Science*. 1st. ed.,

Cham, Switserland: Springer Nature, 2016.

[53] M. Al Hasan and M. J. Zaki, “A survey of Link prediction in Social Networks,” *Social Network Data Analalytics*, vol.3, no. 9, pp. 243–275, 2011.

[54] C. C. Aggarwal, *Data Mining: The Textbook*. 1st. ed., New York: Springer, 2015.

[55] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data mining*. 1st. ed., Boston: Pearson Education. 2005.

[56] R. D. Peng, *R Programming for Data Science.* 1st. ed., *New York: R Foundation.*, 2015.

[57] L. Massaron, J. P. Mueller, *Phyton for Data Science for dummies*. 1st. ed., New Jersey: John Wiley, 2015.

[58] P. S. Foundation. (2017, Aug 29). Phyton Project [Online]. Available: https://www.python.org/.

[59] A. Syropoulos and K. K. C. Loverdos, *Steps in Scala an Introduction to Object-Functional Programming*. 1st. ed., New York: Cambridge Press, 2010.

[60] J. King and R. Magoulas, *2016 Data Science Salary Survey*. 1st. ed.,California: O´Reilly Media, 2016.

[61] SAS Institute Inc. (2017, Aug 30). SAS [Online].Available: https://www.sas.com/es_mx/home.html.

[62] I. SPSS. (2017, Aug 30) IBM SPSS Statistics [Online]. Available: https://www.ibm.com/mx-es/marketplace/spss-statistics. [Accessed: 30-Aug-2017].

Anexo A

Reseña de fuentes de información de Ciencia de Datos

El gran interés que ha generado la Ciencia de Datos en la comunidad científica ha permitido que la cantidad de información relacionada a ella vaya creciendo. Se han escrito artículos y libros, se han creado revistas y programas académicos con la finalidad de generar conocimiento en esta área.

A continuación se presentan algunos de los recursos de información que proporcionan los conceptos fundamentales de la Ciencia de Datos.

1. En el artículo **"Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics"** [48] William S. Cleveland propuso la creación de la Ciencia de Datos como disciplina. Desarrolló un plan para implementar un departamento de Ciencia de Datos en las universidades, el cual tendría seis áreas:
   - Investigaciones multidisciplinarias**:** colaboraciones de análisis de datos en una colección de áreas de importancia,
   - Modelos y métodos para los datos: modelos estadísticos; métodos para la construcción de modelos; métodos de estimación y distribución basados en inferencia probabilística,
   - Cómputo de datos: sistemas de hardware; sistemas de software, algoritmos computacionales,
   - Pedagogía: planeación de curriculum y enfoques de enseñanza para las escuelas primaria, secundaria, bachillerato, universidad y entrenamiento corporativo,
   - Evaluación de herramientas: encuestas de herramientas en uso, encuestas de necesidades percibidas para nuevas herramientas y estudios de los procesos para el desarrollo de nuevas herramientas,
   - Teoría: Fundamentos de Ciencia de Datos, enfoques generales para modelos y métodos, cómputo de datos, enseñanza y evaluación de herramientas.

2. En el libro **"An introduction to Data Science"** [49] se describen las habilidades que deben poseer los científicos de datos. Incluye información acerca del origen y representación de los datos y se presentan ejemplos de análisis de datos básicos utilizando el lenguaje R (lenguaje de programación estadística).
3. El libro **"Data Science: create teams that ask the right questions and deliver real value"** [32] aborda el tema de la Ciencia de Datos desde la perspectiva de las tareas involucradas en el desarrollo de proyectos. Se incluye información de los diferentes tipos de datos. El autor describe el proceso para formar un equipo de Ciencia de Datos y las características que éste debe tener para que sea exitoso.
4. El libro **"Data Science for business"** [6] presenta información detallada de los conceptos de Ciencia de Datos, su relación con paradigmas como Big Data y sus principales objetivos. También se pueden encontrar los principios de Ciencia de Datos. A partir del Capítulo 2 se detallan las técnicas y algoritmos que se utilizan en un proyecto de Ciencia de Datos.
5. El artículo **"Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Managemen"** [2] define los conceptos de Ciencia de Datos y análisis predictivo desde una perspectiva del diseño y gestión de la cadena de suministro (SCM, por sus siglas en inglés), se plantean los retos que surgen por la gran cantidad de datos generados y los beneficios que existen al usar la Ciencia de Datos.
6. En el libro **"Data Science from Scratch"** [50] se describe el aumento en la cantidad de los datos. Se incluyen las tareas y habilidades que necesitan poseer los científicos de datos. También se describen las disciplinas, tareas, técnicas y algoritmos utilizados en el desarrollo de un proyecto de Ciencia de Datos. Incluye ejemplos de análisis de datos utilizando el lenguaje Phyton.
7. En el artículo **"Foundational Methodology for Data Science"** [30]**,** IBM propone una metodología para la Ciencia de Datos, donde se describen 10 etapas para desarrollar un proyecto de Ciencia de Datos.
8. El libro **"Principles of Data Science"** [3] contiene información de las terminologías básicas de Ciencia de Datos: conceptos, objetivos y enfoques.

Se presenta el Diagrama de Venn de la Ciencia de Datos donde se describen las disciplinas que integran la Ciencia de Datos. En el Capítulo 2 se describen a detalle las características de los datos que se utilizan en la Ciencia de Datos. En el Capítulo 3 del libro se presenta el proceso para desarrollar un proyecto de Ciencia de Datos. En los Capítulos subsecuentes se introduce a los temas más importantes de las disciplinas involucradas en la Ciencia de Datos.

En la Tabla A.1 se muestra la comparación entre las diferentes fuentes de información presentadas.

Tabla A.1 Fuentes de información de Ciencia de Datos

No.	Fuente	Año	Tipo fuente	Temas
1	Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics	2001	Artículo	• Propone a la Ciencia de Datos como disciplina independiente
2	An introduction to Data Science	2012	Libro	• Habilidades del científico de datos • Concepto de datos • Ejemplos de análisis de datos
3	Data Science: create teams that ask the right questions and deliver real value	2012	Libro	• Concepto de Ciencia de Datos • Tipos de datos • Creación de equipo de Ciencia de Datos
4	Data Science for business	2013	Libro	• Concepto de Ciencia de Datos • Objetivos de la Ciencia de Datos • Principios de la Ciencia de Datos • Enfoques de la Ciencia de Datos • Tareas y técnicas de la Ciencia de Datos
5	Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management	2013	Artículo	• Concepto de Ciencia de Datos y términos relacionados • Crecimiento de los datos • Aplicaciones de la Ciencia de Datos
6	A comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)	2014	Artículo	• Metodologías de minería de datos
7	Data Science from Scratch	2015	Libro	• Tareas del científico de datos • Disciplinas de la Ciencia de Datos • Técnicas de Ciencia de Datos
8	Foundational Methodology for Data Science	2015	Artículo	• Metodología para Ciencia de Datos
9	Principles of Data Science	2016	Libro	• Conceptos de Ciencia de Datos • Disciplinas de Ciencia de Datos • Tipos de datos • Proceso de Ciencia de Datos

Desde el 2001 se propone a la Ciencia de Datos como una disciplina independiente. A partir de ese año, han surgido muchas fuentes de información que hablan de ella, tales como libros y artículos. Los temas más recurrentes son el concepto, objetivos, aplicaciones y disciplinas que integran la Ciencia de Datos.

Anexo B

Principios de Ciencia de Datos

Hay un conjunto de conceptos fundamentales subyaciendo los principios de extracción de conocimiento de los datos, con apoyo teórico y empírico. Estos conceptos fundamentales de la Ciencia de Datos son extraídos de muchos campos que estudian el análisis de datos. Algunos reflejan la relación entre la Ciencia de Datos y problemas de negocios a resolver. Algunos reflejan los tipos de descubrimientos del conocimiento que se pueden hacer y que son la base para soluciones técnicas [1]. Los principios de la Ciencia de Datos se presentan a continuación:

1. **La extracción de conocimiento útil de los datos para resolver problemas de negocios se puede realizar sistemáticamente al seguir un proceso con etapas definidas razonablemente [1].**

Al mantener un proceso en mente se puede estructurar el pensamiento acerca de los problemas de análisis de datos. Un pensamiento estructurado acerca del análisis enfatiza –en ocasiones- aspectos despreciados de apoyo a la toma de decisiones con los datos. Tal pensamiento estructurado también contrasta con puntos críticos en los que la intuición humana y creatividad son necesarias en contra de puntos en los que las herramientas analíticas de gran potencia pueden llegar a producir.

2. **Evaluar los resultados de Ciencia de Datos requiere consideración cuidadosa del contexto en el cual serán usados [1].**

Si el conocimiento extraído de los datos ayudará a la toma de decisiones depende críticamente de la aplicación en cuestión. Muchos marcos de trabajo de evaluación de Ciencia de Datos se basan en este concepto fundamental.

3. **La relación entre el problema de negocio y la solución de análisis a veces puede ser descompuesta en subproblemas manejables a través de marcos de trabajo de análisis de valor esperado [1].**

Existen varias herramientas para minería de datos, pero los problemas de negocio raramente vienen preparados específicamente para sus aplicaciones. Es ampliamente útil descomponer el problema de negocio en componentes que correspondan a la estimación de probabilidades y cómputo o estimación de valores, junto con una estructura para recombinar los componentes.

4. **En grandes cantidades de datos, las Tecnologías de la Información (TI) se pueden usar para encontrar información de atributos descriptivos de entidades de interés [1].**

Uno de los primeros conceptos de Ciencia de Datos aplicados en los negocios es la *correlación*. La correlación, en ocasiones, se usa libremente para dar a entender que algunos elementos de datos proveen información acerca de otros – específicamente, cantidades conocidas que reducen nuestra incertidumbre sobre las desconocidas. En este concepto fundamental subyace un gran número de técnicas para el análisis estadístico, modelado predictivo y otras técnicas de minería de datos.

5. **Las entidades que son similares con respecto a características o atributos conocidos en ocasiones son similares con respecto a desconocidos [1].**

Computar la similitud es una de las herramientas principales de la Ciencia de Datos. Existen muchas formas para computarla y cada año se inventan más.

6. **Si se observa fijamente un conjunto de datos, se encontrará algo – pero podría no generalizar más allá de los datos que se están observando [1].**

Esto se refiere a como se “sobrealimenta” un conjunto de datos. Las técnicas para la minería de datos pueden ser muy poderosas y la necesidad de detectar y evitar la sobrealimentación es uno de los conceptos más importantes a tomar en cuenta cuando se aplican para resolver problemas reales. La sobrealimentación se trata

de evitar en la mayoría de los procesos, algoritmos y métodos de evaluación de Ciencia de Datos.

7. **Para extraer conclusiones causales, se debe poner mucha atención a la presencia de factores confusos, posiblemente aún no vistos [1].**

En ocasiones, no es suficiente descubrir correlaciones en los datos, tal vez queramos usar nuestros modelos para guiar las decisiones en cómo influenciar el comportamiento al producir los datos. Todos los métodos para extraer conclusiones causales – desde la interpretación de los coeficientes de modelos de regresión hasta experimentos aleatorios controlados- incorporan suposiciones considerando la presencia o ausencia de factores confusos. Al aplicar tales métodos, es importante entender sus suposiciones claramente con la finalidad de entender el alcance de cualquier petición causal.

Anexo C

Disciplinas que integran la Ciencia de Datos

La Ciencia de Datos es un área de trabajo emergente interesada en la recopilación, preparación, análisis, visualización, administración y preservación de grandes cantidades de información. Aunque el nombre de “Ciencia de Datos” pareciera estar más fuertemente relacionado con áreas como bases de datos y ciencias de la computación, se necesitan muchos tipos de habilidades –incluyendo habilidades no matemáticas- para hacer Ciencia de Datos [49].

La Ciencia de Datos es un área multidisciplinaria. Ésta se apoya de varias disciplinas para lograr su principal objetivo: generar soluciones que apoyen la toma de decisiones basados en los datos (DDD). En 2010, Drew Conway desarrolló un diagrama de Venn con las disciplinas que integran la Ciencia de Datos [51] (Ver Figura C.1).

Figura C.1 Disciplinas de la Ciencia de Datos

- **Ciencias de la Computación:** Área que se dedica a la administración adecuada de los recursos computacionales; creación, desarrollo y optimización de herramientas de software. Provee conocimiento para

desarrollar diferentes tipos de bases de datos. Ésta se utiliza para generar soluciones utilizando la computadora.

- **Matemática/Estadística:** La Matemática, en el área de la Ciencia de Datos, utiliza ecuaciones y fórmulas para realizar el análisis de los datos. La Estadística se usa para describir los datos a analizar, de ella también se utilizan principios para determinar relaciones en los datos.

- **Conocimiento del dominio:** Se refiere al hecho de entender el dominio del problema que se está abordando, por ejemplo: salud, finanzas, ciencias sociales, transporte, seguridad, etc.

Como se puede observar en la Figura C.1, hay intersecciones entre las tres áreas. Entre las Ciencias de la Computación y la Matemática/Estadística se encuentra el *aprendizaje automático,* que es el estudio que combina el poder de las computadoras con algoritmos de aprendizaje inteligente con el fin de automatizar las relaciones en los datos y crear modelos de datos poderosos [3]. La intersección entre las Ciencias de la Computación y el Conocimiento del Dominio se encuentra el *software tradicional,* se refiere a la creación, desarrollo y administración de herramientas de software enfocadas a resolver un problema en un dominio específico. En la intersección de las disciplinas de Matemática/Estadística y Conocimiento del Dominio se encuentra la *investigación tradicional,* que es el estudio –desde el punto de vista analítico- de un problema específico. Finalmente, en la intersección de las tres disciplinas, se encuentra la *Ciencia de Datos.*

Anexo D

Técnicas de Ciencia de Datos

Uno de los principales objetivos de la Ciencia de Datos es apoyar la toma de decisiones basadas en el análisis de grandes cantidades de datos. La forma y representación de los datos iniciales no permiten hacerlo. La Ciencia de Datos es una amalgama de métodos analíticos con el propósito de extraer información de los datos [50]. Para poder extraer el conocimiento "escondido" en los datos, es necesario aplicar una serie de transformaciones a los datos para generar *modelos* que representen el comportamiento de esos datos. Un modelo de datos refiere a una relación formal y organizada entre elementos de datos [3]. Existen diversas técnicas en la Ciencia de Datos que, a través del uso de algoritmos, generan los modelos de datos. Las técnicas de Ciencia de Datos están agrupadas en dos enfoques [46]:

1. Predictivo: las técnicas predictivas se utilizan para estimar valores futuros de variables usando valores conocidos,
2. Descriptivo: las técnicas descriptivas caracterizan las propiedades de los datos con la finalidad de conocer el comportamiento de los mismos.

En esta sección mostramos las técnicas más utilizadas en Ciencia de Datos, así como algunos de los algoritmos que pertenecen a ellas.

D.1. Técnicas predictivas

Dentro de las técnicas predictivas se encuentran las siguientes:

1. Clasificación

La clasificación es el proceso de encontrar un modelo (o función) que describe y distingue clases o conceptos de datos, con el propósito de usar el modelo para predecir la clase de objetos cuya etiqueta de clase es desconocida. El modelo derivado se basa en el análisis de un conjunto de datos de entrenamiento (es decir, objetos de datos cuya etiqueta de clase es conocida) [46].

Algunos de los algoritmos utilizados en esta técnica son:

- Reglas de clasificación IF-THEN,
- Arboles de decisión,
- Redes neuronales,
- Clasificación bayesiana,
- Máquinas de Soporte Vectorial (SVM, Support Vector Machines),
- Clasificación de vecinos más cercanos (k-Nearest Neighbors).

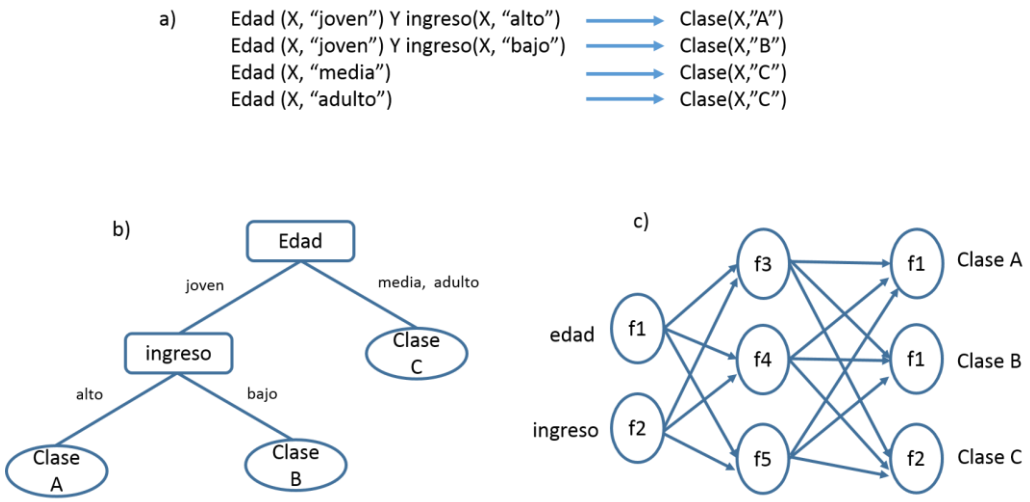

Figura D.1 Técnicas de clasificación

2. Regresión

La regresión intenta estimar o predecir, para cada individuo, el valor numérico de alguna variable para ese individuo. Un procedimiento de regresión produce un modelo que, dado un individuo, estima el valor de la variable particular específica a ese individuo. La regresión es una metodología estadística desarrollada por Sir Frances Galton (1822-1911) [46].

El análisis de regresión puede usarse para modelar la relación entre una o más variables *independientes* o *predictoras* y una variable *dependiente* o *de respuesta* (que es de valor continuo). En el contexto de la minería de datos, las variables predictoras son los atributos de interés que describen la tupla (es decir, constituyen el vector de atributo). En general, se conocen los valores de las variables predictoras. (Existen técnicas para manejar casos en los que pueden faltar dichos

valores.) La variable de respuesta es la que se desea predecir. Dada una tupla descrita por variables predictoras, se desea predecir el valor asociado de la variable de respuesta.

Los algoritmos utilizados en la regresión son:

- Regresión lineal simple (implica una sola variable predictora),
- Regresión lineal múltiple (implica dos o más variables predictoras),
- Modelos lineales generalizados,
- Regresión de Poisson,
- Modelos log-lineales,
- Arboles de regresión.

3. Predicción de enlaces

La *predicción de enlaces* es una técnica importante en el análisis de redes sociales, el cual también tiene aplicaciones en otros dominios como recuperación de información, bioinformática y comercio electrónico [53]. Esta técnica intenta predecir conexiones entre elementos de datos, por lo general sugiriendo que debe existir un enlace, y posiblemente también estimar la fuerza del enlace. La predicción de enlaces es común en los sistemas de redes sociales. Estos enlaces forman la base para las recomendaciones [6].

Los algoritmos utilizados en esta técnica son:

- Clasificación basada en características (feature-based classification),
- Método basado en Kernel (kernel-based method),
- Factorización matricial (matrix factorization).

D.2. Técnicas descriptivas

1. Coincidencia de semejanzas

La similaridad es la base de muchos métodos de Ciencia de Datos y soluciones a problemas de negocios. La *coincidencia de semejanzas* intenta identificar individuos similares basados en datos conocidos acerca de ellos. Se puede usar directamente para encontrar entidades similares. En esta técnica se utiliza la

concordancia de similitud basada en los datos que describen las características de las entidades. La igualdad de concordancia es la base de uno de los métodos más populares para hacer recomendaciones de productos. Las medidas de similitud subyacen en ciertas soluciones a otras técnicas, tales como la clasificación, la regresión y la agrupación [6].

Uno de los métodos usados en esta técnica es Vecinos más cercanos (Nearest-Neighbors).

2. Agrupamiento

Esta técnica intenta agrupar individuos en una población por su similitud, pero no por un propósito específico. El agrupamiento es útil en la exploración preliminar de dominio para ver qué grupos naturales existen porque estos grupos a su vez pueden sugerir otras tareas o enfoques de Ciencia de Datos [6]. Un problema de agrupamiento se puede definir como un problema de optimización, en el cual las variables del problema representan miembros de grupos de puntos de datos y la función objetivo compara la similaridad entre los datos en términos de esas variables [54].

Los algoritmos que pertenecen a esta técnica son:

- K-means,
- Fuzzy C-means,
- DBSCAN (Density-based Spatial Clustering of Applications with Noise, Agrupamiento Espacial de Aplicaciones con Ruido basado en Densidad),
- Agrupamiento Jerárquico Aglomerativo (Agglomerative Hierarchical Clustering), [54]
- Mezcla de Gausianos (Mixture of Gaussians)

3. Agrupamiento de co-ocurrencia

También se conoce como *minería de elementos frecuentes*, *descubrimiento de reglas de asociación* o *análisis de cesta de mercado*. Se utiliza para buscar asociaciones entre entidades basadas en transacciones que las involucren. Mientras que el agrupamiento mira la similitud entre objetos basada en los atributos

de los objetos, el agrupamiento de co-ocurrencia considera la similitud de los objetos basándose en su aparición en las transacciones. La co-ocurrencia de productos en compras es un tipo común de agrupamiento conocido como análisis de cesta de mercado. El resultado del agrupamiento de co-ocurrencia es una descripción de elementos que ocurren juntos. Estas descripciones suelen incluir estadísticas sobre la frecuencia de la co-ocurrencia [6].

Los algoritmos aplicados en esta técnica son:

- Reglas de asociación (Association Rule Mining),
- Algoritmo A priori.

4. Perfilado

Esta técnica también es conocida como *descripción del comportamiento*, intenta caracterizar el comportamiento típico de un individuo, grupo o población. El comportamiento se puede describir generalmente sobre una población entera, o al nivel de pequeños grupos o incluso de individuos. La elaboración de perfiles se utiliza a menudo para establecer normas de comportamiento para las aplicaciones de detección de anomalías, como la detección de fraudes y el monitoreo de intrusiones en sistemas informáticos. Por ejemplo, si se sabe qué tipo de compras suele hacer una persona en una tarjeta de crédito, se puede determinar si un nuevo cargo en la tarjeta se ajusta a ese perfil o no. Se puede usar el grado de desajuste como una puntuación de sospecha y emitir una alarma si es demasiado alta [6].

Anexo E

Herramientas de Ciencia de Datos

En las Ciencias de la Computación es indispensable el uso de herramientas para acelerar el trabajo, desde la administración de sistemas de cómputo, hasta el análisis de grandes cantidades de datos. El área de Ciencia de Datos no es la excepción. Hoy en día hay muchas herramientas (tanto libres como propietarias) que facilitan la extracción de conocimiento de grandes cantidades de datos.

En esta sección se citarán algunos de los lenguajes y herramientas utilizadas en Ciencia de Datos.

E.1. Herramientas gratuitas

1. Lenguaje R

R es un lenguaje de programación estadística de código abierto. Fue desarrollado en 1991 por Ross Ihanka y Robert Gentleman en el Departamento de Estadística de la Universidad de Auckland, concebido como una evolución del lenguaje S el cual fue desarrollado en los Laboratorios Bell (originalmente parte de AT & T Corp, ahora Lucent Technologies) [56]. Es un lenguaje desarrollado por estadísticos para estadísticos, que provee una amplia variedad de técnicas estadísticas y gráficas. Hoy en día es uno de los lenguajes más utilizados en la Ciencia de Datos debido a las siguientes características:

- Provee almacenamiento y manipulación efectiva de datos,
- Posee una amplia, coherente e integrada colección de herramientas para análisis de datos,
- Es un lenguaje de programación bien desarrollado, simple y efectivo,
- Permite la extensión del lenguaje básico a través del desarrollo de extensiones,
- Posee una amplia comunidad que soporta el desarrollo del lenguaje,
- Posee un repositorio de extensiones con documentación,
- El lenguaje está publicado bajo la versión GNU General Public License, por lo que es software libre.

Figura E.1 Lenguaje R

2. Phyton

Python es un lenguaje multipropósito que tiene su origen en 1991. Fue desarrollado por Guido Van Rossem como un reemplazo para el lenguaje ABC [57], con el objetivo de hacer un lenguaje de programación ágil y sencillo, con una curva de aprendizaje muy corta. Desde sus inicios, Phyton ha sido destinado para profesionales procedentes de la Estadística, pero ahora muchos desarrolladores lo han adoptado para generar aplicaciones tanto de *back-end* como de *front-end* debido a su flexibilidad y versatilidad. Al igual que R, Phyton está siendo muy utilizado por la comunidad de Ciencia de Datos debido a las siguientes razones [57]:

- Está desarrollado bajo la versión Open Source Licence, lo cual la hace de libre acceso y distribuible, aun para usos comerciales,
- Posee un grupo que la administra (Phyton Software Foundation),
- Posee un repositorio de módulos (PyPI, Python Package Index),
- Posee módulos para desarrollar actividades de minería de datos,
- Es un lenguaje de propósito general, lo que hace que sea muy utilizado en diferentes ámbitos de la programación.

Figura E.2 Lenguaje Phyton

3. Scala

Scala es un lenguaje de programación que fue diseñado por Martin Odersky y lanzado en 2003. Los rasgos distintivos de Scala incluyen una integración de funciones del lenguaje Java (lenguaje de programación orientado a objetos). Scala

debe su nombre a su capacidad de escalamiento, es decir, es un lenguaje que puede crecer al proporcionar una infraestructura que permite la introducción de nuevos constructos y tipos de datos. Una de las características más fuertes de Scala es que es un lenguaje de programación concurrente, por lo que permite, el análisis de datos en tiempo real. Scala es un lenguaje compilado. Su compilador produce bytecode para la máquina virtual Java, permitiendo así el uso (casi) transparente de las herramientas y construcciones Java desde Scala. Esta infraestructura ha sido utilizada por varias compañías en todo el mundo (por ejemplo, Siemens, Sony Pictures Imageworks, Twitter, etc.) [59].

A continuación se enlistan algunas características de Scala:

- Fuerte afinidad a los datos,
- Estado del arte Orientada a Objetos para la composición de clases,
- Programación Funcional en tiempo real,
- Alto rendimiento en la nube con Akka,
- Posee varios Entornos de Desarrollo Integrado (IDE),
- Soporte de varias empresas líderes en tecnología.

Figura E.3 Lenguaje Scala

4. SQL

SQL (Structured Query Language, Lenguaje de Consulta Estructurado) es un lenguaje para la definición y manipulación de datos mediante bases de datos relacionales. Es un lenguaje declarativo que está basado en el álgebra y cálculo relacional. Fue implementado por primera vez por Oracle en 1979.

A la fecha, un gran número de empresas e instituciones operan de manera transaccional con manejadores de bases de datos relacionales y es previsible que se continúe utilizando durante muchos años más.

Un estudio realizado por [60] acerca del uso de las herramientas en el análisis de datos, mostró que el 70% de las personas encuestadas utilizan manejadores de bases de datos relacionales en sus proyectos mediante el lenguaje SQL.

Varias fuentes enfatizan al lenguaje SQL como una herramienta de apoyo en proyectos de Ciencia de Datos, sin embargo, se refieren al uso de manejadores de bases de datos relacionales.

Figura E.4 Lenguaje SQL

E.2. Herramientas propietarias

1. Excel

Excel es un programa de análisis de datos que forma parte de la paquetería de Microsoft Office. Es ampliamente utilizado en oficinas y trabajos escolares, debido a su facilidad de uso y que no requiere de conocimientos de programación. Aunque resulte difícil de creer, Excel también se utiliza para realizar análisis de alto nivel, sobre todo cuando se necesita acceder a datos que están en formatos .csv, .dbf, entre otros.

De la misma forma que sucede con SQL, Excel también se usa frecuentemente en el análisis de datos, el 70% de los encuestados admitieron estar usándolo. [60]

Figura E.5 Software Excel

2. SAS

SAS (Statistics Analysis System, Sistema de Análisis Estadístico) es un software (también empresa) propietario especializado en el análisis de datos a gran escala. Se inició en la Universidad Estatal de Carolina del Norte como un proyecto para

analizar investigación agrícola. Conforme aumentó la demanda del software, se fundó la empresa SAS en 1976 para ayudar a toda clase de clientes (desde compañías farmacéuticas y bancos hasta entidades académicas y de gobierno).

Algunos de los servicios que ofrece SAS son [61]:

- Analítica avanzada,
- Inteligencia de negocios,
- Inteligencia del cliente,
- Administración de datos,
- Administración de riesgos,
- Inteligencia en fraude y seguridad.

Figura E.6 Software SAS

3. SPSS

SPSS (originalmente acrónimo de Statistical Package for the Social Sciences, Paquete Estadístico para Ciencias Sociales) es un software estadístico muy usado en Ciencias exactas, sociales y aplicadas, además de las empresas de investigación de mercado. La empresa propietaria de SPSS es IBM, de allí su nombre formal (IBM SPSS Statistics). Entre las características con las que cuenta SPSS se encuentran [62]:

- Análisis de datos con fines específicos,
- Pruebas de hipótesis,
- Análisis geoespacial,
- Presentación de informes,
- Gestión de usuarios, licencias, descargas y actualizaciones.

Figura E.7 Software SPSS

F.3. Uso de herramientas en Ciencia de Datos

En 2016 [60], se realizó una encuesta a 900 profesionales en el análisis de datos de una variedad de industrias, los cuales respondieron 64 preguntas vía online, las preguntas incluían información demográfica, tiempo usado en tareas relacionadas con análisis de datos y el uso/no uso de herramientas de software.

En la F.8 se muestran los porcentajes de uso de las herramientas.

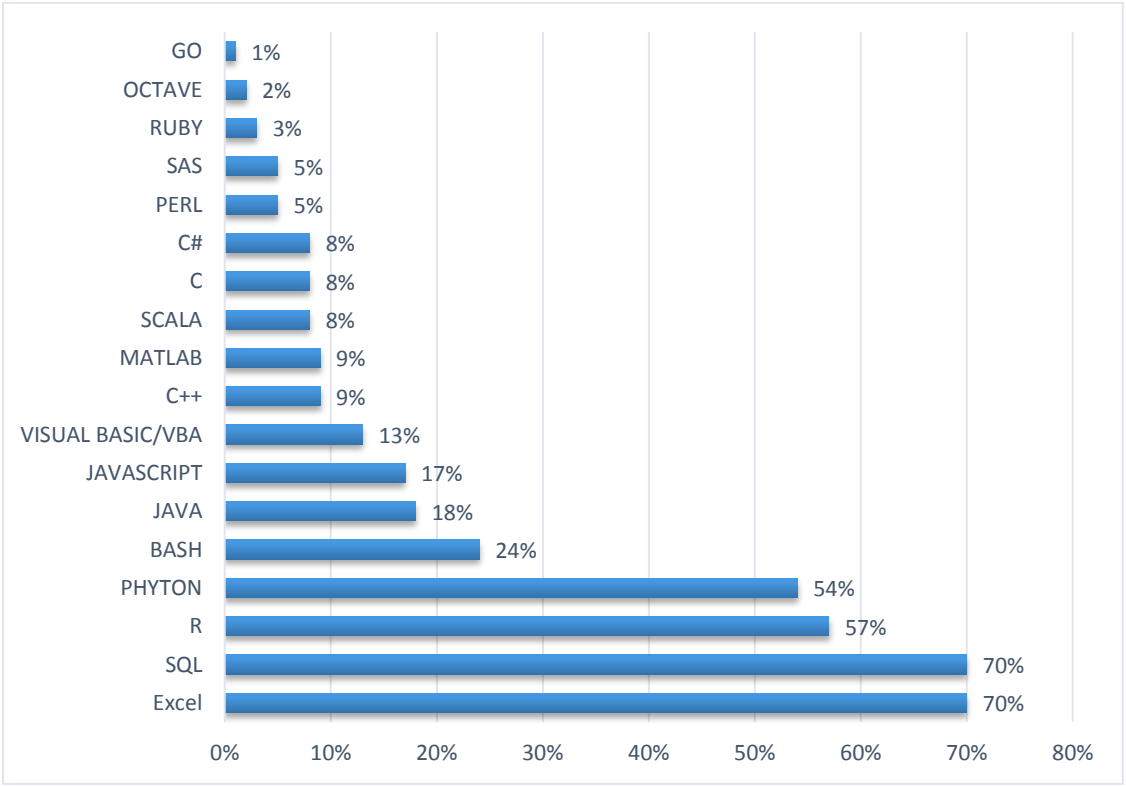

Figura E.8 Uso de herramientas de Ciencia de Datos