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 mortalidad por COVID-19 de México

presentada por

Ing. Gerardo Alfonso Martínez González

como requisito para la obtención del grado de

Maestro en Ciencias Computacionales

Director de tesis

Dr. Joaquín Pérez Ortega

Codirectora de tesis

Dra. María Yasmín Hernández Pérez

Cuernavaca, Morelos, México. Noviembre de 2022.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 18/octubre/2022

OFICIO No. DCC/084/2022
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del C. Gerardo Alfonso Martínez González, con número de control M20CE064, 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 MORTALIDAD POR COVID-19 DE 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.

DR. JOAQUÍN PÉREZ ORTEGA
Director de tesis

DR. JOSÉ MARÍA RODRÍGUEZ LELIS
Revisor 1

DR. JAVIER ORTIZ HERNÁNDEZ
Revisor 2

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
27 OCT 2022
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SERVICIOS ESCOLARES
RECIBIDO
EBN

SEP TecNM CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
Depto. Servicios Escolares
Expediente / Estudiante
Iniciales JD/secretaria
RECIBIDO
SUBDIRECCIÓN ACADÉMICA

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

2022 Ricardo Flores Magón
Año de

Cuernavaca, Mor., 31/octubre/2022
No. De Oficio: SAC/155/2022
Asunto: Autorización de impresión de tesis

GERARDO ALFONSO MARTÍNEZ GONZÁLEZ
CANDIDATO AL GRADO DE MAESTRO(A) 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 MORTALIDAD POR COVID-19 DE 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®
"Educación Tecnológica al Servicio de México"

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
03 ~~OCT~~ NOV 2022
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SERVICIOS ESCOLARES
RECIBIDO
EBN

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

CMAZ/CHG

Dedicatoria

Dedico este trabajo de tesis a la hermosa familia que Dios me ha dado, a ustedes que siempre me han apoyado y han estado ahí cada momento.

A mis maravillosos padres Víctor Manuel Martínez Marentes y María del Socorro González Castañeda, quienes siempre han trabajado arduamente para guiarnos, apoyarnos, alentarnos y amarnos a mis hermanos y a mí en todo momento. Gracias a ustedes llegué al mundo y gracias a ustedes tuve todo lo necesario.

A mis extraordinarios hermanos Víctor Manuel Martínez González y Jessica Alejandra Martínez González quienes siempre han estado en todo momento para apoyarme y cuidarme. Les agradezco por estar siempre presentes en mi vida aportando buenas enseñanzas.

No encuentro las palabras necesarias para decirles cuanto los amo. Son lo que más valoro en esta vida. Gracias por darme tanto de cada uno.

Agradecimientos

A Dios por brindarme sabiduría y porque ha sido mi guía y fortaleza en cada momento.

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

Al Dr. Joaquín Pérez Ortega, por confiar en mí al seleccionarme para ser su asesorado y quien siempre me brindo su apoyo en innumerables veces y gracias a su amplia experiencia y conocimiento dirigió este trabajo de investigación. Simplemente mi más sincero agradecimiento por el aprendizaje brindado, su tiempo y paciencia.

A la Dra. María Yasmín Hernández Pérez por sus consejos y observaciones brindadas en el desarrollo de este trabajo de investigación.

Al comité tutorial: Dr. José María Rodríguez Lelis y Dr. Javier Ortiz Hernández por el tiempo dedicado y contribuir en la revisión de este trabajo. A la Dra. Leticia Sánchez Lima por su apoyo en todo momento, sus enseñanzas y por su ayuda en la revisión en la redacción de este documento.

A Martha Aura Cid Urbiola por estar siempre apoyándome en cada momento.

A mis compañeros y amigos de la línea de investigación; Paco, Carlos, Sandra, Nancy, Gilberto, Andrea.

No tengo palabras para agradecer todas sus muestras de apoyo a cada una de las personas que han contribuido a lo largo de mi vida, formación personal y profesional

Resumen

La Ciencia de Datos ha mostrado ser una herramienta de apoyo para la toma de decisiones en diferentes áreas del conocimiento, siendo la epidemiología una de estas. Sin embargo, debido a que la Ciencia de Datos es un área emergente, aún tiene limitaciones en sus metodologías publicadas, porque son de propósito general y cada investigación requiere condiciones específicas. *BATCH FMDS* es una metodología de Ciencia de Datos de *IBM* orientada al dominio epidemiológico que promete resultados favorables, sin embargo, es necesario un caso práctico con datos reales y con la participación de expertos en el dominio epidemiológico y en ciencia de datos para interpretar y contribuir a la evaluación de la *BFMDS*. Se presenta la selección e implementación computacional de un caso práctico usando BFMDS. El caso práctico consiste en el análisis de datos de mortalidad por COVID-19 de México, a nivel municipal para el año 2020. Se busca contestar la pregunta de investigación ¿Cuáles factores sociodemográficos tienen en común los municipios con tasas de mortalidad por COVID-19 similares? Es destacable que en las fases de análisis de resultados participaron expertos en la interpretación y validación de los resultados. Desde el punto de vista epidemiológico y como resultado del caso práctico se encontraron que los valores de indicadores de densidad poblacional y porcentaje de personas en situación de pobreza tenían una alta correlación con los valores de la tasa de mortalidad por COVID-19. Desde el punto de vista computacional se observó que siguiendo los pasos indicados en la metodología *BFMDS* fue posible el desarrollo del caso práctico y dar respuesta a la pregunta de investigación. Los datos de entrada fueron datos poblacionales provenientes de instituciones oficiales de México.

Abstract

Data Science has proven to be a support tool for decision making in different areas of knowledge, epidemiology being one of these. However, since Data Science is an emerging area, it still has limitations in its published methodologies, because they are of a general purpose and each investigation requires specific conditions. BATCH FMDS is an IBM Data Science methodology oriented to the epidemiological domain that promises favorable results, however, a practical case with real data and the participation of experts in the epidemiological domain and data science is necessary to interpret and contribute to the BFMDS assessment. The selection and computational implementation of a practical case using BFMDS is presented. The practical case consists of the analysis of mortality data from COVID-19 in Mexico, at the municipal level for the year 2020. It seeks to answer the research question: What sociodemographic factors do municipalities with mortality rates from COVID-19 have in common? Similar? The results analysis phases involved experts in the interpretation and validation of the results. From the epidemiological point of view and as a result of the practical case, it was found that the values of the population density indicators and the percentage of people living in poverty had a high correlation with the values of the mortality rate from COVID-19. From the computational point of view, it was observed that following the steps indicated in the BFMDS methodology, it was possible to develop the practical case and answer the research question. The input data were population data from official institutions.

Contenido Pág.

Lista de Tablas **Pág.**

Lista de Figuras **Pág.**

Capítulo 1
Introducción

El misterio es la cosa más bonita que podemos experimentar. Es la fuente de todo arte y ciencia verdaderos.

Albert Einstein

1. Introducción

1.1 Contexto de la investigación

La Ciencia de Datos es una disciplina emergente de naturaleza multidisciplinaria que analiza grandes volúmenes de datos utilizando herramientas y técnicas para encontrar patrones, obtener información significativa y para ayudar a la toma de decisiones, Este análisis permite que los científicos de datos planteen y respondan a preguntas como “qué pasó”, “por qué pasó”, “qué pasará” y “qué se puede hacer con los resultados” [1]. Presenta limitaciones en metodologías propias para el desarrollo de proyectos de diferentes dominios, particularmente en el campo epidemiológico.

Como una disciplina emergente, la Ciencia de Datos ha cobrado relevancia en el campo epidemiológico, debido a su oportuna aplicación para el estudio y análisis de cantidades masivas de datos de enfermedades en el mundo. Un ejemplo de tales aplicaciones es la epidemia del COVID-19.

Existen metodologías de propósito general como la *Team Data Science Process* de *Microsoft* y *Foundational Methodology for Data Science* de *IBM*. Esta última cuenta con 10 etapas que guían de manera general como desarrollar un proyecto de Ciencia de Datos.

En el Centro Nacional de Investigación y Desarrollo Tecnológico se han hecho varios proyectos relacionados con la aplicación de Ciencia de Datos en el campo epidemiológico entre ellos una extensión de la metodología *FMDS* de *IBM* para el dominio epidemiológico. A dicha extensión se le denomina *BATCH FMDS (BFMDS)*. Es prometedora, sin embargo, se detectó la necesidad de mostrar que es factible la aplicación de la metodología *BFMDS* por medio de un caso práctico, que contribuya a la validación.

En la presente investigación, se desarrolló un caso práctico de Ciencia de Datos para el análisis de datos de mortalidad por COVID-19 de México. Se analizaron los datos de mortalidad por COVID-19 del año 2020, que se obtuvieron de fuentes oficiales como: Dirección General de Información en Salud (DGIS), Instituto Nacional de Estadística y Geografía (INEGI), Consejo Nacional de Población (CONAPO) Y Centro Mexicano para la Clasificación de Enfermedades (CEMECE). Se utilizó la metodología *BFMDS*. El enfoque

de la investigación fue descriptivo ya que se identificaron aquellos municipios con mayor tasa de mortalidad en México para el año 2020.

El caso práctico consistió en responder las preguntas de investigación: ¿Qué municipios tienen mayor tasa de mortalidad por COVID-19? y ¿Qué factores sociodemográficos tienen en común los municipios con tasas de mortalidad por COVID-19 similares? Se respondieron aplicando las diez etapas de la metodología *BFMDS* analizando los resultados para identificar aquellas regiones de México a nivel municipal con mayores tasas de mortalidad por COVID-19. Además, de identificar aquellos factores sociodemográficos que tenían relación con la tasa de mortalidad por COVID-19. Como resultado del caso práctico se pudo identificar los factores sociodemográficos como el porcentaje de personas en situación de pobreza y la densidad poblacional. Además de identificar que a menor cantidad de personas en situación de pobreza mayor era la tasa de mortalidad y viceversa a mayor cantidad de personas en situación de pobreza menor es la tasa de mortalidad por COVID-19.

1.2 Descripción del problema

La Ciencia de Datos es una disciplina emergente que se centra en obtener conocimiento de grandes volúmenes de datos para apoyar la toma de decisiones. Se sustenta en principios de diferentes disciplinas como las matemáticas, computación y estadística.

En el Centro Nacional de Investigación y Desarrollo Tecnológico se desarrolló una extensión de la metodología de Ciencia de Datos para el dominio epidemiológico llamada *BFMDS*. Sin embargo, a pesar de que ya se está utilizando se requiere contribuir a la validación de la misma aplicándola a un caso práctico que incorpore la opinión de expertos del área de epidemiología. Es de importancia mostrar la factibilidad de usar la metodología de Ciencia de Datos *BFMDS* mediante un caso práctico. El caso práctico consistió en identificar aquellos municipios con mayor tasa de mortalidad por COVID-19 utilizando agrupamiento (*clustering)* como método. La pregunta de investigación que se planteó fue la siguiente: ¿Qué factores sociodemográficos tienen en común los municipios con tasas de mortalidad por COVID-19 similares? Es importante mencionar, que México es el quinto país

con mayor número de muertes por COVID-19 en el mundo, después de Estados Unidos, Brasil, India y Rusia, con 324,334 muertes registradas a principios de mayo de 2022 [2].

Por esta razón, es importante mostrar que es factible la metodología *BFMDS* para el análisis de datos de mortalidad por COVID-19 de México, para apoyar a la toma de decisiones a las instituciones pertinentes. Además, se describirá la aplicación de cada una de las etapas de la metodología tomando en consideración la opinión de los expertos en el área epidemiológica, computación y Ciencia de Datos.

1.3 Objetivos

1.3.1 Objetivo General

Contribuir a la validación de la metodología *BFMDS* mediante el desarrollo de un caso práctico.

1.3.2 Objetivos Específicos

1. Desarrollar un caso práctico en el área de epidemiología.
2. Aplicar Ciencia de Datos para el análisis de mortalidad por COVID-19 a nivel municipal.
3. Identificar los municipios con mayor y menor tasa de mortalidad por COVID-19.
4. Identificar los índices sociodemográficos que tienen correlación con las altas tasas de mortalidad a nivel municipal.

1.4 Justificación

La metodología de Ciencia de Datos *BFMDS* es prometedora, sin embargo, debido a su reciente desarrollo se requiere validar la metodología aplicándola a un caso práctico.

Aplicar la metodología de Ciencia de Datos para su aplicación en casos reales dentro del dominio epidemiológico es de importancia por la gran cantidad de datos que se generan en este dominio. Tan solo en el año 2013 se generaron alrededor de 153 *exabytes* de datos en el dominio epidemiológico. Durante el 2020 se generaron aproximadamente 2,314 *exabytes* de nuevos datos [3]. Se espera mostrar la factibilidad de aplicar la metodología *BFMDS,* para apoyar al dominio epidemiológico y a las instituciones gubernamentales pertinentes a la toma de decisiones.

1.5 Alcances y limitaciones de la investigación

1.5.1 Alcances

a) Se aplicará la metodología de Ciencia de Datos *BATCH FMDS (BFMDS)* para el desarrollo de un caso práctico.
b) Se implementará computacionalmente.
c) El caso práctico se enfocará en los registros de mortalidad por COVID-19.
d) Se utilizarán bases de datos de fuentes oficiales como: DGIS, INEGI, CONAPO Y CEMECE, del año 2020.

1.5.2 Limitaciones

a) Todos los datos serán oficiales.
b) La validación de resultados será de manera experimental.
c) Se realizarán pruebas de la aplicación únicamente en equipo disponible en el CENIDET.

1.6 Organización del documento

El documento posee cinco capítulos. La tesis está organizada de la siguiente forma: el Capítulo 2, presenta el estado del arte, mediante el cual se presentan aquellos trabajos relacionados, tesis con relación en la investigación y una breve línea del tiempo sobre la Ciencia de Datos, Análisis de Datos y Metodologías. El Capítulo 3, integra las metodologías propuestas para desarrollar aplicaciones de Ciencia de Datos, se realiza una comparación entre ellas y se selecciona una para desarrollar la aplicación que nos permita analizar los datos de mortalidad por COVID-19 de México. En el Capítulo 4, se presenta la resolución del caso práctico utilizando la metodología de Ciencia de Datos seleccionada. En el Capítulo 5, se exponen las conclusiones derivadas del desarrollo de la investigación, así como las aportaciones. Además, se proponen trabajos futuros para dar continuidad al tema de investigación.

Capítulo 2
Antecedentes

Lo que sabemos es una gota de agua; lo que ignoramos es el océano
Isaac Newton

2. Antecedentes

A continuación, se muestran trabajos relacionados, realizados en el departamento de Ciencias Computacionales del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), como así también en otras instituciones. Estos trabajos relacionados hacen referencia a aplicaciones de Ciencia de Datos relacionadas con el dominio epidemiológico y en caso particular con el COVID-19 y en otros dominios. También se anexa en esta sección la definición de caso práctico, en conjunto con sus características.

2.1 Caso práctico

En [4] menciona que es un método de investigación que consiste en la descripción de una situación o problemática real que sucede en una organización, con el objetivo de examinar la problemática. También sirve para realizar un diagnóstico, presentar posibles soluciones fundamentadas con argumentos teóricos y prácticos.

Implementar un caso práctico ayuda a demostrar el análisis, conocimientos y disciplinas estudiadas para la solución de uno o más problemas. La evaluación del caso práctico es importante porque ayuda a evaluar ciertos aspectos como: presentación y contenido y evaluación de los resultados.

Para la elaboración del caso práctico en [4] menciona que requiere de una actualización, ayuda a tomar toda aquella información actual como marco conceptual y referencial. También, el caso práctico debe tener congruencia entre las secciones. Además, debe tener integridad en los puntos de vista como así también las posibles repercusiones sociales y ecológicas. También, debe ser didáctico con el fin de aportar a la enseñanza y aprendizaje de posibles interesados en el tema.

2.2 Trabajo relacionado

En la tesis “Desarrollo de una aplicación de Ciencia de Datos” [5] aborda la asimilación de conceptos de Ciencia de Datos y la creación de una infraestructura de conocimiento para aplicaciones en el campo de la Ciencia de Datos. Se desarrolló un caso práctico con el que se hizo la proyección de las tasas de mortalidad por diabetes mellitus tipos E11-E14, en algunas

regiones de México clasificadas como C24, C08 y C51 para conocer su crecimiento o disminución y la predicción para los próximos años. Se utilizó la metodología de la empresa *IBM*, denominada *Foundational Methodology for Data Science* (*FMDS*, por sus siglas en inglés), en el desarrollo de la implementación se utilizó el lenguaje de programación R. Los datos que se utilizaron fueron obtenidos de repositorios de instituciones oficiales.

En el artículo "Prediction of Diabetes Mortality in Mexico City Applying Data Science" [6], aborda el problema de la proyección de mortalidad en la Ciudad de México. Aplicando Ciencia de Datos, se utilizó la extensión de la metodología de IBM denominada *BFMDS* que está orientada al dominio epidemiológico. Dentro de las aportaciones de la investigación, se encuentran: el desarrollo de un almacén de datos para la diabetes a partir de datos oficiales; el uso de la metodología *BFMDS* que está orientada al dominio epidemiológico. La finalidad fue contribuir a la toma de decisiones de las autoridades de salud pública en México.

Cabe resaltar que debido a la contingencia que provocó la pandemia de COVID-19, se visualizó la relevancia de la Ciencia de Datos dentro de la epidemiología. En [7], se realiza un estudio sobre el estado del arte de los trabajos que toman el escenario pandémico actual. En estas investigaciones, se revisan las prácticas actualizadas resumiendo los principales desafíos en el campo. Además, en [8], se analizan las nuevas técnicas y enfoques de investigación basados en datos relacionados con la COVID-19 ya disponibles, con los cuales pueden ayudar implementando métodos como la inteligencia artificial y el aprendizaje máquina al proceso de facilitar la atención de los pacientes infectados.

En [9,10,11], se revisa el papel que juega la Ciencia de Datos en el dominio epidemiológico. En conjunto con disciplinas como el análisis estadístico, informática, la biología computacional, entre otras. Es posible identificar las diferentes formas para lograr aplicaciones epidemiológicas, descubrimiento de fármacos y el diseño molecular con fines diagnósticos y terapéuticos.

En [12,13], se analiza y se da una descripción general de algunas prácticas como adquisición de datos y manejo de datos dentro de la Ciencia de Datos, los cuales son importantes para el pronóstico de epidemias. Así también, identifica las características y problemas que llegan a tener los datos obtenidos por el COVID-19 y como afectan en el modelado y proyección de la epidemia. Además, proporciona nuevas perspectivas de la

Ciencia de Datos sobre los desafíos de la recopilación, conservación y validación de datos como así también las limitaciones de los modelos.

En [14], abordan la problemática de la propagación de COVID-19 en transporte público (metro) en las ciudades de México, Nueva York y Madrid desde una perspectiva de Ciencia de Datos. Para ello, aplican la metodología *BFMDS*, ampliaron la metodología en su primera etapa con una orientación al dominio epidemiológico en la cual se integran conceptos básicos de epidemiología fusionándose con la Ciencia de Datos. Su principal aportación, es evaluar la eficacia de las políticas públicas para mitigar la pandemia de COVID-19, utilizando la metodología de Ciencia de Datos para la obtención de relaciones entre movilidad del transporte público y la mortalidad por COVID-19.

En [15], propone una metodología para identificar a las personas infectadas por SARS-CoV-2(COVID-19). Mediante la utilización de imágenes de tomografía computarizada y rayos-x del tórax utilizando inteligencia artificial. En la primera etapa se redimensionan las imágenes del conjunto de datos, esto con el fin de tener rapidez en el procesamiento algorítmico. Posteriormente, las imágenes se pasan a formato RGB. La predicción de las imágenes está dividida en dos, entrenamiento y pruebas. Finalmente, se aplica el clasificador de árbol de decisión para predecir si la imagen es positiva a COVID-19 o no. El modelo recomendado para identificar casos positivos de COVID-19 tuvo una precisión del 93% en imágenes de tomografía computarizada, mientras que la precisión en imágenes de rayos-x de tórax obtuvo un 88% de precisión.

Es importante mencionar que también existen aplicaciones de Ciencia de Datos en diferentes dominios como, por ejemplo, ciberseguridad, accidentabilidad, finanzas entre otros. En el artículo “Data Science methodology for cybersecurity projects” [16] propone una metodología de Ciencia de Datos para cubrir y proporcionar los recursos adecuados para el análisis de grandes volúmenes de datos en proyectos de ciberseguridad. Se exponen metodologías como: *Knowledge Discovery in Databases* (*KDD*, por sus siglas en inglés), *Cross Industry Standard Process for Data Mining* (*CRISP-DM*, por sus siglas en inglés), *Team Data Science Process* (*TDSP*, por sus siglas en inglés) y *Foundational Methodology for Data Science* (*FMDS*, por sus siglas en inglés). Un proyecto de Ciencia de Datos en ciberseguridad requiere de cuatro pasos: 1. Definición del problema; 2. Recopilación de

información; 3. Análisis de datos recopilados; 4. Producción (implementación de los módulos relevantes y el sistema que ejecute todo el proceso de forma automática cuando sea necesario). Como resultado del análisis comparativo de las metodologías *KDD*, *CRISP-DM*, *FMDS* y *TDSP* se obtuvo que la metodología *FMDS* cubre con todos los atributos favorables. Además, es una metodología con enfoque general por lo que puede personalizarse para adaptarse a cualquier proyecto de ciberseguridad.

En [17], propone un modelo de Ciencia de Datos para la predicción de precios de acciones en el mercado de valores de Indonesia utilizando datos de *Yahoo finance*. El modelo se basa en computación estadística con lenguaje de programación R y memoria a largo plazo. Se utilizó la Ciencia de Datos para visualizar los datos y simular los precios importantes de variables como: apertura, máximo, mínimo y cierre con diferentes parámetros. El modelo resultó ser útil para predecir datos en corto plazo, con una exactitud del 94.57 %.

En [18], propone una metodología para estudiar datos abiertos sobre accidentes viales (en Medellín, Colombia) usando Ciencia de Datos. La metodología propuesta está compuesta por cuatro macroprocesos: planificación, preparación de datos, análisis automático y visualización de datos. Dentro de éstos, se realizan un total de 15 subetapas. Fueron integrados secuencialmente y automatizados en lenguaje R, bajo el entorno *RStudio*. Como resultado, se obtuvo un sistema por el cual se visualizan datos de eventos de accidentes mes a mes. Aplicaron un caso de uso, tomando como referencia los reportes de accidentes viales en 2016 en Medellín, se reportaron al mes, 56% de casos con personas heridas, 43.4% solo obtuvieron daños no físicos y 0.6% casos de muerte. Además, determinaron que dentro de las 7:00 – 7:59, 11:00 - 11:59, 12:00 - 12:59 y 17:00 - 17:59, fueron las horas con mayor índice de reportes de accidentes.

En [19], *IBM* propone una metodología general para proyectos de Ciencia de Datos con el fin de proporcionar una estrategia de orientación, que se independiente de las tecnologías, volúmenes de datos o los enfoques utilizados para resolver problemas y responder preguntas a través del análisis de datos.

En [20], *Microsoft* se proporciona una metodología ágil e iterativa para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. Ayuda a mejorar la colaboración y el aprendizaje en equipo al sugerir cómo los roles de equipo

funcionan mejor juntos. Incluye procedimientos recomendados y estructuras de *Microsoft* y otros lideres del sector para su implementación.

Los trabajos relacionados aquí descritos ayudan a identificar como están trabajando proyectos en epidemiologia, identificar que técnicas y enfoques están aplicando. Así también aquellas publicaciones que utilizan metodologías de Ciencia de Datos en otros dominios.

Los trabajos relacionados descritos ayudan a identificar como están trabajando proyectos en el área de epidemiología, identificar que técnicas y enfoques están aplicando. Así también aquellas publicaciones que utilizan metodologías de Ciencia de Datos en otros dominios Tabla 2.1.

Tabla 2.1 Relevancia de los trabajos relacionados

Artículo	Aplicación de metodologías de Ciencia de Datos	Aplicación y análisis	Análisis del estado del arte	Análisis de la importancia de la Ciencia de Datos
Prediction of Diabetes Mortality in Mexico City Applying Data Science [6]	x			
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 [14]	x			
Data Science methodology for cybersecurity projects [16]	x			
Metodología Fundamental para la Ciencia de Datos [19]	x			
Proceso de ciencia de datos en equipo [20]	x			
Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19 [15]		x		
Data Science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM) [17]		x		
Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín [18]		x		
Data science in unveiling COVID-19 pathogenesis and diagnosis: evolutionary origin to drug repurposing [7]			x	
Data Science Techniques for COVID-19 in Intensive Care Units [8]			x	
Methods, Challenges, and Practical Issues of COVID-19 Projection:A Data Science Perspective [12]			x	
COVID-19 Is a Data Science Issue			x	
Role of data science in managing COVID-19 pandemic [9]				x
Fast and Efficient Data Science Techniques for COVID-19 Group Testing [10]				x
On the Convergence of Epidemiology, Biostatistics, and Data Science [11]				x

2.3 Análisis de la evolución de la Ciencia de Datos

A continuación, se da un breve contexto sobre la historia consultada en [21,22,23]. Únicamente tomando como referencia las siguientes disciplinas: Ciencias Computacionales, Análisis de Datos y Metodologías que integran a lo que hoy se conoce como Ciencia de

Datos. Además, se realizó una línea del tiempo que contempla los acontecimientos destacados que dieron lugar al término Ciencia de Datos (Figura 2.1).

Hablando del Análisis de Datos fue en el año 1805 cuando Adrien Marie Legendre y Carl Friedrich Gauss aplicaron la regresión para determinar las órbitas de los cuerpos alrededor del sol. Es importante mencionarlo porque es uno de los métodos mayormente utilizados dentro de la Ciencia de Datos. En el año 1956 se resolvió el problema del camino más corto, mediante analítica computacional. Fue hasta el año 1961 que se acuña el término de *Data Analysis* (Análisis de Datos) por JW Turkey. Posteriormente, en el año 1962 cuando el mismo JW Turkey presente la idea de la evolución del análisis de datos; y fue hasta el año 1993 donde surge la versión 1.0 de R.

Dentro de Ciencias Computacionales, en el año 1936 cuando Alan Turing introdujo la idea de una máquina universal. Fue hasta el año 1962, que Kennet Iversion introdujo "Fundation *OLAP* (On-Line Analytical Processing)". En 1963 surge el término "Base de Datos". En el año 1974 cuando se menciona por primera vez el término "Ciencia de Datos" en el libro del científico danés, Peter Naur. Otro acontecimiento importante para el manejo de los datos es la llegada de Excel 1.0 en el año 1985, al igual que el surgimiento del término de "minería de datos" el cual surge en la comunidad de bases de datos en el año 1990. Posteriormente surge el lenguaje de programación Python diseñado por Guido Van Rossun en el año 1991.

Dentro de Metodologías, en el año de 1989 se presenta la metodología *Knowledge Discovery in Database* (*KDD*, por sus siglas en inglés); posteriormente surge la metodología SAS *SEMMA* (*Sample, Explore, Modify, Model, and Acces*) en 1996; fue en 1999 cuando surge la metodología *Cross Industry Standard Process for Data Mining* (*CRISP-DM*, por sus siglas en inglés). Fue hasta el año 2015 cuando *IBM* pública *Foundational Methodology for Data Science* (*FMDS*, por sus siglas en inglés); en 2016 surge la metodología de *Microsoft Team Data Science Process* (*TDSP*, por sus siglas en inglés).

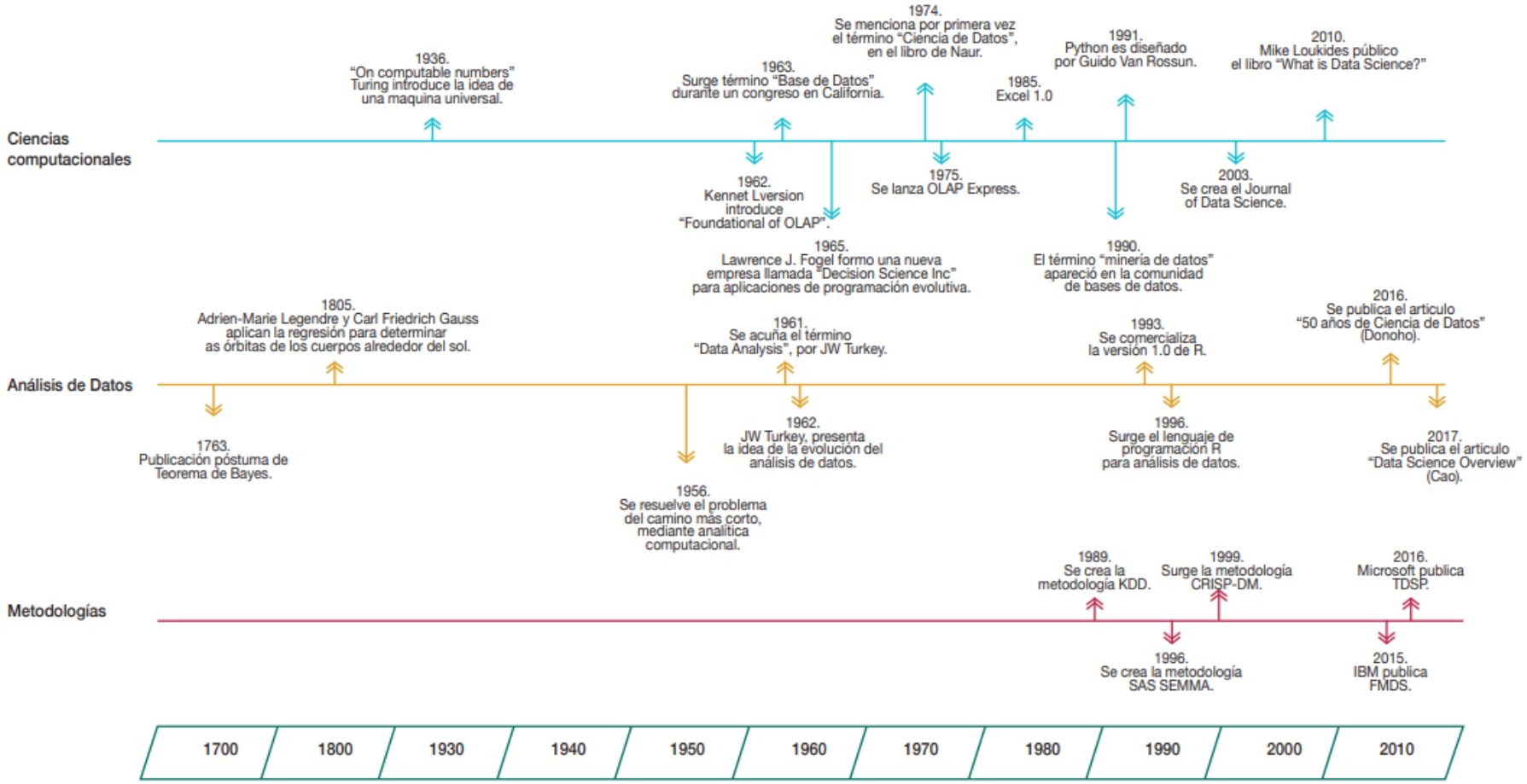

Figura 2.1 Evolución de la Ciencia de Datos

Capítulo 3
Metodologías de Ciencia de Datos

Un hombre sabio buscará más oportunidades de las que se le presentan.
Francis Bacon

3. Metodologías de Ciencia de Datos

En este capítulo se presentan diversas metodologías para apoyar el análisis de casos como la diabetes, COVID-19, entre otros con Ciencia de Datos. Se realizó una comparación entre las fases y actividades de la metodología *FMDS* y *BFMDS*.

Una metodología se define como: "Una estrategia general que guía los procesos y actividades dentro de un dominio dado" [5]. Se puede referir que una metodología de Ciencia de Datos son los métodos a seguir durante el ciclo de vida de un proyecto en particular. También se consideran como una estrategia general que sirve de guía para los procesos y actividades que están dentro de un dominio determinado. Se utilizan para obtener respuestas o resultados [19].

A continuación, se muestran las metodologías que se pudieron identificar dentro de la literatura son mayormente utilizadas en el dominio epidemiológico e incluso otros dominios. Es importante mencionar, que además de las metodologías que se identificaron en la literatura, se anexó la metodología de Ciencia de Datos de *Microsoft TDSP*.

3.1 Metodología Fundamental para la Ciencia de Datos MFCD

La metodología Fundamental para la Ciencia de Datos (*FMDS*, por sus siglas en inglés) de *IBM* consta de 10 etapas, (Figura 3.1). En [19], menciona que tiene similitudes con metodologías de minería de datos *CRISP-DM* [24], porque cuenta con etapas como la comprensión del negocio, comprensión de los datos, análisis de los datos, modelado y despliegue.

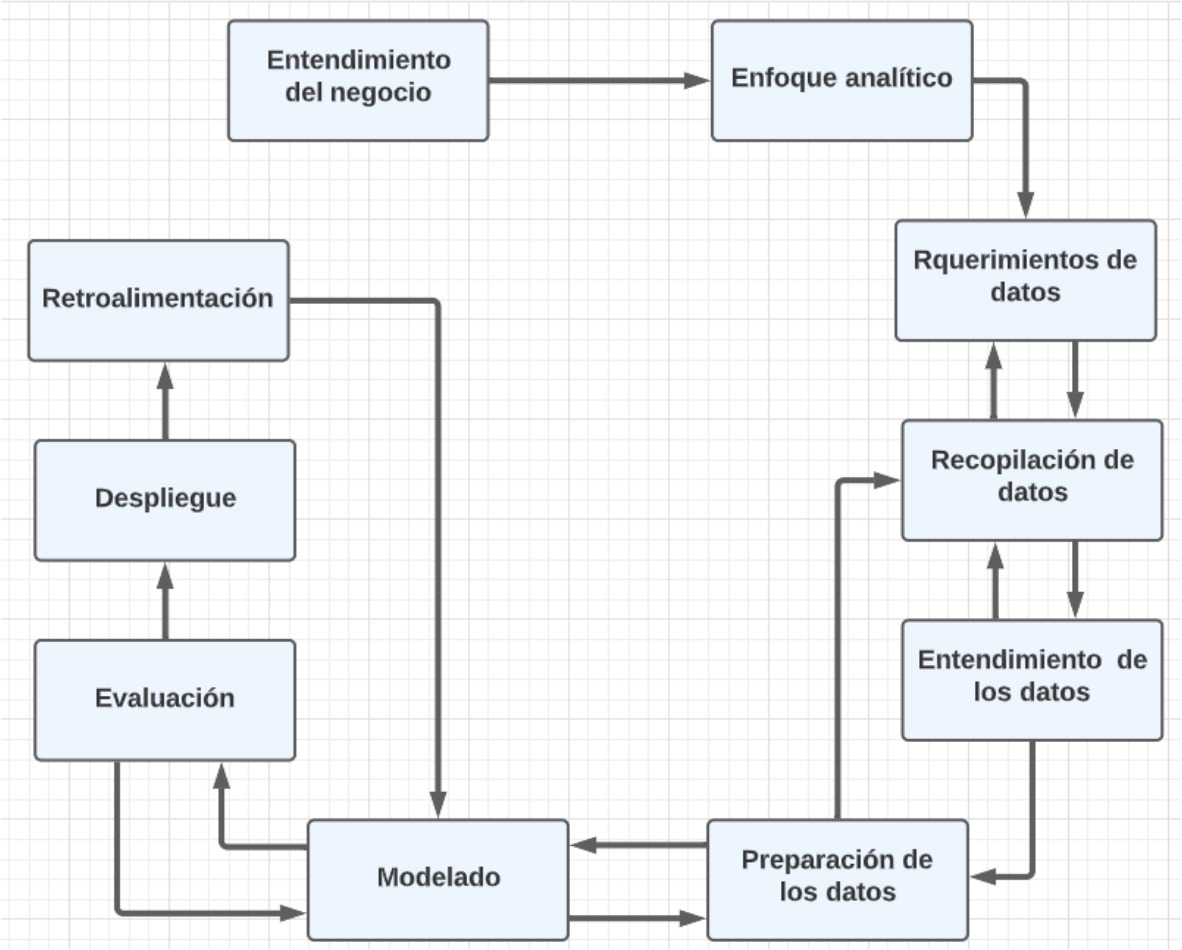

Figura 3.1 Metodología Fundacional para la Ciencia de Datos[19]

3.1.1 Entendimiento del negocio

Se define el problema, los objetivos del proyecto y los requisitos que debe tener la solución desde una perspectiva empresarial. Comienza con dedicar tiempo a buscar aclaraciones, para lograr lo que se puede denominar entendimiento del negocio. Esto permite determinar qué datos se utilizarán para responder la pregunta central.

En [19], menciona que se debe tener una pregunta de investigación claramente definida porque, en última instancia, dirige el enfoque analítico que será necesario para abordar la pregunta de investigación. Establecer una pregunta claramente definida comienza con la comprensión del objetivo de la persona que hace la pregunta. Se debe tener claro cuál

es la meta y descubrir los objetivos que apoyan la meta. De manera general, pretende aclarar los aspectos que ayuden a plantear el problema como los objetivos del proyecto desde una perspectiva empresarial.

3.1.2 Enfoque analítico

Se centra en definir el enfoque analítico que dará solución al problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático, con el objetivo de identificar las más adecuadas para el resultado deseado.

La selección del enfoque analítico correcto depende de la pregunta que se formule. Una vez que se establece una sólida comprensión de la pregunta, se puede seleccionar el enfoque analítico. Esto significa identificar qué tipo de patrones se necesitarán para abordar la pregunta de manera más efectiva. Si la pregunta es determinar las probabilidades de una acción, entonces podría usarse un modelo predictivo, si la pregunta es mostrar relaciones, tal vez se requiera un enfoque descriptivo.

El análisis estadístico se aplica a problemas que requieren conteos. Es decir, si la pregunta requiere una respuesta sí o no, entonces sería adecuado un enfoque de clasificación para predecir una respuesta. En el caso en que la pregunta sea aprender sobre el comportamiento humano, una respuesta adecuada sería utilizar enfoques de asociación o de agrupaciones.

3.1.3 Requisitos de datos

La selección del enfoque analítico determinará los requisitos de los datos. Estos suelen requerir determinados contenidos de datos, formatos y representaciones, orientados por el conocimiento en el dominio.

Si el problema que debe resolverse es la receta y los datos son un ingrediente, entonces el científico de datos debe identificar: cuales ingredientes se requieren, cómo obtenerlos o recolectarlos, cómo entenderlos o trabajarlos y cómo preparar los datos para

alcanzar el resultado deseado. Esto incluye identificar el contenido, los formatos y las fuentes de datos necesarios para la recopilación inicial de datos.

3.1.4 Recopilación de datos

Consta de identificar y reunir los recursos de datos (estructurados, no estructurados y semiestructurados) disponibles y relevantes para el dominio del problema. Esto incluye, de ser necesario, inversiones adicionales para la obtención de elementos informativos menos accesibles. La recopilación de datos requiere que se conozca la fuente de información o se tenga el conocimiento de dónde encontrar los elementos de datos que se necesitan.

Una vez realizada la recopilación inicial, el científico de datos realiza una evaluación para determinar si tienen o no lo que necesitan. En esta fase se revisan los requisitos de datos y se toman decisiones sobre si la recopilación requiere o no más o menos datos. Se pueden aplicar técnicas como la estadística descriptiva y la visualización para evaluar el contenido, la calidad y los conocimientos iniciales.

3.1.5 Comprensión de datos

Abarca todas las actividades relacionadas con la construcción del conjunto de datos, es decir, seleccionar los atributos necesarios que ayuden a responder la pregunta: ¿Los datos que recopiló son representativos del problema que se va a resolver?

3.1.6 Preparación de datos

Se enfoca en construir el conjunto de datos que se utilizará en la etapa de modelado. Las actividades que se deben realizar dentro de esta etapa son: limpieza, combinación de múltiples fuentes (archivos, tablas y plataformas) y transformarlos en variables útiles. La preparación suele ser el paso más largo de los proyectos de Ciencia de Datos.

La preparación de datos consta en eliminar elementos no deseados. Por lo general, [19] toma el setenta por ciento e incluso hasta el noventa por ciento del tiempo total del proyecto.

Específicamente, la etapa de preparación de datos de la metodología responde a la pregunta: ¿Cuáles son las formas en que se preparan los datos? Para trabajar eficazmente con los datos, debe estar preparado de manera que aborde los valores faltantes o no válidos y elimine los duplicados, para garantizar que todo tenga el formato correcto.

Si bien esta fase puede llevar un tiempo, si se hace correctamente, los resultados respaldarán el proyecto. Si se omite esto, entonces el resultado no estará a la altura. Es vital tomarse su tiempo en esta etapa y utilizar las herramientas disponibles para automatizar los pasos comunes para acelerar la preparación de datos.

3.1.7 Modelado

Se enfoca en desarrollar modelos predictivos o descriptivos según el enfoque analítico previamente definido. Es posible probar múltiples algoritmos con sus parámetros para identificar el mejor modelo para las variables disponibles. En esta etapa, el científico de datos jugará con diferentes algoritmos para asegurarse de que las variables en juego sean realmente necesarias, es la etapa en la metodología en la que el científico de datos tiene la oportunidad de probar y determinar si está funcionando el algoritmo seleccionado.

El modelado de datos se centra en el desarrollo de modelos descriptivos o predictivos. Uno de los ejemplos más claros son las encuestas online donde se identifica y mide el impacto de una opinión, actitud o comportamiento dentro de un grupo objetivo sobre un tema determinado. Un modelo predictivo intenta producir resultados de tipo sí / no o detener / continuar como por ejemplo la optimización de campañas de marketing para determinar las respuestas de los clientes a campañas de marketing o patrones de compa. Estos modelos se basan en el enfoque analítico que se adoptó, ya sea estadísticamente o basado en el aprendizaje automático.

Un conjunto de entrenamiento es un conjunto de datos históricos en los que ya se conocen los resultados. El conjunto de entrenamiento actúa como un medidor para determinar si es necesario calibrar el modelo.

El éxito de la recopilación, preparación y modelado de datos depende de la comprensión del problema y del enfoque analítico adecuado que se adopte. En la metodología descriptiva de ciencia de datos de *IBM* hecha por John Rollins [19], el marco está diseñado para hacer tres actividades: primero, comprender la pregunta en cuestión. En segundo lugar, seleccionar un enfoque o método analítico para resolver el problema y, en tercer lugar, obtener, comprender, preparar y modelar los datos.

3.1.8 Evaluación

Se enfoca en el cálculo de varias medidas de diagnóstico y de otros resultados, como tablas y gráficos para identificar la calidad y eficacia del modelo en la resolución del problema. En el caso de los modelos predictivos se utiliza un conjunto de pruebas que es utilizado para evaluar el modelo y para ajustarlo según las necesidades.

Las etapas de modelado y evaluación se realizan de forma iterativa, la evaluación del modelo se realiza durante el desarrollo del modelo y antes de que se implemente. La evaluación, permite evaluar la calidad del modelo, pero también es una oportunidad para ver si cumple con las condiciones iniciales.

La evaluación del modelo puede tener dos fases principales. La primera es la fase de medidas de diagnóstico, que se utiliza para garantizar que el modelo funcione como se esperaba. Si el modelo es un modelo predictivo, se puede utilizar un árbol de decisiones para evaluar si la respuesta que puede generar el modelo está alineada con el diseño inicial. Se puede usar para ver dónde hay áreas que requieren ajustes. Si el modelo es descriptivo, uno en el que se evalúan las relaciones, se puede aplicar un conjunto de pruebas con resultados conocidos y el modelo se puede refinar según sea necesario.

La segunda fase de evaluación que se puede utilizar es la prueba de significación estadística. Este tipo de evaluación se puede aplicar al modelo para garantizar que los datos se manejen e interpreten correctamente dentro del modelo. Esto está diseñado para evitar dudas innecesarias cuando se revela la respuesta.

3.1.9 Implementación

Una vez que el modelo haya sido satisfactorio, esta etapa se enfoca en implementarlo en el entorno de producción o en un entorno de pruebas comparable al real. A menudo se implementa de manera limitada para que su rendimiento sea evaluado.

3.1.10 Retroalimentación

Se recopilan los resultados del modelo implementado y con ello se obtiene retroalimentación sobre el rendimiento. Se monitorea el modelo implementado por un tiempo definido y se evalúan los resultados obtenidos. En el caso de no obtener cambios positivos, es necesario regresar a la etapa de modelado. El proyecto de Ciencia de Datos termina cuando se satisfacen las necesidades y objetivos de la organización.

3.2 Metodología *BFMDS*

En [25], propone una metodología derivada principalmente del análisis de la metodología de Ciencia de Datos de *IBM*, enfocada para su uso en el dominio epidemiológico llamada *BFMDS* (Figura 3.2). Su principal aportación se encuentra integrada por las etapas: entendimiento del negocio y enfoque analítico que son parte de la fase definición del negocio. Por esta razón se explicarán únicamente estas dos etapas ya que las siguientes etapas siguen el mismo proceso que la metodología Fundacional de Ciencia de Datos de *IBM*.

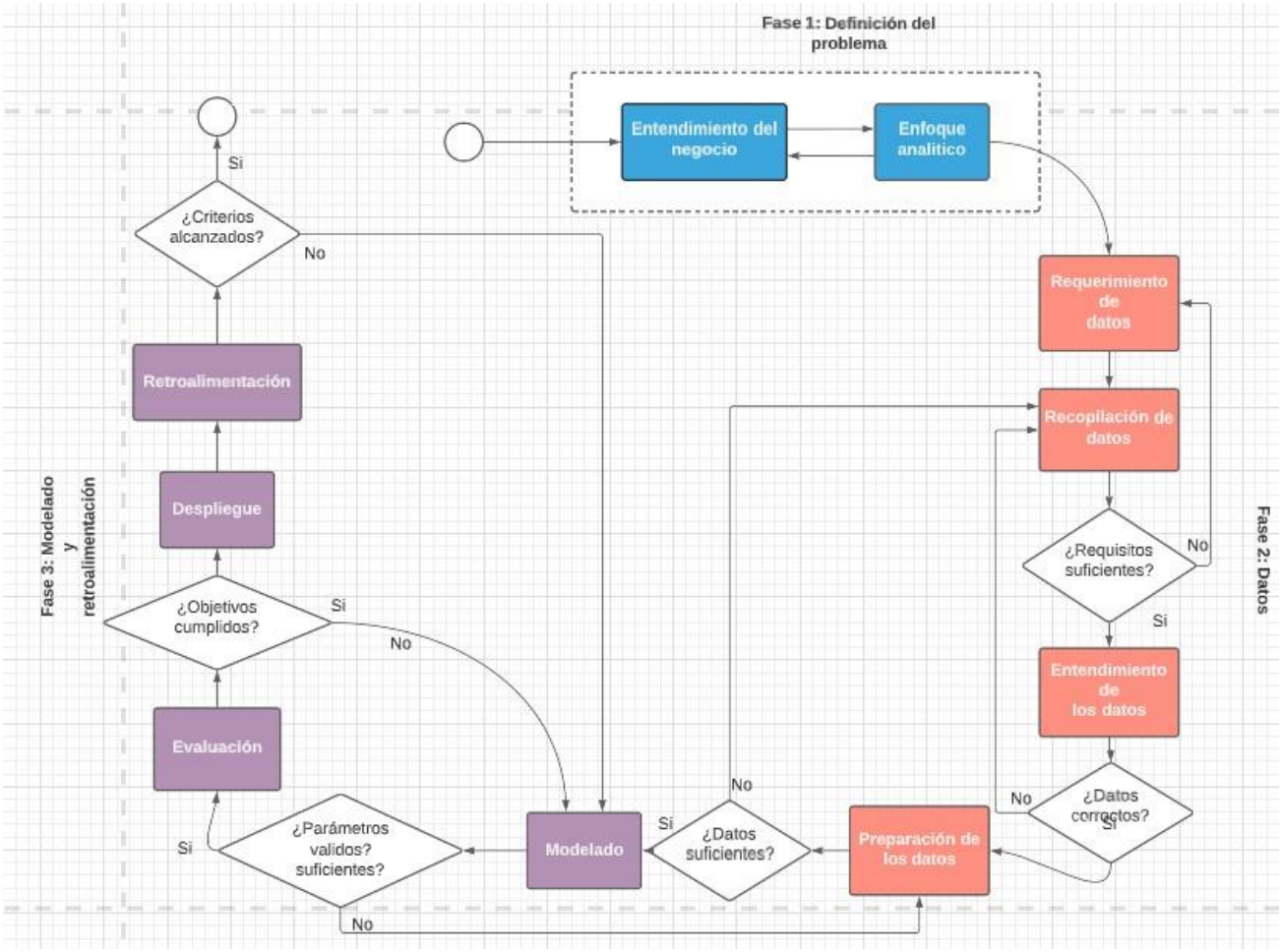

Figura 3.2 Metodología *BFMDS* [25]

Esta fase tiene como objetivo obtener la mayor información posible sobre el problema a resolver, determinar los datos que serán necesarios para alcanzar el objetivo del proyecto de Ciencia de Datos e identificar qué tipo de modelos serán utilizados para resolver el problema. La Figura 3.3, muestra las actividades que se realizan dentro de esta fase, las cuales se describen a continuación.

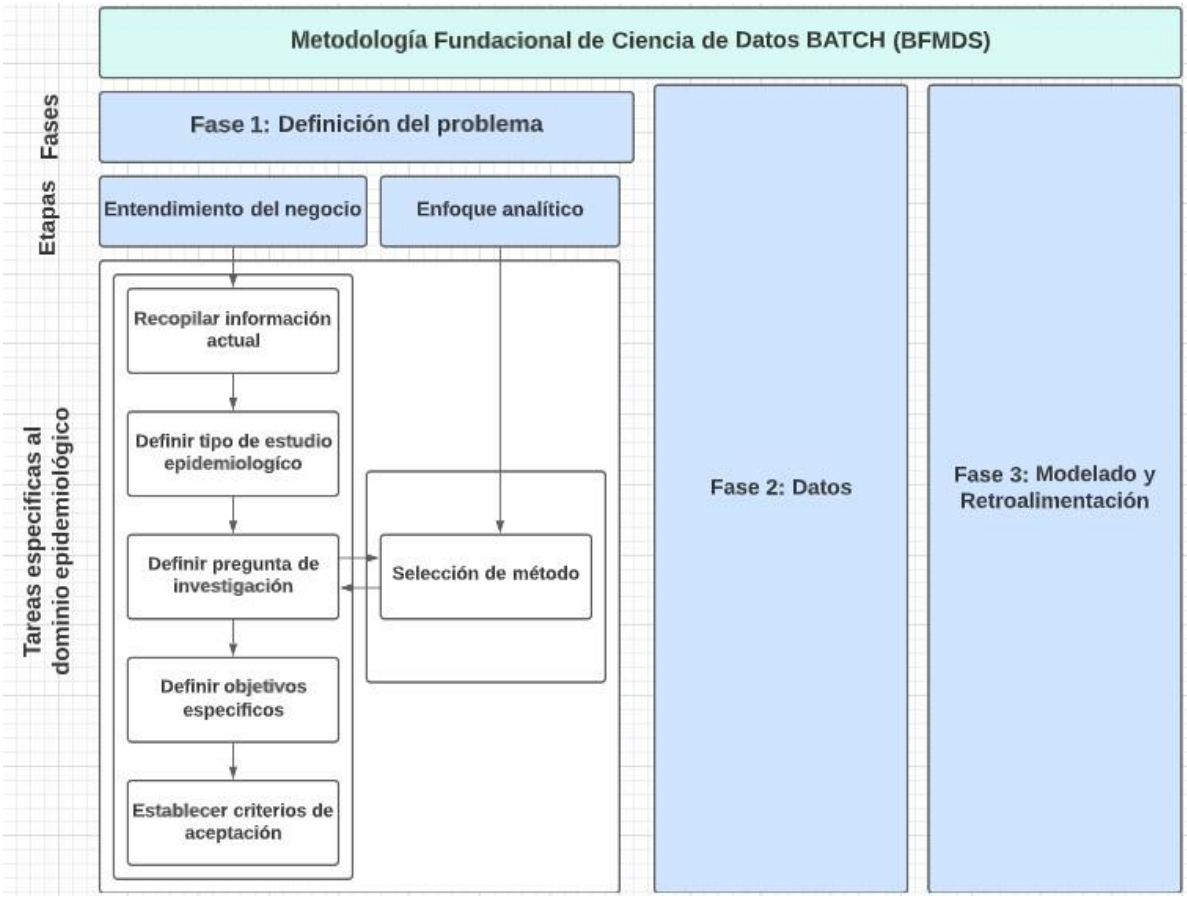

Figura 3.3 Actividades dentro de la etapa entendimiento del negocio [25]

3.2.1 Comprensión del negocio

Esta etapa tiene como objetivo proporcionar una base para definir el problema, objetivos y los requisitos de solución del proyecto. Esta etapa cuenta con cinco subetapas, en las cuales en su mayoría cuentan con tareas específicas dentro de ellas (Figura 3.4).

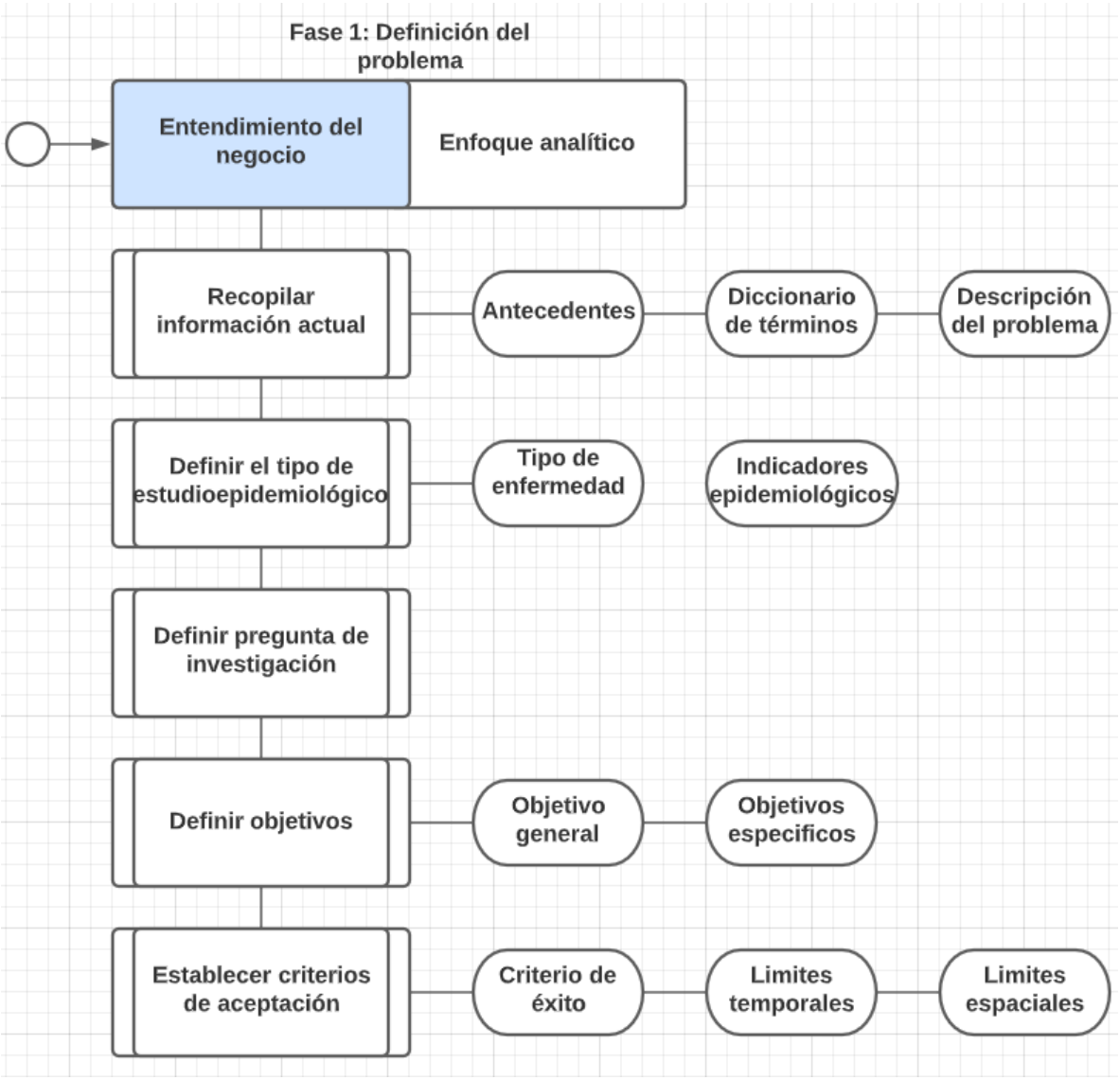

Figura 3.4 Actividades para la etapa de entendimiento del negocio [25].

3.2.1.1 Recopilar información

Antecedentes. Se centra en la comprensión de los antecedentes del proyecto, esto incluye detallar los recursos disponibles para llevarse a cabo como personal disponible, materiales, entre otros.

Diccionarios de términos. Consiste en crear un listado de la terminología propia del área epidemiológica relacionada con el proyecto.

Descripción del problema. Consiste en identificar las causas y consecuencias del problema.

3.2.1.2 Definir tipo de estudio epidemiológico

Tipo de enfermedad. Se centra en definir el tipo de enfermedad epidemiológica que se va a tratar. Existen dos clases principales de enfermedades epidemiológicas transmisibles causadas por un agente infeccioso y crónico degenerativo las que producen alteraciones fisiopatológicas afectando a los órganos y tejidos.

Tipo de estudio epidemiológico. Existen dos tipos de estudios epidemiológicos: analíticos o de observación, consisten en evaluar hipótesis entre posibles exposiciones a ciertos factores de riesgo y los descriptivos, hacen uso de los datos disponibles para examinar cómo las tasas varían de acuerdo con variables demográficas, como las obtenidas en los censos.

3.2.1.3 Definir pregunta de investigación

Consiste en definir de manera correcta la pregunta de investigación.

3.2.1.4 Delimitar objetivos

Objetivo general. Propósito que se quiere alcanzar con el proceso y precisa los resultados que desean ser obtenidos. Este objetivo debe de estar limitado por los recursos con los que se cuenta.
Objetivos específicos. Consisten en los pasos que se deben seguir para poder alcanzar el objetivo general.

3.2.1.5 Establecer criterios de aceptación

Criterios de éxito. Conjunto de principios por los que se puede determinar si se han conseguido los resultados deseados. Permitirán evaluar el resultado de la aplicación de la metodología y garantizar que este sirva como apoyo para la toma de decisiones basadas en él.

Límites temporales. Se encargan de definir el periodo de tiempo que se deberá tomar en cuenta para realizar el proyecto.
Límites espaciales. Definen la región o regiones que el proyecto debe abarcar.

3.2.2 Enfoque analítico

Teniendo el problema del negocio claramente establecido, es necesario determinar un enfoque analítico. Consiste en seleccionar el método o métodos que pueden dar respuesta a la pregunta de investigación antes definida. Esta etapa consta de una fase con actividades (Figura 3.5).

3.2.2.1 Selección del método

Seleccionar el método o combinación de métodos que se adapten para procesar los datos y obtener la información requerida para resolver la pregunta.
Descripción de la categoría. Estas categorías están relacionadas con las posibles preguntas de investigación. Cada una de ellas, tiene objetivos y elementos propios. Una definición de estos ayudará a la mejor comprensión y selección de método.
Descripción del método. Consiste en la descripción y análisis de dicho método.
Identificación de parámetros. Cada método necesita de ciertos parámetros específicos, de entrada y salida, para poder aplicarlo. Identificarlos permitirá una mejor selección de los datos necesarios para resolver el problema.
Variaciones del método. Existen variaciones y mejoras a muchos de los métodos presentados, las cuales generalmente presentan mejor calidad en los resultados o en el tiempo de procesamiento. Esto puede ayudar significativamente en la aplicación y resultados del proyecto.

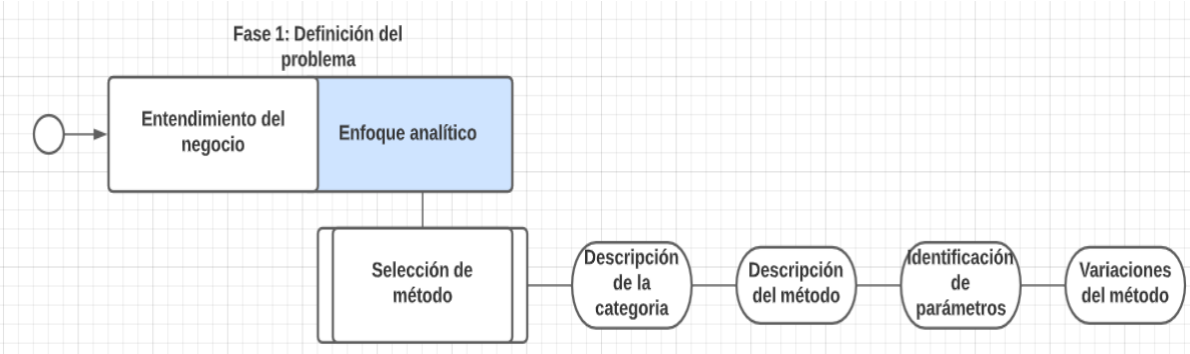

Figura 3.5 Actividades para la etapa de enfoque analítico [25].

3.3 Metodología: Proceso de Ciencia de Datos en Equipo

En [26], menciona que la metodología de Ciencia de Datos *TDSP* (por sus siglas en inglés) de *Microsoft* proporciona soluciones de análisis predictivo y aplicaciones inteligentes, mejora la colaboración y el aprendizaje en equipo al sugerir como los roles de equipo funcionan de una mejor manera. *TDSP* tiene procedimientos recomendados y estructurados recomendados por *Microsoft* y otros líderes en tecnología (Figura 3.6).

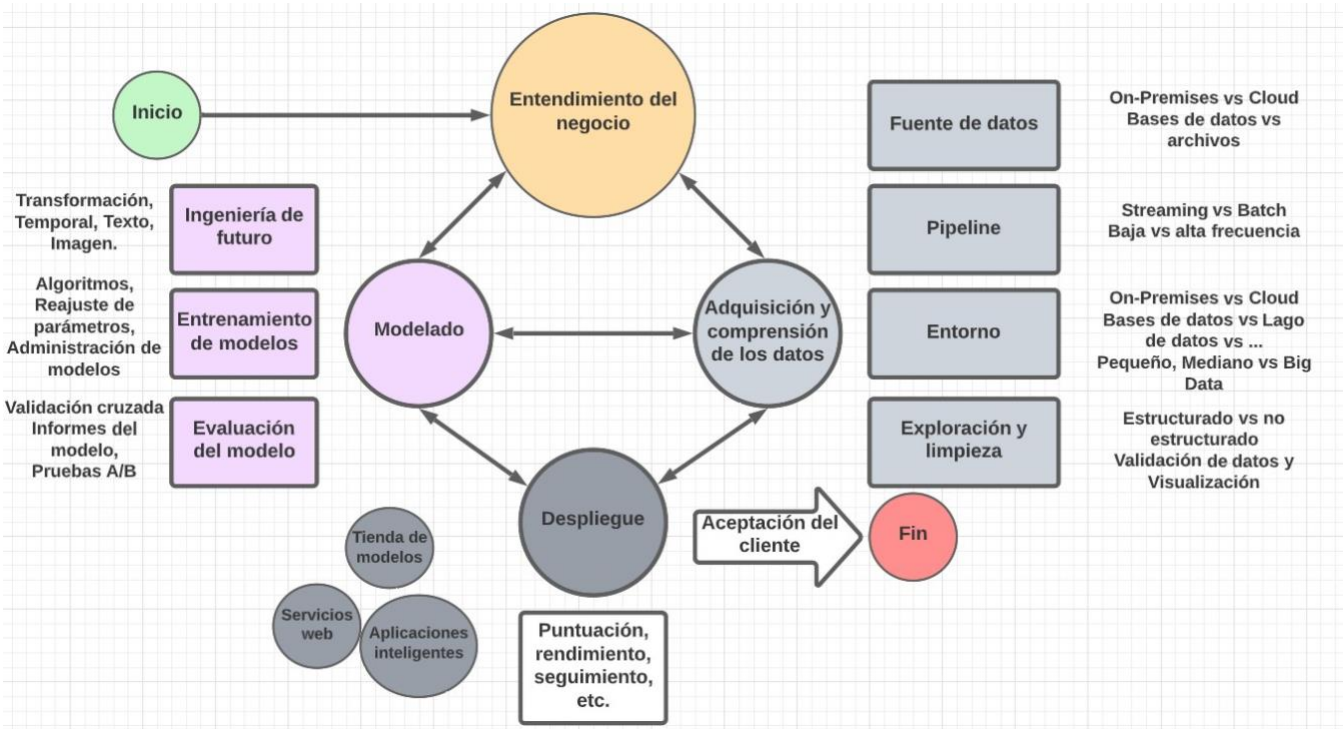

Figura 3.6 Metodología TDSP [20]

La metodología *TDSP* cuenta con cuatro etapas:

3.3.1 Entendimiento del negocio

Se centra en dos tareas: la primera, definición de objetivos que tiene como objetivo trabajar en conjunto con el cliente y las partes interesadas para poder formular preguntas que definan los objetivos empresariales y a las que se puedan aplicar técnicas de Ciencia de Datos; y la segunda, identificar los orígenes de los datos necesarios que ayuden a responder a las preguntas que se definieron anteriormente.

3.3.2 Adquisición y comprensión de los datos

Se enfoca en generar un conjunto de datos limpio y de alta calidad con las variables adecuadas para desarrollar una arquitectura de solución de canalización de datos que actualice y puntúe los datos con regularidad.

3.3.3 Modelado

Se enfoca en determinar las características óptimas de los datos que se ingresarán al modelo de aprendizaje automático. Además, crea un modelo de aprendizaje automático que predice el objetivo con máxima precisión y por último crea un modelo adecuado de aprendizaje automático para entornos de producción.

3.3.4 Despliegue

Se centra en implementar los modelos en un entorno parecido al real, para que el usuario final lo acepte. Por último, se tiene una aceptación del cliente, que consiste en implementar el modelo en un entorno de producción y cumpla con los objetivos del cliente.

3.4 Comparación de las metodologías *BFMDS* y *FMDS*

Se comparó la metodología de *FMDS* de *IBM* contra la extensión de la metodología *BFMDS* y se realizó un análisis comparativo de las mismas (ver Tabla 3.1). Como se puede observar la metodología “Fundacional para Ciencia de Datos” cuenta con 10 etapas con actividades

definidas, tiene un propósito. Sin embargo, cada investigación requiere condiciones específicas.

Por último, la metodología *BFMDS* tiene mayor nivel de especificidad para llevar su aplicación dentro del dominio epidemiológico. Se compone por 10 etapas con actividades referentes al dominio epidemiológico claramente definidas en las primeras dos etapas: entendimiento del negocio y enfoque analítico. En las siguientes etapas se realizan las mismas actividades que en la metodología *FMDS* de *IBM*. Es importante mencionar que es en las primeras dos etapas donde se desarrollan actividades referentes al dominio epidemiológico, porque queremos analizar los datos de mortalidad por COVID-19 de México. Es importante porque el COVID-19 es considerada una enfermedad epidemiológica. Por tanto, se aplicará la metodología *BFMDS* para desarrollar el análisis de datos de mortalidad por COVID-19 de México.

Tabla 3.1 Metodologías *BATCH-FMDS* y *FMDS*

Metodología Fundacional para Ciencia de Datos [1]	Fase	Entendimiento del negocio	Enfoque analítico de solución	Requerimientos de datos	Recopilación de datos	Entendimiento de los datos	Preparación de los datos	Modelado	Evaluación	Despliegue	Retroalimentación
	Actividades	Definir el problema, los objetivosy reuisitos de solución.	Definir las técnicas a usar de acuerdo al problema a resolver.	Determinar los reuisitos de datos (contenido, formato y representación).	Identificar y reunir los recursos de datos.	Comprender el contenido, evaluar su calidad y descubrir caracteristicas iniciales sobre los datos.	Limpieza intgración, transformación y construcción del conjunto de datos	Desarrollo de modelos según el enfoque analítico definido.	Evaluar el modelo de acuerdo a su eficacia.	Implementar el modelo en un entorno de producción	Recopilar los resultados del modelo implementado
BATCH FMDS [3]	Fase	Entendimiento del negocio	Enfoque analítico de solución								
	Actividades	Tareas especificas: - Recopilar información. - Definir el tipo de estudio epidemiológico. Definir pregunta de investigación - Definir objetivos. - Establecer criterios de aceptación	Selección del método con tareas especificas: - Descripción de la categoria. - Descripción del método. - Identificación de parámetros. - Variaciones del método								

Capítulo 4
Caso práctico

El tiempo es la cosa más valiosa que una persona puede gastar.
Theophrastus

4. Caso práctico

Se desarrolló el caso práctico “aplicación de la Ciencia de Datos para el análisis de grupos de mortalidad por COVID-19 según factores sociodemográficos a nivel municipal” utilizando la metodología *BFMDS.* En este capítulo se da una breve descripción de la etapa y se describen las actividades realizadas para el desarrollo del caso práctico en cada una de las etapas de esta metodología.

4.1 Etapa entendimiento del negocio

Es importante entender que una de las enfermedades que actualmente ha sido motivo de mayor preocupación para los gobiernos e instituciones a nivel mundial es la epidemia de la COVID-19. La COVID-19 es un problema importante de salud pública. Unos 14.9 millones de personas murieron en el mundo a causa de la pandemia entre el 2020 y el 2021 [27].

De acuerdo con la OMS (Organización Mundial de la Salud), la COVID-19 es una enfermedad respiratoria transmisible, las personas que llegan a contraer la enfermedad y desarrollan síntomas, alrededor del 80% se recuperan sin recibir tratamiento hospitalario. Alrededor del 15% desarrollan una enfermedad grave y requieren oxígeno y el 5% llegan a un estado crítico y requieren cuidados intensivos. En México, según datos del Instituto Nacional de Estadística y Geografía, el COVID-19 fue una de las principales causas de muerte en el año 2020 [28].

Para el desarrollo de esta primera etapa se debe tener una pregunta claramente definida porque, dirige el enfoque analítico que será necesario para abordar la pregunta. Para lograr realizar la pregunta se debe tener claro cuál es la meta y descubrir los objetivos que apoyan la meta.

Las preguntas que guiaron la presente investigación y que fueron respondidas son las siguientes: a) ¿Qué municipios de México tuvieron mayores tasas de mortalidad por COVID-19 en el año 2020? Y b) ¿Qué factores sociodemográficos tienen en común aquellos municipios con tasas de mortalidad similares?

El objetivo de la investigación consistió en aplicar el enfoque metodológico de la ciencia de datos para formar grupos de municipios de México con valores similares de indicadores sociodemográficos determinantes y tasas de mortalidad por COVID-19, en el año 2020.

Desarrollar el criterio de éxito. Para validar que el proyecto de investigación logre su cometido se elaboraron los criterios de éxito que contemplan un entendimiento de los objetivos, conocimiento del dominio a trabajar y la obtención de la información de fuentes oficiales. Para este caso práctico el criterio de éxito es identificar aquellos municipios de México con mayor tasa de mortalidad con el criterio de los expertos en el área epidemiológico. También, identificar los factores sociodemográficos que tienen relación con aquellos municipios con mayor y menor tasa de mortalidad.

4.2 Etapa de enfoque analítico

Definiendo el problema que se va a abordar, se selecciona el enfoque analítico apropiado para el problema. Esta fase, ofrece preguntas de apoyo para identificar el método adecuado para modelar los datos. Los métodos se dividen en tres categorías, descriptiva, predictiva y prescriptiva.

Atendiendo las preguntas de apoyo y de acuerdo con el tipo de pregunta de investigación, la categoría descriptiva utilizando técnicas de agrupamiento es la que mejor se adapta a nuestras necesidades para identificar los indicadores sociodemográficos que tienen relación con la mortalidad por COVID-19 a nivel municipal. En la actualidad existen diferentes técnicas para realizar agrupamiento.

Se busca agrupar aquellos municipios con factores sociodemográficos similares en cuanto a las tasas de mortalidad por Covid-19. De esta manera, utilizaremos una variante híbrida del algoritmo de agrupamiento OK-means ++ [29] como herramienta para modelar los datos recopilados y alcanzar el objetivo establecido.

4.3 Etapa de requerimiento de datos

En esta etapa se definieron los datos requeridos para desarrollar el caso práctico. Las fuentes para la selección de datos fueron las siguientes:

a) Datos de mortalidad de México ocurridas durante el año 2020.
b) Datos de la población total de México a nivel municipal registrados en el año 2020.
c) Datos del catálogo de enfermedades.
d) Datos de las coordenadas geográficas de México a nivel municipal.
e) Datos de superficie a nivel municipal de México.
f) Datos sobre los indicadores de pobreza a nivel municipal de México.

Los atributos que se utilizarán del conjunto de datos de mortalidad se encuentran especificados en la Tabla 4.1.

Tabla 4.1 Atributos para conjunto de datos de mortalidad

Atributo	Ent_ocurr	Mun_ocurr	Causa_def	Edad
Tipo de dato	Cualitativo	Cualitativo	Cualitativo	Cuantitativo
Descripción del atributo	Clave de la entidad de ocurrencia	Clave del municipio de ocurrencia	Causa de defunción	Edad del fallecido

Un dato cuantitativo es aquel que sus valores son números y representan una cantidad [30]. Un dato cualitativo es aquel que sus valores representan una cualidad, un atributo o una categoría y también se les conoce como datos categóricos [30]. En algunos casos, se puede clasificar un dato como cualitativo con un valor numérico, en aquellos casos en los que no se pueden realizar operaciones matemáticas con éste.

Los atributos que se utilizarán de los conjuntos de datos de población se encuentran en la Tabla 4.2.

Tabla 4.2 Atributos para el conjunto de datos de población

Atributo	Entidad	Mun	Pobtot	P_15mas
Tipo de dato	Cualitativo	Cualitativo	Cuantitativo	Cuantitativo
Descripción del atributo	Clave de la entidad	Clave del municipio o demarcación territorial	Población total	Población de 15 años o más

Los atributos que se utilizarán del conjunto de datos del catálogo de enfermedades se muestran en la Tabla 4.3.

Tabla 4.3 Atributos para conjunto de datos del catálogo de enfermedades

Atributo	Letra	Catalog_key	Nombre
Tipo de dato	Cualitativo	Cualitativo	Cualitativo
Descripción del atributo	Letra inicial del código de enfermedad	Códigos de la CIE-10	Nombre de la enfermedad

Los atributos que se utilizarán del conjunto de datos de de coordenadas geográficas se encuentran en la Tabla 4.4.

Tabla 4.4 Atributos para conjunto de datos de coordenadas geográficas

Atributo	Cve_Ent	Cve_Mun	Lat_Decimal	Lon_Decimal	Altitud
Tipo de dato	Cualitativo	Cualitativo	Cuantitativo	Cuantitativo	Cuantitativo
Descripción del atributo	Clave de la entidad	Clave del municipio	Latitud expresada en decimal	Longitud expresada en decimal	Altitud de la demarcación territorial

Los atributos que se utilizarán del conjunto de datos del sistema nacional de información municipal se encuentran en la Tabla 4.5.

Tabla 4.5 Atributos para el conjunto de datos de información municipal

Atributo	Entidad federativa	Municipio	Superficie
Tipo de dato	Cualitativo	Cualitativo	Cuantitativo
Descripción del atributo	Nombre de la entidad	Nombre del municipio	Superficie del municipio expresado en km^2

Los atributos que se utilizarán del conjunto de datos del Consejo Nacional de Evaluación de la Política de Desarrollo Social se encuentran en la Tabla 4.6.

Tabla 4.6 Atributos para el conjunto de datos de desarrollo social

Atributo	Clave_entidad	Clave_municipio	Pobreza
Tipo de dato	Cualitativo	Cualitativo	Cuantitativo
Descripción del atributo	Clave de la entidad	Clave del municipio	Porcentaje de pobreza

4.4 Etapa de recopilación de datos

Para el estudio se obtuvieron datos de seis fuentes oficiales. La Tabla 4.7 contiene el nombre de la fuente o institución responsable, el nombre del conjunto de datos.

Tabla 4.7 Fuentes

Fuente	**Conjunto de datos**
DGIS- (Dirección General de Información Sanitaria) [31]	Defunciones generales 2020
INEGI- (Instituto Nacional de Estadística y Geografía) [32]	Población de México por municipios 2020
AGEE- (Áreas Geoestadísticas Estatales) [33]	Coordenadas geográficas por municipios 2020
CEMECE- (Centro Mexicano para la Clasificación de Enfermedades) [34]	Catálogo de enfermedades
SNIM- (Sistema Nacional de Información Municipal) [35]	Registros de información municipal
CONEVAL- (Consejo Nacional de Evaluación de la Política de Desarrollo Social) [36]	Registros del porcentaje de pobreza por municipio para el año 2020

4.5 Etapa de entendimiento de los datos

Se realizó una familiarización con los seis conjuntos de datos mencionados en la Tabla 4.7. Se analizaron las características de los datos recopilados, utilizando técnicas descriptivas y de visualización como por ejemplo histogramas para entender el conjunto de datos, evaluar su calidad e identificar información de interés.

En la Tabla 4.8 se presentan los conjuntos de datos, las instituciones donde se obtuvieron y la cantidad de registros.

Tabla 4.8 Conjuntos de datos

Fuente	Conjunto de datos	Número de registros
DGIS- (Dirección General de Información Sanitaria)	Defunciones generales 2020	1,086,743
INEGI- (Instituto Nacional de Estadística y Geografía)	Población de México por municipios 2020	195,662
AGEE- (Áreas Geoestadísticas Estatales)	Coordenadas geográficas por municipios 2020	14,483
CEMECE- (Centro Mexicano para la Clasificación de Enfermedades)	Catálogo de enfermedades	14,485
SNIM- (Sistema Nacional de Información Municipal)	Registros de información municipal	2,469
CONEVAL- (Consejo Nacional de Evaluación de la Política de Desarrollo Social)	Registros del porcentaje de pobreza por municipio para el año 2020	2,469

4.6 Etapa de preparación de los datos

En esta etapa se realizan actividades de preparación de datos que incluyen la limpieza de datos, es decir: tratar con valores no válidos, eliminar datos duplicados, entre otros.

4.6.1 Selección, limpieza y transformación de las bases de datos

Cada una de los conjuntos de datos se transformaron a la extensión *.CSV* (valores separados por coma) para su manipulación con Python.

Es importante mencionar, que teniendo los atributos seleccionados se realizó una búsqueda de valores faltantes o no válidos en cada uno de los atributos, para su correcto procesamiento y evitar errores o alterar los posibles resultados. Para estos conjuntos de datos

en específico no se encontraron datos faltantes o no válidos. Es importante mencionar que todos los conjuntos de datos cuentan con su llave primaria clave entidad y municipio para los conjuntos de datos utilizados.

4.6.1.1 Datos Población INEGI

Para el conjunto de datos del censo poblacional 2020 se realizó un filtro horizontal, se estableció obtener todos aquellos municipios que tuvieran una población mayor o igual a 100 mil habitantes, de esta manera se acotó la cantidad de la muestra de municipios. Se obtuvieron un total de 233 registros. Además, se aplicó un filtro vertical donde se seleccionaron los atributos: ENTIDAD, NOM_ENT, MUN, NOM_MUN, POBTOT y P_15MAS.

4.6.1.2 Datos catálogo de enfermedades CEMECE

Para identificar las muertes por COVID-19 se buscaron los códigos correspondientes por medio de los atributos “LETRA”, “CATALOG_KEY” y “NOMBRE”.

4.6.1.3 Datos Mortalidad DGIS

En el archivo de mortalidad 2020 se realizó la selección vertical (columnas o atributos) de los 233 municipios antes obtenidos. Los atributos seleccionados fueron: ENT_OCURR, MUN_OCURR, CAUSA_DEF y EDAD.

Se incluyeron todas aquellas defunciones cuyo código de defunción fuera U071 (COVID-19, virus identificado) o U072 (COVID-19, virus no identificado) y cuyo domicilio habitual fuera uno de los municipios seleccionados. Además, se excluyeron los registros cuyo atributo de edad era menor de 15 años.

4.6.1.4 Datos de coordenadas geográficas AGEE

En el archivo de datos geográficos se realizó un filtro vertical donde se seleccionaron los siguientes atributos: CVE_ENTIDAD, CVE_MUN, LONGITUD_DEC, LATITUD_DEC, ALTITUD. Estos atributos se utilizaron para realizar experimentos con el modelo de agrupamiento.

4.6.1.5 Datos de información municipal SNIM

Para obtener la superficie en km^2 se realizaron las consultas por cada uno de los 233 municipios obtenidos anteriormente en el sitio web del Sistema Nacional de Información

Municipal (http://www.snim.rami.gob.mx/). Se creó el atributo superficie en el cual se registraron cada una de las superficies de los municipios obtenidos.

4.6.1.6 Datos de desarrollo social CONEVAL

En el archivo de datos de desarrollo social se realizó un filtro vertical donde se seleccionaron los siguientes atributos: CLAVE_ENTIDAD, CLAVE_MUNICIPIO y POBREZA. Se obtuvieron los registros de porcentaje de pobreza de los 233 municipios. Es importante destacar que para esta investigación se selecciono la pobreza que de acuerdo a [36], una persona se encuentra en situación de pobreza cuando tiene al menos una carencia social (en los seis indicadores de rezago educativo, acceso a servicios de salud, acceso a la seguridad social, calidad y espacios de la vivienda, servicios básicos en la vivienda y acceso a la alimentación) y su ingreso es insuficiente para adquirir los bienes y servicios que requiere para satisfacer sus necesidades alimentarias y no alimentarias.

4.6.1.7 Datos de tasa de mortalidad

Para cada municipio, se calculó la tasa bruta de mortalidad por COVID-19 “TM_COVID” por cada 100.000 habitantes para el año 2020, utilizando la ecuación (1)

$$Tasa\ mortalidad = \frac{muertes}{población} * 100{,}000 \qquad (1)$$

4.6.1.8 Datos de densidad poblacional

Para cada municipio, se calculó la densidad poblacional “DENSIDAD_P” con los datos de los atributos población total y superficie del municipio, utilizando la ecuación (2).

$$Densidad\ poblacional\ = \frac{Población}{área\ terrestre} \qquad (2)$$

4.6.2 Normalización de datos

Una vez obtenidos los atributos, es necesario normalizarlos para un fácil análisis y manejo de los datos. Se normalizaron algunos casos como: la tasa de mortalidad, latitud, longitud, altitud, porcentaje de pobreza y densidad poblacional para obtener una escala uniforme.

La normalización es un proceso de ajustar los valores de los datos a una escala definida, se calcula mediante la fórmula [37]:

$$X' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (3)$$

Donde X' es el valor normalizado resultante, X es el valor a normalizar, $Xmin$ y $Xmax$ los valores mínimos y máximos respectivamente.

Los atributos que se normalizaron fueron latitud, longitud, altitud, total de defunciones por COVID-19, tasa de mortalidad, porcentaje de pobreza y densidad poblacional.

4.6.3 Almacén de datos

Como resultado de la preparación de los datos se generó el almacén de datos el cual quedó conformado como se muestra en la Tabla 4.9.

Tabla 4.9 Atributos del almacén de datos

Id_atributo	Atributo	Tipo de datos	Unidad
1	Clave entidad	Cualitativo	NA
2	Entidad	Cualitativo	NA
3	Clave municipio	Cualitativo	NA
4	Latitud	Cuantitativo	Decimal
5	Longitud	Cuantitativo	Decimal
6	Altitud	Cuantitativo	Decimal
7	Superficie	Cuantitativo	Km^2
8	Población total	Cuantitativo	NA
9	Total de defunciones por covid-19	Cuantitativo	NA
10	Edad promedio	Cuantitativo	NA
11	Tasa de mortalidad	Cuantitativo	NA
12	Porcentaje de pobreza	Cuantitativo	NA
13	Densidad de población	Cuantitativo	NA

4.7 Etapa de modelado

Es importante mencionar que se tomó en cuenta el artículo [29] sobre OK-means el cual es un estudio comparativo contra otros algoritmos y el cual muestra tener mejor rendimiento a través de la experimentación informática con instancias sintéticas y reales. El éxito de la recopilación, preparación y modelado de datos depende de la comprensión del problema y el enfoque analítico seleccionado.

Existen técnicas para el análisis de grupos, que han sido utilizadas con éxito para obtener y aumentar el conocimiento sobre la enfermedad del COVID-19 a partir de grandes conjuntos de datos que se han recopilado. Algunas contienen algoritmos jerárquicos y particionales. Algunos ejemplos de algoritmos particionales son Fuzzy C-means, K-medoids y K-means. El algoritmo K-means ha sido preferido por encima de otros algoritmos de agrupamiento por la facilidad que brinda para interpretar los resultados y su fundamento teórico [38]. Es importante resaltar que se ha utilizado en varios estudios de investigación sobre contagio y mortalidad por COVID-19 [39, 40]. Para esta investigación se utilizó una variante híbrida del algoritmo de agrupamiento de K-means, llamada OK-means++ que, según los resultados experimentales, supera a los algoritmos estándar en cuanto tiempo computacional (número de iteraciones).

Se aplicó una variante híbrida del algoritmo de agrupamiento de K-means [41, 42, 43], llamada OK-means++ que, según los resultados experimentales, supera a los algoritmos estándar en cuanto tiempo computacional (número de iteraciones). Integra un algoritmo para la selección optimizada de los centroides iniciales, llamado K++ [44] y un algoritmo que acelera la convergencia del algoritmo K-means, llamado OK-means [45]. El algoritmo OK-means acelera el proceso de convergencia al detener el algoritmo cuando el número de objetos cambian la membresía del grupo en una iteración es menor que un umbral. El valor de un umbral expresa una relación entre el esfuerzo computacional y la calidad de la solución.

El pseudocódigo de la variante híbrida OK-means ++ se muestra en el Algoritmo 1. Dado un conjunto de datos *X* y el valor de *K*, genera el conjunto optimizado de centroides (líneas 1-9) según el algoritmo K++. Desde la línea 10 hasta la 23, se muestra el pseudocódigo del algoritmo OK-means. En la línea 10 se asigna el valor del umbral para el algoritmo OK-means, que para este caso se fijó en 0.72. En la línea 15, γ representa el

porcentaje de objetos que cambian la membresía del clúster en la iteración t, y se calculó de la siguiente manera: $\gamma^t = 100(^{ot}/_n)$ donde ^{ot} es el número de objetos que cambian la membresía del grupo.

Algoritmo 1: OK-means++

Inicialización:
  $X := \{x_1, \ldots, x_n\}$;
  Asignar el valor de k;
  $V := \emptyset$;
  $V := V \cup \{v_1\}$; // Seleccionar aleatoriamente el primer centroide v_1 del conjunto X.
  para $i = 2$ a k hacer
    Selección del i-th centroide v_i de X con probabilidad $D(x_i, v_j) / \sum_{x \epsilon X} D(x_i, v_j)$;
    $V := V \cup \{v_i\}$;
  $V := \{v_1, \ldots, v_k\}$;
  εok := valor del umbral para determinar la convergencia;
Clasificación:
  para $x_i \epsilon X$ y $v_k \epsilon V$ hacer
    Calcular la distancia euclidiana de cada x_i a los k centroides;
    Asignar el objeto x_i al centroide más cercano v_k;
    Calcular $\copyright$;
Cálculo del centroide:
  Calcular los nuevos centroides del conjunto V;
Convergencia:
  si ($\copyright \leq \varepsilon ok$) entonces
    Detener el algoritmo;
  no
    Ir a **Clasificación**
Fin del algoritmo

4.8 Etapa de evaluación

La evaluación del modelo implica realizar tablas y gráficos lo que permite interpretar la calidad del modelo y su eficacia para resolver el problema [19]. Para la valoración de los resultados se tomó en consideración la opinión de un experto en epidemiología, quien fue quien valido los resultados desde el punto de vista epidemiológico.

Se presentan los principales resultados obtenidos a partir del análisis de grupos. En particular, se observó que la densidad poblacional y el porcentaje de personas en situación de pobreza fueron determinantes para generar grupos cuyos elementos tuvieran valores similares de tasa de mortalidad por COVID-19.

Para visualizar la distribución de los municipios según la densidad poblacional y el porcentaje de pobreza, se generó el gráfico de la (Figura 4.1), que muestra los municipios representados por puntos. Los valores de los atributos están normalizados en el rango de 0 a 1. Se puede observar que la mayoría de los puntos tienen valores bajos de densidad poblacional. También, se puede observar que los puntos de pobreza están más dispersos.

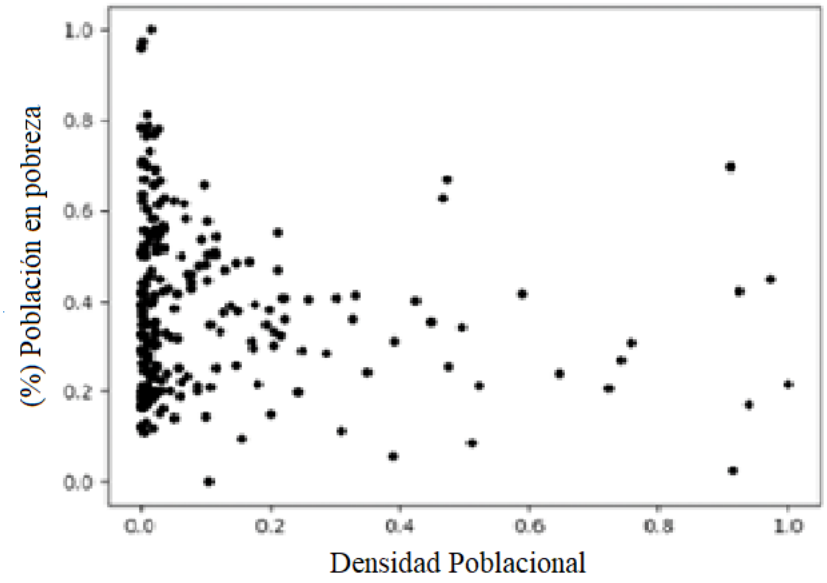

Figura 4.1 Distribución de los municipios

En la Tabla 4.10 se observan los resultados del agrupamiento de 233 municipios divididos en 16 grupos. Se establecieron 16 grupos porque fue la cantidad de grupos que mostraba una mejor distribución de los objetos, se realizaron pruebas con distintas cantidades de grupos (6, 12, 16 y 18). Las tres primeras filas corresponden a la mortalidad más alta y las últimas tres filas corresponden a la mortalidad más baja. A estos grupos se les denomina grupos extremos. La primera columna contiene el identificador del grupo, la segunda y la tercera columnas incluyen los centroides de los grupos, que tienen como atributos la densidad poblacional y el porcentaje de personas en situaciones de pobreza. La cuarta columna muestra el número de municipios en cada grupo. Los valores de las dos últimas columnas se determinaron después de la agrupación.

Tabla 4.10 Resultados de la agrupación

Grupo	Promedio de densidad poblacional	Promedio % de población en pobreza	Número de municipios	Tasa de mortalidad promedio por grupo
0	0.9138	0.0264	3	0.7970
12	0.7223	0.2059	6	0.5524
7	0.1995	0.1509	7	0.2471
9	0.1947	0.3491	21	0.2463
8	0.4734	0.6696	2	0.2420
3	0.4250	0.4042	8	0.2365
11	0.9717	0.4515	2	0.2103
14	0.0345	0.1630	25	0.2037
13	0.1393	0.3910	18	0.1911
1	0.0399	0.2401	30	0.1889
15	0.0032	0.3271	30	0.1571
5	0.0059	0.4185	25	0.1437
6	0.0270	0.6145	33	0.1316
2	0.9108	0.6982	1	0.0714
10	0.0089	0.7676	19	0.0579
4	0.0009	0.9582	3	0.0080

La Figura 4.2 muestra la distribución de los centroides del grupo y los municipios cercanos a los centroides. Algunos de los centroides se traslapan en las áreas de alta densidad de puntos.

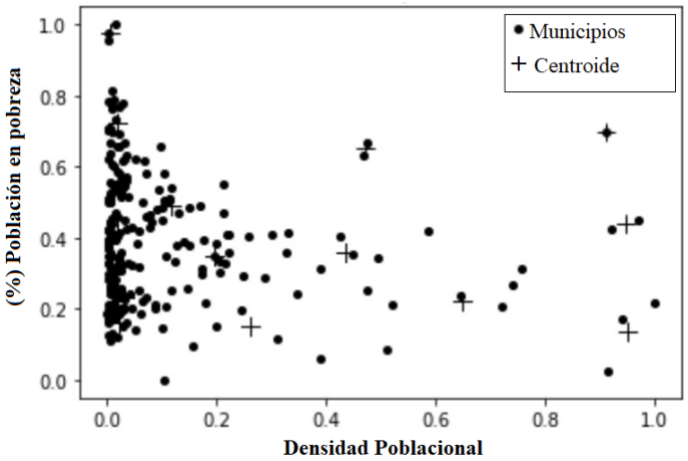

Figura 4.2 Distribución de los centroides en los grupos.

Para visualizar la distribución de los municipios en los grupos extremos, se generó el gráfico de la (Figura 4.3), que muestra los municipios representados por puntos y los centroides

representados por cruces. El color de cada punto corresponde al color de cada grupo del que forma parte el municipio. Cabe resaltar que el grupo con la mayor tasa de mortalidad se encuentra en la esquina inferior derecha, mientras que el grupo con menor tasa de mortalidad se encuentra en la esquina superior izquierda.

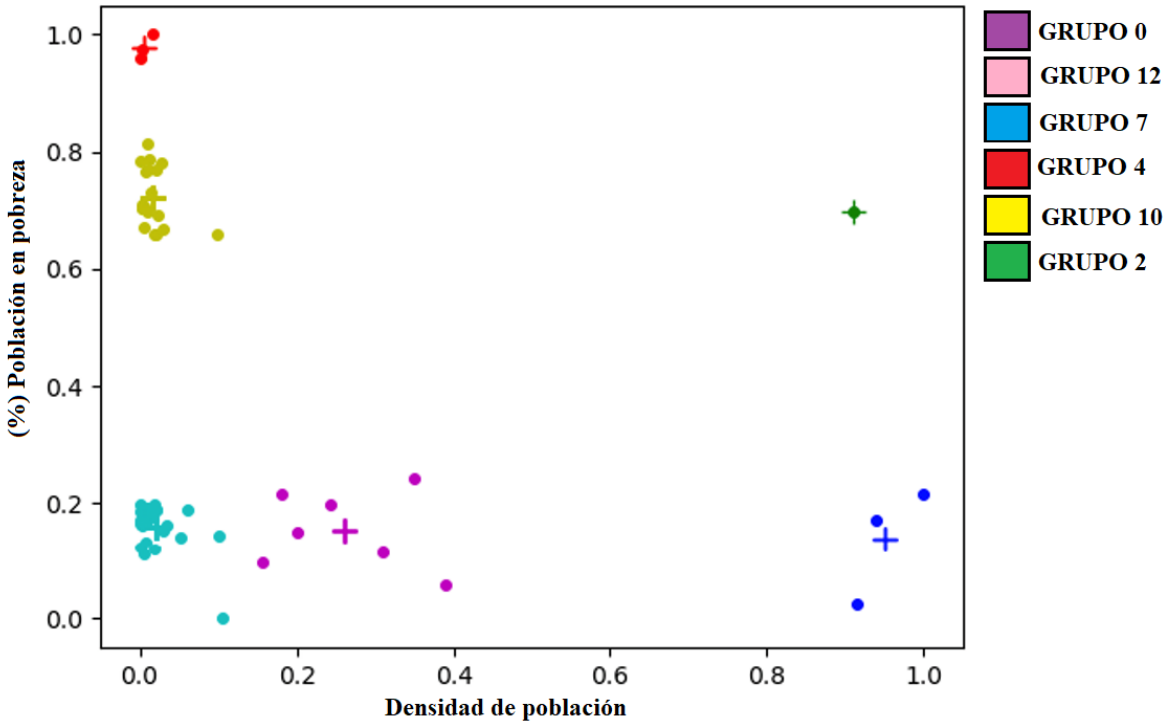

Figura 4.3 Distribución de los municipios de los grupos extremos.

4.9 Etapa de despliegue

Una vez elegido el modelo, se proyectaron los grupos con los municipios con mayor y menor tasa de mortalidad por COVID-19, para el año 2020. La Tabla 4.11 muestra los municipios o alcaldías que son parte de cada uno de los grupos 0 (morado), 12 (rosa), 7 (azul), con los municipios con mayor tasa de mortalidad. La columna uno muestra los municipios o alcaldías, la columna dos muestra la densidad poblacional, la columna tres muestra el porcentaje de población en situaciones de pobreza y la cuarta columna muestra la mortalidad promedio por municipio. La Tabla 4.12 muestra los municipios o alcaldías que son parte de cada uno de los grupos 2 (verde), 10 (amarillo) y 4 (rojo), con los municipios con menor tasa de mortalidad.

Tabla 4.11 Municipios con mayor tasa de mortalidad

Municipios / Alcaldías	Densidad poblacional	% de población en pobreza	Mortalidad promedio por municipio
Benito Juárez	16079.74	7.90	935.845
Iztacalco	17595.43	25.20	696.327
Cuauhtémoc	16541.94	20.90	634.933
Azcapotzalco	12711.91	24.20	915.769
Miguel Hidalgo	9010.22	13.50	858.687
Coyoacán	11378.65	27.10	471.318
Gustavo A. Madero	13333.53	33.80	410.278
Guadalajara	9176.35	24.80	392.962
Venustiano Carranza	13050.12	30.00	95.334
San Nicolás de los Garza	6869.98	10.80	575.935
Ciudad Madero	4290.27	23.40	393.332
Monterrey	3516.90	19.20	379.530
Guadalupe	5450.36	15.80	151.444
Apodaca	2746.71	14.20	113.030
General Escobedo	3186.84	25.00	20.157
San Pedro Tlaquepaque	6135.06	27.40	11.206

Tabla 4.12 Municipios con menor tasa de mortalidad

Municipios / Alcaldías	Densidad poblacional	% de población en pobreza	Mortalidad promedio por municipio
Chimalhuacán	16027.11	68.90	68.634
Huejutla de Reyes	321.78	65.40	131.723
Comitán de Domínguez	169.92	68.80	125.769
Taxco de Alarcón	162.19	75.00	113.651
Ixtlahuaca	476.60	76.40	105.533
San Felipe del Progreso	392.75	75.40	84.872
Macuspana	65.29	69.30	76.923
San Martín Texmelucan	1730.42	65.30	62.926
Chilapa de Álvarez	164.96	75.20	54.154
San Andrés Tuxtla	169.73	79.30	48.637
Huauchinango	414.13	68.40	47.140
San Cristóbal de las Casas	547.90	66.10	45.397
Palenque	45.80	69.90	38.559
Centla	40.00	76.80	38.058
Villaflores	57.62	69.50	21.911
Almoloya de Juárez	1269.55	26.60	19.475
San José del Rincón	205.09	77.00	18.984
Papantla	109.83	69.70	11.882
Hidalgo	109.98	66.30	8.750
Villa Victoria	255.18	71.90	6.470

Ocosingo	24.73	92.50	14.063
Las Margaritas	46.78	94.10	10.636
Chamula	295.56	96.30	0.981

La Figura 4.4 muestra un mapa de México donde se destaca cada municipio de acuerdo al grupo al que pertenece. El cuadro (a) incluye varios municipios del estado de Nuevo León. Es importante resaltar que estos municipios tienen altos valores de mortalidad. El cuadro (b) contiene el municipio de Guadalajara, que también tiene un alto nivel de mortalidad. El cuadro (c) está integrado por las tres alcaldías de la Ciudad de México en el grupo con las mayores tasas de mortalidad y los menores porcentajes de personas en situación de pobreza. Se puede observar el contraste con los municipios en el cuadro (d), donde los grupos en color naranja son los que tienen la menor tasa de mortalidad, la menor densidad poblacional y el mayor porcentaje de población en situación de pobreza. En la Figura 4.5 se muestra un acercamiento de las imágenes.

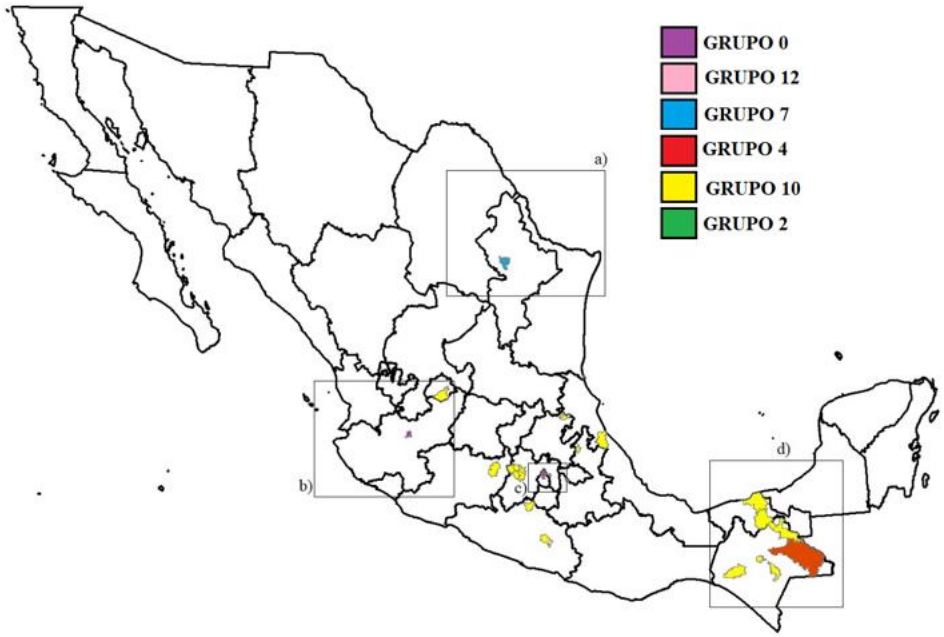

Figura 4.4 Distribución espacial de los municipios en los grupos extremos.

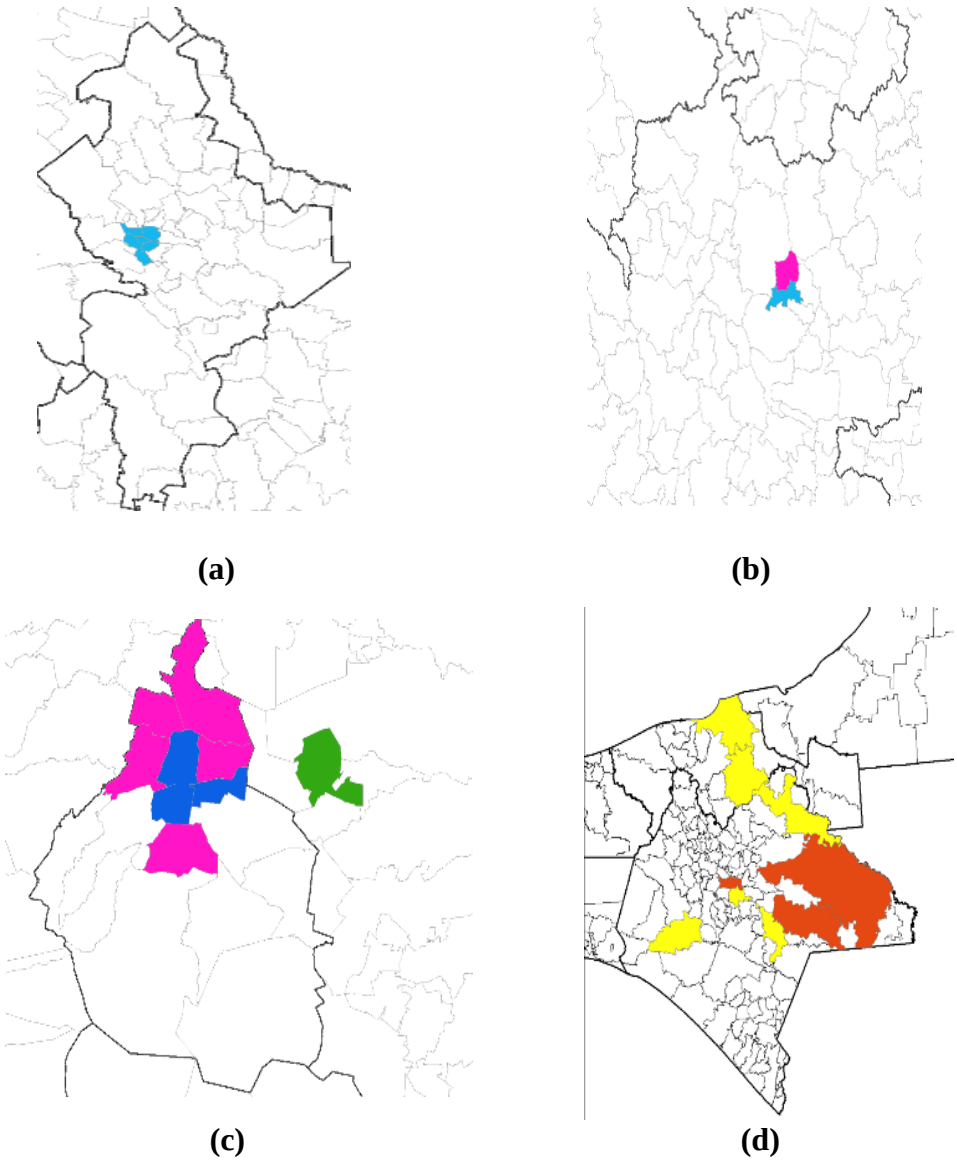

Figura 4.5 Acercamiento de municipios con mayor y menor tasa de mortalida

4.10 Etapa de retroalimentación

Esta etapa consiste en implementar la solución dentro del entorno organizacional. Para ello, se debe conocer el dominio en el que se aplica la solución propuesta. Esto implica relacionarse estrechamente con la organización y sus roles. La organización debe monitorear un tiempo (tres meses, seis meses o un año) el rendimiento del modelo y determinar si se están alcanzando los objetivos de la organización. Cada rol de la organización adquirirá el conocimiento de lo que se debe mejorar en el modelo, para que se realicen los cambios pertinentes y se despliegue de nueva forma.

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

Las actividades propuestas son:

1. Monitorear cada año las estadísticas de las tasas promedio de mortalidad por COVID-19 a nivel municipal en México.
2. Registrar las diferencias entre las tasas promedio de mortalidad por COVID-19 a nivel municipal actuales y las anteriores.
3. Determinar y ejecutar los cambios al modelo de acuerdo a los registros obtenidos.
4. Desplegar la técnica computacional con el nuevo modelo generado.

Capítulo 5
Conclusiones

La felicidad está en la alegría del logro y la emoción del esfuerzo.
Franklin D. Roosevelt

5. Conclusiones

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

5.1 Conclusiones

Con base en los resultados del caso práctico, se muestra que es factible aplicar la metodología de Ciencia de Datos *BFMDS* para el desarrollo de aplicaciones de Ciencia de Datos en el dominio epidemiológico. De esta manera, se da respuesta a la pregunta de investigación.

La metodología *BFMDS* propuesta por [25] es una extensión de la metodología de Ciencia de Datos de IBM [19]. Esta investigación contribuyó a validar dicha metodología documentando la aplicación de la misma en el caso práctico describiendo cada una de sus etapas. El caso práctico consistió en el análisis de grupos de mortalidad por COVID-19 de acuerdo a factores sociodemográficos a nivel municipal en México. Los datos poblacionales usados en la investigación provinieron de instituciones oficiales: DGIS, INEGI, CONAPO Y CEMECE.

Para validar los resultados del caso práctico, se integraron expertos en el área de epidemiología, cómputo y Ciencia de Datos y como resultado de la documentación se elaboró el artículo "*Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico*", dichos resultados pueden verse en el anexo A. Los expertos valoraron el resultado del modelo resuelto con la técnica de agrupamiento, en particular evaluaron la agrupación y los municipios con mayor y menor tasa de mortalidad.

El principal hallazgo en esta investigación, fue encontrar la relación entre las defunciones por COVID-19 y dos factores sociodemográficos de los municipios en el año 2020. Se identificó que la mayor tasa de mortalidad se encontraba en aquellos municipios con una mayor densidad poblacional y un menor porcentaje de personas en situación de pobreza, esto se vio reflejado en municipios de Nuevo León, Guadalajara, Alcaldías de la Ciudad de México como Benito Juárez, Cuauhtémoc e Iztacalco entre otros. También se identificó que la menor tasa de mortalidad se registró en aquellos municipios con una menor

densidad poblacional y un porcentaje de personas en situación de pobreza alto, esto se vio mayormente reflejado en algunos municipios del estado de Chiapas.

Desde el punto de vista computacional, los resultados son destacables, tomando en cuenta la calidad de la información que se obtiene después de realizar el proceso de Ciencia de Datos con la metodología *BFMDS*. Esta investigación muestra que es factible utilizar la metodología *BFMDS* en el dominio epidemiológico. Desde el enfoque computacional, es importante destacar que se utilizó una metodología de Ciencia de Datos con enfoque en el dominio epidemiológico.

Los beneficios tecnológicos que aporta esta investigación son: Desarrollo de aplicación relacionada con la COVID-19 aplicando conocimientos de Ciencia de Datos y la factibilidad al usar la metodología *BFMDS* en el dominio epidemiológico.

Desde una perspectiva social, es necesario enfatizar que la COVID-19 es una enfermedad que ha causado un alto índice de mortalidad a nivel mundial. Por lo que contar con este tipo de herramientas que ayuden a los gobiernos a destinar recursos económicos para detectar, atender y prevenir, cuestiones de salud pública con alto impacto social es de suma importancia.

Como trabajo futuro se sugiere aplicarla a otros casos de estudio en el dominio epidemiológico, por ejemplo: identificar aquellas regiones con mayor tasa de mortalidad por diabetes, cáncer, entre otras. Además, se sugiere desarrollar este mismo caso práctico con otra metodología de Ciencia de Datos para contrastar los resultados.

Referencias

[1] J. G. Moreno, "Científico de datos: codificando el valor oculto e intangible de los datos," Revista Digital Universitaria UNAM, vol. 18, pp. 1-16, 2017. Accessed: Nov. 2021. [Online]. Available: http://www.revista.unam.mx/vol.18/num7/art53/index.html

[2] L. Torres, A. Basto, M. Carnalla, et al, "SARS-CoV-2 infection fatality rate the first epidemic wave in Mexico," *Int J Epidemiol,* vol. 51, no. 2, pp. 429-439, 2022.

[3] C. Stewart. "Total amount of global healthcare data generated in 2013 and a projection for 2020," https://www.statista.com/statistics/1037970/global-healthcare-data-volume/#statisticContainer (accessed Sep. 24, 2020).

[4] C. Garnica, "Características generales de un caso práctico como opción de titulación en la mestria en administración," https://tauniversity.org/sites/default/files/documentos/guia-para-caso-practico.pdf (accessed Oct, 8, 2022)

[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] J. Pérez-Ortega, A. Vega, N. Almanza, et al, "Prediction of Diabetes Mortality in Mexico City Applying Data Science," in Progress in Artificial Intelligence and Pattern Recognition, pp. 211-218, 2021.

[7] J. Kumar, G. Tradigo, P. Veltri, et al, "Data science in unveiling COVID-19 pathogenesis and diagnosis: evolutionary origin to drug repurposing," *Brief Bioinform,* vol. 22, no. 2, pp. 855-872, 2021.

[8] S. Muñoz, F. López and A. Corbi, "Data Science Techniques for COVID-19 in Intensive Care Units", *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 6, no. 4, pp. 8-17, 2020.

[9] N. Saxena, P. Grupta, R. Raman, et al, "Role of data science in managing COVID-19 pandemic," *Indian Chemical Egineer,* vol. 62, no. 4, pp. 385-395, 2020.

[10] V. Kutateladze and E. Seregina, "Fast and Efficient Data Science Techniques for COVID-19 Group Testing," *Journal of Data Science,* vol. 19, no. 3, pp. 390-408, 2021.

[11] N. D. Goldstein, M. LeVasseur, and L. A. McClure, "On the Convergence of Epidemiology, Biostatistics, and Data Science," *Harvard Data Science Review,* vol. 2 no. 2, pp. 1-22, 2020.

[12] M. Kim, Z. Gu, S. Yu, et al, “Methods, Challenges, and Practical Issues of COVID-19 Projection: A Data Science Perspective,” *Journal of Data Science*, vol. 19, no. 2, pp 219-242, 2021.

[13] S. Callaghan, “COVID-19 Is a Data Science Issue,” *Patterns (N Y)*, vol. 1, no. 2, pp. 1-3, 2020.

[14] A. Vega-Villalobos, N. Almanza-Ortega, K. Torres-Poveda, et al, “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, no. 3, pp.1-14, 2022.

[15] D. N. Vinod and S. R. S. Prabaharan, “Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19,” *Chaos, Solitons & Fractals*, vol. 140, pp. 1-7, 2020.

[16] F. Foroughi and P. Luksch, “Data Science methodology for cybersecurity projects,” *Institute of Computer Science University of Rostock (ICSU)*, vol. 10, pp. 01-14, 2018.

[17] W. Budiharto. “Data Science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM),” *Journal of Big Data*, vol. 8, pp. 1 – 9, 2021.

[18] J. Pérez, J. C. Correa and F. González, “Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín,” *Ingeniare*, vol. 27, no 3, pp. 495-509, 2019.

[19] J. B. Rollins, "Metodología Fundamental para la Ciencia de Datos," IBM, Armonk, NY, Technology, IMW14828-COES-01, 2015.

[20] Microsoft, “Proceso de ciencia de datos en equipo (TDSP),” https://docs.microsoft.com/es-es/azure/architecture/data-science-process/overview (accessed Nov, 3, 2020)

[21] G. Piatestsky. “History of Data Science Infographic in 5 strands” https://www.kdnuggets.com/2015/02/history-data-science-infographic.html (accessed Nov. 8, 2020)

[22] R. Li. “History of Data Mining” https://www.kdnuggets.com/2016/06/rayli-history-data-mining.html (accessed Nov, 8, 2020)

[23] OLAP.com. “OLAP and business intelligence history” https://olap.com/learn-bi-olap/olap-business-intelligence-history/ (accessed Nov, 8, 2020).

[24] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining,

or data science projects," https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, (accessed Feb. 21, 2022).

[25] A. Vega, “Desarrollo de una Metodología de Ciencia de Datos orientada a la solución de problemas en Epidemiología,” Ph.D. in process, Comput. Sci., Centro Nacional de Investigación y Desarrollo Tecnológico, Morelos, México.

[26] J. S. Saltz, I. Shamshurin and K. Crowston, “Comparing Data Science Project Management Methodologies via a Controlled Experiment,” in Proceding in Hawaii International Conference on Systen Sciences, pp. 1013-1021, 2017.

[27] E. Maloletka. “Las muertes por COVID-19 sumarían 15 millones entre 2020 y 2021.” https://news.un.org/es/story/2022/05/1508172 (accessed Nov. 24, 2021)

[28] INEGI. “Características de las defunciones registradas en México durante 2020,” https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/DefuncionesRegistradas2020preliminar.pdf (accessed Nov. 28, 2021)

[29] J. Pérez-Ortega, N. N. Almanza-Ortega, and D. Romero, "Balancing effort and benefit of K-means clustering algorithms in Big Data realms," *PLOS ONE*, vol. 13, no. 9, pp. 1-19, 2018.

[30] L. Rincon, “Conceptos elementales,” in estadística descriptiva, 1st ed. CDMX, 2017, pp. 4-7.

[31] DGIS. “Defunciones.” http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_defunciones_gobmx.html. (accessed Feb. 21, 2022)

[32] INEGI. “Censo de población y vivienda 2020.” https://www.inegi.org.mx/programas/ccpv/2020/ (accessed Feb. 21, 2022)

[33] AGEE. “Catálogo Único de Claves de Áreas Geoestadística Estatales Municipales y Localidades.” https://www.inegi.org.mx/app/ageeml/ (accessed Feb. 21, 2022)

[34] CEMECE. “Clasificación de Enfermedades.” https://www.gob.mx/salud/acciones-y-programas/menu-clasificacion-de-enfermedades-dgis?state=published (accessed Feb. 22, 2022)

[35] SNIM. “Sistema Nacional de Información Municipal.” http://snim.rami.gob.mx/ (accessed Mar. 16, 2022)

[36] CONEVAL. “Medición de la pobreza.” https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipio-2010-2020.aspx (accessed Mar. 16, 2022)

[37] J. Hernández, M. Ramirez and C. Ferri, Introducción a la Minería de Datos. 1st ed. España, pp. 49-93, 2001.

[38] K. Gohari, A. Kazemnejad, A. Sheidaei, et al, "Clustering of countries according to the COVID-19 incidence and mortality rates," *BMC Public Health*. vol. 22, no. 1, pp. 1-12, 2022.

[39] J. P. Gutierrez and S. M. Bertozzi, "Non-communicable diseases and inequalities increase risk of death among COVID-19 patients in Mexico," *Plos One*. vol. 15, no. 10 pp. 1-11, 2020.

[40] E. Cornelius, O. Akman and D. Hrozencik, "COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty," *Mathematics*, vol. 9, no. 17, pp. 1-22, 2021.

[41] M. Sánchez-Montañés, P. Rodríguez-Belenguer, A. J. Serrano-López, et al, "Machine Learning for Mortality Analysis in Patients with COVID-19," *Int J Environ Res Public Health,* vol. 17, no. 22, pp. 1-20, 2020.

[42] A. Kiaghadi, H. S. Rifai and W. Liaw, "Assessing COVID-19 risk, vulnerability and infection prevalence in communities," *Plos One*, vol.15, no. 10, pp. 1-21, 2020.

[43] A. Dorosshenko, "Analysis of the distribution of COVID-19 in italy using clustering algorithms," in Proceding in IEEE Third International Conference on Data Stream Mining & Processing (DSMP), pp. 325–328, 2020.

[44] R. Jancey, "Multidimensional group analysis," *Australian Journal of Botany*. vol. 14, no. 1, pp. 127-130, 1996.

[45] J. MacQueen, ". Some methods for classification and analysis of multivariate observations," In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.

Anexo A: Publicación de artículo derivado de la investigación

mathematics

MDPI

Article

Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico

Joaquín Pérez-Ortega [1,*], Nelva Nely Almanza-Ortega [2,*], Kirvis Torres-Poveda [3,4], Gerardo Martínez-González [1], José Crispín Zavala-Díaz [5] and Rodolfo Pazos-Rangel [6]

[1] Tecnológico Nacional de México/CENIDET, Cuernavaca 62490, Mexico; gerasmtz93@gmail.com
[2] Tecnológico Nacional de México/IT de Tlalnepantla, Tlalnepantla de Baz 54070, Mexico
[3] Centro de Investigación Sobre Enfermedades Infecciosas, Instituto Nacional de Salud Pública, Cuernavaca 62100, Mexico; kjtorres@insp.mx
[4] CONACyT-Instituto Nacional de Salud Pública, Cuernavaca 62100, Mexico
[5] Administración e Informática, Facultad de Contaduría, Universidad Autónoma de Morelos, Cuernavaca 62209, Mexico; crispin_zavala@uaem.mx
[6] Tecnológico Nacional de México/IT de Cd. Madero, Madero 89440, Mexico; r_pazos_r@yahoo.com.mx
* Correspondence: jpo_cenidet@yahoo.com.mx (J.P.-O.); nnaortega@outlook.com (N.N.A.-O.)

Abstract: Mexico is among the five countries with the largest number of reported deaths from COVID-19 disease, and the mortality rates associated to infections are heterogeneous in the country due to structural factors concerning population. This study aims at the analysis of clusters related to mortality rate from COVID-19 at the municipal level in Mexico from the perspective of Data Science. In this sense, a new application is presented that uses a machine learning hybrid algorithm for generating clusters of municipalities with similar values of sociodemographic indicators and mortality rates. To provide a systematic framework, we applied an extension of the International Business Machines Corporation (IBM) methodology called Batch Foundation Methodology for Data Science (FMDS). For the study, 1,086,743 death certificates corresponding to the year 2020 were used, among other official data. As a result of the analysis, two key indicators related to mortality from COVID-19 at the municipal level were identified: one is population density and the other is percentage of population in poverty. Based on these indicators, 16 municipality clusters were determined. Among the main results of this research, it was found that clusters with high values of mortality rate had high values of population density and low poverty levels. In contrast, clusters with low density values and high poverty levels had low mortality rates. Finally, we think that the patterns found, expressed as municipality clusters with similar characteristics, can be useful for decision making by health authorities regarding disease prevention and control for reinforcing public health measures and optimizing resource distribution for reducing hospitalizations and mortality.

Keywords: clustering; COVID-19; Data Science; Data Science methodology; epidemiology; machine learning; pandemic; unsupervised learning

MSC: 62H30; 62R07; 68T09; 91C20

check for updates

Citation: Pérez-Ortega, J.; Almanza-Ortega, N.N.; Torres-Poveda, K.; Martínez-González, G.; Zavala-Díaz, J.C.; Pazos-Rangel, R. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico. *Mathematics* **2022**, *10*, 2167. https://doi.org/10.3390/math10132167

Academic Editor: Victor Leiva

Received: 14 May 2022
Accepted: 19 June 2022
Published: 22 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The public health impact of the ongoing COVID-19 pandemic has been estimated globally by the number of reported COVID-19 deaths and estimates of excess mortality in different populations and locations [1].

Given the availability of public epidemiological data on COVID-19 in many countries, several studies have focused on the analysis of patterns of similarity in incidence and mortality rates of COVID-19 and clustering by geographical areas [2,3]. Some of the approaches of these studies have been the analysis of temporal trends of mortality rates [4–7] as well