EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Aplicación de Ciencia de Datos para el análisis de datos de una enfermedad transmisible en México

presentada por

Lic. Eduardo Velasco Ramírez

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Directora de tesis

Dra. María Yasmín Hernández Pérez

Codirector de tesis

Dr. Joaquín Pérez Ortega

Cuernavaca, Morelos, México. Febrero de 2025.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 20/enero/2025
OFICIO/DCC/011/2025

Asunto: Aceptación de documento de tesis
CENIDET-AC-M14-OFICIO

CARLOS MANUEL ASTORGA ZAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes del Comité Tutorial de EDUARDO VELASCO RAMÍREZ, con número de control M22CE055, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Aplicación de ciencia de datos para el análisis de datos de una enfermedad transmisible en méxico"**, y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

ATENTAMENTE
Excelencia en Educación Tecnológica
"Conocimiento y Tecnología al Servicio de México"

Dra. María Yasmin Hernández Pérez
Directora de Tesis

Dr. Joaquín Pérez Ortega
Codirector de Tesis

Dr. Javier Ortiz Hernández
Revisor 1

Dr. Jonathan Villanueva Tavira
Revisor 2

Educación
Secretaría de Educación Pública

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo tecnológico
Subdirección Académica

Cuernavaca Mor, 27/enero/2025
Oficio No. SAC/036/2025
Asunto: Autorización de impresión de tesis

EDUARDO VELASCO RAMÍREZ
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Aplicación de ciencia de datos para el análisis de datos de una enfermedad transmisible en México"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

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

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

EDUCACIÓN | Tecnológico Nacional de México
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4104,
e-mail: acad_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet

100% Libre de Plástico

Agradecimientos

Con un especial agradecimiento al Consejo Nacional de Humanidades Ciencia y Tecnología (CONAHCYT) y al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por el apoyo que me brindaron para realizar mis estudios de maestría.

Se agradece al Dr. Joaquín Pérez Ortega y a la Dra. Leticia Sánchez Lima por su gran apoyo en la realización de esta tesis y sus puntuales observaciones a lo largo de las actividades necesarias que permitieron cumplir con todos los objetivos planteados.

Resumen

La Ciencia de Datos ha cobrado gran relevancia en los últimos años en diversos ámbitos a lo largo de todo el mundo. En este trabajo se contextualiza dentro del dominio epidemiológico a la Ciencia de Datos, la cual se considera aún una disciplina emergente. Es por ello que la investigación que aquí se presenta, contribuye a su crecimiento y adopción en nuevos dominios.

El problema a resolver en este trabajo de tesis radica en la generalidad existente en las metodologías de Ciencia de Datos. Las descripciones que presentan dichas metodologías no están dirigidas a ningún dominio particular y carecen de especificidad en la descripción de sus tareas, lo que limita su aplicación a problemas específicos. Extender una metodología de Ciencia de Datos existente y dirigirla a un dominio específico es de gran importancia.

En esta tesis, se atiende el problema mencionado mediante la extensión de una metodología de Ciencia de Datos aplicada al dominio epidemiológico, así como su validación con un caso práctico tomando los datos de la reciente pandemia por COVID-19 en México.

Dicha extensión forma parte de un proyecto mayor en el que otras etapas de la metodología ya han sido extendidas y validadas exitosamente mediante casos de estudio tomando datos reales de enfermedades transmisibles.

Se desarrolló una extensión a la metodología de Ciencia de Datos "*Batch Foundational Methodology for Data Science*", la cual es a su vez una extensión de la metodología de Ciencia de Datos de IBM.

Las etapas extendidas correspondieron a las referentes a la preparación de datos. La extensión se conformó por nueve tareas específicas, las cuales van desde la toma de fuentes de datos oficiales hasta la generación de un almacén de datos funcional para las etapas posteriores.

A esta extensión se le denominó metodología BFMDS Extendida. El tema abordado para dicha validación es la reciente pandemia por COVID-19, de la cual se tomó la mortalidad para un conjunto de municipios que se consideraron más influyentes en el comportamiento de toda la República Mexicana.

Los datos empleados para la experimentación fueron tomados de fuentes oficiales del gobierno de México, específicamente de la Dirección General de Epidemiología (DGE) y de la Dirección General de Información Sanitaria (DGIS). Gracias a los resultados obtenidos, fue posible obtener respuesta a las distintas preguntas de investigación plateadas. Se respondió principalmente a ¿Cómo se distribuyó la mortalidad por COVID-19 en los municipios con factores sociodemográficos similares y con la mortalidad más alta del país?

Entre los hallazgos más relevantes, se observó la existencia de una correlación entre la mortalidad por COVID-19 en municipios de México y su porcentaje de población en pobreza. Dicha correlación mostró en la etapa de evaluación que los municipios con menor índice de pobreza y mayor densidad poblacional, se vieron afectados con una tasa de mortalidad por COVID-19 más alta y viceversa. Se aplicó la técnica de agrupamiento para tener un criterio de separación de acuerdo con la densidad poblacional y el porcentaje de población en pobreza.

Desde el punto de vista computacional, se observó que siguiendo los pasos indicados en la metodología BFMDS Extendida fue posible el desarrollo del caso práctico para la extracción de conocimiento y apoyo a la toma de decisiones.

Abstract

Data Science has gained great relevance in recent years in various fields throughout the world. This paper contextualizes Data Science within the epidemiological domain, which is still considered an emerging discipline. That is why the research presented here contributes to its growth and adoption in new domains.

The problem to be solved in this thesis work lies in the existing generality of Data Science methodologies. The descriptions presented in these methodologies are not directed to any particular domain and lack specificity in the description of their tasks, which limits their application to specific problems. Extending an existing Data Science methodology and targeting it to a specific domain is of great importance.

In this thesis, the aforementioned problem is addressed by extending a Data Science methodology applied to the epidemiological domain, as well as its validation with a practical case taking data from the recent COVID-19 pandemic in Mexico.

This extension is part of a larger project in which other stages of the methodology have already been successfully extended and validated through case studies using real data on communicable diseases.

An extension to the Data Science methodology "Batch Foundational Methodology for Data Science" was developed, which in turn is an extension of the IBM Data Science methodology.

The extended stages corresponded to those related to data preparation. The extension consisted of nine specific tasks, ranging from the collection of official data sources to the generation of a functional data warehouse for the subsequent stages.

This extension was called the Extended BFMDS methodology. The topic addressed for this validation is the recent COVID-19 pandemic, from which mortality was taken for a set of municipalities that were considered to be the most influential in the behavior of the entire Mexican Republic.

The data used for the experimentation were taken from official sources of the Mexican government, specifically from the General Directorate of Epidemiology (GDE) and the General Directorate of Health Information (GDHI). Thanks to the results obtained, it was possible to obtain answers to the different research questions posed. The main answer was: How was COVID-19 mortality distributed in the municipalities with similar sociodemographic factors and highest mortality in the country?

Among the most relevant findings was the existence of a correlation between COVID-19 mortality in Mexican municipalities and the percentage of the population living in poverty. This correlation showed in the evaluation stage that municipalities with a lower poverty index and higher population density were affected by a higher COVID-19 mortality rate and vice versa. The grouping technique was applied to have a separation criterion according to population density and the percentage of population in poverty.

From the computational point of view, it was observed that following the steps indicated in the Extended BFMDS methodology, it was possible to develop the case study for knowledge extraction and decision support.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

1.1 Contexto de la investigación

Existen varias definiciones de Ciencia de Datos, todas ellas convergen en el propósito de buscar el aprovechamiento de los datos para generar conocimiento, resolver preguntas de investigación y contribuir al área de la toma de decisiones. Al igual que en esta investigación, las metodologías tienen una jerarquía en su contenido, ya que suelen estar conformadas por fases, las cuales a su vez están constituidas por una o más etapas y estas a su vez por tareas. Sumado a lo anterior, debe mencionarse que la extensión propuesta en esta tesis está orientada a ciertas etapas de una metodología de Ciencia de Datos y forma parte de un proyecto mayor, el cual se enfoca en la extensión de todas las etapas.

Las metodologías de Ciencia de Datos actuales son diversas. De acuerdo con la literatura, son las de IBM y Microsoft las más reconocidas. Dichas metodologías se conocen con el nombre de *Foundational Methodology for Data Science* (FMDS) y *Team Data Science Process* (TDSP). En el Centro Nacional de Investigación y Desarrollo Tecnológico se ha trabajado anteriormente para desarrollar una extensión de la metodología de IBM, a la cual se le conoce como *Batch* FMDS. La extensión anterior ha sido orientada al dominio epidemiológico y se han obtenido resultados prometedores, los cuales han llevado a la publicación de diversos artículos de investigación en revistas indexadas. Ahora bien, a pesar de que se vea prometedora, sólo se ha logrado extender hasta la etapa previa a la preparación de datos, por lo que ha surgido la necesidad de extenderla en esta etapa.

En torno al objeto de aplicación, se conoce que el COVID-19 es una enfermedad que fue identificada por primera vez en Wuhan, China en diciembre de 2019. Esta enfermedad se difundió a gran velocidad por todo el mundo, lo que provocó que la Organización Mundial de la Salud (OMS) declarara al COVID-19 como una pandemia para todos los países del mundo [1]. El 5 de mayo de 2023 se hizo oficial el final de dicha pandemia, por lo que este trabajo pretende contemplar toda su duración.

Los conjuntos y bases de datos empleados para esta investigación provienen de dependencias gubernamentales como lo son: la Dirección General de Información Sanitaria (DGIS), el Instituto Nacional de Estadística, Geografía e Informática (INEGI), la Dirección General de Epidemiología (DGE), el Centro Mexicano para la Clasificación de Enfermedades (CEMECE) y el Sistema Nacional de Información Municipal (SNIM). En caso de que existieran cambios administrativos por el término de la pandemia, algunas de estas fuentes podrían cambiar.

Este trabajo consistió en proponer tareas en las etapas de la fase de preparación de datos de la metodología BFMDS y su validación mediante un caso de estudio del dominio epidemiológico. En particular, se tomó la reciente problemática debida al COVID-19 para estudiarla y proporcionar información. Con los resultados, se dio respuesta a las siguientes preguntas de investigación: ¿Cuál fue el comportamiento de los municipios con mayor mortalidad en México en comparación con todo el país? y ¿Existió una relación entre la mortalidad por COVID-19 y la localización geográfica? Dichas preguntas se resolvieron haciendo uso de la metodología extendida elaborada a lo largo de esta tesis, contemplando tanto las etapas extendidas como las genéricas.

1.2 Descripción del problema

El principal problema que aborda este trabajo es la generalidad existente en las metodologías de Ciencias de Datos actuales, lo que limita su uso. De las metodologías de Ciencia de Datos más utilizadas, no existe alguna que funcione como guía precisa que pueda adoptarse para un dominio en específico, lo que ocasiona un panorama difuso sobre qué hacer para todo el que quiera adoptar la metodología. Esta generalidad puede deberse a que la Ciencia de Datos es una disciplina relativamente emergente que aún no ha adoptado una única manera de llevarse a cabo, por lo que es de gran importancia contribuir a su avance.

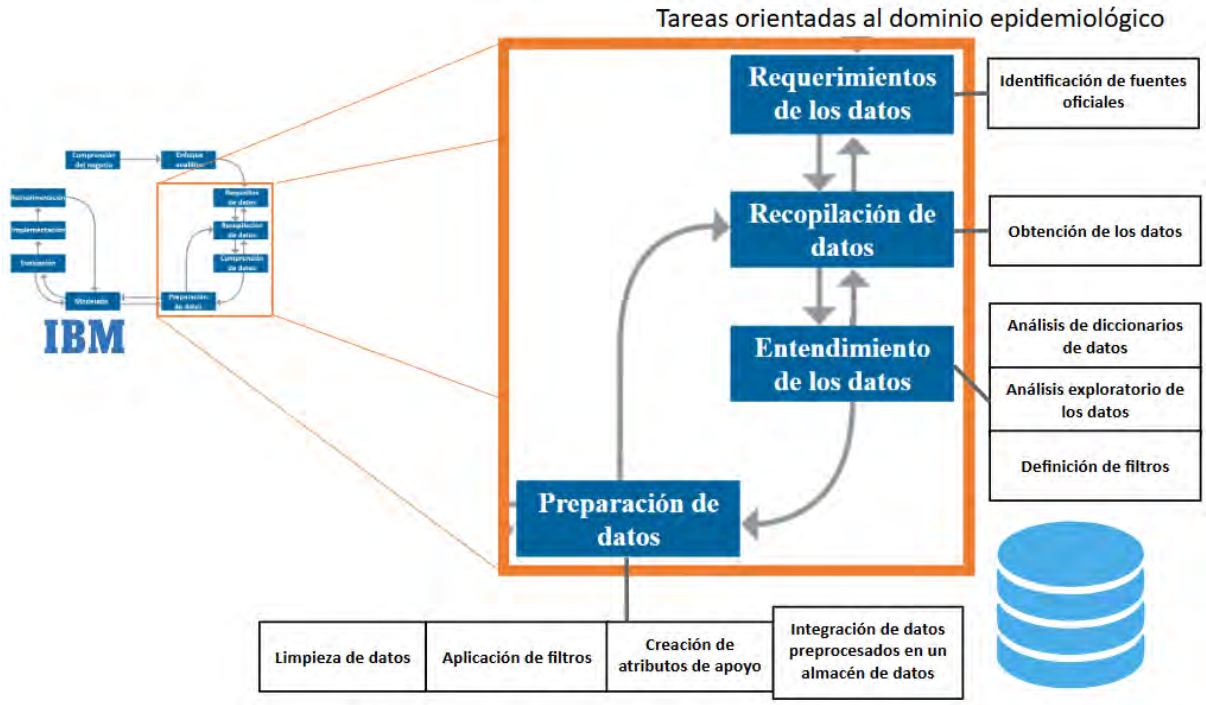

Figura 1.1: Esquema general de la problemática de este trabajo.

Por otro lado, el crecimiento de la cantidad y la diversidad de la información ha dado lugar a conjuntos de datos más grandes de lo que se puede gestionar con las herramientas de gestión convencionales actuales, es por ello que es vigente la necesidad de detallar las metodologías de Ciencia de Datos para la mayor cantidad de dominios.

La problemática debida a la pandemia por SARS-CoV-2 expuso en todo el mundo debilidades de las que se adolece para controlar un fenómeno de esta magnitud [2]. De lo anterior, emerge la necesidad latente para contribuir a la toma de decisiones basadas en datos que puedan sustentar con mayor fuerza cada decisión que pretenda tomarse.

La pandemia se dio por terminada en México el 9 de mayo del 2023 y los trabajos que documentan toda su duración y proceso son pocos o inexistentes [3, 4], lo que de igual forma otorga interés a una investigación como la presente, que la integre en su totalidad.

En la Figura 1.1 se ve hace un acercamiento a las etapas extendidas de la metodología de IBM, en las cuales se indica el título de las tareas que las comprenden.

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar una extensión de la metodología BFMDS en las etapas de preparación de datos y validarla mediante un caso práctico para analizar los datos obtenidos de la epidemia por COVID-19 en México.

1.3.2 Objetivos específicos

a) Extender la metodología BFMDS para el dominio epidemiológico en la fase de preparación de datos.
b) Validar la extensión de la metodología mediante el desarrollo de un caso práctico basado en datos oficiales de la pandemia por COVID-19 en México durante la pandemia.
c) Mostrar los hallazgos obtenidos al resolver el caso práctico con base en la extensión de la metodología BFMDS Extendida.

1.4 Justificación

El presente estudio forma parte de un proyecto mayor, en el que se han extendido y validado etapas previas a la fase de preparación de datos con la metodología FMDS. Por esta razón, a pesar de que se tengan resultados prometedores en la previa extensión de esta metodología, aún restan etapas que siguen siendo generales y que carecen de especificidad con las actividades a realizar. Contribuir con la extensión de la metodología *Batch* FMDS (extensión de FMDS) será de gran apoyo para todo aquel que quiera aplicarla dentro del dominio epidemiológico. Estas contribuciones son importantes para poder extender las etapas posteriores a la preparación de datos con futuros estudios.

1.5 Alcances y limitaciones de la investigación

1.5.1 Alcances

Como alcance central, se contempla aportar una guía que pueda replicarse para aplicar Ciencia de Datos dentro del dominio epidemiológico.

Presentar al menos una publicación enfocada a la problemática de la COVID-19 aplicando Ciencia de Datos.

1.5.2 Limitaciones

La extensión de la metodología de Ciencia de Datos únicamente cubre etapas relacionadas con la preparación de datos, por lo que no se pretende ser específico con las demás etapas. La validación de la metodología es llevada a cabo con el caso práctico solo con datos relacionados con COVID-19.

Otra limitación que se presenta en este trabajo corresponde a la disponibilidad de los datos y su actualización en las fuentes de acceso público, debido a que existen conjuntos y bases de datos dentro de los sitios web de la Secretaría de Salud que no tienen una fecha de actualización definida. De manera similar, la veracidad de los datos y la forma en la que fueron registrados en las dependencias oficiales excede los alcances de este trabajo.

1.6 Organización del documento

El presente documento de tesis está organizado de la siguiente manera: En el Capítulo 2 se presentan los antecedentes y el contexto en el que se ubica el problema por resolver, desde su planteamiento y contexto, hasta la propuesta de solución. Se muestra un contexto de algunas de las metodologías de Ciencia de Datos más relacionadas con esta investigación en el Capítulo 3. En el Capítulo 4, se desarrolla la extensión propuesta como resultado de la investigación. Posteriormente, en el Capítulo 5, se aplica la metodología propuesta a un caso práctico con datos reales de la pandemia por COVID-19 en México. Finalmente, en el Capítulo 6, se presentan las conclusiones derivadas del estudio realizado, así como las principales aportaciones y propuestas de trabajos para el futuro.

Capítulo 2

Antecedentes

En este capítulo se presentan las investigaciones relacionadas con el tema de esta tesis, que se han llevado a cabo dentro del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) y fuera de él, ya sea dentro y fuera de México. Entre estas investigaciones existe una relación, ya sea en los aspectos del uso de las metodologías de Ciencia de Datos, la pandemia por COVID-19, dentro y fuera de México o bien el uso de algoritmos de agrupamiento para apoyar a la aplicación de la Ciencia de Datos en sí.

2.1. Trabajo relacionado

Una de las primeras tesis de posgrado desarrolladas al respecto en el CENIDET fue "Desarrollo de una aplicación de Ciencia de Datos" [5], en la cual se tratan conceptos importantes de la Ciencia de Datos. En esta tesis, se investiga la viabilidad de asimilar conceptos de ciencia de datos y crear una infraestructura de conocimiento para apoyar el desarrollo de aplicaciones específicas en este campo. Utilizando la metodología "*Foundational Methodology for Data Science*" de IBM y el lenguaje de programación estadística R, se llevó a cabo un caso práctico centrado en la proyección de las tasas de mortalidad por *diabetes mellitus* (tipos E11-E14) en regiones específicas de México durante el período 2016-2020.

El estudio se enfocó en tres regiones, identificadas como C24, C08 y C51, las cuales comprenden 25 municipios con las tasas de mortalidad más altas por esta enfermedad. Se

utilizó una combinación de datos poblacionales de fuentes oficiales como SINAIS, INEGI, CONAPO y CEMECE para llevar a cabo el análisis.

Los hallazgos de la investigación indican que, de mantenerse las tendencias actuales, se espera una reducción en las tasas de mortalidad por diabetes en las regiones C24 y C08 para el año 2020, con descensos estimados entre 6.2 % y 11.1 % en C24 y entre 18.8 % y 21.9 % en C08, en comparación con los picos de mortalidad registrados en 2003 y 2002, respectivamente. Contrariamente, en la región C51, la tasa de mortalidad podría variar entre una disminución del 1.3 % y un aumento del 16.6 % respecto a 2015.

Desde un enfoque computacional, la tesis destaca el impacto de la Ciencia de Datos en el sector académico, empresarial y gubernamental, y subraya la importancia de seleccionar una metodología adecuada y construir modelos de proyección utilizando herramientas estadísticas avanzadas como R. Tecnológicamente, el estudio proporciona una base para futuras investigaciones en Ciencia de Datos en el CENIDET y el desarrollo de aplicaciones relacionadas con problemas de salud pública de gran relevancia en México.

Socialmente, la tesis resalta la gravedad de la *diabetes mellitus* en México, una enfermedad con altas tasas de mortalidad a pesar de los esfuerzos gubernamentales significativos en prevención y tratamiento. En conclusión, el estudio no solo demuestra la aplicabilidad de la Ciencia de Datos en la evaluación de problemas de salud pública, sino que también sugiere cómo las proyecciones pueden informar y mejorar las estrategias de salud pública.

Otra tesis relacionada con la presente es "Aplicación de Ciencia de Datos para el análisis de datos de mortalidad por COVID-19 de México" [6]. La relación que tiene con este trabajo se debe a que se extiende la metodología FMDS de IBM en sus dos primeras etapas: Entendimiento del negocio y Enfoque analítico.

En este estudio, se explora la efectividad de la metodología BATCH FMDS, una técnica de Ciencia de Datos específicamente orientada al ámbito epidemiológico, propuesta por IBM. Esta metodología se aplicó en un estudio práctico que analizó los datos de mortalidad por COVID-19 a nivel municipal en México durante el año 2020. El objetivo principal fue identificar los factores sociodemográficos comunes en los municipios con tasas de mortalidad similares por COVID-19.

La investigación se basó en datos poblacionales de instituciones oficiales mexicanas como DGIS, INEGI, CONAPO, y CEMECE. Uno de los principales hallazgos fue que la densidad poblacional y el porcentaje de personas en situación de pobreza mostraron una alta correlación con las tasas de mortalidad por COVID-19. Específicamente, se encontró que los

municipios con mayor densidad poblacional y un menor porcentaje de personas en situación de pobreza, como algunos en Nuevo León, Guadalajara y ciertas alcaldías de la Ciudad de México, tenían las tasas de mortalidad más altas. Por otro lado, los municipios con menor densidad poblacional y un alto porcentaje de pobreza, principalmente en Chiapas, mostraron las menores tasas de mortalidad.

Desde el punto de vista computacional, se confirmó la viabilidad de implementar la metodología BFMDS para desarrollar aplicaciones de Ciencia de Datos en el contexto epidemiológico. Los resultados fueron validados y apreciados por expertos en epidemiología y Ciencia de Datos, y documentados en el artículo *Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico*.

Esta investigación subraya la importancia de herramientas de Ciencia de Datos en la gestión de la salud pública, especialmente en el contexto de la pandemia de COVID-19. Los beneficios tecnológicos incluyeron el desarrollo de aplicaciones específicas para abordar problemas de salud pública significativos y la demostración de la utilidad de la metodología BFMDS en estudios epidemiológicos. Se sugiere que es posible la aplicación de esta metodología a otros estudios de mortalidad por enfermedades como la diabetes y el cáncer, y comparar los resultados utilizando diferentes metodologías de Ciencia de Datos.

Algunos de los artículos que se han publicado por profesores y estudiantes del CENIDET, relacionados con esta área, son los siguientes:

A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases [7]. A pesar de que no se utilice el término de Ciencia de Datos en este artículo, resulta de gran relevancia mencionarlo ya que se habla con detalle de la fase de preparación de datos, etapa que comparte la Minería de Datos. La investigación propone dividir la preparación de datos en dos etapas: Preparación General de los Datos (GDP, por sus siglas en inglés) y Preparación Específica de Datos (SDP, por sus siglas en inglés). En la primera etapa se incluye lo que es la limpieza de datos y la selección de datos más relevantes, mientras que en la segunda etapa se le da formato, construcción e integración a los datos. Al final se validó la propuesta con un caso de estudio tomando en cuenta los datos de mortalidad de México del 2000.

En la publicación *Correlation between mobility in mass transport and mortality due to COVID-19: A comparison of Mexico City, New York, and Madrid from a data science perspective* [8], se analizó la correlación existente entre la movilidad y la mortalidad por COVID-19 en tres ciudades de distintas partes del mundo con densidades poblacionales similares. Se encontró, durante el periodo estudiado, una correlación entre la disminución de

uso del transporte público y las muertes por COVID-19, dando resultados en periodos distintos de tiempo para cada ciudad. En esta investigación fue empleada la metodología FMDS de IBM y fueron mostradas las medidas no farmacéuticas más relevantes implementadas en cada país.

El artículo *Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico* [9]. En esta investigación se extendió la metodología FMDS de IBM en sus dos primeras etapas, dando lugar a la metodología Batch FMDS. Relacionado con el presente trabajo de tesis, fue validada la extensión mediante un caso práctico tomando el análisis de la mortalidad por COVID-19 en municipios con una población mayor a 100 mil habitantes.

Fuera del CENIDET existen estudios como *Data preparation for data mining* [10], el cual fue de los primeros artículos que expuso la etapa de preparación de datos. Este estudio subraya la importancia de la preparación de los datos, proponiendo futuras direcciones como el desarrollo de entornos de minería de datos interactivos e integrados, teorías de preparación de datos, y algoritmos eficientes para múltiples fuentes de datos. Se menciona que la preparación de datos es crucial en el análisis de datos y la minería de datos, enfocándose en transformar datos de baja calidad en formatos limpios para múltiples usos. Se afirma que esta etapa es esencial debido a la impureza de los datos reales, la necesidad de sistemas de minería de alto rendimiento y la obtención de patrones concentrados a partir de datos de calidad.

En *An analysis to identify the important variables for the spread of COVID-19 using numerical techniques and data science* [11], se utilizan técnicas numéricas y Ciencia de Datos para identificar variables socioeconómicas y meteorológicas cruciales en la propagación de COVID-19, encontrando que la población total, ingresos del hogar, ocupación y transporte son especialmente relevantes. Se demostró que combinaciones de hasta cinco variables pueden capturar con alta precisión la dinámica de propagación del virus, con coeficientes de correlación de hasta 0.985, sugiriendo potencial para el desarrollo de modelos predictivos.

En *The data science of COVID-19 spread: Some troubling current and future trends* [12], se resaltan las dificultades metodológicas en el estudio de la COVID-19 usando series temporales de casos y muertes, advirtiendo sobre la cautela necesaria para consumidores y productores de conocimiento sobre la enfermedad. Aunque se han logrado avances en diagnóstico, tratamiento clínico y seguimiento filogenético, la investigación es crítica con el análisis macro-observacional y sugiere una consideración crítica de las limitaciones y sesgos potenciales, especialmente en el contexto estadounidense, señalando mejores prácticas en otros países como Corea del Sur.

El estudio *Simple correlation between weather and covid-19 pandemic using data mining algorithms* [13], destaca la correlación significativa entre las condiciones meteorológicas y el número de casos confirmados de COVID-19 en Surabaya, Indonesia. Variables como la duración promedio de la luz solar, la temperatura media y la humedad promedio demostraron influir en el número de casos, con la luz solar mostrando la correlación más fuerte. La precisión del modelo alcanzó el 96.77 %.

En el estudio titulado *Spatiotemporal patterns of the COVID-19 epidemic in Mexico at the municipality level* [14], se analizan los patrones espaciotemporales de la epidemia de COVID-19 en México a nivel municipal. Se identifican grupos que evolucionan en espacio y tiempo como epidemias paralelas. Utilizan un modelo de distancia gravitacional para predecir el contagio entre municipios. Este enfoque supera a otros métodos geográficos tradicionales. Los resultados proporcionan una base para mapear puntos críticos de COVID-19 y diseñar estrategias efectivas de control y prevención.

En *Socio-demographic inequalities and excess non-COVID-19 mortality during the COVID-19 pandemic: a data-driven analysis of 1,069,174 death certificates in Mexico* [15], se realizó un análisis detallado de 1,069,174 certificados de defunción en México para investigar la mortalidad excesiva durante la pandemia de COVID-19, distinguiendo entre muertes por COVID-19 y causas no relacionadas. Mediante comparaciones con datos promedio de 2015 a 2019, se estimó la mortalidad excesiva, analizando las causas principales de muerte, ubicaciones (dentro y fuera de hospitales) y distribución geográfica. Se emplearon modelos de regresión para evaluar factores asociados a la mortalidad no COVID-19 a nivel individual y municipal, destacando la influencia de desigualdades socio-demográficas. El estudio concluyó que hubo un aumento significativo en la mortalidad, exacerbado por factores como la baja cobertura de seguridad social y condiciones socioeconómicas precarias, y llamó a mejorar el acceso y la cobertura de la atención sanitaria.

En la Tabla 2.1, ubicada al final del capítulo, se presenta una clasificación de la relevancia que tienen las investigaciones anteriormente mencionadas para este trabajo de tesis.

2.2 La Ciencia de Datos

La ciencia de datos, nacida en 1967 como evolución del análisis de datos, ha capturado el interés creciente de la comunidad científica por su potencial para mejorar la toma de decisiones y cambiar nuestra percepción del mundo [16]. Son diversas las definiciones de lo

que es la Ciencia de Datos. A continuación, se incluye información que es de apoyo para este trabajo y que se consideran puntuales.

De acuerdo con *Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management* [17], la ciencia de datos representa una convergencia crítica de habilidades analíticas y conocimientos específicos del dominio, enfocada en la gestión y análisis de datos de gran volumen y diversidad. Son utilizados métodos cuantitativos y cualitativos para abordar desafíos complejos y prever tendencias, siendo fundamental en diversas áreas. A medida que la cantidad de datos crece exponencialmente, la ciencia de datos se vuelve esencial para descubrir *insights*[1] valiosos y desarrollar soluciones innovadoras, subrayando la necesidad de investigación y de formación especializada en este ámbito interdisciplinario.

La educación en ciencia de datos es un campo académico transdisciplinario en crecimiento que combina conocimientos de informática, matemáticas y estadísticas. Aunque la integración tecnológica en la enseñanza mejora las habilidades y el conocimiento sobre conjuntos de datos reales, los rápidos cambios en prácticas organizacionales y tendencias han creado brechas de conocimiento, potencialmente llevando a interpretaciones erróneas de datos y decisiones equivocadas [18].

Además, de acuerdo con *Data science and its relationship to big data and data-driven decision making* [19], la ciencia de datos ha emergido como una disciplina clave para los proyectos que buscan obtener ventajas competitivas mediante el análisis de grandes volúmenes y variedades de datos, superando las capacidades de los métodos de análisis manual y de las bases de datos convencionales. Este campo aprovecha el poder de computación avanzada, la omnipresencia de las redes y algoritmos innovadores para realizar análisis más amplios y profundos. Sin embargo, existe confusión sobre su definición exacta, lo que puede llevar a una percepción errónea de su valor y fundamentos.

2.3 Pandemia por COVID-19 en México

El COVID-19 es una enfermedad respiratoria causada por el contagio del virus SARS-CoV-2, el cual trastornó gravemente la infraestructura sanitaria a nivel mundial [20]. Los primeros casos por COVID-19 se dieron en la ciudad de Wuhan, China en 2019, dando lugar

[1] Término ampliamente utilizado en la Ciencia de Datos que se refiere a un descubrimiento revelador obtenido a partir del análisis profundo de datos.

a que posteriormente la Organización Mundial de la Salud (OMS) declarase de manera oficial el inicio de la pandemia por este virus el 30 de enero de 2020.

En México, la primera muerte por COVID-19 se registró el 18 de marzo de 2020, dando lugar a que el gobierno mexicano declarara el inicio de la pandemia cinco días después, el 23 de marzo [21].

Durante el primer año de pandemia, México presentó altos índices de mortalidad por COVID-19, lo que alarmó a la población y al gobierno del país, ocasionando el surgimiento de estudios que criticaron el manejo de la pandemia [22]. Fue hasta el 24 de diciembre del 2020, una vez probada la vacuna Pfizer, que se inició con la vacunación masiva, la cual comenzó en la Ciudad de México.

Finalmente, el cinco de mayo de 2023, la OMS declaró oficialmente el término de esta pandemia y el gobierno de México hizo lo propio cuatro días después, el nueve de mayo del mismo año.

La aplicación de Ciencia de Datos en el caso de la pandemia por COVID-19 ha permitido realizar numerosos estudios de acceso público que pueden proporcionar información de manera muy clara, sin la necesidad de tener conocimientos acerca de Ciencia de Datos [23].

Tabla 2.1: Comparación entre algunos artículos relacionados con Ciencia de Datos y su aplicación.

Artículo	Pandemia por COVID-19	En México	Fuera de México	Uso explícito de una metodología de Ciencia de Datos
A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases [7]		✓		✓
Correlation between mobility in mass transport and mortality due to COVID-19: A comparison of Mexico City, New York, and Madrid from a data science perspective [8]	✓	✓		✓
Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico [9]	✓	✓		✓
Data Preparation for Data Minig [10]			✓	✓
An analysis to identify the important variables for the spread of COVID-19 using numerical techniques and data science [11]	✓		✓	✓
The Data Science of COVID-19 Spread: Some Troubling Current and Future Trends [12]	✓		✓	
Simple Correlation Between Weather and COVID-19 Pandemic Using Data Mining Algorithms [13]	✓		✓	
Spatiotemporal patterns of the COVID-19 epidemic in Mexico at the municipality level [14]	✓	✓		
Socio-demographic inequalities and excess non-COVID-19 mortality during the COVID-19 pandemic: a data-driven analysis of 1,069,174 death certificates in Mexico [15]	✓	✓		

Capítulo 3

Metodologías de Ciencia de Datos

En este capítulo se explicarán algunas de las metodologías más populares de Ciencia de Datos. Estas metodologías son: Metodología Fundamental de Ciencia de Datos (FMDS, por sus siglas en inglés) de IBM [24], y el Proceso de Ciencia de Datos por Equipos (TDSP, por sus siglas en inglés) de Microsoft [25]. Además, de la primera surge la Metodología Fundamental de Ciencia de Datos por Lotes (*Batch* FMDS) [9], una extensión de la FMDS realizada en un grupo de trabajo del CENIDET.

Es necesario entender que las metodologías tienen la función de guiar el proceso de un dominio general, ya sea de forma poco o muy detallada a través de etapas constituidas por una o más tareas. Para un proyecto de Ciencia de Datos cobran gran relevancia debido a que se requieren varios pasos para obtener uno o varios resultados que sean de interés para quienes están involucrados en un grupo de trabajo.

A continuación, se expondrá la descripción de cada una de las partes de las metodologías mencionadas, lo que servirá como antecedente para mencionar las características propias de la extensión propuesta. Es necesario enfatizar que la extensión que se propone en este trabajo toma como base a la metodología *Batch* FMDS, la cual a su vez proviene de la metodología FMDS de IBM.

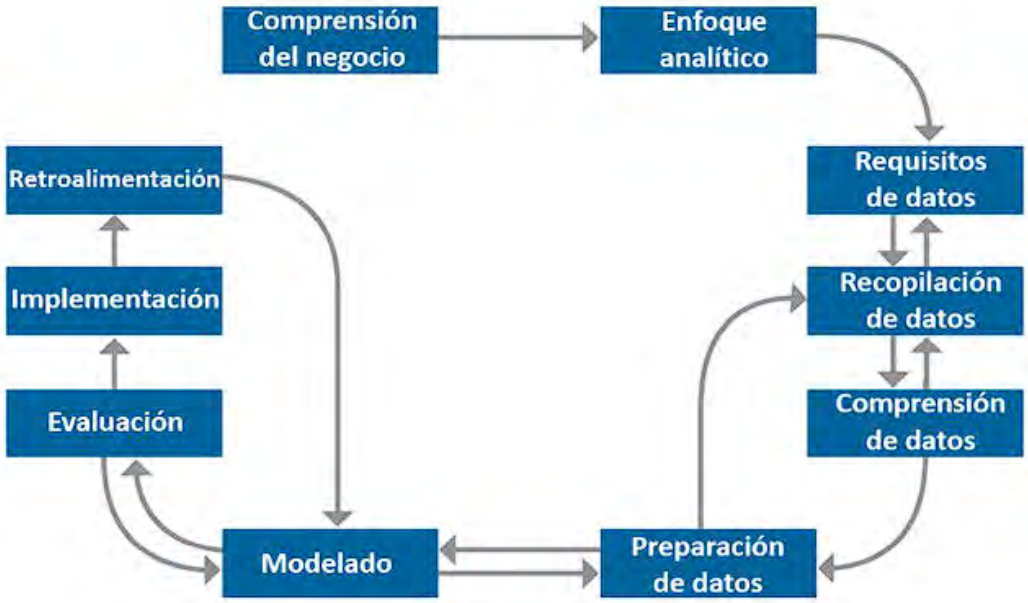

Figura 3.1: Diagrama general de la metodología FMDS de IBM.

3.1. Metodología Fundamental para la Ciencia de Datos (FMDS)

La Metodología Fundamental para la Ciencia de Datos (FMDS, por sus siglas en inglés) fue desarrollada por la conocida corporación IBM. Como puede verse en la Figura 3.1, está constituida por 10 etapas, las cuales se detallarán a continuación. Es destacable que la descripción está basada en el documento de IBM escrito por John B. Rollins [24].

3.1.1. Comprensión del negocio

En esta etapa se define el problema, los objetivos y los requisitos que se hayan acordado por todos los miembros del proyecto. Al elaborar una pregunta se podrán conocer los objetivos que se quieren alcanzar con el proyecto, lo cual ayuda a cada miembro del equipo a aclarar lo que se quiere hacer. La pregunta puede modificarse hasta cierto punto a lo largo del proyecto o incluso podrán surgir más.

Cada proyecto inicia con un entendimiento claro del negocio. En esta fase inicial, quienes impulsan la iniciativa y requieren de análisis son elementos clave, porque ellos definen los problemas a resolver, los objetivos del proyecto y los requerimientos de la solución desde un enfoque empresarial. Esta etapa preliminar es fundamental para la resolución exitosa del problema de negocio. Para asegurar el éxito del proyecto, es crucial que estos impulsores permanezcan involucrados a lo largo del desarrollo del proyecto, aportando su conocimiento del sector, evaluando avances y asegurando que el proyecto se mantenga en ruta para lograr la solución esperada.

Lo anterior es todo lo que la metodología menciona; como puede verse, no hay tareas o pasos que orienten al usuario para trasladar dicha etapa a un problema en específico. Se verá en las etapas restantes que este problema es el mismo.

3.1.2. Enfoque analítico

Con el enfoque analítico se pretende identificar las técnicas estadísticas y de modelado más adecuadas para conseguir los objetivos. Dependiendo del problema que se espera resolver, será el método seleccionado.

Una vez que el problema del negocio esté claramente definido, el científico de datos podrá determinar el método analítico adecuado para abordarlo. Esta fase consiste en reformular el problema utilizando el lenguaje de técnicas estadísticas y/o de *machine learning*, lo que permite a la organización seleccionar las que considere óptimas para alcanzar los resultados deseados. Por ejemplo, si el propósito es prever una respuesta de tipo "sí" o "no", entonces el método analítico seleccionado podría ser el diseño, evaluación e implementación de un modelo de clasificación.

3.1.3. Requisitos de datos

La etapa anterior sugiere los requisitos que deberán poseer los datos a utilizar. Se deben identificar los formatos, contenidos y representaciones que estén acompañados por el dominio.

El método analítico seleccionado establece las necesidades específicas de los datos. En detalle, las técnicas analíticas que se aplicarán dictan requerimientos particulares sobre el contenido de los datos, sus formatos y sus formas de representación, los cuales están guiados por el entendimiento del campo específico.

3.1.4. Recopilación de datos

Durante la fase de recopilación de datos, los científicos de datos buscan y compilan los recursos de datos existentes (estructurados, no estructurados y semiestructurados) que sean pertinentes al ámbito del problema. Frecuentemente, deben decidir si vale la pena realizar inversiones adicionales para acceder a información que no está fácilmente disponible. A menudo es prudente posponer las decisiones sobre posibles inversiones hasta tener una comprensión más profunda de los datos y del modelo. Si emergen deficiencias en la

recopilación de datos, será necesario que el científico revise los requisitos y busque mayor información o nuevos datos.

Aunque técnicas como el muestreo y la segmentación de datos siguen siendo cruciales, las modernas plataformas de alto rendimiento y las capacidades analíticas integradas en las bases de datos permiten a los científicos trabajar con conjuntos de datos mucho más amplios, que abarquen la mayoría o todos los datos disponibles. Al integrar mayores volúmenes de datos, los modelos predictivos pueden capturar mejor los eventos que son poco frecuentes, como la aparición de una enfermedad o una falla en sistemas.

3.1.5. Entendimiento de los datos

Al disponer ya de los datos, se sugiere hacer un análisis descriptivo de los mismos, así como aplicar técnicas de visualización para analizar su comportamiento y evaluar su calidad. En caso de que sea necesario un mayor número de datos, se puede regresar a la etapa anterior.

Tras la recopilación inicial, es común que los científicos de datos empleen estadísticas descriptivas y métodos de visualización para analizar su contenido, verificar su calidad y obtener las primeras percepciones útiles. Si se identifican lagunas en la información, puede ser esencial recoger más datos con el fin de completar estos vacíos.

3.1.6. Preparación de datos

Esta etapa es una de las más conocidas e incluye todas las actividades necesarias para desarrollar el conjunto de datos que se utilizarán en la etapa posterior de modelado. Las tareas de preparación de datos comprenden su limpieza (abordar valores inválidos o faltantes, eliminar duplicados y formatear adecuadamente), la integración de diversas fuentes (archivos, tablas y plataformas), y su transformación en variables más útiles.

Los científicos de datos realizan un proceso denominado ingeniería de características para generar variables explicativas adicionales, también conocidas como indicadores o características, utilizando una combinación de conocimiento del dominio y variables estructuradas existentes. En presencia de datos en texto, como registros de atención al cliente o notas médicas en formatos no estructurados o semiestructurados, se emplea la analítica de texto para derivar nuevas variables estructuradas y enriquecer el conjunto de indicadores, mejorando así la precisión del modelo.

La preparación de datos generalmente representa la fase más extensa en los proyectos de Ciencia de Datos. En muchos campos, ciertos pasos de la preparación de datos son comunes para distintos problemas. La automatización temprana de algunos de estos pasos puede agilizar el proceso, reduciendo el tiempo necesario para preparaciones personalizadas. Con sistemas de alto rendimiento y paralelismo masivo, así como funcionalidades analíticas integradas en donde se almacenan los datos, los científicos pueden preparar grandes conjuntos de datos de manera más eficiente y rápida.

3.1.7. Modelado

La fase de modelado comienza utilizando la primera versión del conjunto de datos ya preparado. Se centra en el desarrollo de modelos predictivos o descriptivos, según el enfoque analítico establecido anteriormente. Para los modelos predictivos, los científicos de datos emplean un conjunto de entrenamiento (datos históricos cuyos resultados ya se conocen) para construir el modelo. Este proceso es habitualmente iterativo, puesto que a medida que se obtienen nuevos *insights* intermedios, podrán surgir necesidades de ajustes tanto en la preparación de los datos como en la especificación del modelo. Para una técnica específica, los científicos de datos pueden experimentar con diversos algoritmos y ajustar sus parámetros para identificar el modelo más adecuado a las variables disponibles.

Esta etapa suele ser iterativa, debido a que los modelos mencionados se deberán ajustar para obtener mejores resultados.

3.1.8. Evolución

Durante el desarrollo del modelo y antes de proceder a su implementación, el científico de datos deberá llevar a cabo una evaluación detallada del mismo, con el fin de asegurar su calidad y efectividad para abordar completamente el problema empresarial. Esta evaluación incluye el cálculo de diversas medidas diagnósticas, además del análisis de resultados visuales como tablas y gráficos, que facilitan al científico de datos interpretar la eficacia del modelo en la solución del problema. En el caso de modelos predictivos, se emplea un conjunto de pruebas, independiente del conjunto de entrenamiento, pero con la misma distribución de probabilidad y resultados conocidos, para verificar y ajustar el modelo según sea necesario. A menudo, el modelo final también se prueba en un conjunto de validación para realizar una última evaluación.

Además, los científicos de datos podrán realizar pruebas de significancia estadística al modelo como una verificación adicional de su calidad. Esta prueba es crucial especialmente en contextos en los cuales las decisiones basadas en dicho modelo tienen consecuencias significativas, como en protocolos médicos costosos.

3.1.9. Puesta en operación o implementación

Esta etapa también es conocida como Implementación; sin embargo, se considera que el nombre adecuado para esta investigación debe ser Puesta en operación. Cuando se ha desarrollado un modelo satisfactorio y ha recibido la aprobación de los promotores del negocio, éste se pone en marcha en un entorno de producción o en un entorno de prueba equivalente. Generalmente, la implementación se realiza de manera limitada inicialmente, hasta que el rendimiento del modelo ha sido completamente evaluado. La implementación puede variar en complejidad, desde algo tan simple como la generación de un informe con recomendaciones hasta algo más complejo como la integración del modelo en procesos operativos detallados y flujos de trabajo gestionados por aplicaciones personalizadas. Implementar un modelo en un proceso operativo de negocio suele requerir la colaboración de varios grupos dentro de la empresa, así como habilidades y tecnologías adicionales.

3.1.10 Retroalimentación

Una vez implementado el modelo, la organización comienza a recoger resultados que proporcionen retroalimentación sobre el rendimiento del modelo y su efecto en el ambiente operativo. Por ejemplo, esta retroalimentación puede manifestarse en los porcentajes de respuesta a una campaña de *marketing* dirigida a un segmento de clientes que el modelo ha identificado como de alto potencial. Los científicos examinan estos datos de retroalimentación para realizar ajustes en el modelo, mejorando así su precisión y utilidad. Es posible automatizar algunos o todos los procesos de evaluación del modelo, recopilación de retroalimentación, ajuste y reimplementación del modelo, lo que permite acelerar su actualización para optimizar los resultados obtenidos.

En otras palabras, los integrantes del equipo deberán reunirse para analizar y evaluar el rendimiento del modelo, así como discutir posibles áreas de mejora.

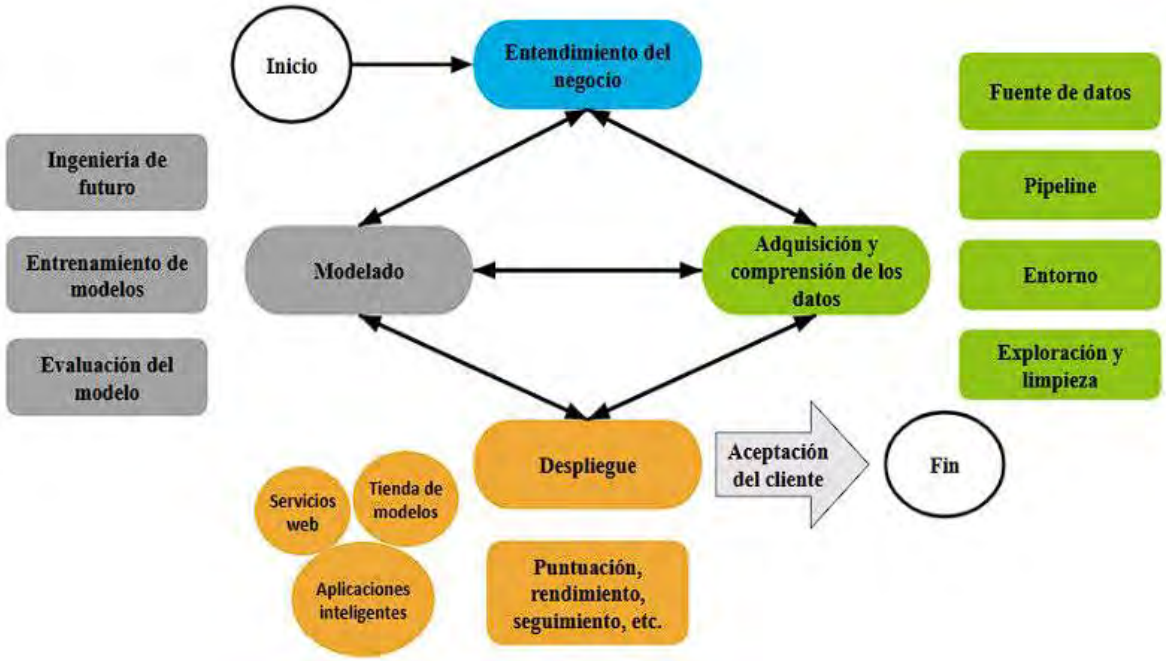

Figura 3.2: Diagrama general de la metodología TDSP de Microsoft

3.2. Proceso de Ciencia de Datos en Equipo (TDSP)

El Proceso de Ciencia de Datos en Equipo (TDSP, por sus siglas en inglés), es una metodología de Microsoft que está orientada principalmente a mejorar la colaboración dentro de un grupo de trabajo, aprovechando los roles de cada miembro del equipo. Los componentes principales que integran esta metodología son: la definición del ciclo de vida que tendrá el proyecto de ciencia de datos, la estandarización de la estructura del proyecto, la infraestructura y recursos recomendados para el proyecto, y las herramientas para su ejecución.

Ahora bien, dentro del ciclo de vida de un proyecto de Ciencia de Datos en la metodología TDSP, se identifican las siguientes etapas: Entendimiento del negocio, Adquisición y comprensión de datos, Modelado y Despliegue. La representación visual de este ciclo de vida se muestra en la Figura 3.2.

A continuación, se describe cada una de las cuatro etapas de la metodología TDSP.

3.2.1. Entendimiento del negocio

Conocer el dominio en el que se está trabajando es importante para todos los integrantes del equipo, característica que no solo está presente en las metodologías de Ciencia de Datos, sino en cualquier grupo de trabajo con un objetivo en común. En esta etapa, se definen los objetivos del proyecto en conjunto con todos los integrantes que corresponda, así como

compartir la visión en la mayor medida posible y, si es el caso, con el cliente. Además, es necesario conocer el origen de los datos que serán empleados a lo largo del proyecto.

3.2.2. Adquisición y comprensión de los datos

Es una etapa que comprende la generación de un conjunto de datos a partir de toda la información recolectada. Este conjunto deberá estar adecuadamente caracterizado para que sea de utilidad en el proyecto. Como se verá más adelante, esta etapa es muy general en comparación con las demás metodologías, ya que está contenida en una sola.

3.2.3. Modelado

En esta etapa se determinan las condiciones necesarias dentro de los datos para aplicar las técnicas de aprendizaje automático. Se pretende encontrar el modelo que tenga mayor precisión a la hora de inferir los objetivos establecidos, para que sea trasladado a un entorno de producción.

3.2.4. Despliegue

Terminadas las etapas anteriores, es necesario aplicar los modelos en un entorno real para que tanto el grupo de trabajo como el cliente los acepten. Igualmente, pueden ser aplicadas métricas de evaluación u otras herramientas que mejoren el desempeño del producto final.

3.3. Metodología Fundamental para la Ciencia de Datos por Lotes (BFMDS)

Esta metodología corresponde a la primera extensión de la metodología FMDS de IBM, la cual se lleva a cabo en las dos primeras etapas. Dicha extensión está orientada a un dominio epidemiológico.

En la Figura 3.3 se muestra un esquema similar al de la metodología FMDS inicial. Sin embargo, se identifica en un recuadro verde la fase que es extendida en la metodología BFMDS. En el recuadro naranja, se muestra la fase que será extendida en esta investigación, extensión de la cual se hablará más adelante.

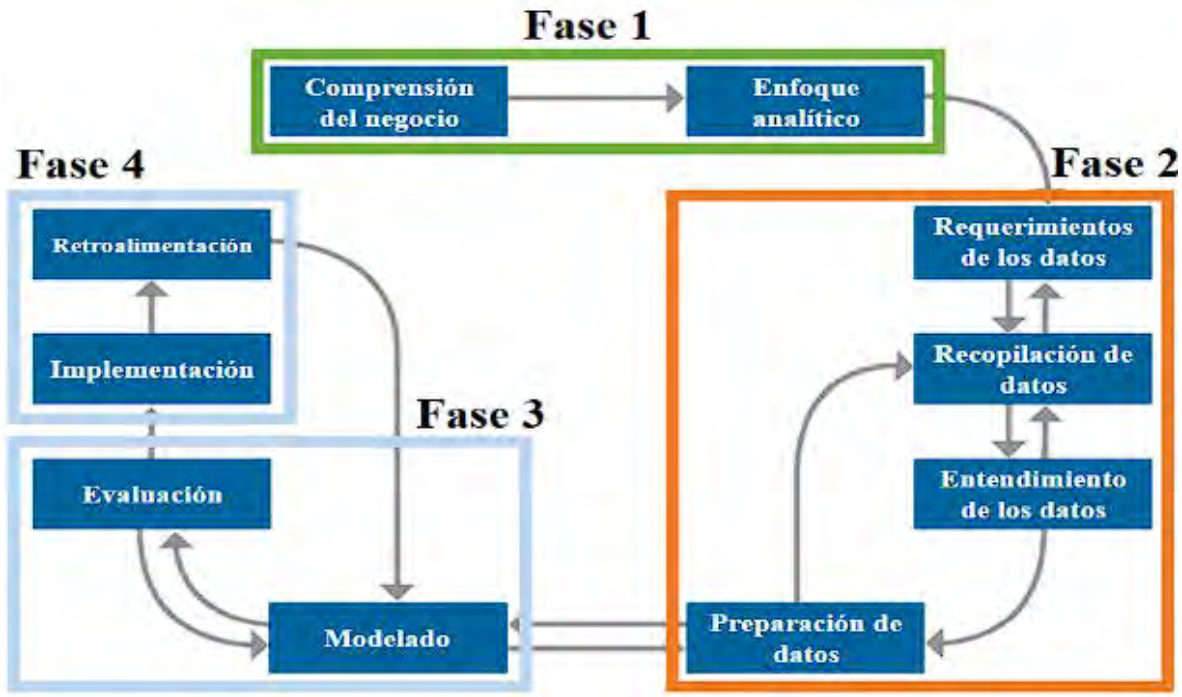

Figura 3.3: Fases extendidas en las metodologías *Batch FMDS* y *Batch FMDS* Extendida.

A partir de la etapa de requisitos de datos, las descripciones son las mismas que en la metodología *FMDS*, ya que no fueron extendidas en la investigación. Por lo tanto, solo se hablará sobre las primeras dos, las únicas etapas extendidas en *FMDS*.

3.3.1. Comprensión del negocio

Esta etapa tiene la finalidad de proporcionar una base para definir el problema, los objetivos y los requisitos de solución del proyecto. Comprende cinco tareas, las cuales cuentan con subtareas específicas para el dominio epidemiológico. A continuación, se detalla cada una de las tareas.

3.3.1.1. Recopilación de información

Antecedentes: se busca información acerca de los antecedentes del proyecto. Aquí se analizan los recursos que serán necesarios, tanto materiales como personales.

Diccionarios de términos: se hace un listado de la terminología referente al área epidemiológica adoptada en el proyecto.

Descripción del problema: se identifican las causas y consecuencias del problema.

3.3.1.2. Definición del tipo de estudio epidemiológico

Tipo de enfermedad: se establece la enfermedad epidemiológica sobre la que se va a trabajar. Son dos los tipos de enfermedades que existen: transmisibles y crónico-degenerativas. Las primeras son causadas por un agente infeccioso y se contagian por un medio, mientras que las segundas son alteraciones fisiopatológicas que se originan en los órganos y tejidos de un ser vivo.

Tipo de estudio epidemiológico: los dos tipos de estudio epidemiológicos son: analíticos y descriptivos. Los primeros se encargan de evaluar hipótesis entre exposiciones a factores de riesgo, mientras que los segundos utilizan los datos disponibles para analizar su comportamiento y extraer conocimiento del mismo.

3.3.1.3 Definición de preguntas de investigación

Se plantean la o las preguntas de investigación que serán resueltas al finalizar la investigación. Algunos ejemplos de estas preguntas pueden ser: ¿Qué otros factores se vieron afectados en las poblaciones con ciertas características?, ¿En qué regiones se aconseja realizar una vacunación?, ¿Qué sector de la población se vio más afectado por la enfermedad?, por mencionar algunas.

3.3.1.4. Delimitación de objetivos

Objetivo general: propósito que se quiere alcanzar con el proceso y que precisa los resultados que se desean obtener. Este objetivo deberá estar limitado por los recursos con los que se cuenta.

Objetivos específicos: comprenden los pasos que se deben seguir para alcanzar el objetivo general.

3.3.1.5. Establecimiento de criterios de aceptación

Criterios de éxito: puntos que determinan si se ha cumplido con el objetivo general y los específicos. Con estos criterios se podrá evaluar si es posible responder a las preguntas de investigación planteadas. Límites temporales: se define el periodo de tiempo que es de interés para el trabajo. Límites espaciales: se delimita el espacio geográfico que se pretende estudiar.

3.3.2. Enfoque analítico

Al delimitar el problema por todos los integrantes del grupo de trabajo, se debe determinar un enfoque analítico. Este enfoque tiene la finalidad de seleccionar el o los métodos que darán lugar a una solución a las preguntas de investigación establecidas. La única fase que compone esta etapa se conoce como *selección del método*.

3.3.2.1. Selección del método

Principalmente se seleccionan el o los métodos para procesar los datos y analizar el resultado de los mismos. Específicamente, se deberá:

a) Describir la o las categorías relacionadas con la pregunta de investigación.
b) Describir y analizar el método.
c) Identificar los parámetros de entrada y salida de los métodos.
d) Variar el método para analizar posibles mejoras.

Capítulo 4

Metodología Fundamental para la Ciencia de Datos por Lotes Extendida (MFCDLE)

El planteamiento de la metodología MFCDL Extendida surge de la necesidad de contar con una metodología que sea clara en cada etapa de la misma. Como se ha mencionado a lo largo de este documento, las metodologías de Ciencia de Datos actuales sufren de ser muy generales en cada una de sus etapas, es por ello que en este caso se extenderán las etapas de la fase de preparación de datos.

Ya que esta investigación es parte de un proyecto mayor, con el cual se pretende extender todas las diez etapas de la metodología de IBM, se remarca que únicamente se enfatizará en las etapas comprendidas dentro de la Fase 2. Sumado a lo anterior, la extensión realizada en este trabajo de tesis está orientada al dominio epidemiológico, por lo que no se garantiza su funcionamiento en dominios ajenos al establecido. El dominio epidemiológico se caracteriza principalmente por la naturaleza de los datos, los cuales corresponden a una población de individuos que, a su vez, podrá comprender una cantidad moderada o grande de humanos. Algunos ejemplos de datos poblacionales son censos regionales, municipales, estatales, nacionales o internacionales.

A pesar de que la metodología de IBM esté compuesta de 10 etapas, en este estudio, se agrupan en tres fases para hacer énfasis en la extensión propuesta. Es necesario mencionar que la metodología BFMDS fue la primera extensión que se hizo a la metodología original de IBM,

la cual se encargó de extender la Fase 1 y fue validada con un caso de estudio [9]. En la presente investigación se extiende únicamente la Fase 2, la cual se focaliza en el preprocesamiento de los datos. En la Figura 4.1, se muestra un esquema general de la metodología de IBM, en el que se destacan, en color naranja, las etapas que se extendieron.

Figura 4.1: Esquema general de la metodología BFMDS Extendida

Por otro lado, en la Figura 4.2, se muestra el esquema correspondiente a cada una de las dos etapas extendidas en la metodología *Batch FMDS* [9], remarcadas en color verde. Estas etapas dieron lugar a un total de seis tareas específicas orientadas a un dominio epidemiológico y fueron validadas con un caso práctico usando datos poblacionales de México.

Figura 4.2: Desglose de tareas de las etapas extendidas en la metodología *Batch FMDS.*

En la Figura 4.3, se muestra el esquema correspondiente a cada una de las cuatro etapas extendidas con esta investigación, las cuales están remarcadas en color naranja y se extienden con un total de nueve tareas específicas orientadas al dominio epidemiológico.

Las etapas extendidas son: Requerimientos de los datos, Recolección de datos, Entendimiento de los datos y Preparación de datos. En las secciones siguientes se detalla cada una.

Es necesario recordar que únicamente se están extendiendo las cuatro etapas correspondientes al tratamiento y preparación de los datos. Las primeras dos etapas están tomadas de la Metodología *Batch FMDS*, y las últimas cuatro permanecen igual que las de la metodología *FMDS* inicial, por lo que no se explicarán, únicamente se detallarán las representadas en la Figura 4.3.

Figura 4.3: Desglose de tareas de las etapas extendidas en la metodología BFMDS Extendida.

4.1. Requerimientos de los datos

La tarea que fue identificada para cumplir con la etapa de requisitos de los datos para un dominio epidemiológico fue la investigación de las fuentes de datos, de lo cual se explicará con más detalle a continuación.

4.1.1. Identificación de fuentes oficiales

Para identificar los requisitos con los que deben contar los datos, es necesario tener claro el problema que se pretende resolver con la aplicación de la metodología de Ciencia de Datos. Ya que el dominio sobre el cual se orienta esta metodología es el epidemiológico, fue incluida la tarea explicada a continuación.

Las fuentes deben ser oficiales, de acceso público y deben reportar su última fecha de actualización. Dado que el robo y falsificación de información se ha vuelto una práctica común en la red, es de suma importancia que los datos contemplados pertenezcan a fuentes oficiales. Dentro de México, para el dominio epidemiológico, una de las fuentes más importantes es la Dirección General de Epidemiología, por lo que será consultada. A partir de una fuente, y

dependiendo de las consideraciones realizadas en las etapas anteriores, se tendrá una base para elegir las restantes.

Algunas de las consideraciones que deben tomarse en cuenta son las siguientes:

a) En caso de que se quiera hacer uso de datos del censo poblacional, se recomienda utilizar los datos del Instituto Nacional de Estadística, Geografía e Informática (INEGI) [26].
b) Para obtener un diccionario con las claves oficiales de cada enfermedad, se consulta el portal del Centro Mexicano para la Clasificación de Enfermedades (CEMECE) [27].
c) Para la consulta del estado socioeconómico de cada municipio, se dispone del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) [28].
d) Para la obtención de información geográfica de cada municipio, se recomienda recurrir a la fuente del Sistema Nacional de Información Municipal (SNIM) [29].

4.2. Recopilación de datos

4.2.1 Obtención de los datos

Para esta etapa, deberá contemplarse el periodo infeccioso que se desea estudiar, con lo que se procede a descargar y estandarizar los conjuntos de datos correspondientes. Es una buena práctica asegurarse de contar con una copia física de los datos recolectados. Lo anterior se recomienda debido a que, en ocasiones, las fuentes pueden retirarlos o realizar algún tipo de mantenimiento en el sitio, dejando los datos inaccesibles. Esta tarea estará completa una vez que se cuente con una copia de todos los conjuntos de datos en un repositorio, así como la fecha de su adquisición.

4.3. Entendimiento de los datos

4.3.1. Análisis de diccionarios de datos

Las dependencias que proporcionan sus datos suelen aportar un diccionario de todos los atributos (columnas) comprendidos. En el dominio epidemiológico, es necesario revisar dichos diccionarios para encontrar las claves de cada enfermedad. En el campo de la epidemiología, existen claves universales para cada enfermedad, como por ejemplo el COVID-

19, cuyas claves son “U071” y “U072”. Para el caso de los atributos, algunos ejemplos pueden ser la clave de un municipio o entidad, cuyos nombres serían “CVE_MUN” y “CVE_ENT”, respectivamente.

Hay casos en los que dos atributos pueden estar conjugados para formar uno solo, por ejemplo, tomando el caso de las claves de municipio o entidad, al unirse, se obtiene la clave única de un municipio en todo el país. Es necesario saber diferenciar estos casos porque, aunque provengan de otros dos atributos, el conjugado puede tener funcionalidades muy distintas, contribuyendo al éxito de los objetivos planteados. Es una buena práctica documentar todos los atributos de interés para el proyecto, lo que ahorrará tiempo en tareas futuras.

Teniendo en cuenta el dominio epidemiológico, se considerará que esta tarea está completada si los integrantes que trabajarán con los datos cuentan con un diccionario sin ambigüedades y saben de forma precisa a qué se refieren los atributos de interés.

4.3.2. Análisis exploratorio de los datos

Con motivo de tener un mejor entendimiento de los conjuntos, se recomienda caracterizar la distribución de los datos. La obtención de tablas de frecuencia e histogramas es de gran ayuda para caracterizar los datos y conocer sus alcances y limitaciones.

En el dominio epidemiológico, es común hacer distinciones orientadas al tipo de población a analizar. Tener clara esta distinción en la investigación ayuda a entender mejor los datos con los que se cuentan. Por ejemplo, si el conjunto de datos corresponde a un periodo de tiempo específico, al importar los datos a una herramienta de manipulación, ésta puede determinar de forma sencilla el número de elementos existentes, lo que de entrada otorgaría el número de casos en la población analizada.

El uso de histogramas o tablas de frecuencia es muy común, ya que su implementación y entendimiento son muy intuitivos, lo cual podría aprovecharse en un dominio epidemiológico, determinando qué grupos de la población cuentan con una característica en común y en qué magnitud.

Se sabrá que la tarea ha concluido correctamente si, al realizar el análisis exploratorio, se cuenta con los documentos necesarios y se tiene documentado qué tipos de datos se tienen, así como la comprensión de su clasificación para ubicarlos de forma inmediata cuando sean requeridos.

4.3.3. Definición de filtros

Una vez definidos los alcances y limitaciones, se establecen los criterios de inclusión y exclusión que permitan delimitar los datos base del estudio. Por ejemplo, en estudios epidemiológicos se recomienda tomar municipios con más de 100 mil habitantes, pero también pueden definirse filtros de restricción por edad, sexo, o incluso si la población tiene algún tipo de seguro o nivel académico.

Muchos de los conjuntos de datos en el dominio epidemiológico contienen una gran cantidad de atributos y registros, lo que muchas veces hace que el conjunto sea innecesariamente grande. Es por ello que, al haber transitado las primeras dos etapas de la metodología, el usuario tendrá una idea clara de los objetivos del proyecto y los atributos que se requieren. Por esta razón, en esta tarea se definen los filtros que permitirán generar un almacén de datos específico para cumplir con los objetivos propuestos.

El usuario sabrá que esta tarea está completa cuando identifique qué atributos necesita y cuáles no, así como los rangos de valores de algún atributo en específico y si será necesario crear nuevos atributos a partir de los ya existentes.

4.4. Preparación de datos

4.4.1. Limpieza de datos

Uno de los pasos más conocidos es la limpieza de datos, en el cual se eliminan los registros (filas) con anomalías y se aplican técnicas estadísticas para resolver problemas específicos. En el dominio epidemiológico, se suele trabajar con varios conjuntos de datos, por lo que esta etapa debe aplicarse a cada uno.

A continuación, se listan algunas de las técnicas más comunes de limpieza de datos:

a) Eliminación de duplicados: Identificar y eliminar registros repetidos en el conjunto de datos . Los duplicados pueden llevar a resultados engañosos en el análisis.
b) Manejo de valores faltantes: Los datos incompletos pueden convertirse en un problema significativo. Las técnicas para manejarlos incluyen:
    - Eliminación de registros: Si un registro tiene demasiados valores faltantes, es mejor eliminarlo.

- Imputación: Los valores faltantes pueden ser reemplazados por un valor estimado (media, mediana, moda) o mediante métodos más complejos como la imputación múltiple o el uso de modelos predictivos.

c) Corrección de errores de formato y tipografía: Resolver problemas de formato (fechas mal escritas o números mal puntuados) y errores tipográficos en datos textuales.
d) Normalización y estandarización:
    - Normalización (*Min-Max Scaling):* Escala los datos entre un rango definido (típicamente 0 a 1).
    - Estandarización (*Z-score Scaling*): Los datos se reescalan para tener una media de cero y una desviación estándar de uno.
e) Detección y tratamiento de valores atípicos (*outliers*): Los valores extremadamente altos o bajos pueden distorsionar el análisis. Se pueden manejar mediante truncamiento, transformación de datos (logarítmica, por ejemplo), o mediante métodos robustos de análisis.
f) Validación de la consistencia de datos: Asegurarse de que los datos cumplan con las reglas de negocio o lógica interna (por ejemplo, que una fecha de inicio sea anterior a la fecha de fin).
g) Conversión de tipos de datos: Cambiar el tipo de datos de una columna según sea necesario para el análisis (por ejemplo, convertir una columna de texto a fechas o números).
h) Desambiguación y estandarización de categorías: Unificar categorías similares bajo un solo nombre (por ejemplo, “SARS-CoV-2” y “COVID-19”).
i) Integración de datos: Unificar datos de múltiples fuentes, resolviendo inconsistencias en la estructura o etiquetas de los datos.
j) Segmentación de texto y corrección de errores: En datos textuales, segmentar texto en unidades lógicas, corregir errores ortográficos o eliminar palabras irrelevantes.

El usuario habrá completado la tarea cuando se cuente con conjuntos de datos aptos para ser manipulados sin problemas, y que puedan ser tratados con el *software* seleccionado.

4.4.2 Aplicación de filtros

Resuelto lo anterior, se procede a filtrar vertical y horizontalmente de acuerdo con los criterios de la población, desechando así los atributos que no se requieran y los registros que no cumplan con los criterios de aceptación definidos.

En esta tarea, es necesario hacer uso de herramientas computacionales, como *software* de manipulación de datos. La mayoría de este *software* especializado ya cuenta con herramientas de filtrado de datos simples, pero en caso de que algún filtro sea muy específico y no esté contenido en dicho *software*, este tendrá que ser creado o programado para lograr el cometido. En el caso del dominio epidemiológico, es posible que se requiera elaborar filtros en cascada, es decir, aplicar uno sobre otro para un mismo conjunto de datos. En estos casos, se recomienda tener presente la memoria del equipo de cómputo, porque si no se administra correctamente la manipulación de los datos, el proceso de filtrado puede demorarse.

La tarea estará completada cuando se cuenten únicamente con los registros y atributos acordados para el cumplimiento de los objetivos. Es necesario aclarar que se puede regresar a esta tarea en caso de ser necesario, ya sea porque se descubrió que se necesitaba otro atributo o porque algún atributo resultó innecesario.

4.4.3. Creación de atributos de apoyo

Esta tarea hace referencia a la creación de atributos que no se encuentran en los conjuntos de datos, pero que pueden obtenerse a partir de los ya existentes. Para lograrlo, es necesario hacer uso de las herramientas computacionales elegidas para realizar las acciones pertinentes y obtener dichos atributos, como puede ser la generación de código para crear funciones que automaticen dicha tarea para todos los valores necesarios.

Al tratar con diversos conjuntos de datos, como en el caso del dominio epidemiológico, es necesario crear identificadores foráneos para el acople de atributos entre los conjuntos. Por ejemplo, la densidad poblacional puede obtenerse a partir de los atributos de población total y área territorial. Otro ejemplo es la clave única de municipio, que se puede obtener mediante el acople de la clave de entidad más la clave de municipio dentro de esta, generando un identificador único que puede usarse para conectar distintos conjuntos de datos.

Esta tarea estará completada cuando ya no sean necesarios más atributos que se consideren indispensables para el cumplimiento de los objetivos establecidos en las primeras etapas de la metodología.

4.4.4. Integración de datos preprocesados en un almacén de datos

Los datos preprocesados se integran en una estructura denominada *almacén de datos*, la cual debe cumplir con los formatos necesarios para su explotación con las herramientas de modelado, así como para su uso en etapas posteriores, como la evaluación, implementación y retroalimentación.

El usuario sabrá que la tarea está completada cuando disponga de un único conjunto de datos con todos los atributos necesarios para proceder sin dificultades a la siguiente etapa de la metodología: el modelado.

4.5. Comparación de las metodologías FMDS, BFMDS y BFMDS Extendida

Aunque está fuera del alcance del presente estudio, se tiene planeado extender todas y cada una de las etapas de la metodología FMDS inicial. Esta investigación forma parte de un proyecto más amplio, el cual tiene ese objetivo. En la siguiente página se presenta una comparativa que muestra un panorama del avance, tomando en cuenta lo mostrado en este capítulo (Tabla 4.1).

Tabla 4.1: Comparación de las metodologías FMDS, BFMDS y BFMDS Extendida

Metodología			
Etapa	FMDS	Batch FMDS	Batch FMDS Extendida
Entendimiento del negocio	Definición del problema, los requisitos y objetivos de solución.	Adicionalmente: Recopilación de información, Definición del tipo de estudio epidemiológico, definición de pregunta de investigación, Definición de objetivos y Establecimiento de criterios de aceptación.	
Enfoque analítico	Definición de técnicas a utilizar de acuerdo con la problemática.	Selección del método mediante la descripción de la categoría y del método, identificación de parámetros y variaciones del método.	
Requerimientos de los datos	Determinación de los requisitos de datos.		Identificación de fuentes oficiales de los datos.
Recopilación de los datos	Identificación y reunión de los conjuntos de datos.		Identificación de periodo infeccioso de interés y resguardo una copia física de los datos.
Entendimiento de los datos	Comprensión de los conjuntos, evaluación de su calidad y descubrimiento de características iniciales sobre los datos.		Revisión de diccionarios de datos, análisis exploratorio de los datos y definición de filtros.
Preparación de datos	Limpieza, integración, transformación y contracción del conjunto de datos.		Limpieza de datos, aplicación de filtros, creación de atributos de apoyo e integración de los datos preprocesados.
Modelado	Implementación de los modelos de acuerdo con el enfoque analítico definido.		
Evaluación	Evaluación del modelo de acuerdo a su eficacia.		
Implementación	Implementación del modelo en un entorno de producción		
Retroalimentación	Recopilación de los resultados del modelo implementado.		

Capítulo 5

Desarrollo del Caso Práctico

Para validar la extensión de las etapas mencionadas en el capítulo anterior, se propuso desarrollar un caso práctico enfocado en analizar la reciente pandemia por COVID-19 en México, la cual sacó a la luz problemas relacionados con la toma de decisiones en todo el mundo. Por esta razón, se considera que este trabajo puede ser un gran aporte al dominio epidemiológico del país, contribuyendo a la toma de decisiones en caso de una posible pandemia futura.

Se muestra a continuación, la forma en la que se abordó cada etapa que conforma la metodología MFCDL Extendida, haciendo énfasis en las etapas correspondientes a la preparación de datos, enfocadas en el dominio epidemiológico.

5.1. Comprensión del negocio

De acuerdo con la metodología BFMDS [30], en esta etapa se comienza con la recopilación de información, que consiste en la búsqueda de fuentes oficiales dentro del dominio epidemiológico para la obtención de los datos pertinentes. Para la fase de definición del tipo de estudio epidemiológico, se identificó el tipo de enfermedad como transmisible, lo que implica que se transfiere de un infectado a un no infectado, ya sea de forma directa (ambos presentes) o indirecta (vectores de transmisión, artículos infectados). Con base en estos criterios, se definió el tipo de estudio como descriptivo [31].

Como definición de la pregunta de investigación, se respondió a: ¿Cómo se distribuyó la mortalidad por COVID-19 en los municipios que promediaron la menor y mayor tasa de mortalidad?

El objetivo general de este caso práctico fue aplicar una metodología de ciencia de datos para analizar el comportamiento de la pandemia por COVID-19 en México.

Para terminar esta etapa, el criterio de aceptación se estableció como la capacidad de responder a la pregunta de investigación planteada, mediante el cumplimiento del objetivo general al visualizar gráficamente los datos de COVID-19 provenientes de fuentes oficiales.

5.2. Enfoque analítico

En esta etapa, se analizó la pregunta de investigación y se definieron los métodos más adecuados para responderla. Con este propósito y con base en los motivos descritos en el capítulo 5.7, se decidió utilizar el algoritmo de agrupamiento *K-means* en los municipios que cumplan con los criterios de aceptación definidos. Hecho esto, se emplearán gráficas de mortalidad para analizar el comportamiento de los grupos que arrojen más información al evaluar los resultados obtenidos del agrupamiento.

5.3. Requerimientos de datos

5.3.1 Identificación de fuentes oficiales

Se identificaron cinco fuentes oficiales de datos, las cuales se presentan en la Tabla 5.1. Los conjuntos de estas fuentes de datos fueron los únicos necesarios para la realización de este caso práctico.

Dentro de los conjuntos de datos de cada una de estas fuentes oficiales, fue posible realizar lecturas del contenido. En ocasiones, los datos pueden estar encriptados o encapsulados para leerse únicamente con un *software* específico. De ser el caso, será necesario revisar la documentación de las herramientas necesarias para acceder a dichos datos. En este caso práctico, únicamente fueron necesarias funciones específicas para los archivos CSV, TXT y DAT, gracias a que el software de programación Python cuenta con herramientas que permiten leer esos tipos de archivos. Por lo tanto, solo fue necesario revisar la documentación para aplicar correctamente dichas herramientas. Las extensiones más comunes en las cuales los datos suelen estar guardados se muestran en la Tabla 5.2.

Tabla 5.1: Fuentes oficiales de datos consultadas.

Fuente oficial	Descripción del conjunto de datos	Número de registros	Número de atributos
DGIS (Dirección General de Información Sanitaria) [32]	Registros de muertes 2020-2021	2,208,992	59
INEGI (Instituto Nacional de Estadística y Geografía) [26]	Censo poblacional y de vivienda 2020	195.662	286
CEMECE (Centro Mexicano para la Clasificación de Enfermedades) [27]	Catálogo internacional de enfermedades	600,689	---
CONEVAL (Consejo Nacional de Evaluación de la Política de Desarrollo Social) [28]	Indicadores de pobreza 2020	2,469	161
SNIM (Sistema Nacional de Información Municipal) [29]	Registros de información municipal	2,469	5

5.4. Recopilación de datos

5.4.1. Obtención de los datos

Se definió el periodo de estudio desde la primera muerte registrada en México el 18 de marzo de 2020 hasta el último registro reportado en el conjunto de datos de la DGIS el 31 de diciembre de 2021. Cabe mencionar que, en el dominio epidemiológico, al requerir de varios conjuntos de datos, no siempre se dispone de aquellos que coincidan exactamente con el periodo de estudio. Por ejemplo, en el caso de los censos, los datos se actualizan cada 10 años, por lo que se deben tomar los más cercanos al periodo definido. Para el caso de los datos de la DGIS, se obtuvieron los correspondientes al periodo de 2020 a 2021, mientras que los conjuntos de las fuentes oficiales restantes se tomaron del año 2020. Es altamente recomendable contar con una copia física de los datos en los equipos donde se realice el estudio de Ciencia de Datos, debido a que la información en línea puede ser modificada o retirada en cualquier momento. Por esta razón, como parte del proyecto, se realizaron resguardos

periódicos de todos los conjuntos de datos, desde los descargados directamente de la fuente hasta los generados a lo largo de las etapas.

Tabla 5.2: Tipos de archivos de almacenamiento de datos.

Extensión del archivo	Descripción
CSV	Ideal para datos tabulares simples. Excelente interoperabilidad entre sistemas. Fácilmente legible y editable con texto plano o herramientas de hoja de cálculo. No admite estructuras complejas ni tipado de datos.
JSON	Formato ligero, basado en texto para representar estructuras de datos simples a complejas. Altamente interoperable con tecnologías web. Facilita el intercambio de datos entre servidores y aplicaciones web. No es ideal para grandes volúmenes de datos.
XML	Estructura los datos en un formato que es tanto humano como máquina legible. Altamente personalizable, soporta esquemas y espacios de nombres. Más verboso que JSON, lo que puede aumentar el tamaño del archivo.
DAT	Un formato genérico para datos en formato específico o binario. Su interpretación depende del programa que lo crea. Versátil para almacenar datos binarios o texto, pero carece de estandarización, lo que puede complicar la interoperabilidad.
TXT	Máxima simplicidad para almacenar datos en texto plano. Fácil de crear y editar con cualquier editor de texto. No admite estructuras de datos complejas ni metadatos. Ideal para notas, configuraciones sencillas o datos lineales.

5.5. Entendimiento de los datos

5.5.1. Análisis a diccionarios de datos

En esta tarea se consultaron los diccionarios de datos para conocer qué nombres de columnas hacen referencia a los atributos de interés. Algunos ejemplos de estos atributos fueron los referentes a la clave de entidad y municipio, la población total de cada municipio, su área territorial y el porcentaje de su población en pobreza, por mencionar algunos.

Por el lado del entendimiento de los valores en los registros, únicamente se identificó que las claves que corresponden a defunción por COVID-19 son las equivalentes a “U071” y “U072”, por lo que únicamente son de interés las defunciones con dichas claves.

Ya que desde las primeras etapas se tuvo claro qué es lo que se quería extraer de los datos y por ende ya se sabía qué buscar, el entendimiento de los mismos fue inmediato. Cabe

mencionar que aspectos como éste, en el que desde un principio se llevan a cabo las etapas de manera adecuada, se contribuye posteriormente a agilizar el proceso de la metodología, ya que algunas etapas pueden resultar más inmediatas.

5.5.2. Análisis exploratorio de los datos

Se realizó un análisis estadístico exploratorio con la finalidad de caracterizar los valores de los conjuntos de datos. Como parte de este análisis, se identificaron algunos municipios que no reportaron mortalidad por COVID-19 y otros con el mayor número de fallecidos por COVID-19. Además, como parte de un resumen corto de los conjuntos de datos que se usaron, se aplicaron algunas funciones para determinar las dimensiones de los conjuntos. En la Figura 5.1 se puede ver que los conjuntos de datos tienen dimensiones de 200,249 y 238,677 registros respectivamente, así como 6 columnas en ambos casos.

```
Número de atributos del dataframe: 6; cada uno con 200249 datos.
Número de atributos del dataframe: 6; cada uno con 238677 datos.
```

Figura 5.1: Ejemplo de visualización de algunas características de los datos.

Ya que se conoce que cada conjunto de datos corresponde a un año en específico, es posible obtener de forma inmediata el número de muertes por COVID-19 para cada año, siendo 200,249 para el 2020 y 238,677 para el 2021, de acuerdo con la fuente de la DGIS.

5.3.3. Definición de filtros

Con base en el análisis anterior, se determinó que los municipios/alcaldías que participarían en el estudio serían aquellos cuya población fuera mayor de 100 mil habitantes y que hubieran tenido al menos un fallecido por COVID-19. Estos son parámetros que ya se han utilizado y establecido en el dominio epidemiológico. Posteriormente, para los datos municipales se determinó que serían filtrados todos los atributos que no correspondieran a los listados a continuación:

a) Clave de entidad
b) Clave de municipio
c) Nombre de municipio

d) Población total
e) Área territorial
f) Porcentaje de población en pobreza

5.6. Preparación de los datos

5.6.1. Limpieza de datos

Como parte de esta actividad, se identificaron los datos con valores anómalos. Dentro del conjunto de datos de mortalidad, se encontraron registros de fechas de defunción con valores igual al símbolo "?", lo que indica que dichos registros no habían sido ingresados correctamente debido a un error humano, por lo que fueron desechados. Todos los demás conjuntos no presentaron problemas con valores anómalos y se procedió directamente con la siguiente tarea.

5.6.2 Aplicación de filtros

Esta actividad consiste en la implementación de instrucciones que permitan seleccionar los registros y atributos que serán utilizados como base para generar el almacén de datos. Los atributos seleccionados son los listados en la sección 5.5.3, por lo que se desecharon todos aquellos que no correspondían con dichos atributos.

En esta parte, es común el empleo de herramientas computacionales, como algunos lenguajes de programación. Para este estudio en particular, se utilizó Python en su versión 3.7 con el fin de programar las funciones que permitieran filtrar los atributos mencionados anteriormente.

En las primeras cuatro filas de la Tabla 5.3, se encuentran las funciones de los filtros utilizados para esta tarea. En dichas funciones, se entiende como marco de datos al conjunto de datos en proceso de preparación. El detalle de estas funciones está fuera del alcance de este documento de tesis.

Tabla 5.3: Funciones implementadas para el filtrado de datos y generación de atributos.

Función	Descripción
vert_f ilt(df rame, valores)	Filtro vertical de valores en un marco de datos.
hor_f ilt(df rame, var, restriccion, valor)	Filtro horizontal de acuerdo con el tipo de restricción indicada para un marco de datos.
hor_f _rang(df rame, var, minval, maxval)	Filtro horizontal para un marco de datos de acuerdo con un intervalo dado.
ver_f ilt2(df rame, atributo, claves)	Realiza un filtro vertical de acuerdo con un conjunto de claves. (utilizado para detectar las muertes por COVID-19 en DGIS)
gen_id(df rame, atributo1, atributo2)	Genera identificadores únicos para cada municipio. Estos también funcionan como identificadores foráneos.
mes_dia(df rame, anio, mes, dia)	Genera un formato de fecha año-mes-día para los registros de mortalidad. Lo que también ayuda a su análisis temporal.
acople_col(df rame, lista, nombre)	Incorpora un atributo nuevo con todas las instancias a un marco de datos.
cont_enf (idd, iddgis)	Devuelve el conteo de las muertes de todos los municipios de acuerdo con el conjunto de mortalidad de DGIS.
den_pob(df rame)	Devuelve la densidad poblacional del marco de datos.
tasa_mort(df rame)	Devuelve la tasa de mortalidad del marco de datos.
normm(df rame, atributo)	Normaliza los valores de un atributo del marco de datos.
prom_mort_mem(df rame, atributo, grupos)	Devuelve el promedio de mortalidad por COVID-19 de cada grupo.
extremos(df rame, atributo, memb, grupos)	Devuelve los tres grupos con menor mortalidad y los tres grupos con mayor mortalidad por COVID-19.

5.6.3. Creación de atributos de apoyo

Debido a que algunos de los atributos necesarios provienen de distintos conjuntos de datos, fue necesario crear nuevos atributos para utilizarlos como llaves.

Primero, para identificar a cada municipio de forma única, se creó una nueva llave a partir de la clave de entidad y la clave municipal. La clave de entidad contiene un número del 1 al 32, que indica el estado correspondiente, mientras que la clave municipal indica un número que varía desde el 1 hasta la cantidad de municipios que tenga el estado. Por esta razón, no es posible usar solo la clave municipal como identificador de un municipio, ya que el municipio con clave municipal igual a 1 se repetiría 32 veces. Al crear una función que acople la clave de

entidad junto con la clave municipal, se genera una clave única e irrepetible para cada municipio.

Una vez que cada municipio fue adecuadamente identificado en todos los conjuntos de datos, se trasladaron los atributos relacionados al número de muertes en cada municipio, así como su área territorial y porcentaje de población en pobreza, desde los distintos conjuntos de datos a los que estos atributos pertenecían.

Al tener un solo conjunto de datos con todos los atributos necesarios, se crearon nuevos atributos a partir de los existentes debido a los requerimientos del estudio.

Los atributos que fue necesario definir fueron los siguientes: Densidad poblacional (a partir del número de habitantes de cada municipio y su área territorial) Tasa de mortalidad por 100 mil habitantes (a partir de las muertes por COVID-19 y la población total). Las expresiones para el cálculo de dichos atributos son la Ec. 5.1 y la Ec. 5.2, respectivamente.

$$densidad\ poblacional = \frac{población}{área\ territorial\ [km^2]} \tag{5.1}$$

$$mortalidad = \frac{muertes}{población} \cdot 100{,}000 \tag{5.2}$$

Estos nuevos atributos fueron creados para cada uno de los municipios seleccionados para el estudio e incorporados al conjunto de datos que contiene todos los atributos filtrados hasta el momento.

5.6.4. Integración de datos preprocesados en un almacén de datos

La tarea final previa al modelado es la conjunción de un almacén de datos con los atributos y registros necesarios para todas las etapas posteriores. Es importante que, aunque no se ingresen todos los atributos en la técnica de modelado seleccionada, se tomen en cuenta para identificar y caracterizar adecuadamente cada hallazgo resultante.

El almacén de datos quedó conformado con los siguientes atributos:

a) Clave única municipal,
b) Nombre del municipio,
c) Densidad poblacional,
d) Porcentaje de población en pobreza,
e) Tasa de mortalidad por COVID-19,

Como puede verse, ya no son necesarios los atributos que dieron lugar a la clave única de municipio, así como también fueron desechados los correspondientes al área territorial, población total y total de muertes por COVID-19, ya que estos dieron lugar a la densidad poblacional y a la tasa de mortalidad por cada 100 mil habitantes.

5.7. Modelado

La etapa de modelado se abordó implementando técnicas ampliamente conocidas dentro del grupo de trabajo. Una de estas técnicas de modelado es el agrupamiento. Específicamente, se utilizó el algoritmo *K-means* debido a su fácil aplicación y la claridad de sus resultados [33].

Este trabajo utiliza el algoritmo *K-means* estándar, que es un método iterativo que consiste en la partición de un conjunto de n objetos en k ≥ 2 grupos, de modo que los objetos de un grupo sean similares entre ellos y diferentes de los elementos de otros grupos. La descripción matemática del algoritmo *K-means* es la siguiente:

Sea $X = \{x_1, ..., x_n\}$ el conjunto de n objetos a particionar de acuerdo con un criterio de similitud, donde $X_i \in R^d$ para $i = 1, ..., n$ y $d \geq 1$ es el número de dimensiones. Además, sea $k \geq 2$ un número entero y $K = \{1, ..., k\}$. Para una k-partición $P = \{G(1), ..., G(k)\}$ de X, v_j denota el centro del grupo $G(j)$, para $j \in K$, y sea $V = \{v_1 \ldots v_k\}$ y $W = \{W_{11} ..., W_{ij}\}$.

En la Ecuación 5.3, se muestra el problema de agrupamiento como un problema de optimización.

$$P: minimizar\ Z(W,V) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} D(x_i, v_j) \tag{5.3}$$

donde $w_{ij} = 1$ si y solo si el objeto x_i es miembro del grupo G(j), y $D(x_i, v_j)$ denota la distancia Euclídea entre x_i y v_j para i = 1, ..., k.

El pseudocódigo del algoritmo se muestra en el Algoritmo 1.

Algoritmo 1 K-means

1: **Inicialización**:
2: Asignar el valor a k;
3: X := {x_1, . . . , x_n;
4: V := {v_1, . . . , v_k;
5: **Clasificación**:

6: **for** $x_i \in N$ & $v_k \in V$ do
7: Cálculo de la distancia de cada xi a los centroides k;
8: Asignar el objeto x_i al centroide más cercano v_k;
9: end **for**
10: **Cálculo de centroide**:
11: Calcular el centroide v_k;
12: **Convergencia**:
13: **if** $V := v_i, \ldots, v_k$ no cambia en dos iteraciones consecutivas then
14: Parar el algoritmo
15: **else**: **Ir a Clasificación**
16: **end if**
17: **Fin**

La implementación del algoritmo *K-means* fue realizada en lenguaje C en un equipo con las siguientes características: (i) SO: Windows 10 Home, (i) RAM: 8 *Gigabytes*, y (i) Procesador: *Intel® Core™* i5-9300. Al realizar el agrupamiento con el algoritmo *K-means*, considerando 18 grupos y usando los atributos de densidad poblacional y porcentaje de población en pobreza, se obtuvo el gráfico mostrado en la Figura 5.2.

En la Figura 5.2, se puede observar que los municipios del grupo con mayor tasa de mortalidad por COVID-19 están señalados en color anaranjado, y los que presentan menor tasa de mortalidad en color verde. En este trabajo, al grupo con mayor tasa de mortalidad por COVID-19 se le denominó Grupo CDMX, mientras que al grupo con menor tasa de mortalidad se le llamó Grupo Chiapas.

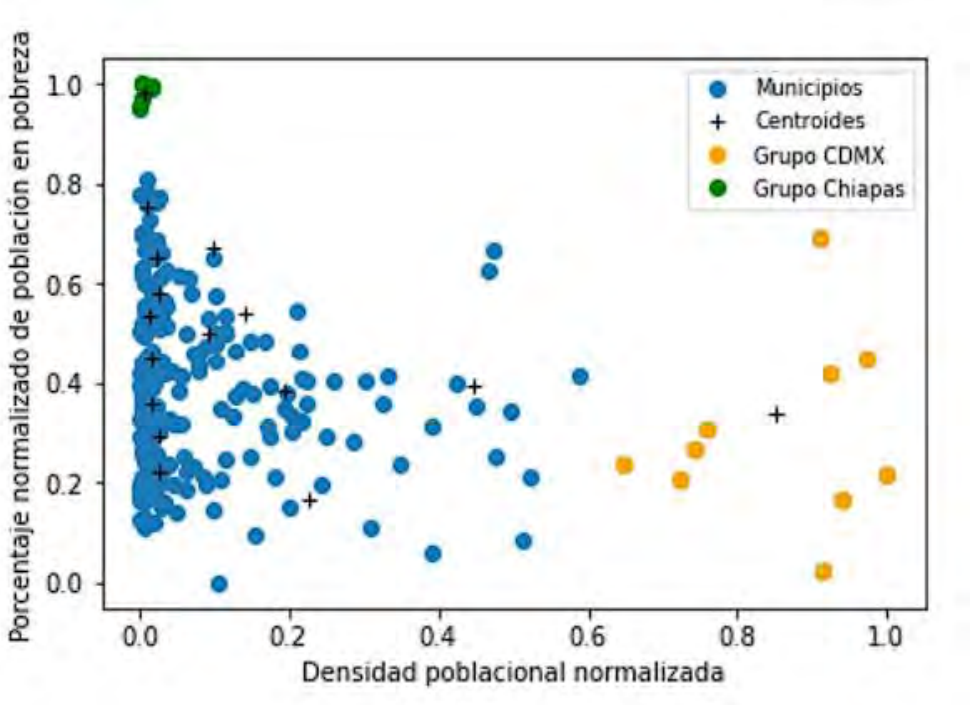

Figura 5.2: Distribución de municipios y centroides mediante agrupamiento.

5.8. Evaluación

La generación de recursos visuales es de gran ayuda para evaluar la coherencia de los resultados obtenidos en el modelado. Las Figuras 5.3 y 5.4 servirán de apoyo para evaluar la etapa de modelado, ya que en ellas pueden observarse los municipios de interés en sus respectivos estados.

Primero, en la Figura 5.3, se ubican las alcaldías del grupo que presentó la tasa de mortalidad promedio por COVID-19 más alta de México. Puede verse que estas alcaldías se encuentran principalmente en la región norte de la Ciudad de México.

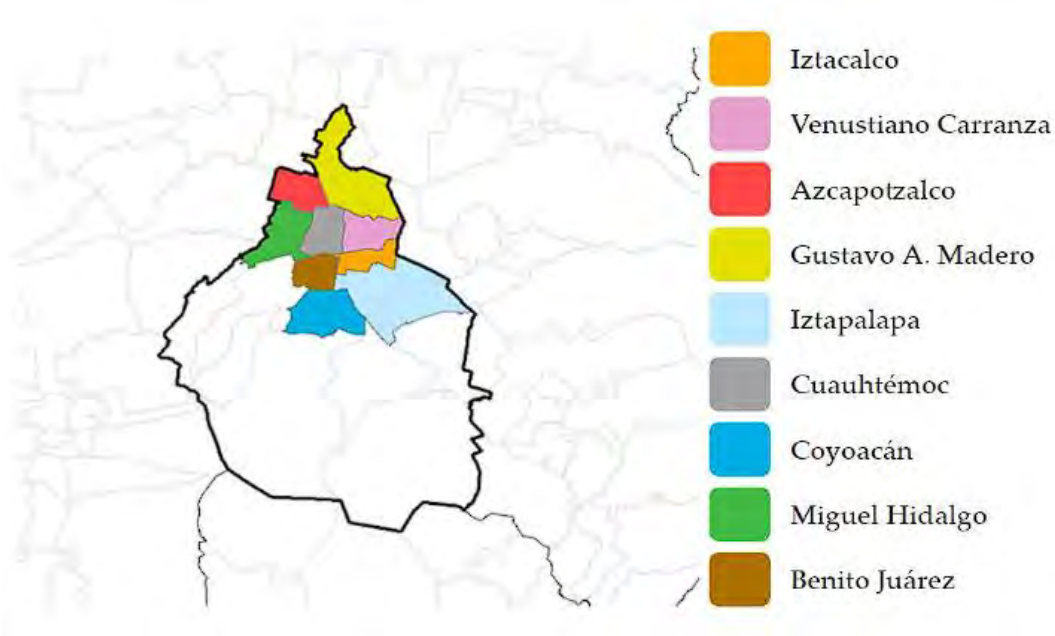

Figura 5.3: Alcaldías con mayor tasa de mortalidad por COVID-19.

Para el Grupo Chiapas, se puede ver la Figura 5.4, en la cual se ubican los municipios pertenecientes al grupo con la tasa promedio de mortalidad por COVID-19 más baja del país. Especialistas en el área epidemiológica del Instituto de Salud Pública en el estado de Morelos analizaron los resultados del modelado y validaron lo obtenido.

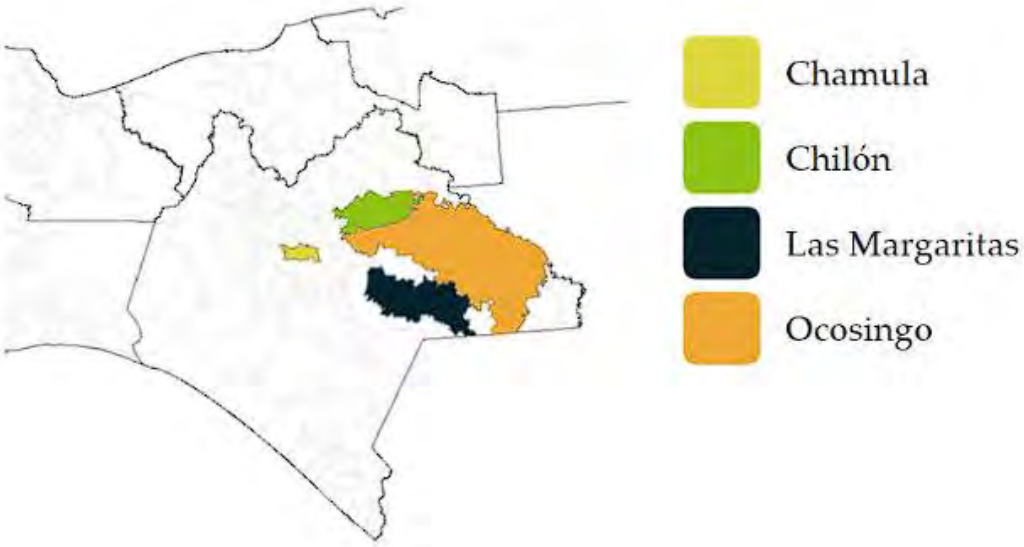

Figura 5.4: Municipios con menor tasa de mortalidad por COVID-19 en México.

Al analizar cada alcaldía del Grupo CDMX, se obtuvieron los valores presentados en la Tabla 5.4, en los cuales se muestra la tasa de mortalidad por COVID-19 por 100 mil habitantes para cada alcaldía del grupo, su densidad poblacional por kilómetro cuadrado, y el porcentaje de su población en situación de pobreza.

Tabla 5.4: Datos de mortalidad por COVID-19 y población de las alcaldías del grupo CDMX.

Alcaldía	Mortalidad por 100 mil habitantes (2020-2021)	Densidad poblacional (población/km^2)	Porcentaje de población en pobreza (%)
Iztacalco	841.62	17,595	25.2
Venustiano Carranza	825.09	13,05	30.0
Azcapotzalco	780.64	12,712	24.2
Gustavo A. Madero	749.73	13,333	33.8
Iztapalapa	706.29	16,243	43.9
Cuauhtémoc	680.54	16,542	20.9
Coyoacán	626.57	11,378	27.1
Miguel Hidalgo	544.25	9,01	13.5
Benito Juárez	492.91	16,08	7.9
Promedio	694.18	13,994	25.1

Puede verse en la Tabla 5.2 que las tasas de mortalidad por 100 mil habitantes más altas se encuentran en Iztacalco, Venustiano Carranza y Azcapotzalco para el periodo de 2020 y 2021. A diferencia de Iztapalapa, todas las alcaldías del grupo están por debajo del 33.8 % de población en situación de pobreza, lo que en contraste con todo el país indica que entre menor pobreza, existe una mayor tasa de mortalidad por COVID-19.

5.9. Implementación

La forma en la que se abordó la implementación para este caso de estudio fue la redacción de un informe técnico para presentarse a los especialistas en el área epidemiológica. Esto fue importante ya que permitió redactar posteriormente un artículo de investigación que fue sometido a una revista indexada.

Cabe mencionar que, al ser un grupo de trabajo dedicado a la investigación, la etapa de implementación se suele abordar de esta forma, ya que en el sector industrial, por ejemplo, se suele externar con algún prototipo o un producto que pueda ser probado por agentes externos.

En artículos como [34], las técnicas de visualización reflejan de manera clara lo que se pretende con una buena implementación de las mismas, ya que posibilitan comunicar la Ciencia de Datos detrás del COVID-19 a una audiencia amplia, facilitando la comprensión de diversos aspectos de la pandemia mediante gráficas ilustrativas.

Otro ejemplo de implementación es el trabajo de [35], en el que, analizando los datos, se destaca cómo la implementación de intervenciones no farmacológicas (NPIs) en Yucatán, México, logró mantener bajos y estables los niveles de transmisión de SARS-CoV-2, incluso sin un programa de vacunación COVID-19.

5.10. Retroalimentación

Al llegar a esta etapa, fueron discutidas las áreas de mejora que se pudieron rescatar. El primer punto discutido fue la aplicación de esta metodología extendida a otra enfermedad epidemiológica, como lo puede ser el dengue o la hepatitis tipo C, enfermedades que son relativamente comunes en México. El segundo punto resultó ser la extensión de las cuatro etapas restantes de la metodología. Ya que el grupo de trabajo cuenta con una vasta experiencia en la etapa de modelado, su extensión tiene un panorama prometedor.

5.11. Discusión

De acuerdo con el estudio de [36], en abril de 2020, investigadores de salud pública sugirieron tres lecciones de la pandemia del VIH para COVID-19: anticipar desigualdades de salud, apoyar el cambio de comportamiento y fomentar esfuerzos multidisciplinarios. Después de dos años, revisan cómo se han manifestado estas lecciones, destacando la importancia del diálogo y la colaboración multidisciplinaria para enfrentar futuros brotes de enfermedades infecciosas.

El estudio de [37] analiza cinco olas de COVID-19 en México hasta agosto de 2022, registrando 3,396,375 casos confirmados. Se observó un aumento en detecciones tras la introducción de pruebas rápidas. Aunque el número de casos aumentó, las métricas de gravedad, como hospitalizaciones e intubaciones, disminuyeron con el tiempo. La tasa de mortalidad hospitalaria fluctuó, alcanzando su punto más alto durante la segunda ola. Factores como comorbilidades, edad y género masculino incrementaron el riesgo de enfermedad grave y muerte, pero este riesgo disminuyó en olas posteriores y con la vacunación.

Durante la pandemia de COVID-19, la tasa de suicidios en la Ciudad de México aumentó significativamente en 2020, especialmente de enero a junio. Aunque en 2021 las tasas volvieron a niveles anteriores, el estudio revela una prevalencia mayor en hombres y jóvenes, subrayando la necesidad de estrategias específicas de prevención y cuidado ante esta problemática multifactorial [38].

Similar a este trabajo, en el estudio de [39] se aplica una técnica de agrupamiento para clasificar grupos de datos de COVID-19 en estados y territorios de la India basándose en su alta similitud. Los resultados optimizan las técnicas de monitoreo, apoyando a gobiernos y profesionales de la salud en la mejora de políticas y tratamientos para reducir el número de infectados y fallecidos.

Por otro lado, en el sureste de México, la pandemia de COVID-19 impactó gravemente el sector alimentario, en una región maya caracterizada por altos niveles de pobreza y malnutrición [40]. Tanto este estudio, como las observaciones de los párrafos anteriores, resultaron de la aplicación de alguna metodología de Ciencia de Datos, ya sea de forma explícita o implícita.

Capítulo 6

Conclusiones

Mediante esta investigación de maestría, se muestra que es posible extender la metodología de Ciencia de Datos *Batch* FMDS en la fase de preparación de datos con buenos resultados. La principal contribución de este trabajo es la incorporación de nueve tareas que no estaban definidas previamente en la metodología *Batch* FMDS. Dichas tareas sirven como una guía más detallada para el desarrollo de aplicaciones de Ciencia de Datos en el área de epidemiología. Para validar esta propuesta, se desarrolló un caso práctico sobre el análisis de datos de mortalidad durante la pandemia de COVID-19 en México.

Las tareas que fueron desarrolladas para la extensión de la metodología estuvieron orientadas a un dominio epidemiológico, dominio sobre el cual no existe una metodología de Ciencia de Datos establecida con la suficiente especificidad para su aplicación directa en la solución de algún problema. Esta orientación contribuye a su posible uso en caso de futuras epidemias en México, lo que aportaría conocimiento al área respectiva y a la toma de decisiones.

Dentro de la extensión, las tareas definidas en la metodología propuesta van desde la consulta exclusiva de fuentes de datos de dependencias oficiales hasta la generación de un almacén de datos con las características suficientes para proceder con las etapas posteriores de manera exitosa.

Las tareas propuestas para cada etapa se resumen de la siguiente forma:

1. Requerimientos de los datos: Identificación de fuentes oficiales.

2. Recopilación de datos: Obtención de los datos.
3. Entendimiento de los datos: Análisis de diccionarios de datos, Análisis exploratorio de los datos y Definición de filtros.
4. Preparación de datos: Limpieza de datos, Aplicación de filtro, Creación de atributos de apoyo e Integración de datos preprocesados en un almacén de datos.

Al extender la metodología con tareas específicas en las etapas de interés, se apuntó a la realización de un caso práctico con el uso de datos de una enfermedad transmisible; dicha enfermedad fue la que protagonizó la pandemia iniciada en el año 2020 en México, el COVID-19.

Se conoce que existen estudios en los cuales se analizan datos poblacionales relacionados con la pandemia por COVID-19, sin embargo, presentan una ventana de tiempo reducida. En contraste, en este trabajo se analizan los datos generados de manera oficial durante un periodo más amplio de la pandemia.

Los principales hallazgos obtenidos durante el caso práctico fueron que se identificaron regiones de municipios/alcaldías con altas tasas de mortalidad. Por ejemplo, se encontró que un grupo de nueve alcaldías en el norte de la Ciudad de México tuvo la tasa promedio de mortalidad más alta del país, con 694 muertes por cada 100 mil habitantes. Al ubicar geográficamente a dichas alcaldías, se observó que se encuentran de forma muy marcada en el norte de la Ciudad de México, lo que puede sugerir el análisis de una correlación geográfica en futuros estudios. En contraste, el grupo de municipios con la menor tasa de mortalidad se encontró en el estado de Chiapas.

Se observó que al aplicar la primera propuesta a un caso práctico, se contribuyó a una sistematización a lo largo de todo el proceso. Tareas relevantes como el uso exclusivo de fuentes de datos oficiales o el uso de identificadores de apoyo propios del ámbito epidemiológico contribuyen a que esta pueda ser una primera metodología de Ciencia de Datos para este dominio. Es importante destacar que las actividades realizadas durante el desarrollo de esta investigación fueron supervisadas y validadas por investigadores expertos en el campo de la epidemiología en México.

Finalmente, el presente trabajo puede ser continuado en las siguientes vertientes: a) aplicaciones a otros casos prácticos de otras enfermedades epidemiológicas; b) extensión de otras etapas de la metodología.

Referencias

[1] A. Althniana, A. Elwafab, N. Aloboudc, H. Alrasheeda, and H. Kurdi. “Prediction of COVID-19 individual susceptibility using demographic data: A case study on Saudi Arabia,” *Procedia Computer Science*, vol. 117, pp. 379–386, 2020.

[2] O. Yaxmehen, N. Antonio, S. Valdés, C. A. Fermín, L. Fernández, and A. Vargas. “Effectiveness of a nationwide COVID-19 vaccination program in Mexico against symptomatic COVID-19, hospitalizations, and death: a retrospective analysis of national surveillance data,” *International Journal of Infectious Diseases*, vol. 129, pp. 188–196, 2023.

[3] J. Pamplona da Costa, A. L. Sica de Campos, P. Cintra, L. Greco, and J. Hendrik. “The nature of rapid response to COVID-19 in Latin America: an examination of Argentina, Brazil, Chile, Colombia and Mexico,” *Online Information Review*, vol. 45, pp. 729–750, 2021.

[4] V. Suárez, M. Suarez, S. Oros, and E. Ronquillo. “Epidemiology of COVID-19 in Mexico: From the 27th of February to the 30th of April 2020,” *Revista Clínica Española*, vol. 220, pp. 463–471, 2020.

[5] L. Sánchez, “Desarrollo de una aplicación de Ciencia de Datos,” M.S. thesis, Dept. Comput. Sci., Centro Nacional de Investigación y Desarrollo Tecnológico, Morelos, México, 2018.

[6] G. Martínez, “Aplicación de Ciencia de Datos para el análisis de datos de mortalidad por COVID-19 de México” M.S. thesis, Dept. Comput. Sci., Centro Nacional de Investigación y Desarrollo Tecnológico, Morelos, México, 2022.

[7] J. Pérez, E. Iturbide, V. Olivares, M. Hidalgo, A. Martínez, and N. Almanza. “A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases,” *Systems-Level Quality Improvement*, vol. 39, pp. 1-6, 2015.

[8] A. Vega, N. N. Almanza, K. Torres, J. Pérez, and I. Barahona. “Correlation between mobility in mass transport and mortality due to COVID-19: A comparison of Mexico City, New York, and Madrid from a data science perspective,” *PLoS ONE*, vol. 17, pp. 1-14, 2020.

[9] J. Pérez, N. N. Almanza, K. Torres, G. Martínez, J. C. Zavala, and R. Pazos. “Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico,” *Mathematics*, vol. 10, pp. 1-16, 2022.

[10] S. Zhang, C. Zhang, and Q. Yang. “Data Preparation for Data Mining,” *Applied Artificial Intelligence*, vol. 17, pp. 375–381, 2003.

[11] F. Pasha, A. Lundeen, D. Yeasmin, and M. Fayzul. “An analysis to identify the important variables for the spread of COVID-19 using numerical techniques and data science,” *Case Studies in Chemical and Environmental Engineering*, vol. 3, pp. 1-7, 2020.

[12] W. Douglass, T. L. Scherer, and E. Gartzke. “The Data Science of COVID-19 Spread: Some Troubling Current and Future Trends,” *Peace Economics, Peace Science and Public*, vol. 26, pp. 1-7, 2020.

[13] A. Fadli, A. W. Widhi, M. Syaiful, A. Taryana, Y. I. Kurniawan, and W. H. Purnomo. "Simple Correlation Between Weather and COVID-19 Pandemic Using Data Mining Algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 982, pp. 1-10, 2020.

[14] J. François and A. Pérez. "Spatiotemporal patterns of the COVID-19 epidemic in Mexico at the municipality level," *PeerJ*, vol. 9, pp. 1-6, 2021.

[15] N. Antonio, O. Bello, J. Pisanty, A. Gonzalez, A. Vargas, S. Barquera, L. M. Gutiérrez, and J. Seiglie. "Socio-demographic inequalities and excess non-COVID-19 mortality during the COVID-19 pandemic: a data-driven analysis of 1,069,174 death certificates in Mexico," *International Journal of Epidemiology*, vol. 51, pp. 1711-1721, 2022.

[16] F. Ruiz, Y. Hernandez, J. Perez, J. Ortiz, and S. Saenz. "Systematic Review of Methodologies in Data Science," *2021 Mexican International Conference on Computer Science (ENC)*, vol. 1, pp. 1-6, 2021.

[17] A. Waller and E. Fawcett. "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management," *Journal of Business Logistics*, vol. 34, pp. 77–84, 2013.

[18] N. Theoren, T. Mawela, and H. Twinomurinzi. "Transdisciplinary teaching practices for data science education: A comprehensive framework for integrating disciplines," *Social Sciences and Humanities*, vol. 8, pp. 20-28, 2023.

[19] F. Provost and T. Fawcett. "Data Science and it´s Relationship to Big Data and Data-Driven Decision Making," *Mary Ann Liebert, Inc*, vol. 1, pp. 51–59, 2015.

[20] V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, and R. Chadaga. "A machine learning and explainable artificial intelligence triage-prediction system for COVID-19," *Decision Analytics Journal*, vol. 7, pp. 39-46, 2023.

[21] Gobierno de México. "Se declara como emergencia sanitaria la epidemia generada por COVID-19." https://www.gob.mx/cjef/documentos/se-declara-como-emergencia-sanitaria-la-epidemia-generada-por-covid-19?idiom=es. (Consultada el 27 de noviembre de 2023)

[22] I. Ibarra, J. A. Cardenas, R. E. Ruiz, and G. Salazar. "Mexico and the COVID-19 Response," *Disaster Medicine and Public Health Preparedness*, vol. 14, pp. 9-17, 2020.

[23] S. Callaghan. "COVID-19 Is a Data Science Issue," *Patterns*, vol. 1, pp. 1–3, 2020.

[24] J. B. Rollins. "Foundational Methodology for Data Science," *IBM Analytics*, 2015.

[25] Microsoft. "What is the Team Data Science Process?" https://learn.microsoft.com/en-us/azure/ architecture/data-science-process/overview. (Consultada el 2 de marzo de 2024)

[26] INEGI. "Instituto Nacional de Estadística y Geografía." https://www.gob.mx/salud/ documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia. (Consultada el 30 de abril de 2023)

[27] CEMECE. "Centro Mexicano para la Clasificación de Enfermedades." https://www.gob.mx /salud/acciones-y-programas/menu-clasificacion-de-enfermedades-dgis?state=published. (Consultada el 30 de abril de 2023)

[28] CONEVAL. "Consejo Nacional de Evaluación de la Política de Desarrollo Social." https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipio-2010-2020.aspx. (Consultada el 7 de marzo de 2023)

[29] SNIM. “Sistema Nacional de Información Municipal.” http://snim.rami.gob.mx/. (Consultada el 7 de marzo de 2023)

[30] J. Pérez, A. Vega, N. N. Almanza, R. A. Pazos, J. C. Zavala, J. M. Rodríguez, and Y. Hernández. “Prediction of Diabetes Mortality in Mexico City Applying Data Science,” Int. *Workshop Artif. Intell. Pattern Recognit*, vol. 1, pp. 221–218, 2021.

[31] J. Sheng, J. Amankwah, Z. Khan, and X. Wang. “COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions,” *British Journal of Management*, vol. 0, pp. 1–20, 2020.

[32] DGIS. “Dirección General de Información Sanitaria.” http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_defunciones_gobmx.html. (Consultada el 30 de abril de 2023)

[33] J. Pérez, R. Pazos, V. Olivares, M. Hidalgo, J. Ruiz, A. Martínez, N. Almanza, and M. Gonzáles. “Optimization of the K-means Algorithm for the Solution of High Dimensional Instance,” *AIP Conference Proceedings*, vol. 1738, pp. 301-310, 2015.

[34] L. D. Comba. “Data Visualization for the Understanding of COVID-19,” *California Institute of Technology*, vol. 1, pp. 81–86, 2020.

[35] G. Ayora, P. Granja, M. Sauri, C. I. Hernandez, I. P. Hennessee, I. Lopez, G. Barrera, A. Che, P. Manrique, J. Clennon, H. Gomez, and G. Vazquez. “Impact of layered non-pharmacological interventions on COVID-19 transmission dynamics in Yucatan, Mexico,” *Preventive Medicine Reports*, vol. 28, pp. 1-12, 2022.

[36] D. Auerbach, D. Forsyth, C. Davey, and R. Hargreaves. “Living with COVID-19 and preparing for future pandemics: revisiting lessons from the HIV pandemic,” *Lancet HIV*, vol. 10, pp. 62–68, 2022.

[37] I. J. Ascencio, O. D. Ovalle, R. A. Rascón, V. H. Borja, and G. Chowell. “Comparative epidemiology of five waves of COVID-19 in Mexico, March 2020–August 2022,” *BMC Infectious Diseases*, vol. 22, pp. 72-81, 2022.

[38] F. García, H. Tendilla, F. Flores, L. Carbajal, R. Mendoza, L. Gomez, A. Vazquez, F. De la Cruz, A. Genis, and G. Flores. “Increased suicide rates in Mexico City during the COVID-19 pandemic outbreak: An analysis spanning from 2016 to 2021,” *Heliyon*, vol. 9, 2023.

[39] S. Kumar. “Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis,” *Annals of Data Science*, vol. 7, pp. 417–425, 2020.

[40] A. Nadal and D. Austreberta. “COVID-19: Solidarity initiatives for food security in the Mayan indigenous region of south-southeast Mexico,” *Global Food Security*, vol. 37, pp. 91-97, 2023.