UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE

CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

DESCUBRIMIENTO DEL CONOCIMIENTO EN BASES DE DATOS PARA LA PREVENCIÓN DE LA MORTALIDAD MATERNA

TESIS PARA OBTENER EL TÍTULO DE:

MAESTRO EN ADMINISTRACIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

PRESENTA:

L.C. FREDY LÓPEZ MENESES

BAJO LA DIRECCIÓN DE:

DR. PABLO PAYRÓ CAMPOS

BAJO LA CODIRECCIÓN DE

DR. EDDY ARQUÍMEDES GARCÍA ALCOCER

CUNDUACÁN, TABASCO. AGOSTO, 2025.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

DESCUBRIMIENTO DEL CONOCIMIENTO EN BASES DE DATOS PARA LA PREVENCIÓN DE LA MORTALIDAD MATERNA

TESIS PARA OBTENER EL TÍTULO DE:

MAESTRO EN ADMINISTRACIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

PRESENTA:

L.C. FREDY LÓPEZ MENESES

BAJO LA DIRECCIÓN DE:

DR. PABLO PAYRÓ CAMPOS

BAJO LA CODIRECCIÓN DE

DR. EDDY ARQUÍMEDES GARCÍA ALCOCER

CUNDUACÁN, TABASCO. AGOSTO, 2025.

Declaración de Autoría y Originalidad

En la ciudad de Cunduacán el día once del mes agosto del año 2025, el que suscribe **Fredy López Meneses**, alumno del Programa de Maestría en Administración en Tecnologías de la Información, con número de matrícula **102H11010** adscrito a la **División Académica de Ciencias y Tecnologías de la Información,** de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención del Grado de Maestría y titulada **"DESCUBRIMIENTO DEL CONOCIMIENTO EN BASES DE DATOS PARA LA PREVENCIÓN DE LA MORTALIDAD MATERNA"**, dirigido por el Dr. Pablo Payró Campos y el Dr. Eddy Arquímedes García Alcocer.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando y aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de: originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 12 de agosto de 2025.

Estudiante: Fredy López Meneses

UJAT
UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA, ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

2025 AÑO DE LA Mujer Indígena

Cunduacán, Tabasco, a 11 de agosto de 2025
Oficio No. 1386/2025/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Fredy López Meneses
Egresado de la Maestría en Administración de Tecnologías de la Información

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Descubrimiento del conocimiento en bases de datos para la prevención de la mortalidad materna"***, para presentar examen y obtener el Grado de Maestro en Administración de Tecnologías de la Información.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

MTE. Óscar Alberto González González
Director

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Dr. Eddy Arquimedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.

M.T.E. OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Carta de Cesión de Derechos

Villahermosa, Tabasco a 12 de agosto de 2025.

Por medio de la presente manifestamos haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Descubrimiento del Conocimiento en Bases de Datos para la Prevención de la Mortalidad Materna.**

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención, de que tenemos el derecho al reconocimiento como autores de la obra, y la Universidad Juárez Autónoma de Tabasco mantendrá en un 100% la titularidad de los derechos patrimoniales por un periodo de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Fredy López Meneses

TESTIGOS

Dr. Pablo Payró Campos

Dr. Eddy Arquímedes García Alcocer

Dedicatoria

Dedico este trabajo, con profundo respeto y gratitud, a quienes han sido pilares fundamentales en mi camino académico y personal.

A mi familia, por su amor incondicional, paciencia y apoyo constante. Gracias por creer en mí incluso en los momentos difíciles. Mi amada esposa: **BLANCA LUCERO CASAS GARCÍA,** HIJO E HIJAS: **FREDY, ANGELICA Y NORAH LUCERO.**

A mis queridos padres: **SALATIEL LÓPEZ RAMÍREZ Y BARTOLA MENESES BROCA** por siempre motivarme, aconsejarme, apoyo y amor infinitos. Y por estar orgullosos de mi y siempre desearme el mejor de los éxitos en todo lo que emprendo.

A mis maestros, jurados y directores, por compartir su conocimiento y por sembrar en mí la semilla del pensamiento crítico y la pasión por el saber.

A mis compañeros y compañeras de estudio, por los desafíos compartidos, las ideas debatidas y el compañerismo que hizo más llevadero este proceso.

Y finalmente, me lo dedico a mí mismo, por la perseverancia, el esfuerzo y la determinación de seguir creciendo, incluso ante la adversidad. Este logro no es un punto final, sino el inicio de nuevas metas y aprendizajes. Y guardo en silencio la clave de mi proyecto mayor, una semilla de inteligencia universal que solo germinará en el tiempo adecuado.

“…y si el adiós fuera el principio de un perpetuo comienzo”.

(Acción Poética)

Agradecimientos

A Dios por ser dador de vida, sabiduría y los medios necesarios para que concluya con éxito este Gran logro. Eternamente agradecido y a ÉL sea la gloria por los siglos de los siglos Amen.

A mi familia esposa, hijos, padres y hermanos, por su amor incondicional, por las palabras de aliento en los momentos difíciles y por estar siempre presente en mi vida personal y profesional. Mi gratitud infinita y bendición para todos en unidad, amor y fraternidad.

Agradezco especialmente a mis directores de tesis **DR. PABLO PAYRÓ CAMPOS** Y **DR. EDDY ARQUÍMEDES GARCÍA ALCOCER**, por su apoyo incondicional, orientación, paciencia y valiosas sugerencias a lo largo de este proceso. Su experiencia y compromiso académico fueron fundamentales para alcanzar este logro.

Agradezco también a la **Universidad Juárez Autónoma de Tabasco** (*mi alma* mater) y a la División Académica De Ciencias Y Tecnologías De La Información (**DACYTI**), por brindarme la oportunidad de formarme en un ambiente que estimula el pensamiento crítico, la investigación y el compromiso con la mejora de nuestra sociedad.

Agradezco profundamente a todas las personas que confiaron en mí y apoyaron mi formación académica. De manera especial, reconozco el invaluable apoyo económico de mis padres, **Salatiel López Ramírez** y **Bartola Meneses Broca**, cuyo esfuerzo ha sido la base de este logro. Extiendo también mi gratitud a quienes, aun en la distancia, se solidarizaron con mi propósito académico: mi tío **Víctor Manuel Meneses Broca**, el **Sr. John Tarsitana**, mis primos **Rober, Jony, Fernando, Rubén, Sergio, Andrés y Mayreni**, así como mis amigos **Delmer, Alex y Javier**. Su apoyo, sin fronteras, quedará siempre en mi corazón.

Gracias.

Índice General

Índice de tablas

Índice de figuras

Título.

DESCUBRIMIENTO DEL CONOCIMIENTO EN BASES DE DATOS PARA LA PREVENCIÓN DE LA MORTALIDAD MATERNA.

Resumen.

La mortalidad materna constituye un problema prioritario de salud pública y uno de los principales retos para alcanzar los Objetivos de Desarrollo Sostenible. El presente estudio analiza los registros abiertos de la Dirección General de Información en Salud (DGIS) correspondientes a las defunciones maternas ocurridas en Tabasco entre 2002 y 2022. Se aplicó la metodología de Descubrimiento del Conocimiento en Bases de Datos (KDD) con un enfoque cuantitativo, exploratorio y descriptivo, empleando el algoritmo no supervisado Expectation–Maximization (EM) en la herramienta WEKA.

El análisis permitió identificar tres perfiles diferenciados de riesgo de mortalidad materna: (1) mujeres jóvenes con escolaridad media y seguridad social, donde el principal factor es la calidad de la atención hospitalaria; (2) mujeres con baja o alta escolaridad, mayormente sin seguridad social, con muertes más frecuentes en fines de semana, asociadas a desigualdades de cobertura; y (3) un grupo crítico compuesto por adolescentes y adultas mayores, con baja escolaridad, sin seguridad social y residentes en zonas rurales dispersas.

Estos hallazgos evidencian que la mortalidad materna no se distribuye al azar, sino que responde a patrones sociodemográficos, territoriales y de acceso a servicios. El estudio demuestra la utilidad de la minería de datos como herramienta para descubrir perfiles ocultos, aportar evidencia para la toma de decisiones y diseñar estrategias diferenciadas de prevención y atención obstétrica que contribuyan a reducir la mortalidad materna en el estado de Tabasco y, potencialmente, en México.

Palabras clave: Mortalidad materna, Minería de datos, KDD, EM, Tabasco.

Title.

KNOWLEDGE DISCOVERY IN DATABASES FOR THE PREVENTION OF MATERNAL MORTALITY.

Abstract.

Maternal mortality remains a priority public health problem and one of the main challenges for achieving the Sustainable Development Goals. This study analyzes open data from the General Directorate of Health Information (DGIS) corresponding to maternal deaths that occurred in Tabasco, Mexico, between 2002 and 2022. The Knowledge Discovery in Databases (KDD) methodology was applied with a quantitative, exploratory, and descriptive approach, using the unsupervised Expectation–Maximization (EM) algorithm in the WEKA tool.

The analysis identified three differentiated risk profiles of maternal mortality: (1) young women with medium education and social security, where the main factor is the quality of hospital care; (2) women with low or high education, mostly without social security, with deaths more frequent on weekends, associated with inequalities in coverage; and (3) a critical group composed of adolescents and older women, with low education, without social security, and living in dispersed rural areas.

These findings show that maternal mortality is not randomly distributed but responds to sociodemographic, territorial, and healthcare access patterns. The study demonstrates the usefulness of data mining as a tool to uncover hidden profiles, provide evidence for decision-making, and design differentiated prevention and obstetric care strategies that contribute to reducing maternal mortality in Tabasco and, potentially, across Mexico.

Keywords: Maternal mortality, Data mining, KDD, EM, Tabasco.

Capítulo I. Introducción

1.1 Planteamiento del problema

El avance de las tecnologías de la información en nuestro mundo actual ha generado grandes volúmenes de datos constantemente, en diversos sectores, incluido el de la salud. Lo que representa una gran oportunidad de procesar dicha información y extraer conocimiento útil mediante técnicas de minería de datos por medio del descubrimiento del conocimiento en bases de datos (KDD, por sus siglas en ingles), con el fin de mejorar la toma de decisiones en el sector salud (Ferraris et al., 2023).

La salud es una de las necesidades primordiales de una población, que los gobiernos y autoridades deben darle importancia, ya que se trata de vidas humanas y además es un derecho humano fundamental, plasmado artículo 4º de la Constitución Política de los Estados Unidos Mexicanos (Última reforma publicada DOF 15-04-2025). Así que es necesario contar con las herramientas de análisis que aprovechen el basto universo de datos e información, con el uso de las tecnologías de ciencia de datos. Es por ello que el Sector Salud en México (SSA, IMSS, ISSSTE, SEDENA, PEMEX, etc) cuenta con sistemas que concentran toda la información en bases de datos que se envían de niveles inferiores a superiores, llegando finalmente a la Organización Mundial de la Salud.

En México existen sistemas de información que se crean en los gobiernos estatales y a nivel federal tales como: SAEH(Egresos), URGENCIAS, SEED(Defunciones), PLATAFORMA SINAVE, SINAC, LESIONES, SIS-DGIS, EPIG, MORTALIDAD MATERNA E INFANTIL, ETC. La mayoría de la información que se concentra de todas las instituciones del país son alojadas en la página web http://sinais.salud.gob.mx (Sistema Nacional de Información en Salud) y en la DGIS (dirección General de

Información en Salud), en formato de base de datos comprimidas de Microsoft Access o cubo dinámico.

Por otro lado, en la Unión Americana (USA), la integración de sistemas de información modernos de informática médica ha mejorado significativamente la administración y utilización de datos relacionados con la salud. La **minería de datos**, un actor clave de estos sistemas, permite extraer conocimientos valiosos de grandes conjuntos de datos, apoyando la toma de decisiones informadas en el sector salud. Según Wu et al. (2021), la minería de datos en grandes datos clínicos incluye pasos y modelos metodológicos, utilizando bases de datos disponibles como SEER (Epidemiologia), NHANES (salud y nutrición), TCGA (Genoma del Cáncer), y MIMIC (Cuidados Intensivos), esenciales para evaluar riesgos de pacientes y asistir en la toma de decisiones clínicas.

Además, la aplicación de herramientas y técnicas de minería de datos en el sector salud a demostrados ser muy beneficiosa para la gestión de los servicios de salud. Como lo señalan Romero Zaldívar et al. (2022), “las técnicas de minería de datos se aplican para apoyar la toma de decisiones de los médicos en el diagnóstico, pronóstico y tratamiento, al procesar grandes volúmenes de información clínica para detectar patrones y asociaciones entre múltiples factores”, lo que justifica el uso de este tipo de herramientas analíticas en problemas prioritarios como la mortalidad materna. Esto no solo ayuda a mejorar la salud de los pacientes, sino que también permite reducir costos a las unidades médicas y la detección temprana de enfermedades o riesgos a la salud.

1.2 Pregunta de investigación

En todos los centros de salud, la información que proporcionan los sistemas, solo aporta información estadística a ciertas áreas, jefaturas y/o departamentos, limitando los sistemas a un periodo de un año estadístico para la captura de información o emisión de reportes. Dichos sistemas son creados de acuerdo a las necesidades de información, de los programas de salud estatal y federal, y no proporcionan información útil a la unidad

médica. El conocimiento necesario para que el personal médico pueda tomar decisiones de forma acertada y con fundamento, permanece latente en las bases de datos de las instituciones de salud.

Uno de los principales problemas en el manejo de grandes volúmenes de datos es que se llegan a perder los objetivos iniciales de los sistemas de información; ya que los reportes no son flexibles y no permiten el análisis inteligente de grandes volúmenes históricos de datos. Limitando su uso y aprovechamiento del todo el potencial que contienen las bases de datos si aplicamos correctamente los procesos de descubrimiento del conocimiento en bases de datos (*KDD por sus siglas en ingles*).

Este es el caso en muchas instituciones del gobierno estatal y federal. Los datos se pierden en las bases de datos o repositorios que nunca se usan y que pocas veces son sometidas a procesos de extracción de conocimiento o minería de datos, por parte de la institución. Perdiendo la oportunidad de obtener información o conocimiento útil para mejorar la práctica médica y dar servicios de salud de calidad.

De lo anterior expuesto se plantea la siguiente pregunta de investigación:

¿QUÉ PATRONES O PERFILES PUEDEN DESCUBRIRSE EN LOS CASOS DE MORTALIDAD MATERNA EN TABASCO? MEDIANTE TECNICAS DE MINERIA DE DATOS NO SUPERVISADAS, QUE CONTRIBUYAN A SU PREVENCIÓN Y ATENCIÓN OPORTUNA

1.3 Hipótesis

El análisis de los registros de mortalidad materna mediante técnicas de minería de datos no supervisadas **permitirá** identificar perfiles o agrupamientos que aporten conocimiento útil para apoyar estrategias de prevención y toma de decisiones en salud pública.

1.4 Objetivos

1.4.1 Objetivo general

Descubrir patrones y perfiles asociados a la mortalidad materna en Tabasco mediante técnicas de minería de datos no supervisadas, aplicadas sobre los registros del sistema de vigilancia epidemiológica de la Dirección General de Información en Salud (DGIS.

1.4.2 Objetivos específicos

- Explorar y caracterizar estadísticamente los registros de mortalidad materna ocurrida en Tabasco entre los años 2002 y 2022.
- Integrar y depurar variables relevantes a partir de fuentes obtenidas como datos sociodemográficos, estacionales y sociales.
- Aplicar técnicas de minería de datos no supervisadas (**Expectation-Maximization**) para descubrir agrupamientos o perfiles con características comunes.
- Interpretar los patrones descubiertos y proponer líneas de acción o hipótesis futuras que puedan ser integradas a los sistemas de información en salud o guías de atención.

1.5 Justificación

Los centros de salud en todos los niveles de atención generan y almacenan grandes volúmenes de información diariamente. Sin embargo, la mayor parte de estos datos se utiliza muy poco para transformar la calidad de los servicios de salud, que reciben los pacientes y sus familias. Esto representa una oportunidad para extraer conocimiento útil que sirva como base para la toma de decisiones informadas y que aporte beneficios sociales significativos, especialmente en un problema tan sensible como la mortalidad materna, cuyas repercusiones trascienden a la madre y afectan a todo su entorno cercano y a la sociedad en general.

En este sentido, existe una necesidad creciente de analizar la información a través de técnicas de minería de datos que permitan comprender mejor los patrones que subyacen en los datos y que pueden ayudar a salvar vidas. Como lo describen Raghupathi y Raghupathi (2014), la analítica de **Bigdata** en salud ofrece un enorme potencial para mejorar la atención, reducir costos y tomar decisiones basadas en evidencias, gracias a la posibilidad de integrar y procesar grandes cantidades de datos de manera eficiente y efectiva. Por ello, al elegir correctamente los datos, las metodologías y las técnicas de minería de datos, y al interpretar adecuadamente los resultados, es posible obtener información clara y valiosa para actuar oportunamente y, en última instancia, contribuir a una mejor calidad de la atención y a la reducción de la mortalidad materna en nuestro país.

Esta investigación también conlleva implicaciones económicas significativas, ya que se permite hacer uso de los datos y sistemas que no tenían ningún uso previo. Recuperando así parte de los recursos invertidos en su implementación. Es así que, a través del conocimiento útil descubierto mediante herramientas y modelos analíticos, el personal médico se le permitirá tomar decisiones informadas y detectar de forma oportuna a las pacientes embarazadas con mayor riesgo de mortalidad materna. Esto ayudará en la aplicación de estrategias o medidas preventivas que eviten complicaciones graves y desenlaces fatales, contribuyendo a la optimización de recursos y a la reducción del costo social asociado a la mortalidad materna.

1.5.1 Alcance del Proyecto

La presente investigación se realizó con información proveniente de los registros de defunciones maternas ocurridas en el estado de **Tabasco**, captadas por el Subsistema Epidemiológico y Estadístico de Defunciones (SEED) y publicadas por la Dirección General de Información en Salud (DGIS), ambos componentes del Sistema Nacional de Información en Salud (SINAIS). Esta información forma parte del proceso de vigilancia epidemiológica realizado por la Secretaría de Salud a nivel nacional. Los resultados

obtenidos podrán ser evaluados para determinar su aplicabilidad a los sistemas de Expediente Clínico Electrónico y proponer estrategias de mejora en la atención materno-infantil, como parte de un trabajo futuro.

Se utilizaron bases de datos de la Dirección General de Información en Salud (DGIS) sobre Mueres Materna de los años 2002-2022. Los datos están disponibles como Datos Abiertos en el sitio: http://www.dgis.salud.gob.mx, dichas bases de datos contienen campos o columnas con datos la persona fallecida, localidad, unidad de salud, lugar de ocurrencia, causa principal de la muerte, quién certifica, si se considera para razón de mortalidad materna o no, entre otros.

Cabe destacar que esta base de datos está compuesta exclusivamente por registros de defunciones maternas, por lo que no se cuenta con información sobre casos no fatales o sobrevivientes. Debido a esta característica, el análisis no permite aplicar técnicas predictivas **supervisadas**, sino que se limita al uso de técnicas de minería de datos **no supervisadas**. Esta restricción representa una limitación en cuanto a la capacidad de predicción, pero también una oportunidad metodológica importante, ya que permite generar **perfiles y patrones de riesgo** dentro del conjunto de casos positivos, los cuales pueden ser utilizados para orientar futuras investigaciones con enfoques más amplios y bases de datos balanceadas.

1.6 Metodología

1.6.1 Enfoque de Investigación

La presente investigación se inscribe en el enfoque cuantitativo, con un diseño de tipo exploratorio y descriptivo, orientado a la identificación de patrones ocultos en los registros de mortalidad materna en Tabasco. A través del uso de técnicas de minería de datos no supervisadas, como el agrupamiento (clustering), que busca descubrir perfiles o

agrupaciones relevantes que contribuyan a la comprensión del fenómeno y sirvan como base para futuras estrategias de prevención.

Según De Jesús, C. (2024), la investigación cuantitativa busca cuantificar variables, identificar patrones y establecer relaciones causales entre los fenómenos estudiados. Este enfoque se basa en la recopilación de datos objetivos y cuantificables, que pueden ser analizados utilizando técnicas estadísticas.

1.6.2 Fuentes de Información

Para esta investigación se utilizaron datos abiertos proporcionados por la Dirección General de Información en Salud (DGIS), en particular la base de datos de mortalidad materna registrada entre los años 2002 y 2022 en el estado de Tabasco. Esta base forma parte del Sistema Nacional de Información en Salud (SINAIS) y se alimenta del Subsistema Epidemiológico y Estadístico de Defunciones (SEED).

Asimismo, se revisaron fuentes secundarias como artículos científicos, tesis académicas, estadísticas epidemiológicas, reportes técnicos y estudios previos relacionados con la mortalidad materna y el uso de minería de datos en salud pública, los cuales sirvieron para contextualizar el fenómeno, sustentar teóricamente la metodología empleada y orientar el análisis exploratorio de patrones.

1.6.3 Técnicas de recolección de información

Los datos se descargaron del sitio de datos abiertos de la **Dirección General de Información en Salud (DGIS)** en formato de archivo comprimido ZIP. Los datos están en formato de archivo texto delimitado por comas “,”, por lo que se obtienen la base de datos solo con pacientes del estado de Tabasco y de pacientes que fueron atendidas por alguna causa de embarazo (CIE-001 O001-O019). La Información se concentró en Microsoft Excel en formato CVS para su manejo y exploración inicial en WEKA; ya que nos permitirá obtener un conocimiento general, explorando y transformando los datos,

para posteriormente aplicar las técnicas no supervisadas y modelado de datos dentro de programa de WEKA; dándonos la ventaja de realizar todo el proceso en un mismo software, permitiéndonos realizar las gráficas iniciales que nos darán una visión descriptiva de los datos.

1.6.4 Metodología

Para la presente investigación se empleará el proceso de **Descubrimiento del conocimiento en Bases de Datos (Knowledge Discovery in Databases o KDD por sus siglas en ingles)** que consiste en el procesamiento y análisis sistemático de datos mediante distintas técnicas para encontrar patrones de comportamiento que sean útiles para la toma de decisiones el cualquier organización (Ferraris, Gabbanelli, Mileta, & Seija, 2023). Esta metodología es ampliamente usada por su flexibilidad y aplicabilidad en proyectos de minería de datos, que es precisamente el enfoque de este trabajo de investigación.

El proceso KDD incluye varias fases iterativas que pueden repetirse las veces que se requiera para refinar resultados. Estas fases son:

1. **Selección de datos**: Identificación y obtención del conjunto de datos relevante para la investigación.
2. **Preprocesamiento**: Limpieza y preparación de datos.
3. **Transformación**: Estandarización, reducción o proyección de datos a un formato adecuado para su análisis.
4. **Minería de datos**: Aplicación de técnicas analíticas, como métodos de clasificación, arboles de decisión, clustering, entre otros, para descubrir patrones.
5. **Evaluación e interpretación**: Análisis de los resultados para determinar su validez y su relevancia en el contexto del problema.

Cabe señalar que esta investigación se basa exclusivamente en registros de defunciones maternas (casos positivos), por lo que **no es posible aplicar modelos supervisados de clasificación o predicción** basados en la comparación entre sobrevivientes y no sobrevivientes. Esta limitación de diseño puede mejorarse si se combinan bases de datos de Egresos Hospitalarios de mujeres embarazadas que nos aportarían casos negativos o de pacientes que no fallecieron, aun con el riesgo de mortalidad materna.

En cambio, se optó por técnicas **no supervisadas** como el **clustering**, que permiten descubrir **perfiles o patrones internos dentro del conjunto de muertes**, sin necesidad de una variable objetivo.

Esta decisión metodológica impone una **limitación en la capacidad predictiva del modelo**, pero representa una **oportunidad exploratoria valiosa** para caracterizar los distintos tipos de mortalidad materna y proponer hipótesis que podrán ser evaluadas en futuros estudios con muestras mixtas.

1.6.5 Población de Estudio

La población de estudio de esta investigación está compuesta por los registros de mujeres que fallecieron por diversas causas y dentro de ellas las obstétricas (embarazo, parto y puerperio) en el estado de Tabasco, México, entre los años 2002 y 2022, de acuerdo con la información contenida en la base de datos oficial de mortalidad materna publicada por la Dirección General de Información en Salud (DGIS). Esta base de datos forma parte del sistema nacional de vigilancia epidemiológica y está disponible como dato abierto en el portal institucional del Gobierno de México.

A nivel nacional, la base contiene un total de 23,278 registros de defunciones maternas. Para delimitar el estudio a Tabasco, se consideraron tres posibles filtros: la entidad de residencia de la persona fallecida, la entidad donde ocurrió la defunción y la entidad

donde se registró legalmente el evento. Los resultados de estos filtros fueron los siguientes:

- 456 registros con residencia en Tabasco.
- 544 registros cuya defunción ocurrió en Tabasco.
- 544 registros registrados oficialmente en Tabasco.

Dado que el interés de esta tesis es analizar el contexto clínico y hospitalario dentro del estado de Tabasco, se seleccionaron los **544 casos** cuya defunción ocurrió en esa entidad federativa. Este criterio garantiza que los patrones descubiertos estén directamente relacionados con los servicios de salud brindados en Tabasco, independientemente del lugar de origen o de registro administrativo de la persona.

Esta población representa aproximadamente el 2.34 % del total de muertes maternas registradas en México durante el periodo 2002–2022, y constituye la muestra objetivo para aplicar técnicas de minería de datos orientadas a la detección de perfiles y patrones de riesgo asociados a la mortalidad materna en el contexto estatal.

Capítulo II. Marco teórico

2.1 Marco referencial

2.1.1 Trabajos relacionados de minería de datos y salud

Existen trabajos similares que nos darán una idea más clara del potencial y de las bondades de la minería de datos para el descubrimiento del conocimiento en las bases de datos. Datos que fueron generados en sistemas de información Hospitalarios que si se diseñan según estándares internacionales pueden ayudarnos a aplicar KDD a las bases de datos generadas y extraer su conocimiento para mejorar los servicios de salud.

2.1.2 Aplicación de minería de datos para el pronóstico de la evolución de la diabetes en México (Tesis de maestría)

Flores Guerrero (2019) desarrolló un prototipo de minería de datos para pronosticar la evolución de la mortalidad por diabetes mellitus tipo 2 en México, implementando la metodología CRISP-DM para comprender, procesar y analizar datos poblacionales obtenidos del INEGI, el Sistema Nacional de Información en Salud (SINAIS) y otras fuentes oficiales. Como parte del modelado, empleó técnicas de agrupamiento (algoritmo K-Means) para identificar municipios con mayores tasas de mortalidad y luego estimó su evolución a cinco años mediante regresión polinomial en R. La investigación demuestra que es posible extraer conocimiento útil y producir estimaciones confiables a partir de registros sanitarios, un planteamiento que es especialmente relevante para el presente estudio, ya que mediante minería de datos también se busca comprender y predecir desenlaces fatales —en este caso, la mortalidad materna.

2.1.3 Prototipo de sistema de información hospitalaria con base en el estándar HL7 homologado (Tesis de Maestría)

Una parte fundamental para aplicar minería de datos en el ámbito hospitalario es contar con sistemas que registren la información en formatos estandarizados e interoperables. En este sentido, Valencia Ramón (2022) desarrolló un prototipo de sistema de información hospitalaria o expediente clínico electrónico, basado en el estándar HL7, el cual permite integrar y homologar los datos clínicos entre distintas áreas del hospital de manera estructurada. Esta clase de sistemas es relevante como referencia, ya que establece la base para que los datos sean consistentes y accesibles, facilitando su posterior procesamiento mediante técnicas de Descubrimiento del Conocimiento en Bases de Datos (**KDD**). Por ello, la presente investigación parte de un entorno hospitalario que ya cuenta con infraestructura de este tipo, lo que hace viable la construcción y entrenamiento de modelos que ayuden a identificar patrones y factores que contribuyen a la mortalidad materna, y que sirvan para la toma de decisiones en los servicios de salud.

2.1.4 Modelo de minería de datos para la detección de enfermedades en pacientes de primer nivel de atención médica

Zárate Hernández et al. (2024), en su trabajo sobre un modelo de minería de datos para la detección de enfermedades en pacientes de primer nivel de atención médica, muestran la aplicación del algoritmo de agrupamiento K-Means para descubrir patrones en los datos clínicos. Su estudio destaca que la minería de datos es una herramienta clave en salud para reducir errores diagnósticos, comprender mejor las afecciones más comunes y su tratamiento, además de analizar posibles efectos secundarios a partir de variables como afección principal y secundaria. Este tipo de enfoque es relevante para la presente investigación, ya que evidencia que técnicas **no supervisadas** como el clustering pueden servir como punto de partida para comprender mejor los datos antes de aplicar métodos predictivos orientados a la reducción de la mortalidad materna.

2.1.5 Técnicas supervisadas y no supervisadas de minería de datos

Pérez Trujillo (2021) aplica técnicas de minería de datos supervisadas en su tesis para predecir condiciones en adultos mayores, empleando algoritmos como árboles de decisión, máquinas de vectores de soporte (SVM), regresión logística y bosques aleatorios.

Por otro lado, en su trabajo “Aplicación de minería de datos para la identificación de factores de riesgo asociados a la muerte fetal”, Toscano de la Torre et al. (2016) analizan datos clínicos utilizando técnicas de clustering jerárquico y FarthestFirst (una variante de k-means). Su propósito fue identificar patrones entre características maternas y fetales sin asumir previamente categorías, lo que permitió extraer conocimiento útil en salud reproductiva. Cabe destacar que en los casos del dataset solo se incluyen eventos de muerte fetal (casos positivos) y no se cuenta con casos negativos (sobrevivientes), ya que la finalidad de la base es concentrar únicamente muertes, lo cual representa un desafío para este tipo de estudios exploratorios. Este tipo de abordaje es útil para comprender mejor los perfiles de riesgo en estudios donde únicamente se tienen registros de eventos fatales.

2.2 Marco conceptual

2.2.1 Mortalidad Materna

La **muerte materna** se define como el fallecimiento de una mujer durante el embarazo, parto o puerperio (hasta 42 días después del término del mismo), por causas relacionadas con este o su atención, excluyendo los eventos accidentales o incidentales (OMS, 2025; NAP, 2000). Las defunciones maternas se clasifican en tres tipos (OMS, 2025):

- **Directas**: Originadas por complicaciones obstétricas, como hemorragias, infecciones o trastornos hipertensivos.

- **Indirectas**: Provienen de condiciones preexistentes que se agravan por el embarazo, por ejemplo, cardiopatías, diabetes o VIH/SIDA.
- **Tardías**: Ocurren entre los 42 días y un año después del parto por causas obstétricas directas o indirectas.

2.2.2 Epidemiología de la Mortalidad Materna

La Organización Mundial de la Salud (OMS) define la **mortalidad materna** como la muerte de una mujer durante su embarazo, parto o dentro de los 42 días posteriores a su terminación, independientemente de la duración, localización del embarazo y las causas de defunción (OMS, 2012, p. 40).

A nivel mundial, entre los años 2000 y 2023, la razón de mortalidad materna disminuyó cerca de un 40 %, sin embargo, la mayor parte de las muertes ocurrieron en países de ingresos bajos y medianos bajos y, en muchos casos, estas defunciones pudieron haberse prevenido con atención oportuna por parte de personal médico. Asimismo, en 2023, más de 700 mujeres perdieron la vida cada día, por complicaciones relacionadas con el embarazo, parto o puerperio, es decir, casi una muerte cada dos minutos a nivel global. Esta situación ha llevado a la Organización Mundial de la Salud a señalar que la mortalidad materna sigue siendo inaceptablemente alta (OMS, 2024).

Un ejemplo que muestra la problemática en el análisis de datos hospitalarios, es el trabajo de Sierra-Juárez et al. (2024), quienes validaron un modelo de inteligencia artificial para predecir la mortalidad por sepsis a partir de expedientes clínicos electrónicos. Compararon redes neuronales, máquinas de soporte vectorial y bosques aleatorios, logrando con redes neuronales un AUC cercano a 0.80. Esto demuestra que, incluso con muestras bajas o moderadas de datos, es posible desarrollar modelos confiables, reforzando la importancia de la minería de datos y el modelado predictivo para apoyar a los profesionales de la salud en problemas críticos como la mortalidad materna.

En este sentido, la evidencia sugiere que iniciar el tratamiento antibiótico y la reanimación temprana en la primera hora tras la identificación de la sepsis grave o el shock séptico reduce la mortalidad hospitalaria (Rhodes et al., 2017). De manera similar, la Organización Panamericana de la Salud (OPS, 2018) enfatiza que la vigilancia oportuna y el uso de herramientas analíticas en los servicios de salud son claves para actuar a tiempo y salvar vidas, lo que hace imprescindible integrar técnicas de minería de datos en los sistemas hospitalarios.

2.2.3 Principales causas de mortalidad

La Organización Mundial de la Salud (2019) señala que La mayoría de las muertes maternas son resultado de **complicaciones** durante el embarazo, parto y puerperio. Las principales complicaciones de mortalidad materna representan el 80% de todas las muertes maternas, que pueden prevenirse y tratarse. Estas son llamadas principales causas de mortalidad y se clasifican en dos grandes grupos:

Causas directas: Son aquellas que resultan directamente de complicaciones obstétricas del embarazo, parto o puerperio y de las intervenciones médicas realizadas para su manejo. Por ejemplo:

- **Hemorragia obstétrica**: Pérdida excesiva de sangre antes, durante o después del parto que compromete la estabilidad circulatoria y puede llevar a shock y muerte.
- **Sepsis puerperal**: Infección del aparato genital que puede diseminarse y producir shock séptico y daño multiorgánico.
- **Trastornos hipertensivos del embarazo**: Alteraciones como preeclampsia y eclampsia que causan presión arterial elevada y daño a órganos vitales como el cerebro, riñones e hígado.
- **Aborto inseguro**: Interrupción del embarazo en condiciones sin las medidas sanitarias adecuadas, con riesgo de hemorragia grave, infecciones y lesiones internas.

- **Obstrucción del parto**: Dificultad para la salida del feto por desproporción entre el canal del parto y el tamaño del bebé, que puede producir ruptura uterina y muerte materna.

Causas indirectas: Son aquellas que resultan del agravamiento de una enfermedad preexistente o que apareció durante el embarazo pero que no es directamente obstétrica. Por ejemplo:

- **Enfermedades cardiovasculares**: Patologías del corazón que empeoran por los cambios fisiológicos del embarazo, aumentando el riesgo de insuficiencia cardíaca y arritmias.
- **VIH/SIDA**: Alteración inmunitaria que aumenta la susceptibilidad a infecciones oportunistas y complicaciones graves en la gestación.
- **Malaria**: Infección transmitida por mosquitos que produce anemia y afecta el transporte de oxígeno a la madre y al feto.
- **Diabetes mellitus**: Alteración metabólica que incrementa el riesgo de preeclampsia, infecciones y complicaciones vasculares maternas.

2.2.4 Razón de mortalidad materna

En este sentido, la **Razón de Mortalidad Materna** (RMM) es un indicador que mide el número de muertes maternas por cada 100 000 nacidos vivos en un año determinado y es fundamental para conocer la magnitud del problema y evaluar las estrategias en salud (OMS GHO, 2025; POP, 2014)

2.2.5 Nivel de Información

Según García-Marco (2011), los **datos** constituyen los elementos básicos sin organización ni significado; sin embargo, cuando estos datos son estructurados y contextualizados, se transforman en **información**. A continuación, la información, al ser procesada e interpretada mediante experiencias y procesos cognitivos, da

lugar al **conocimiento**. Por último, la **sabiduría** emerge cuando dicho conocimiento se utiliza de manera contextualizada, ética y reflexiva para tomar decisiones acertadas.

2.2.6 Minería de Datos o Data Mining

La minería de datos es un área clave dentro de la ciencia de datos y tiene gran relevancia en el análisis automatizado de información. Según Acevedo Morales (2024), “la minería de datos es un campo fascinante que combina la informática, la estadística y el análisis de datos para descubrir patrones, tendencias y relaciones dentro de conjuntos de datos grandes y complejos” (p. 20). Esta definición refleja su carácter interdisciplinario y su capacidad para convertir grandes volúmenes de datos en conocimiento útil, lo que la convierte en una herramienta esencial en múltiples ámbitos, desde el sector empresarial hasta la investigación científica.

2.2.7 Aprendizaje supervisado y Aprendizaje no supervisado

En el campo del aprendizaje automático, los algoritmos se clasifican comúnmente en dos grandes categorías: supervisados y no supervisados. Los primeros se caracterizan por requerir datos de entrenamiento que incluyan tanto las entradas como las salidas deseadas, lo que permite al modelo aprender a predecir o clasificar a partir de ejemplos previamente etiquetados. Este enfoque es útil en tareas como el reconocimiento de caracteres manuscritos, donde los algoritmos identifican patrones visuales en imágenes etiquetadas (Amazon Web Services, 2024). En contraste, los algoritmos no supervisados operan sin etiquetas previas y tienen la capacidad de identificar estructuras ocultas en los datos, como grupos o categorías emergentes, siendo aplicables, por ejemplo, al agrupamiento automático de noticias según su contenido temático (Amazon Web Services, 2024). Ambas técnicas son esenciales para resolver distintos tipos de problemas en inteligencia artificial y minería de datos.

2.2.8 Tipos de Técnicas de minería de datos

Dentro del campo de la minería de datos, existen diversas técnicas que permiten extraer conocimiento útil a partir de grandes volúmenes de información. Estas técnicas se aplican según la naturaleza del problema y del tipo de datos disponibles, y pueden clasificarse como supervisadas o no supervisadas. Entre las más utilizadas se encuentran la clasificación, el agrupamiento y las reglas de asociación.

La clasificación es una técnica supervisada ampliamente usada en inteligencia artificial y minería de datos. Su función principal es predecir la clase a la que pertenece una instancia a partir de un conjunto de datos previamente etiquetado. A través de modelos que aprenden de ejemplos conocidos, permite resolver problemas como el filtrado de correo no deseado o el reconocimiento de imágenes. Ochoa, Rosas y Baluarte (2017) explican que esta técnica consiste en construir modelos que distinguen entre clases ya definidas, con el propósito de predecir correctamente la categoría de datos aún no clasificados.

El agrupamiento, también conocido como clustering, es una técnica no supervisada que organiza datos sin etiquetar en grupos basados en su similitud. A través de medidas como la distancia euclidiana, es posible encontrar patrones y estructuras ocultas, categorizando instancias similares sin necesidad de una guía externa. Según IBM (junio 2025), esta técnica es fundamental para identificar relaciones desconocidas en los datos y facilitar la segmentación o detección de anomalías en diversas aplicaciones.

Las reglas de asociación representan otra técnica no supervisada que permite descubrir relaciones significativas entre variables dentro de grandes conjuntos de datos. Es especialmente útil en contextos comerciales, como el análisis de canastas de compra, donde se busca entender qué productos suelen adquirirse juntos. Tanto IBM (junio 2025) como Amazon Web Services (2024) coinciden en que algoritmos como Apriori permiten

establecer patrones de comportamiento en los consumidores, revelando vínculos útiles para estrategias de marketing y venta cruzada.

2.2.9 Algoritmos de minería de datos

En la minería de datos, los **algoritmos supervisados** se utilizan para construir modelos predictivos a partir de datos etiquetados previamente por humanos. Esta técnica de aprendizaje automático permite que el sistema identifique relaciones entre variables de entrada y salidas esperadas, de modo que pueda generalizar y realizar predicciones precisas sobre datos nuevos. Aunque requiere un esfuerzo considerable en la preparación y etiquetado de los datos, sus resultados suelen ser altamente precisos en tareas como la clasificación o regresión. Entre los algoritmos más representativos se encuentran la regresión lineal, la regresión logística, el algoritmo de Bayes ingenuo, el método de vecinos más cercanos (KNN) y los bosques aleatorios (IBM, 2024).

Por otro lado, los **algoritmos no supervisados** operan sin datos etiquetados, explorando conjuntos de datos complejos para descubrir estructuras subyacentes, patrones o agrupamientos sin guía externa. Este tipo de aprendizaje automático es esencial en procesos de análisis exploratorio, segmentación de clientes o recomendación de productos, ya que permite encontrar relaciones ocultas que no han sido previamente definidas. Sus técnicas principales incluyen la agrupación en clusteres (como k-means o modelos de mezcla gaussiana), el aprendizaje de reglas de asociación (como Apriori), y la reducción de dimensionalidad mediante técnicas como el Análisis de Componentes Principales (PCA) o los autocodificadores (IBM, 2021).

Es importante comparar los algoritmos de clustering ya que para nuestros datos no supervisados son ideales y tenemos que decidir cual usar para obtener un modelo más confiable.

- *K-Means*: Aquí se requiere elegir a priori el número de cluster y supone que cada grupo debe tener forma esférica y un tamaño igual o similar. También es importante que los datos de las variables o atributos sean nominales.
- *DBSCAN*: Funciona bien con los cluster de forma arbitraria, aunque hay que tomar en cuenta que maneja parámetros sensibles y de eso depende su eficiencia. Tiende a clasificar muchos puntos, esto pasa si los datos están dispersos.
- EM: Este algoritmo modela datos con una mezcla de distribuciones. Esto le permite capturar clusters de mezclas continua. Este algoritmo elige el número de k-means.

2.2.10 Proceso KDD como marco metodológico de referencia

En esta sección se refieren las principales metodologías de minería de datos que constituyen el marco metodológico de referencia de la presente investigación. Cabe aclarar que estos modelos no representan la metodología aplicada directamente en este trabajo, sino que sirven para sustentar conceptualmente la elección del proceso KDD como base metodológica, la cual se desarrolla en detalle en el Capítulo III.

Hoy en día se cuenta con grandes volúmenes de información que aumentan exponencialmente esto debido a que nuestras capacidades de recolectar y almacenar datos son mayores cada día. Esto se debe a la gran cantidad de procesamiento de datos de las computadoras, como su bajo coste de almacenamiento. E incluso se habla de Big data que se encarga del almacenamiento y procesamiento de grandes cantidades de información.

El proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD, por sus siglas en inglés) constituye una metodología sistemática que permite extraer conocimiento útil y comprensible a partir de grandes volúmenes de datos. Su núcleo está conformado por técnicas de minería de datos, las cuales aplican algoritmos matemáticos y computacionales para desarrollar modelos capaces de detectar patrones relevantes y

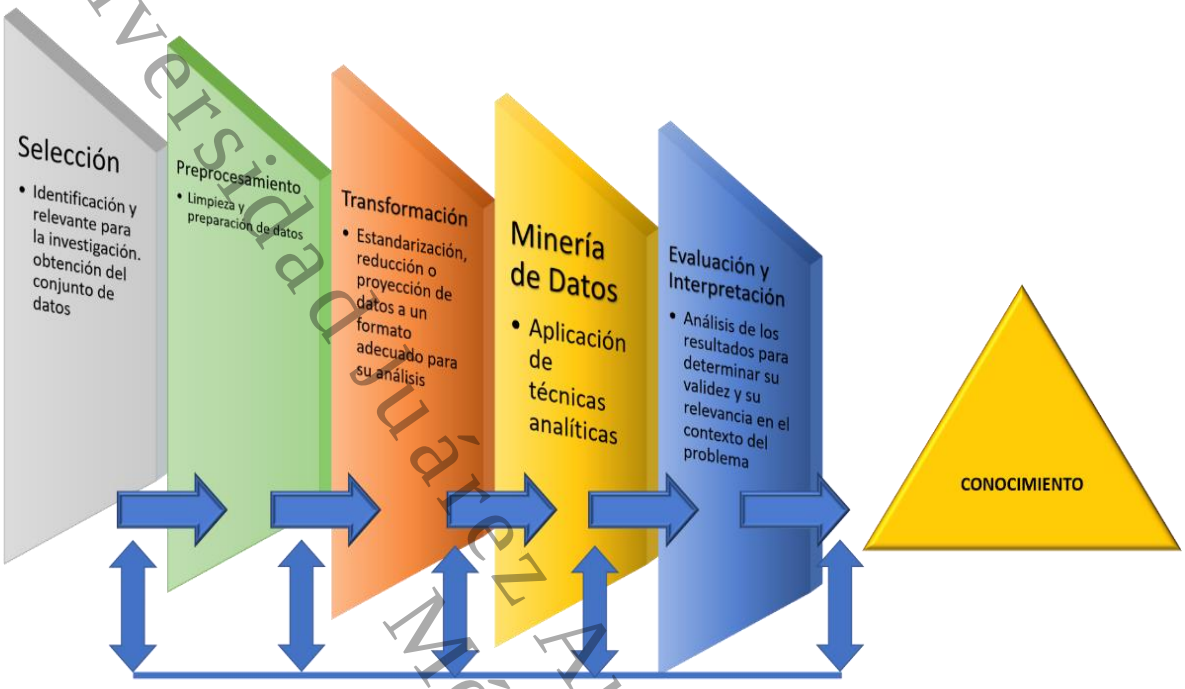
predecir comportamientos futuros. KDD no se limita a una sola técnica, sino que abarca un conjunto de fases interdependientes que permiten convertir datos en bruto en conocimiento accionable, siendo una herramienta fundamental en contextos como la analítica educativa, el diagnóstico médico o la inteligencia de negocios (Ochoa, Rosas Paredes & Baluarte Araya, jun-2025).

Estas fases inician con la **recolección de datos** desde múltiples fuentes; continúa con el **preprocesamiento**, donde los datos se limpian y transforman para asegurar su calidad; sigue la **exploración de datos**, en la que se analizan visual y estadísticamente las características principales del conjunto; posteriormente se pasa al **modelado**, que utiliza algoritmos para construir representaciones predictivas o descriptivas del fenómeno en estudio; y finaliza con la **evaluación e interpretación**, validando el modelo construido y extrayendo conclusiones útiles. Cada etapa aporta valor específico al proceso y su correcta aplicación influye directamente en la calidad de los resultados obtenidos (Acevedo Morales, 2024).

Fases del Proceso KDD (resumen en lista):

- Recolección de datos
- Preprocesamiento de datos
- Exploración de datos
- Modelado o minería de datos
- Evaluación e interpretación

Figura 1 *Proceso de Descubrimiento del Conocimiento en Bases de Datos (KDD)*



Nota: Fuente: Elaboración propia

2.2.11 Metodología CRISP-DM como referencia comparativa

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) ha sido adoptado ampliamente como una metodología de referencia para el desarrollo de proyectos de minería de datos en distintos sectores. Su origen se remonta a finales de los años noventa cuando un consorcio liderado por IBM lo estructuró con el objetivo de estandarizar el proceso de análisis de datos. Este enfoque se caracteriza por su flexibilidad y su estructura en seis fases, las cuales pueden recorrerse de manera iterativa y adaptativa, lo que lo convierte en una herramienta útil para alinear el análisis técnico con las metas de negocio de forma coherente y estratégica (IDECA, jun-2025).

Cada fase de la metodología CRISP-DM cumple un propósito específico dentro del ciclo de vida del análisis. Inicia con la comprensión del negocio, donde se clarifican los

objetivos empresariales; luego sigue la comprensión de los datos, centrada en explorar y validar la información disponible. Posteriormente se realiza la preparación de los datos, asegurando que estén listos para el modelado, donde se aplican algoritmos para encontrar patrones significativos. Tras esto, la fase de evaluación verifica si el modelo cumple los objetivos propuestos. Finalmente, se llega a la implementación, donde el modelo se despliega para generar valor real en el entorno operativo. Esta metodología permite mantener una visión integral y ordenada durante el proceso analítico, facilitando la toma de decisiones informadas y sostenibles (IDECA, jun-2025).

Fases de la metodología CRISP-DM:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Implementación

2.2.12 Comparación conceptual entre KDD y CRISP-DM

Elegí proceso KDD como base metodológica para mi proyecto de investigación debido a que ofrece un enfoque centrado en el descubrimiento de conocimiento a partir de los datos, el cual se alinea mejor con el propósito de mi tesis: extraer patrones útiles, novedosos y comprensibles desde un conjunto de datos específico en el contexto hospitalario/educativo/etc. Mientras que CRISP-DM enfatiza el ciclo completo del negocio con un enfoque muy aplicado a la industria, KDD se enfoca en el corazón científico del

proceso: la extracción de conocimiento como objeto de estudio en sí mismo, lo cual es más coherente con un trabajo académico de investigación.

A diferencia de CRISP-DM, que incorpora fases organizacionales y de implementación empresarial (como “Comprensión del negocio” o “Deployment”), KDD pone más peso en las fases analíticas: selección, preprocesamiento, transformación, minería y evaluación del conocimiento, haciendo énfasis en el proceso intelectual y técnico de descubrir regularidades en los datos. Esta orientación es más adecuada cuando el objetivo del proyecto es generar conocimiento generalizable, más allá de un producto operativo puntual. En resumen, KDD proporciona un marco más científico, estructurado y neutral frente a contextos de negocio, lo que lo convierte en una opción más rigurosa y apropiada para una tesis académica.

2.3 Marco Tecnológico

2.3.1 Python – Programación de alto nivel

Python es un lenguaje de programación de alto nivel, multipropósito y de código abierto que permite leer, procesar y transformar grandes cantidades de información gracias a su amplia biblioteca de paquetes especializados en manipulación y análisis de datos, como pandas, numpy y dask (Morán Linares & Casas Anaya, 2024). Además, su facilidad de uso, sintaxis clara y robusto ecosistema de librerías lo convierten en la herramienta ideal para integrar y depurar datos heterogéneos, eliminando inconsistencias, uniendo tablas mediante claves comunes y estandarizando los campos que posteriormente alimentarán los modelos predictivos.

La riqueza del entorno Python también se traduce en mayor reproducibilidad y flexibilidad para iterar sobre los datos a gran escala, ya que puede ejecutarse tanto en entornos locales como en la nube.

2.3.2 Herramientas de minería de datos: WEKA

WEKA (Waikato Environment for Knowledge Analysis) es una herramienta gratuita y de código abierto desarrollada por la Universidad de Waikato, que permite cubrir todo el ciclo del descubrimiento del conocimiento en bases de datos. De acuerdo con Moujahid e Inza (2010), WEKA ofrece una interfaz sencilla y cuatro entornos principales para que los usuarios realicen desde el preprocesamiento hasta la visualización de resultados: el Simple CLI para quienes prefieren trabajar desde línea de comandos, el Explorer que es su entorno gráfico más utilizado, el Experimenter para la comparación automatizada de algoritmos y el KnowledgeFlow para diseñar procesos mediante bloques visuales (Moujahid & Inza, 2010).

En el entorno Explorer es posible recorrer paso a paso las fases del KDD: primero, se importan los datos en distintos formatos como ARFF o CSV; luego, se limpian y transforman mediante una amplia gama de filtros para eliminar atributos irrelevantes, estandarizar unidades o generar nuevos campos derivados. Esto es clave cuando se trabaja con grandes volúmenes de información del sector salud, como los egresos hospitalarios que utilizamos en este proyecto, donde las herramientas ofimáticas convencionales resultan insuficientes para manejar millones de registros (Moujahid & Inza, 2010).

Asimismo, WEKA incluye un extenso catálogo de algoritmos para minería de datos supervisada y no supervisada. Para clasificación supervisada, por ejemplo, es común emplear métodos bayesianos (NaiveBayes), redes neuronales (MultilayerPerceptron), árboles de decisión (J48) o máquinas de vectores de soporte (SVM). Por su parte, para aprendizaje no supervisado se pueden utilizar técnicas como el clustering jerárquico o K-Means, entre otras (Moujahid & Inza, 2010). Gracias a la flexibilidad de WEKA, es posible comparar automáticamente el desempeño de múltiples algoritmos mediante validación cruzada y métricas como exactitud, sensibilidad y especificidad, obteniendo gráficos que facilitan la interpretación de los resultados (Moujahid & Inza, 2010).

En suma, WEKA es una plataforma robusta que concentra herramientas para **todo** el ciclo KDD, facilitando el procesamiento, modelado y evaluación de datos en investigaciones como la presente, donde es necesario comprender patrones en los datos hospitalarios para apoyar la toma de decisiones en salud (Moujahid & Inza, 2010).

Tabla 1 *Comparación entre WEKA y Python para minería de datos*

Aspecto	WEKA	Python
Enfoque	Herramienta visual enfocada en exploración, modelado y evaluación interactiva de datos.	Lenguaje de programación versátil para procesamiento, transformación y modelado de datos a gran escala.
Facilidad de uso	Interfaz gráfica sencilla, sin necesidad de programar, ideal para usuarios menos técnicos.	Requiere conocimientos de programación, pero ofrece máxima flexibilidad y control del flujo de trabajo.
Escalabilidad	Adecuado para datasets pequeños o medianos, puede presentar problemas con millones de registros.	Escalable a grandes volúmenes de datos gracias a librerías como *pandas*, *numpy* o *dask*.
Preprocesamiento	Opciones básicas a través de filtros predefinidos en su entorno Explorer.	Totalmente personalizable (limpieza, unión, transformación, filtrado) mediante código y librerías especializadas.
Modelado	Amplio catálogo de algoritmos supervisados y no supervisados preimplementados.	Ecosistema de bibliotecas (*scikit-learn*, *tensorflow*, *xgboost*) para modelos avanzados y personalizados.
Evaluación	Métricas básicas y visualización inmediata de resultados (precisión, recall, AUC).	Métricas personalizables, generación de reportes, gráficos e informes automatizados.
Integración	Foco en el trabajo experimental; menos flexible para integrar en entornos productivos.	Puede integrarse fácilmente a flujos automatizados, servicios web y despliegues en la nube.

Nota: Fuente: Elaboración propia.

2.3.3 Justificación para usar WEKA en la investigación

La elección de WEKA para aplicar el modelado de minería de datos es esencial; ya que su entorno gráfico facilita la exploración, construcción y evaluación inicial de modelos predictivos sin necesidad de codificar desde cero. Esto resulta clave para iterar rápidamente entre distintas técnicas de minería no supervisada, evaluando los modelos y algoritmos que ofrecen los mejores resultados antes de implementar una solución final.

En resumen, esta combinación de herramientas (Excel y WEKA) es consistente con las etapas del KDD y proporciona eficiencia, claridad y agilidad en el desarrollo del proyecto. WEKA es el espacio visual para comprender los datos y prototipar los modelos que después podrán llevarse a una implementación automatizada y escalable. Se eligió por su enfoque educativo, su interfaz amigable y su capacidad de experimentar rápidamente con diferentes algoritmos bajo un entorno visual.

2.3.4 Infraestructura computacional y versiones utilizadas

Para el desarrollo de la presente investigación se requiere una infraestructura computacional capaz de ejecutar tareas de análisis, limpieza, modelado y validación de datos de manera eficiente. En este sentido, se emplea un equipo de cómputo personal con recursos suficientes para soportar el ciclo completo de minería de datos en escala mediana, así como el uso de herramientas de software especializadas en distintos puntos del proceso.

La Tabla 1 muestra las especificaciones técnicas del equipo que será utilizado a lo largo del proyecto y que su uso es indispensable para que se obtengan los resultados en tiempo y forma. Dicho equipo se considera un recurso indispensable, pues asegura que las fases del proceso KDD se realizaran de forma óptima y que cumple con los requerimientos del software de minería de datos WEKA.

Tabla 2 *Especificaciones del equipo de cómputo del investigador*

Componente	Descripción técnica
Procesador	AMD FX-9830P RADEON R7, 12 compute cores (4 CPU + 8 GPU), 3.00 GHz
Memoria RAM	8 GB (4 GB originales + 4 GB en expansión)
Almacenamiento	932 GB HDD (modelo ST1000LM048-2E7172)
Sistema Operativo	Windows 10 Pro, 64 bits
Gráficos integrados	AMD Radeon R7 Graphics (62 MB dedicados)
Red de cómputo	Conectividad a internet vía red Wi-Fi y acceso a Google Colab

Nota: Fuente: Elaboración propia

Dado que el volumen de datos a tratar es de tamaño medio, este equipo resulta suficiente para el tratamiento de bases de datos estructuradas, visualización exploratoria, ejecución de algoritmos de clasificación, regresión y agrupamiento, y tareas básicas de evaluación de modelos.

Este diseño permite optimizar recursos tanto humanos como computacionales, asegurando una implementación práctica, reproducible y académicamente robusta del enfoque de minería de datos.

2.4 Marco legal

2.4.1 Constitución Política de los Estados Unidos Mexicanos

El Artículo 4° constitucional establece que “toda persona tiene derecho a la protección de la salud” y que la ley definirá las bases y modalidades para el acceso a los servicios de salud. Esta disposición constitucional sitúa el derecho a la salud como un pilar fundamental y una obligación del Estado, sentando las bases para el tratamiento legal de datos hospitalarios.

2.4.2 Ley General de Salud

Con posterioridad, la Ley General de Salud (1984) desarrolla el marco constitucional, estableciendo en su artículo 2 que se debe proteger el bienestar físico y mental, mejorar la calidad de vida, garantizar servicios eficientes y promover la investigación científica en salud. Así, el uso de bases de datos hospitalarios encuentra sustento en la normatividad que busca optimizar la atención médica mediante evidencia y tecnología.

2.4.3 Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados (LGPDPPSO)

La LGPDPPSO (DOF, 20 mar. 2025) establece principios obligatorios para el manejo de datos personales por entidades públicas, asegurando la licitud, finalidad, calidad, confidencialidad e integridad de la información. Para el tratamiento de datos hospitalarios —considerados sensibles— se exige implementar medidas administrativas, técnicas y físicas para proteger la información.

2.4.4 NOM-004-SSA3-2012 (Expediente Clínico)

De acuerdo con La *Norma Oficial Mexicana NOM-004-SSA3-2012 Del Expediente Clínico* (secretaria de Salud, 2012), establece que el expediente clínico es responsabilidad de la institución de salud, aunque su confidencialidad corresponde al paciente. Fija también criterios sobre la conservación —mínimo 10 años—, autenticidad e integridad de los datos, incluyendo los electrónicos.

2.4.5 NOM-024-SSA3-2012 (Interoperabilidad de sistemas)

La interoperabilidad de los sistemas de información en salud está regulada por la *Norma Oficial Mexicana NOM-024-SSA3-2012, Sistemas de información de registro electrónico para la salud. Intercambio de información en salud* (secretaria de Salud, 2012). Esta norma señala criterios para la integración y operación de sistemas de expediente clínico

electrónico, protegiendo la confidencialidad, disponibilidad y seguridad en la transferencia de datos entre plataformas interoperables.

2.4.6 Lineamientos de Protección de Datos Personales del Sector Público

Emitidos por la Secretaría de Salud, estos lineamientos regulan aspectos como el consentimiento informado, derechos ARCO (Acceso, Rectificación, Cancelación y Oposición), transparencia y seguridad de los datos clínicos, además de establecer protocolos estrictos de custodia y tratamiento de la información.

2.4.7 Implicaciones legales para el proyecto

Principalidad Ley Constitucional: el proyecto se fundamenta en el Artículo 4° y la Ley General de Salud, garantizando el acceso universal a la protección de la salud.

Protección de datos personales: se respetan los principios de la LGPDPPSO, implementando datos anónimos, encriptación y controles de acceso para garantizar integridad y confidencialidad.

Gestión ética y segura del expediente clínico: se apega a la NOM-004-SSA3-2012 y NOM-024-SSA3-2012, asegurando conservación, interoperabilidad y seguridad técnica.

Derechos de los titulares: se reconoce el derecho al acceso, rectificación y transparencia conforme a la LGPDPPSO, con mecanismos formales para defenderlo.

Responsabilidad institucional: las instituciones responsables (DGIS, SINAIS) estarán involucradas en el cumplimiento normativo y la supervisión del tratamiento de datos.

Capítulo III. Aplicación de la Metodología

A continuación, describiremos la aplicación del proceso de Descubrimiento del Conocimiento en Bases de Datos (KDD, por sus siglas en inglés) al problema de la muerte o mortalidad materna, donde utilizamos datos disponibles en fuentes institucionales u oficiales del sector salud. El objetivo es identificar patrones y factores asociados que nos permitan crear un modelo u algoritmo útil para prevenir muertes maternas en contextos clínicos y/o administrativos.

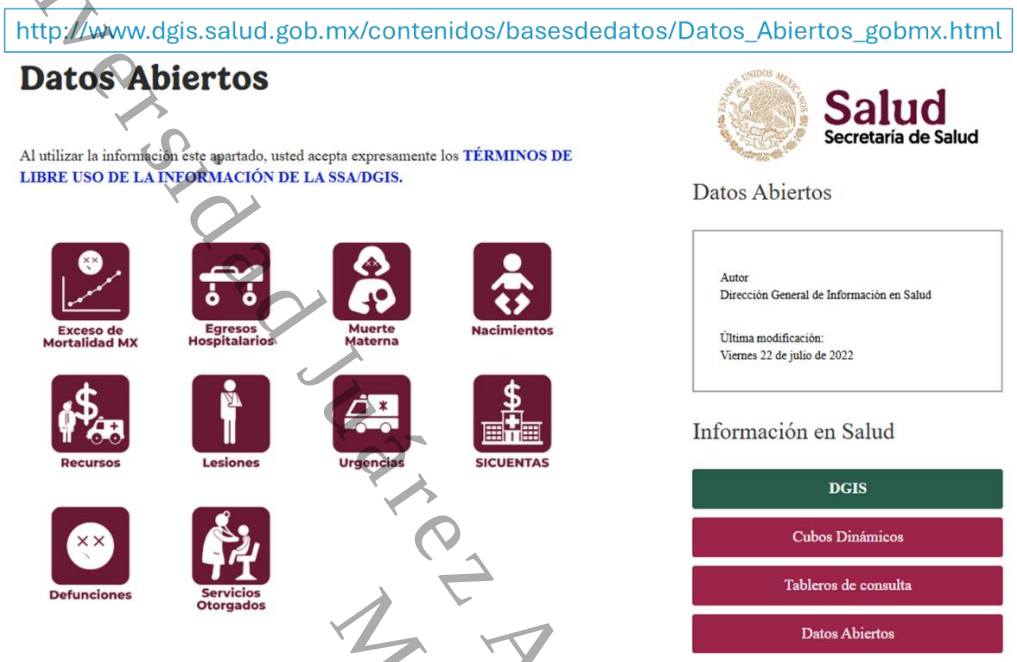
Como se mencionó en el capítulo anterior el proceso KDD consta de las siguientes partes: selección de los datos, preprocesamiento de los datos, transformación de los datos, minería de datos, evaluación del conocimiento descubierto e interpretación y uso del conocimiento. Mismo que será descritos a continuación.

3.1 Selección de los datos.

Los datos son descargados del sitio oficial de la DGIS en el apartado de datos abiertos; ya que son recopilados de los sistemas de información nacionales y la información proporcionada es información ya validada por las autoridades de salud. Para obtener la base de datos entramos a la siguiente dirección: http://www.dgis.salud.gob.mx/contenidos/basesdedatos/Datos_Abiertos_gobmx.html la información disponible esta de los años 2002 al 2022 en formato de archivo .ZIP.

Como se observa en la Figura 1, la plataforma DGIS permite el acceso y descarga a la información o datos abiertos de mortalidad materna; permitiendo su uso siempre que se cite la fuente de información correctamente y respetando los TÉRMINOS DE LIBRE USO DE LA INFORMACIÓN DE LA SECRETARÍA DE SALUD/DIRECCIÓN GENERAL DE INFORMACIÓN EN SALUD (DGIS).

Figura 2 *Página web de la DGIS con datos abiertos de mortalidad materna.*



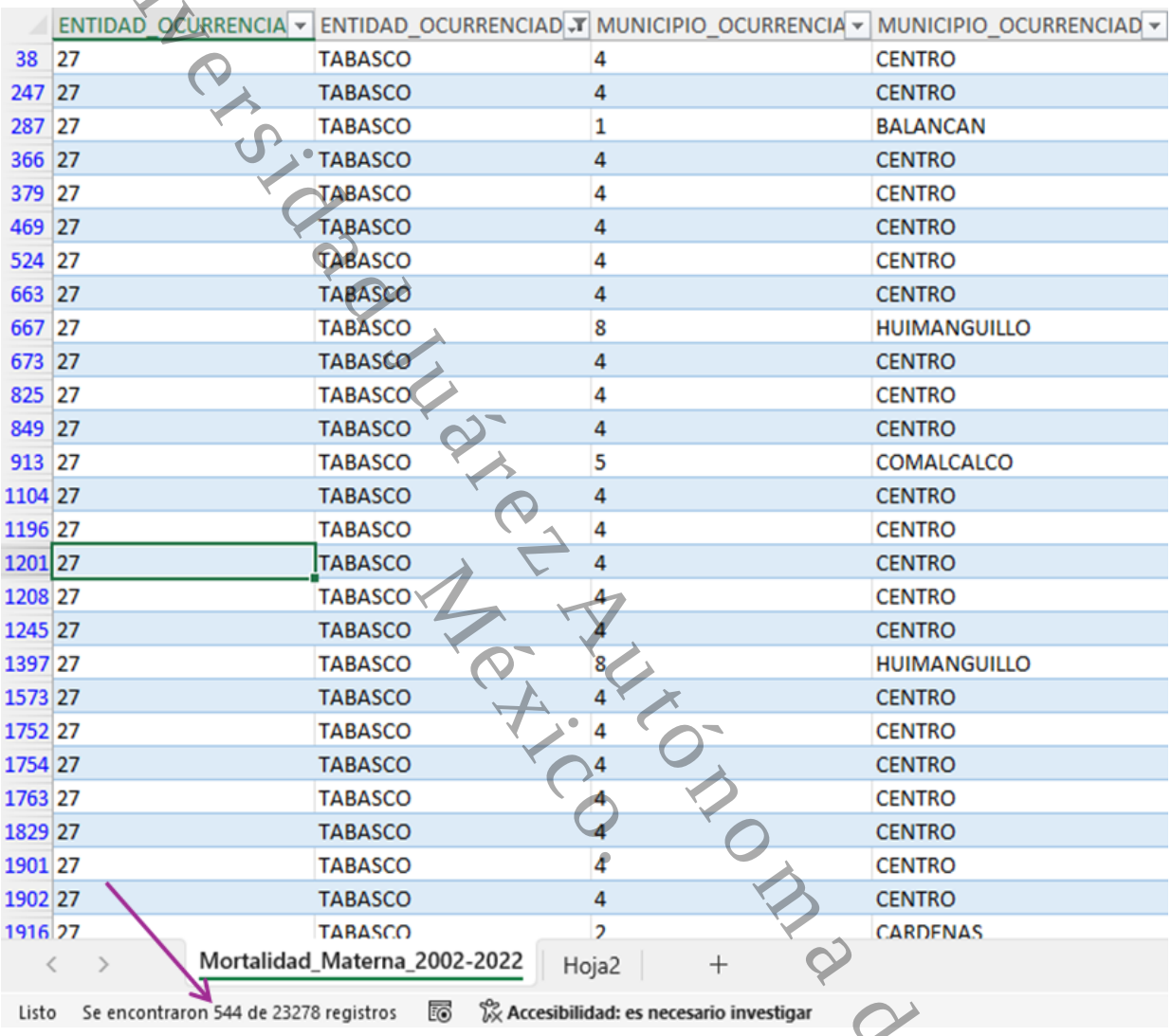
Nota: Fuente: **Adaptado de Datos Abiertos - DGIS** [Sitio web], recuperado de http://www.dgis.salud.gob.mx/contenidos/basesdedatos/Datos_Abiertos_gobmx.html

Los datos descargados se descomprimieron quedando en formato CSV (delimitado por comas) y se abrieron en Microsoft Excel para su selección de registros útiles para nuestra investigación.

El objetivo de esta fase fue identificar los registros de mortalidad materna exclusivamente del estado de **Tabasco**; ya que la base de datos original contenía todos los casos de mortalidad materna en **México** (DGIS), por lo que realizó un filtrado de los datos, quedando un total de 544 registros correspondiente a **ENTIDAD_OCURRENCIA = TABASCO (27)**, tal como lo mostramos en la Tabla 1.

Tabla 3 *Datos de mortalidad materna del estado de Tabasco (2002-2022).*

	ENTIDAD_OCURRENCIA	ENTIDAD_OCURRENCIAD	MUNICIPIO_OCURRENCIA	MUNICIPIO_OCURRENCIAD
38	27	TABASCO	4	CENTRO
247	27	TABASCO	4	CENTRO
287	27	TABASCO	1	BALANCAN
366	27	TABASCO	4	CENTRO
379	27	TABASCO	4	CENTRO
469	27	TABASCO	4	CENTRO
524	27	TABASCO	4	CENTRO
663	27	TABASCO	4	CENTRO
667	27	TABASCO	8	HUIMANGUILLO
673	27	TABASCO	4	CENTRO
825	27	TABASCO	4	CENTRO
849	27	TABASCO	4	CENTRO
913	27	TABASCO	5	COMALCALCO
1104	27	TABASCO	4	CENTRO
1196	27	TABASCO	4	CENTRO
1201	27	TABASCO	4	CENTRO
1208	27	TABASCO	4	CENTRO
1245	27	TABASCO	4	CENTRO
1397	27	TABASCO	8	HUIMANGUILLO
1573	27	TABASCO	4	CENTRO
1752	27	TABASCO	4	CENTRO
1754	27	TABASCO	4	CENTRO
1763	27	TABASCO	4	CENTRO
1829	27	TABASCO	4	CENTRO
1901	27	TABASCO	4	CENTRO
1902	27	TABASCO	4	CENTRO
1916	27	TABASCO	2	CARDENAS



Nota: Fuente: Elaboración Propia.

Variables Seleccionadas:

1. ESTADO_CONYUGALD
2. ESCOLARIDADD
3. DERECHOHABIENCIAD
4. MUNICIPIO_OCURRENCIAD
5. DIA_DEFUNCION
6. EDAD_QUINQUENALD

Estas variables fueron seleccionadas por su relevancia en la literatura médica y de salud pública, así como por su capacidad de revelar desigualdades de acceso y factores de riesgo. En la Tabla 4 mostramos el diccionario de variables de la base de datos con el fin de mostrar de donde obtuvimos nuestra selección de variable para el presente estudio. Así mismo las variables seleccionadas serán transformadas para la creación de otros atributos o variables útiles, tales como Agrupación por edad (**GpoEdad**), Día de la semana (DiaSemana), Sub regiones de Tabasco (**SubRegion**), etc.

Tabla 4 *Diccionario de variables de mortalidad materna (DGIS 2002-2022)*

NÚM	VARIABLE	TIPO DE VARIABLE	NÚM.2	VARIABLE3	TIPO DE VARIABLE
1	ANIO_NACIMIENTO	Cuantitativa discreta	30	ANIO_DEFUNCION	Cuantitativa discreta
2	MES_NACIMIENTO	Cuantitativa discreta	31	MES_DEFUNCION	Cuantitativa discreta
3	MES_NACIMIENTOD	Cualitativa nominal	32	MES_DEFUNCIOND	Cualitativa nominal
4	DIA_NACIMIENTO	Cuantitativa discreta	33	DIA_DEFUNCION	Cuantitativa discreta
5	EDAD	Cuantitativa discreta	34	HORA_DEFUNCION	Cuantitativa discreta
6	ESTADO_CONYUGAL	Cualitativa nominal	35	MINUTOS_DEFUNCION	Cuantitativa discreta
7	ESTADO_CONYUGALD	Cualitativa nominal	36	ASISTENCIA_MEDICA	Cualitativa nominal
8	ENTIDAD_RESIDENCIA	Cualitativa nominal	37	ASISTENCIA_MEDICAD	Cualitativa nominal
9	ENTIDAD_RESIDENCIAD	Cualitativa nominal	38	CAUSA_CIE_4	Cualitativa nominal
10	MUNICIPIO_RESIDENCIA	Cualitativa nominal	39	CAUSA_CIE_4D	Cualitativa nominal
11	MUNICIPIO_RESIDENCIAD	Cualitativa nominal	40	CERTIFICO	Cualitativa nominal
12	LOCALIDAD_RESIDENCIA	Cualitativa nominal	41	CERTIFICOD	Cualitativa nominal
13	LOCALIDAD_RESIDENCIAD	Cualitativa nominal	42	ENTIDAD_REGISTRO	Cualitativa nominal
14	TAMANIO_LOCALIDAD	Cualitativa ordinal	43	ENTIDAD_REGISTROD	Cualitativa nominal
15	TAMANIO_LOCALIDADD	Cualitativa ordinal	44	MUNICIPIO_REGISTRO	Cualitativa nominal
16	OCUPACION_HABITUAL	Cualitativa nominal	45	MUNICIPIO_REGISTROD	Cualitativa nominal
17	OCUPACION_HABITUALD	Cualitativa nominal	46	ANIO_REGISTRO	Cuantitativa discreta
18	ESCOLARIDAD	Cualitativa ordinal	47	MES_REGISTRO	Cuantitativa discreta
19	ESCOLARIDADD	Cualitativa ordinal	48	MES_REGISTROD	Cualitativa nominal
20	DERECHOHABIENCIA	Cualitativa nominal	49	DIA_REGISTRO	Cuantitativa discreta
21	DERECHOHABIENCIAD	Cualitativa nominal	50	ANIO_CERTIFICACION	Cuantitativa discreta
22	ENTIDAD_OCURRENCIA	Cualitativa nominal	51	MES_CERTIFICACION	Cuantitativa discreta
23	ENTIDAD_OCURRENCIAD	Cualitativa nominal	52	MES_CERTIFICACIOND	Cualitativa nominal
24	MUNICIPIO_OCURRENCIA	Cualitativa nominal	53	DIA_CERTIFICACION	Cuantitativa discreta
25	MUNICIPIO_OCURRENCIAD	Cualitativa nominal	54	ANIO_BASE_DATOS	Cuantitativa discreta
26	LOCALIDAD_OCURRENCIA	Cualitativa nominal	55	RAZON_MORTALIDAD_MATERNA	Cualitativa nominal
27	LOCALIDAD_OCURRENCIAD	Cualitativa nominal	56	RAZON_MORTALIDAD_MATERNAD	Cualitativa nominal
28	SITIO_DEFUNCION	Cualitativa nominal	57	EDAD_QUINQUENAL	Cualitativa ordinal
29	SITIO_DEFUNCIOND	Cualitativa nominal	58	EDAD_QUINQUENALD	Cualitativa ordinal

Nota: Fuente: Elaboración propia.

3.2 Preprocesamiento de los datos

Una vez seleccionados los registros y las variables, se procedió a realizar el preprocesamiento de los datos con el objetivo de asegurar su calidad, coherencia y utilidad para las posteriores fases del proceso KDD.

Limpieza de los datos

El archivo de trabajo fue filtrado previamente para conservar solo los casos cuya defunción ocurrió en el estado de Tabasco, obteniendo un total de 544 registros. Posteriormente, se aplicaron las siguientes acciones de limpieza:

- Eliminación de registros completamente vacíos o duplicados (no se identificaron duplicados).

- Corrección de errores en campos numéricos o categóricos con valores fuera de rango lógico.

- Revisión de campos con datos faltantes, sin realizar imputación, dado que el análisis no supervisado no requiere obligatoriamente la completitud de todos los atributos si estos son excluidos del modelado.

Encontramos variables que tenían valores NO ESPECIDIFICADOS (CODIGO=99) y se les colocó el carácter “?” para que Weka los interpretara como **missing values** o valores perdidos; por lo que a dichos valores les asignamos dicho carácter.

En esta fase también eliminamos las columnas sobrantes ya que solo nos quedamos con 6 variables indispensables para usarlas o transformarlas en otras nuevas que aporten mayor control o mejores resultados con los algoritmos de minería de datos sin afectar la esencia y fiabilidad de los datos que nos proporcionó la DGIS.

La frecuencia distribución de variables como edad, escolaridad, estado conyugal y año de defunción se muestran en las Figuras A2.1–A2.7, incluidas en el Anexo 2 para una mejor descripción y entendimiento de los datos.

Exploración y análisis descriptivo de los datos

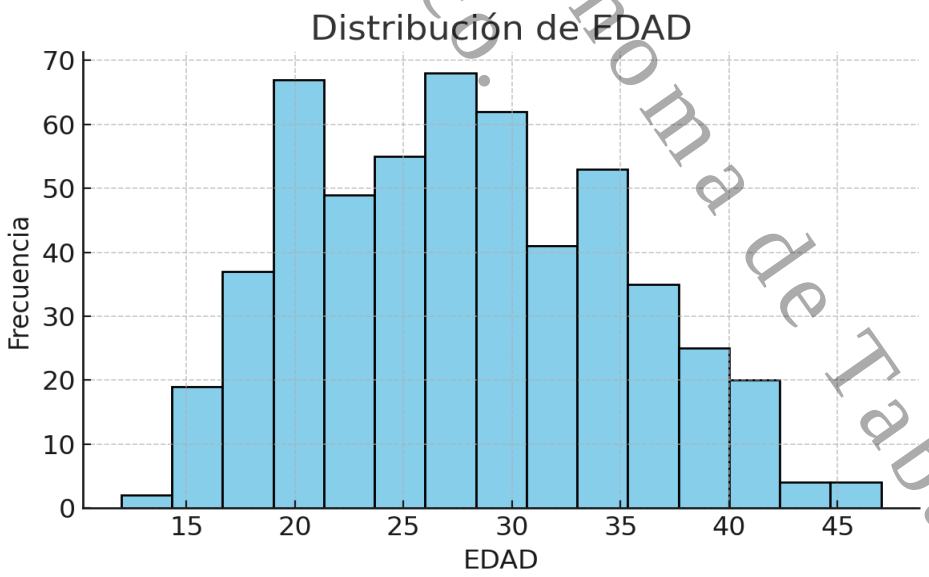
Se realizó una exploración inicial de la base de datos para comprender su estructura, distribución y posibles detalles que pudieran afectar o influir en las fases posteriores de proceso KDD y tener una visión general de cómo es nuestra población de estudio. Este análisis permitió ver patrones preliminares, agrupaciones de casos, tendencias, así como valores inconsistentes o atípicos.

Se elaboraron gráficos de barras e histogramas para una mejor comprensión y ver cuáles son las características de las personas más afectadas visualmente. Lo anterior nos

permitirá formular hipótesis y reorientar nuestros algoritmos de minería de datos y ajustar los ajustes de la fase de preprocesamiento.

La edad materna es un factor de riesgo para afecciones o complicaciones maternas que ocurren desde el embarazo, parto y puerperio y, por ende, pueden causar muerte materna a cualquier edad, por condición social, lugar de residencia y muchos factores más involucrados. Es de observar que en nuestra base de datos seleccionada del estado de Tabasco de los años 2002 al 2022, existen edades que rondan de los 12 a los 47 años de edad. En la Grafica 1 muestra la distribución de edades de mujeres fallecidas por causas obstétricas, lo cual permite ver donde hay mayor concentración de casos y su posible relación con las etapas reproductivas de las mujeres; ya que existen grupos de riesgo que deben ser atendidos con especial cuidado y ahí es donde nuestro estudio permitirá obtener dicha información o conocimientos de las bases de datos analizadas.

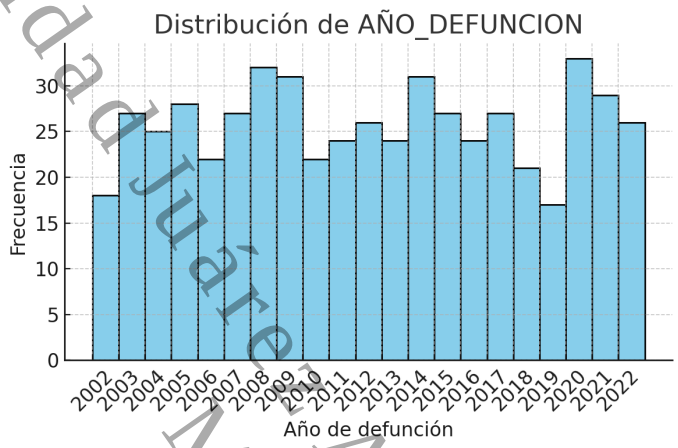
Figura 3 *Distribución de mortalidad materna por edad en Tabasco (2002–2022)*



Nota: Fuente: Elaboración propia.

En la gráfica 2 podemos ver la evolución de la mortalidad materna en el periodo de estudio, con ella podemos observar que existen picos, relacionados con eventos como las inundaciones en 2007 en Tabasco, así como la pandemia COVID-19 en 2020.

Figura 4 *Distribución de mortalidad maternas en Tabasco (2002–2022)*



Nota: Fuente: Elaboración propia. Cambiar por gráfica con línea vertical 2

3.3 Transformación de los datos

Una vez terminada la parte de preprocesamiento, se procedió a la transformación de los datos, con el fin de obtener un conjunto único y depurado. Los datos proporcionados ya están estandarizados y las variables categóricas ya están codificadas. Aunque como vimos en la fase anterior no están exentas de errores en registros por lo que debemos corregirlos para obtener mejores resultados.

Conversión y codificación

Para facilitar el procesamiento en herramientas como WEKA, las variables categóricas fueron convertidas a formato nominal compatible, manteniendo su significado original. Las variables numéricas no requirieron estandarización en esta fase, ya que los algoritmos de clustering podrán ser aplicados de ser necesarios en WEKA y este software ya maneja internamente la normalización.

En esta etapa, se optó hacer las adecuaciones a las variables ya que se detectó que presentaban un número elevado de categorías (*por ejemplo, MUNICIPIO_OCURRENCIA=17 Municipios)* por lo que para facilitarle el trabajo al algoritmo EM y obtener mejores resultados se optó reagrupar las categorías en unos mejor entendibles.

Por esta razón, las variables seleccionadas las recategorizamos para así poder trabajar de forma más optima con el algoritmo EM. Por ejemplo, en la Tabla 5, podemos observar los 17 municipios que se agruparon en las 5 subregiones (Centro, Chontalpa, Sierra, Pantanos y Los Ríos), tal y como están conformados en el estado de Tabasco, según la paginas oficial del gobierno del estado. Este procedimiento permitió reducir la dispersión de los datos, mejorar la interpretación de los resultados y facilitar el trabajo al algoritmo EM, sin perder valor epidemiológico ni la valides de la información obtenida. Lo anterior no permite leer y transmitir de manera más natural los resultados de la investigación a los expertos en salud pública del estado de Tabasco.

Tabla 5 *Subregiones de Tabasco: Centro, Chontalpa, Sierra, Pantanos y Los Ríos.*

Municipio	Etiqueta
BALANCAN	LosRios
CARDENAS	Chontalpa
CENTLA	Pantanos
CENTRO	Centro
COMALCALCO	Chontalpa
CUNDUACAN	Chontalpa
EMILIANO ZAPATA	LosRios
HUIMANGUILLO	Chontalpa
JALAPA	Sierra
JALPA DE MENDEZ	Centro
JONUTA	Pantanos
MACUSPANA	Pantanos

NACAJUCA	Centro
PARAISO	Chontalpa
TACOTALPA	Sierra
TEAPA	Sierra
TENOSIQUE	LosRios

Nota: Fuente: Elaboración propia.

Se detectaron que la variable ESCOLARIDAD tenía múltiples categorías con bajas frecuencias, lo que complicaba el análisis y la aplicación de algoritmos de minería de datos (**EM**). Con el fin de reducir la dispersión y mejorar los resultados de los algoritmos de minería de datos se optó por reagrupara los niveles educativos en tres categorías: Baja, Media y Alta, tal como lo mostramos en la siguiente Tabla.6, que nos permite entender mejor el nivel educativo de la paciente ya que las múltiples categorías no permitían tener una visión más clara de los datos.

Tabla 6 *Variable **ESCOLARIDAD** en categorías: Baja, Media y Alta.*

ESCOLARIDADD	ETIQUETA
BACHILLERATO O PREPARATORIA COMPLETA	Media
BACHILLERATO O PREPARATORIA INCOMPLETA	Media
NINGUNA	Baja
NO ESPECIFICADA	?
POSGRADO	Alta
PREESCOLAR	Baja
PRIMARIA COMPLETA	Baja
PRIMARIA INCOMPLETA	Baja
PROFESIONAL	Alta
SECUNDARIA COMPLETA	Media
SECUNDARIA INCOMPLETA	Baja
SIN ESCOLARIDAD	Baja

Nota: Fuente: Elaboración propia.

De mismo modo se reorganizaron las categorías de la variable DERECHOHABIENCIA en dos grandes grupos: Con seguridad social y Sin Seguridad Social; ya que ello nos permitirá medir el nivel de acceso a los servicios de salud. Esto lo podemos observar en la Tabla 7, donde múltiples categorías las reducimos a 2 para facilitar el trabajo del algoritmo EM seleccionado. Es decir, nos muestra si la persona o paciente femenina tuvo acceso a los servicios de salud independientemente si tiene un trabajo que le aporte seguridad social que incluye entre otros los servicios de salud, vivienda, acceso a servicios básicos, vida digna, libre esparcimiento y desarrollo de la personalidad.

Tabla 7 *Variable **DERECHOHABIENCIA**: Con Seguridad Social y Sin Seguridad Social.*

DERECHOHABIENCIA	ETIQUETA
IMSS	ConSeguridadSocial
IMSS OPORTUNIDADES	ConSeguridadSocial
IMSS PROSPERA	ConSeguridadSocial
ISSFAM	ConSeguridadSocial
ISSSTE	ConSeguridadSocial
NINGUNA	SinSeguridadSocial
NO ESPECIFICADA	?
OTRA	ConSeguridadSocial
PEMEX	ConSeguridadSocial
SECRETARIA DE LA DEFENSA NACIONAL	ConSeguridadSocial
SECRETARIA DE MARINA	ConSeguridadSocial
SEGURO POPULAR	ConSeguridadSocial

Nota: Fuente: Elaboración propia.

Del mismo modo se realizó la trasformación de las categorías de las variables ESTADO_CONYUGAL (Con pareja y Sin pareja) y EDAD_QINQUENAL (Adolescente, Adulta Joven y Adulta Mayor) con el objetivo de simplificar, mejorar la interpretación y análisis de resultados. Por lo que el algoritmo EM trabajara mejor con un número reducido de categorías, facilitando la detección de patrones y la conformación de clusters más sólidos. En casi todas las variables sus atributos contenían valores inválidos, que se

identificaron con el valor de “?” para poder tratarlos como valores perdidos y no afectar los resultados de los algoritmos.

Así mismo las variables seleccionadas y después de las etapas de limpieza, se obtuvo un dataset final en formato tabular (Tabla 8) el cual fue grabado o exportado en formato CSV, para asegurar la compatibilidad de lectura con el software de análisis Weka.

En este proceso se aplicó el filtro **ReplaceMissingValues** de Weka, con el fin de obtener valores validad para el algoritmo EM, en este caso como son variables nominales fueron reemplazados por la moda **(media para variables nominales en WEKA)** lo que permitió una entrada con menos ruido y consistente para el modelado.

Tabla 8 *Variables y categorías transformadas para facilitar el trabajo al algoritmo EM.*

GpoEdad	EstadoCivil	Escolaridad	Derechohabiencia	DiaSemana	SubRegion
AdultaJoven	ConPareja	?	SinSeguridadSocial	lunes	Centro
AdultaJoven	ConPareja	Baja	SinSeguridadSocial	domingo	Centro
Adolescente	ConPareja	Baja	SinSeguridadSocial	sabado	LosRios
AdultaJoven	ConPareja	Media	SinSeguridadSocial	jueves	Centro
AdultaJoven|	ConPareja	Baja	SinSeguridadSocial	domingo	Centro
AdultaJoven	ConPareja	Baja	SinSeguridadSocial	miercoles	Centro
AdultaJoven	ConPareja	Baja	ConSeguridadSocial	jueves	Centro
AdultaJoven	ConPareja	Media	SinSeguridadSocial	domingo	Centro
AdultaJoven	SinPareja	Media	SinSeguridadSocial	lunes	Chontalpa
AdultaJoven	ConPareja	Alta	ConSeguridadSocial	sabado	Centro
AdultaJoven	ConPareja	Baja	SinSeguridadSocial	jueves	Centro
AdultaJoven	ConPareja	Alta	SinSeguridadSocial	miercoles	Centro
AdultaMayor	ConPareja	Baja	SinSeguridadSocial	miercoles	Chontalpa
AdultaJoven	SinPareja	Media	SinSeguridadSocial	martes	Centro
AdultaMayor	ConPareja	Baja	SinSeguridadSocial	jueves	Centro
AdultaJoven	ConPareja	Media	SinSeguridadSocial	lunes	Centro
AdultaJoven	SinPareja	Baja	SinSeguridadSocial	martes	Centro

Nota: Fuente: Elaboración propia.

3.4 Minería de datos

En método EM es el más adecuado para nuestra investigación; ya que nuestros datos solo son casos positivos (defunciones maternas) y tenemos sólo 544 registros relativamente pequeño y disperso. Dicho lo anterior el mejor algoritmo o método probabilístico es EM (Expectation-Maximitation). Así mismo EM maneja mejor los valores faltantes y las categorías transformadas, además con ayuda de WEKA se determinó de manera automática el número de cluster, lo cual aumenta la confianza estadística.

La elección de EM no es arbitraria sino sustentada en la naturaleza del dataset o base de datos (no balanceado, no supervisado, con variables mixtas, y categorías recategorizadas) y además cumple o en la misma línea de nuestra investigación: identificar perfiles de riesgo diferenciados.

En esta fase se aplicó el algoritmo **no supervisado** EM (**Expectation-Maximitation**) en Weka, el cual tiene su base en modelos probabilísticos mezclando distribuciones. En dicha ejecución se ajustaron los siguientes parámetros:

- Numero de iteraciones: 100
- Validación cruzada: 10 folds
- Selección del número de clustes: automático (cross validation)
- Numero de clusters determinados: 3

Como mencionamos en la fase anterior las variables finales fueron: Grupo de edad, Estado Civil, Escolaridad, Derechohabiencia, Dia de la semana y Subregiones. Con las transformaciones aplicadas se pudo enriquecer la caracterización de los perfiles, integrando dimensiones sociodemográficas, territoriales y temporales.

Después de la aplicación del algoritmo EM, se obtuvo un modelo con 3 clusters diferenciados con distribuciones del 45%, 48% y 7% de los casos respectivamente. Los resultados de este modelo se presentarán en la siguiente sección de evaluación del conocimiento descubierto, por lo que aquí solo describimos la parte técnica de la minería de datos. Los detalles completos del proceso de ejecución y la salida del software Weka se dará a conocer en el **Apéndice A**.

3.5 Evaluación del modelo e Interpretación del conocimiento

El modelo de agrupación obtenido por medio del algoritmo Expectation Meximization (**EM**) fue evaluado tanto en consistencia estadística como en su utilidad interpretativa. De forma estadística para tener la confianza en los resultados y de utilidad interpretativa para poder aplicar el conocimiento en las áreas de la salud pública.

En términos de análisis estadísticos Weka seleccionó de manera automática la selección de 3 clusteres como la más adecuada, a través de la validación cruzada. Presentando un log-Likelihood (ajuste probabilístico del modelo) de **-5-55**. Mostrando que se acerca al cero aun siendo un numero negativo; por lo que tenemos un clara diferenciación y convergencia entre clusteres.

Es de notar que ellos 3 clusteres muestran una gran diferenciación entre sí y corresponden a patrones que se observan y que tienen sustento en la literatura médica y epidemiológica, mayormente el 3er cluster; por lo que mostramos a continuación los perfiles de cada cluster:

- Cluster 0: Mujeres adultas jóvenes, con escolaridad media y seguridad social
- Cluster 1: Mujeres con escolaridad alta o baja, con mayor frecuencia sin seguridad social.

- Cluster 2: Grupo critico o de riesgo compuesto por Adolescentes y adultas mayores con baja escolaridad y sin seguridad social, dispersas territorialmente.

Lo anterior va en concordancia con la literatura internacional y nacional por lo que refuerza la validez del conocimiento descubierto. A demás, se obtiene el hallazgo de un grupo minoritario y critico (7%) que nos permitirá orientar políticas públicas de prevención focalizadas en los descubrimientos de conocimientos.

El software determino de forma automática que la mejor solución eran 3 cluster, determinado por el método de validación cruzada, que nos da mayor confianza aun siendo valores negativos, lo importante es que estén cercanos al cero. Los tres perfiles representan agrupaciones diferenciadas de mortalidad materna en Tabasco durante el periodo 2002-2022. Las características de cada perfil o grupo se resumen en la siguiente tabla 9, la cual constituye una base sólida para las siguientes fases de evaluación e interpretación de los datos.

Tabla 9 *Clusters de Mortalidad Materna en Tabasco (2002-2022) obtenidos mediante EM*

Variable	Cluster 0 (45%) – Perfil Base	Cluster 1 (48%) – Perfil Vulnerable	Cluster 2 (7%) – Perfil Crítico
Edad	Adulta joven y adolescentes	Adulta joven y adulta mayor	Adolescentes y adultas mayores
Estado civil	Mayoría con pareja	Mayoría con pareja	Con pareja, pero más vulnerabilidad
Escolaridad	Predominio **media**	Predominio **baja/alta**	Principalmente **baja**
Derechohabiencia	Con seguridad social	Mayor frecuencia sin seguridad social	Mayoría sin seguridad social
Día de defunción	Lunes–miércoles (laborales)	Domingo y martes	Sábado y domingo

Subregión	Centro y Chontalpa	Centro	Centro, Chontalpa y dispersión (Los Ríos, Pantanos, Sierra)
Interpretación	Muertes pese a acceso y escolaridad media: foco en **calidad de atención hospitalaria**	Desigualdad educativa y de cobertura, con riesgo en fines de semana	Perfil más vulnerable: adolescentes/adultas mayores rurales y sin seguridad social

Nota. Fuente: Elaboración propia a partir del modelo EM en WEKA (datos de mortalidad materna en Tabasco, 2002–2022).

Estos resultados evidencian que la mortalidad materna en Tabasco se distribuye de manera homogénea, sino que muestras diferentes perfiles para cada grupo de riesgo: El primer **cluster 0**, refleja mujeres con acceso a los servicios de salud de manera formal, lo que apunta a la idea de mejorar la **calidad de los servicios proporcionados y de la atención hospitalaria** y toda la red de centros de salud, hospitales y centro de especialidades médicas o de alta complejidad. El segundo **cluster 1**, representa desigualdades de cobertura de la atención y **retrasos en la atención obstétrica** oportuna, especialmente los fines de semana. Y finalmente el tercer **cluster 2**, identifica claramente un grupo mínimo crítico, compuesto por adolescentes y adultas mayores, sin seguridad social y en localidades rurales, con mayores vulnerabilidades y exclusión.

Por lo que de manera más puntual se requiere por cada cluster lo siguiente:

- Cluster 0 orientar acciones diferenciadas mejorando la calidad de la atención hospitalaria,
- Cluster 1 fortalecer la cobertura en fines de semana y en mujeres sin seguridad social y,
- Cluster 2 diseñar estrategias de intervención focalizadas en comunidades rurales y adolescentes / adultas mayores.

Capítulo IV. Resultados y discusión

4.1 Resultados

En análisis de los registros de mortalidad materna ocurridos en el estado de Tabasco entre los años 2002-2022 (544 registros), permitió la aplicación del algoritmo no supervisado **EM (Expectation-Maximization)** de la aplicación del proceso y/o metodología KDD, con el fin de descubrir perfiles de riesgo a partir de variables sociodemográficas, territoriales y de acceso a los servicios de salud de las pacientes embarazadas que lamentablemente fallecieron.

Tabla 10 *Distribución de los clusters obtenidos mediante EM de mortalidad materna en Tabasco, 2002–2022*

Cluster	Perfil identificado	Frecuencia	Porcentaje
0	Perfil Base	245	45 %
1	Perfil Vulnerable	261	48 %
2	Perfil Crítico	38	7 %
Total	—	544	100 %

Nota. Fuente: Elaboración propia con datos de la Dirección General de Información en Salud (DGIS, 2002–2022).

El modelo seleccionado automáticamente por WEKA, determinó que la solución más adecuada consistía en **3 clusteres**, con distribuciones de 45%, 48% y 7% respectivamente (Tabla 10). En dichos agrupamientos podemos observar diferencias muy significativas en las características de las mujeres fallecidas y se alinean con os hallazgos previos reportados en la literatura sobre las desigualdades en la mortalidad materna en México y el Mundo.

El descubrimiento de estos patrones concuerda con estudios previos nacionales e internacionales (OMS, 2024; Romero Zaldivar et al., 2022) que destacan la influencia o afección a factores estructurales (educación, acceso a la salud y territorio) en la mortalidad materna. De esta manera, los resultados obtenidos confirman que el uso del proceso KDD y las técnicas de minería de datos no supervisadas, es una alternativa valida para caracterizar o agrupar perfiles de riesgo y orientar intervenciones de salud pública basadas en evidencias.

4.2 Discusión

Al profundizar en el **Cluster 0**, podemos observar y discutir el porqué aun con seguridad social o derechohabiencia ocurrió la muerte en este grupo de mujeres. Lo que sugiere es que exista problemas en la calidad de la atención hospitalaria y se debe atender de inmediato ya que si no se atiende a tiempo seguirán muriendo cada día sin que se pueda detener. Para ello se requieren que las acreditaciones, certificaciones y auditorias se mantengan todo el tiempo actualizadas y vigentes; ya que muchas veces solo se cuenta con recursos en las fechas de auditorías y lo demás del año se tienen carencias.

Así mismo el **Cluster 1**, donde se plantea la discusión en torno a un perfil vulnerable y a porque ocurren las muertes de este grupo los fines de semana y pudiera estar relacionado al transporte público que no opera regularmente como entre semana y en algunos centros de salud públicos y privados no tienen horarios de atención los fines de semana o tienen poco personal o si existe ginecólogo no hay anestesiólogo; es decir no se tiene el equipo de respuesta inmediata completa para las emergencias. Se requiere mayor cobertura obstétrica los fines de semana en los centros de salud y hospitales.

Por último, el **Cluster 2**, se discute el porqué es un perfil critico; ya que refleja una inequidad estructural: educación baja, ruralidad y falta de seguridad social. Aquí son muchos los factores que afectan a estas pacientes ya que por la falta de oportunidades no tienen acceso a los servicios de salud de calidad y no pueden trasladarse de manera

óptima y en su momento a un centro de salud. Se requiere implementar políticas de equidad educativa y acceso sanitario en zonas rurales.

Se clasificaron los perfiles de Bajo, Mediano y Alto Riesgo con el fin de clasificar a las mujeres que cumplen con dichos atributos del Cluster en cuestión, dichos niveles o escalas de riesgo se muestran en la Tabla 11. Al clasificarlos logramos que los hallazgos sean más comprensibles para gestores, médicos y tomadores de decisiones.

Tabla 11 *Clasificación de riesgo de mortalidad materna según perfiles identificados mediante clustering (Tabasco 2002-2022)*

Cluster	Perfil	Clasificación de riesgo	Principales factores
0 (45 %)	Perfil Base	**Bajo**	Jóvenes, escolaridad media, con seguridad social → problemas de calidad hospitalaria
1 (48 %)	Perfil Vulnerable	**Medio**	Sin seguridad social, escolaridad baja/alta, muertes en fines de semana → desigualdad de cobertura
2 (7 %)	Perfil Crítico	**Alto**	Adolescentes y adultas mayores, baja escolaridad, sin seguridad social, rurales → exclusión múltiple

Nota: Fuente: Elaboración propia a partir del modelo EM aplicado en WEKA con datos de mortalidad materna de la Dirección General en Salud (DGIS, 2002-2022).

En base a las características de cada perfil queda de la siguiente manera:

Clasificación de riesgo por cluster

- **Riesgo Bajo → Perfil Base (Cluster 0, 45 %)**

  Mujeres jóvenes, con escolaridad media y acceso a seguridad social.

Razonamiento: Tienen condiciones sociales relativamente favorables; sin embargo, la mortalidad en este grupo indica que los problemas se deben más a la calidad de la atención hospitalaria que a factores estructurales.

- **Riesgo Medio → Perfil Vulnerable (Cluster 1, 48 %)**

  Mujeres con escolaridad baja o alta, en su mayoría sin seguridad social, con mayor riesgo en fines de semana.

  Razonamiento: Existe una clara desigualdad en el acceso y oportunidad de la atención; no son los más excluidos, pero enfrentan barreras de cobertura y tiempos críticos de atención.

- **Riesgo Alto → Perfil Crítico (Cluster 2, 7 %)**

  Adolescentes y adultas mayores, con baja escolaridad, sin seguridad social y residentes en comunidades rurales dispersas.

  Razonamiento: Este grupo concentra las máximas vulnerabilidades (edad, educación, ruralidad, exclusión social y sanitaria). Son el foco prioritario de intervención.

Capítulo V. Conclusiones y Recomendaciones

5.1 Conclusiones

El presente estudio de investigación permitió aplicar todo el proceso KDD y más específico técnicas de minería de datos no supervisadas (Expectatión-Maxomization) sobre los registros almacenados por la DGIS sobre mortalidad materna ocurridas en el estado de Tabasco de los años 2002 al 2022.

Los resultados y agrupación confirmaron que la mortalidad materna no se distribuye de manera aleatoria, sino que responde a **perfiles diferenciados de riesgo**, que fueron descubiertos a través del proceso KDD.

Las principales conclusiones son:

1. El algoritmo EM agrupó todos los casos de mortalidad materna en tres grupos bien diferenciados:

    - Perfil Base (45%) **Riesgo Bajo**: Mujeres jóvenes, con escolaridad media y acceso a seguridad social, lo que refleja que la calidad de la atención hospitalaria sigue siendo un factor crítico.

    - Perfil Vulnerable (48%) **Riesgo Medio**: Mujeres con escolaridad baja o alta, en su mayoría sin seguridad social, con riesgo elevado en fines de semana, evidenciando desigualdad en la cobertura y la oportunidad de atención obstétrica.

    - Perfil Crítico (7%) **Riesgo Alto**: Adolescentes y adultas mayores, con baja escolaridad, sin seguridad social y residentes en comunidades rurales y dispersas, lo que representa el grupo más vulnerable y excluido.

2. **La calidad del a atención es un determinante clave.**

   Aun en mujeres con acceso formal a servicios de salud o seguridad social la mortalidad materna persiste, lo que indica que no basta con la cobertura total; es necesario y urgente mejorar los procesos de atención clínica y hospitalaria en todos los niveles de atención, es decir fortalecer la atención primaria a la salud.

3. **Las desigualdades sociales y territoriales aumentan la mortalidad materna.**

   Los resultados obtenidos refuerzan que la falta de seguridad social, la baja escolaridad y vivir en zonas rurales son factores estructurales que incrementan la mortalidad materna, en concordancia con la literatura médica y epidemiológica nacional e internacional.

4. **El uso de la minería de datos y el proceso KDD aporta valor agregado.**

   La aplicación de técnicas y algoritmos no supervisados permitió descubrir patrones invisibles que estaban ocultos en las bases de datos; permitiendo ver más allá de los reportes estadísticas tradicionales, demostrando el potencial de la ciencia de datos para fortalecer la vigilancia epidemiológica y la toma de decisiones en salud publica basada en evidencia.

5.2 Recomendaciones

1. **Sistema de salud:**

   - Implementar auditorias permanentes de calidad en los hospitales y centros de salud para reducir las muertes maternas en el **Perfil Base**.

   - Fortalecer la atención obstétrica los fines de semana y garantizar la cobertura continua en el **Perfil Vulnerable**.

- Desarrollar programas focalizados de prevención y acceso a servicios para adolescentes, mujeres mayores y comunidades rurales sin seguridad social del **Perfil Crítico**.

2. **Gestión de la Información.**

    - Incorporar algoritmos de minería de datos en los Sistemas Nacionales de Vigilancia Epidemiológica (SEDD, DGIS, SINAIS, etc), con el fin de identificar tempranamente (en el momento que se crea la información) perfiles de riesgo.

    - Promover la interoperabilidad y el uso del expediente clínico electrónico con estandarización de datos, facilitando análisis automatizados en tiempo real.

3. **Futuras Investigaciones.**

    - Ampliar el análisis hacia bases de datos mixtas que incluyan casos no fatales (Egresos Hospitalarios, consultas prenatales), permitiendo aplicar **modelos predictivos supervisados**.

    - Explorar técnicas de minería de datos hibridas o mixtas (clustering + reglas de asociación) para profundizar en las relaciones entre factores clínicos y sociales.

Los conocimientos o resultados obtenidos son una base sólida para diseñar e implementar **estrategias diferenciadas de prevención y atención obstétrica en Tabasco**, con potencial de replicarse a nivel nacional. Además, la investigación muestra que la integración de minería de datos al sector no solo es viable, sino necesaria para enfrentar problemas complejos como la mortalidad materna, contribuyendo a salvar vidas y garantizar el derecho a la salud de las mujeres mexicanas.

Referencias Citadas

Acevedo Morales, S. I. (2024). *Minería de datos: Descubriendo tesoros en montañas de información*. *Revista Digital UDEMEX, 5*(2), 20–21.

Amazon Web Services. (2024). *Supervised vs. unsupervised learning*. AWS Docs. https://aws.amazon.com/es/what-is/machine-learning/

Cámara de Diputados. (2017). *Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados*. DOF, México.

De Jesús, C. (2024). *La investigación cuantitativa*. Bogotá, D.C.: Corporación Universitaria de Asturias.

Dirección General de Información en Salud. (2024). *Cubos dinámicos: egresos hospitalarios*. Secretaría de Salud. http://www.dgis.salud.gob.mx/cubos/

Ferraris, L., Gabbanelli, I., Mileta, M., & Seija, G. (2023). *Aplicación del proceso KDD en la predicción de eventos en salud*. *Journal of Health Informatics, 12*(2), 101–115.

Flores Guerrero, D. (2019). *Aplicación de minería de datos para el pronóstico de la evolución de la diabetes en México* [Tesis de licenciatura]. Tecnológico Nacional de México. https://rinacional.tecnm.mx/handle/TecNM/1377

García-Marco, F.-J. (2011). *La pirámide de la información revisitada: enriqueciendo el modelo desde la ciencia cognitiva*. *El Profesional de la Información, 20*(1), 11–24. https://doi.org/10.3145/epi.2011.ene.02

IBM. (2024). *What is data mining?*. IBM. https://www.ibm.com/analytics/data-mining

IDECA (Bogotá). (2019). *Metodología para la analítica de datos. Guía de buenas prácticas*. Bogotá: IDECA.

Instituto Nacional de Estadística y Geografía. (2020). *Censo de población y vivienda 2020*. https://www.inegi.org.mx/programas/ccpv/2020/

Lengua-Cantero, C., Lambraño Pérez, L., Solorzano Peralta, N., García Medina, M., & Acosta Meza, D. (2024). *Inteligencia artificial y la gestión de requerimientos de software educativo: SGR*. *Revista Ibérica de Sistemas y Tecnologías de Información, E71*, 1–8. https://doi.org/10.17013/risti.e71.1-8

Morán Linares, D. M., & Casas Anaya, F. (2024). *La importancia de contar con conocimientos en cuanto a adquisición de datos con Arduino y procesamiento con Python*. *Revista Digital UDEMEX*, *5*(2), 43–47.

Moujahid, M., & Inza, I. (2010). *Aplicación de WEKA en bioinformática*. *Revista Iberoamericana de Inteligencia Artificial*, *14*(47), 65–75.

National Academies Press. (2000). *The Consequences of Maternal Morbidity and Mortality*. Washington, DC: NAP. https://nap.nationalacademies.org/catalog/9800/the-consequences-of-maternal-morbidity-and-mortality

Ochoa, L. L., Rosas Paredes, K., & Baluarte Araya, C. (2017, July 19–21). *Exploración de datos académicos utilizando herramientas de minería de datos* [Conferencia]. *15th LACCEI International Multi-Conference for Engineering, Education, and Technology*, Boca Ratón, FL, United States. LACCEI. https://doi.org/10.18687/LACCEI2017.1.1.116

Organización Mundial de la Salud. (2012). *Tendencias en mortalidad materna: 1990 a 2010.* Ginebra: OMS. Recuperado de https://apps.who.int/iris/bitstream/handle/10665/44874/9789241503631_eng.pdf

Organización Mundial de la Salud. (2019). *Mortalidad materna: Informe de evidencia.* Ginebra: OMS, Departamento de Salud Reproductiva e Investigaciones Conexas (WHO/RHR/19.20). Recuperado de https://iris.who.int/bitstream/handle/10665/329886/WHO-RHR-19.20-eng.pdf

Organización Mundial de la Salud. (2024). *Mortalidad materna: Hoja informativa.* Ginebra: OMS. Recuperado de https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

Organización Panamericana de la Salud. (2018). *Indicadores básicos 2018: Situación de salud en las Américas*. Washington, DC: OPS. https://iris.paho.org/handle/10665.2/49511

Pérez Trujillo, C. (2021). *Aplicación de minería de datos para predecir condiciones en adultos mayores* [Tesis de maestría]. Universidad Veracruzana.

Population Reference Bureau. (2014). *World Population Data Sheet*. PRB. https://www.prb.org/resources/2014-world-population-data-sheet/

Raghupathi, W., & Raghupathi, V. (2014). *Big data analytics in healthcare: Promise and potential*. *Health Information Science and Systems, 2*(1), 3.

Rhodes, A., Evans, L. E., Alhazzani, W., et al. (2017). *Surviving sepsis campaign: International guidelines*. *Critical Care Medicine, 45*(3), 486–552. https://doi.org/10.1097/CCM.0000000000002255

Romero Zaldívar, Y., Ramírez Pérez, J. F., & Soto Pelegrín, L. (2022). *La minería de datos en apoyo a la toma de decisiones clínicas*. *Revista Cubana de Transformación Digital, 3*(2), e136.

Secretaría de Salud. (2012). *Norma Oficial Mexicana NOM-004-SSA3-2012, del expediente clínico.* Diario Oficial de la Federación. México. Recuperado de https://www.dof.gob.mx/nota_detalle.php?codigo=5272787&fecha=15/10/2012

Secretaría de Salud. (2012). *Norma Oficial Mexicana NOM-024-SSA3-2012, del expediente clínico electrónico. Intercambio de información en salud.* Diario Oficial de la Federación. México. Recuperado de https://www.dof.gob.mx/nota_detalle.php?codigo=5280847&fecha=30/11/2012

Toscano de la Torre, J., et al. (2016). *Aplicación de minería de datos para la identificación de factores de riesgo asociados a la muerte fetal*. *Revista Mexicana de Ginecología y Obstetricia*.

Valencia Ramón, J. (2022). *Prototipo de sistema de información hospitalaria con base en el estándar HL7 homologado* [Tesis de Maestría]. Universidad Autónoma del Estado de México.

Wu, W.-T., Li, Y.-J., Feng, A.-Z., Li, L., Huang, T., Xu, A.-D., & Lyu, J. (2021). *Data mining in clinical big data: the frequently used databases, steps, and methodological models*. *Military Medical Research, 8*(1), 44. https://doi.org/10.1186/s40779-021-00338-z

Anexos

Anexo 1 Salida de WEKA para el modelo EM (3 clusters, 7 atributos)

A continuación, mostramos la salida del algoritmo EM que aplicamos a los registros que abordamos en nuestro estudio en la base de datos de la DGIS de los años 2002 al 2022.

```
=== Run information ===



Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-
6 -K 10 -num-slots 1 -S 100

Relation:                              Algoritmo     EM2     TAB     mortalidad_materna_2002_2022-
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-
cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Instances:    544

Attributes:   7

              ï»¿GpoEdad

              EstadoCivil

              Escolaridad2

              Derechohabiencia2

              DiaSemana

              SubRegion

              cluster

Test mode:    evaluate on training data
```

```
=== Clustering model (full training set) ===


EM

==


Number of clusters selected by cross validation: 3

Number of iterations performed: 36


                         Cluster
Attribute                     0        1        2
                          (0.45)   (0.44)   (0.11)
=================================================
GpoEdad
  AdultaJoven           185.9157 160.4486  17.6357
  Adolescente            41.9162  24.3092  17.7746
  AdultaMayor            19.0018  58.3453  24.6529
  [total]               246.8337 243.1031  60.0631
EstadoCivil
  ConPareja             208.9303 213.3982  51.6715
  SinPareja              36.9034  28.7049   7.3917
```

[total]	245.8337	242.1031	59.0631
Escolaridad2			
Baja	1.0049	175.2687	34.7264
Media	243.8762	1.8183	23.3056
Alta	1.9527	66.0161	2.0312
[total]	246.8337	243.1031	60.0631
Derechohabiencia2			
SinSeguridadSocial	39.0116	82.4887	36.4998
ConSeguridadSocial	206.8222	159.6144	22.5634
[total]	245.8337	242.1031	59.0631
DiaSemana			
lunes	44.9353	28.8087	2.256
domingo	31.028	45.5961	17.3759
sabado	23.0067	27.9869	19.0064
jueves	36.9619	28.7821	7.2561
miercoles	38.9462	34.6207	8.4331
martes	40.9744	45.1656	1.86
viernes	34.9812	36.1431	7.8757
[total]	250.8337	247.1031	64.0631
SubRegion			
Centro	188.943	203.5962	34.4608
LosRios	6.9943	6.7678	5.2379
Chontalpa	32.9359	23.2712	16.7929

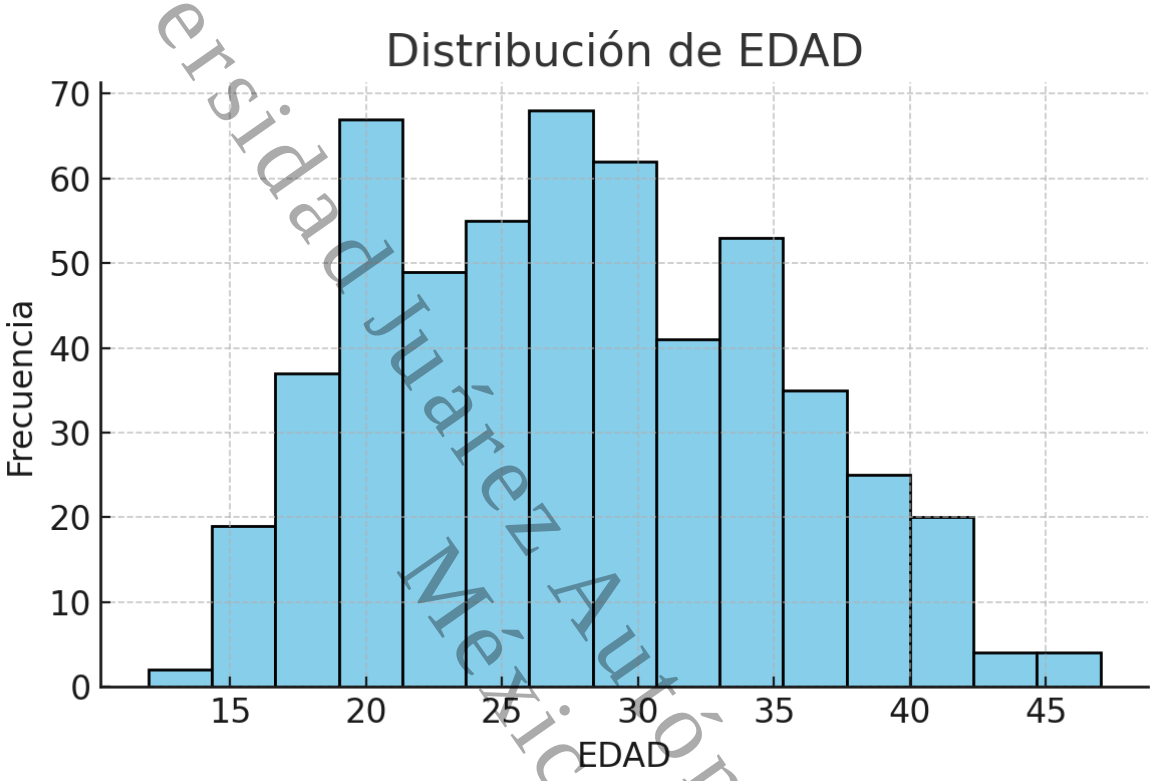
```
  Pantanos                10.9675   9.0663   3.9662
  Sierra                   8.9931   2.4015   1.6054
  [total]                248.8337 245.1031  62.0631
cluster
  cluster1                 1.168 241.0409  57.7911
  cluster2               244.6658   1.0622   1.2721
  [total]                245.8337 242.1031  59.0631



Time taken to build model (full training data) : 1.4 seconds


=== Model and evaluation on training set ===


Clustered Instances


0      244 ( 45%)
1      261 ( 48%)
2       36 (  7%)



Log likelihood: -5.55222
```

Anexo 2 Gráficas descriptivas de los datos

Con el propósito de complementar la salida del algoritmo EM presentada en el anexo 1 y entender mejor los datos a continuación mostramos una graficas descriptivas, elaboradas a partir de la base de datos de mortalidad materna (DGIS, 2002-2022). Estas graficas nos permite tener una mejor visualización de la distribución y las tendencias de las variables incluidas en el análisis, tales como edad, escolaridad, estado conyugal, año y mes de defunción.

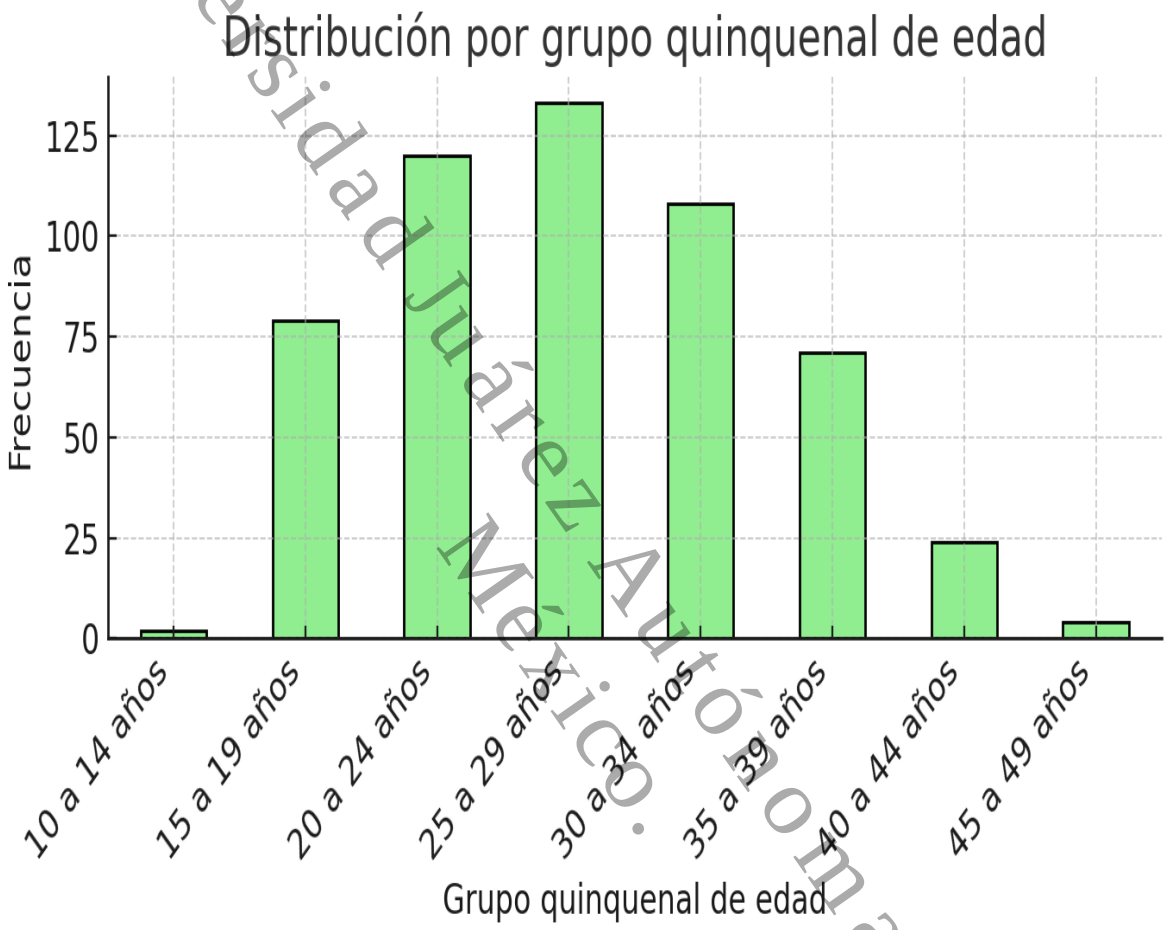
Este análisis resulta fundamental para el desarrollo del proceso KDD; ya que permite la detección de patrones preliminares, la identificación preliminar de patrones y la comprensión de características de la población seleccionada.

Figura A2.1 Distribución de la edad de las mujeres fallecidas por causas maternas en Tabasco (2002–2022).



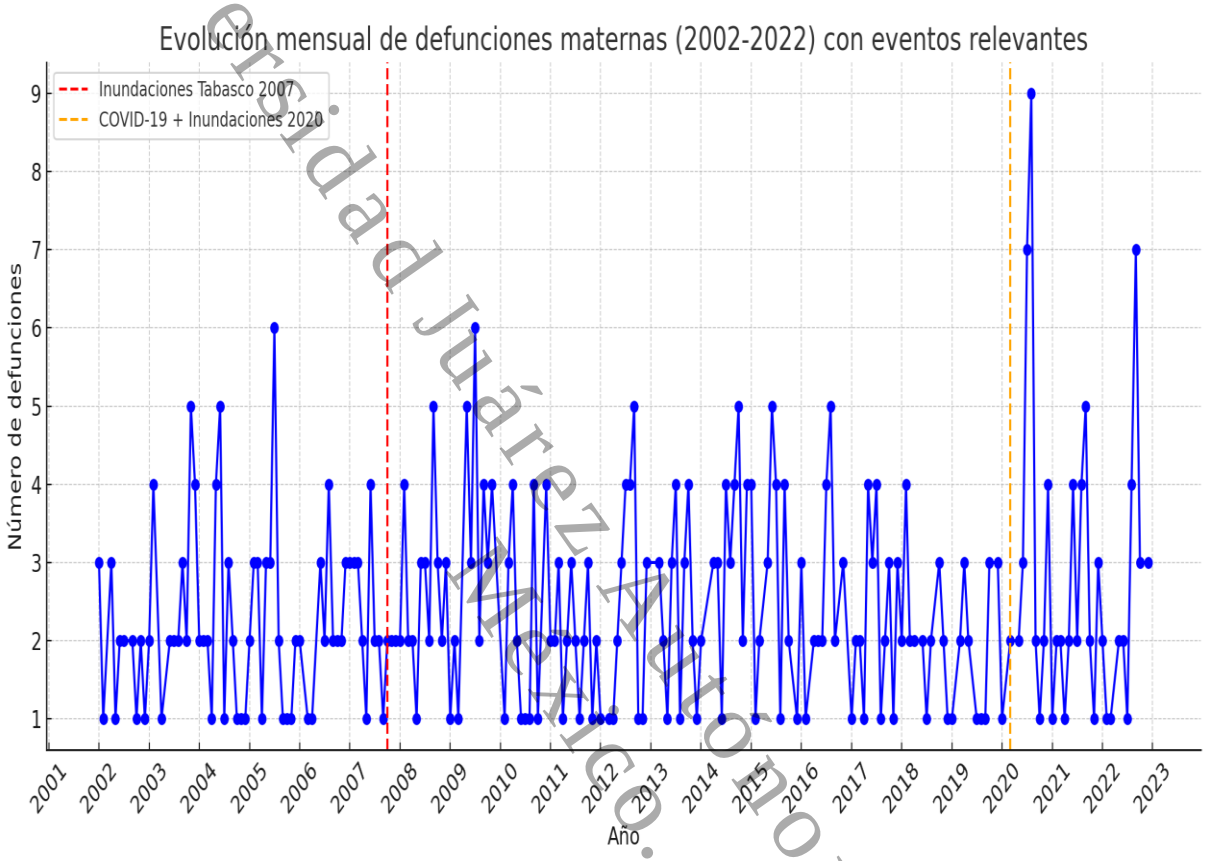
Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Figura A2.2 Distribución por grupo quinquenal de edad de las defunciones maternas en Tabasco (2002–2022).



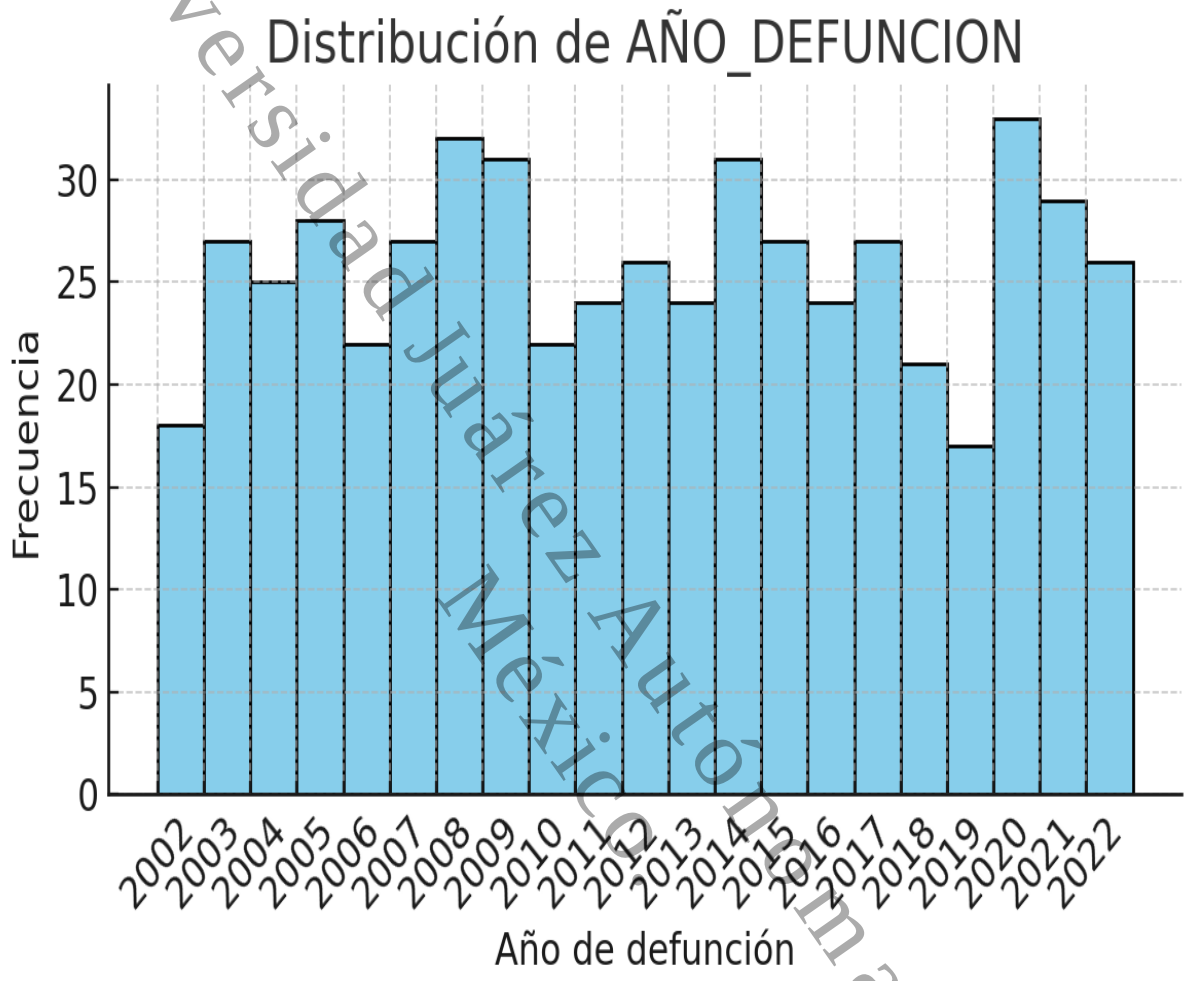
Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Figura A2.3 Evolución mensual de defunciones maternas en Tabasco (2002–2022), con eventos relevantes (inundaciones 2007, pandemia COVID-19 2020).



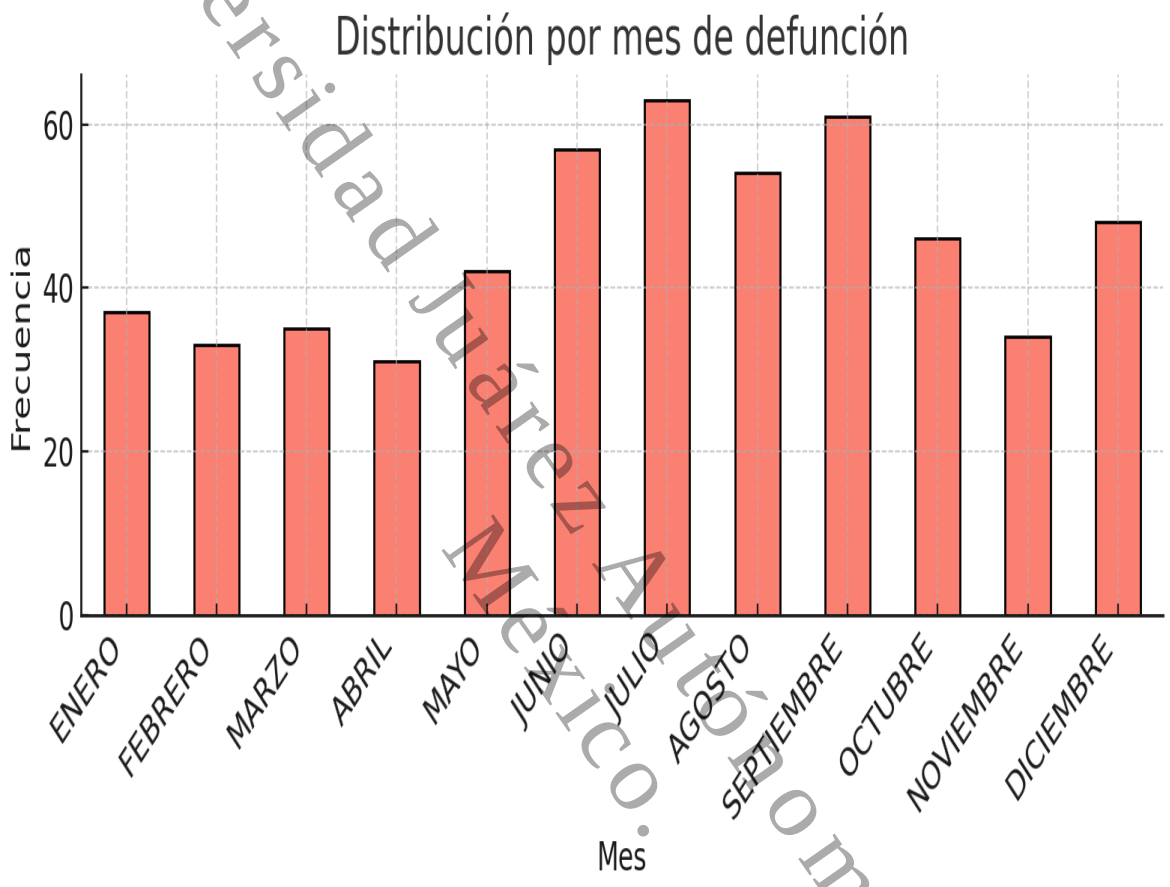
Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Figura A2.4 Distribución anual de las defunciones maternas en Tabasco (2002–2022).



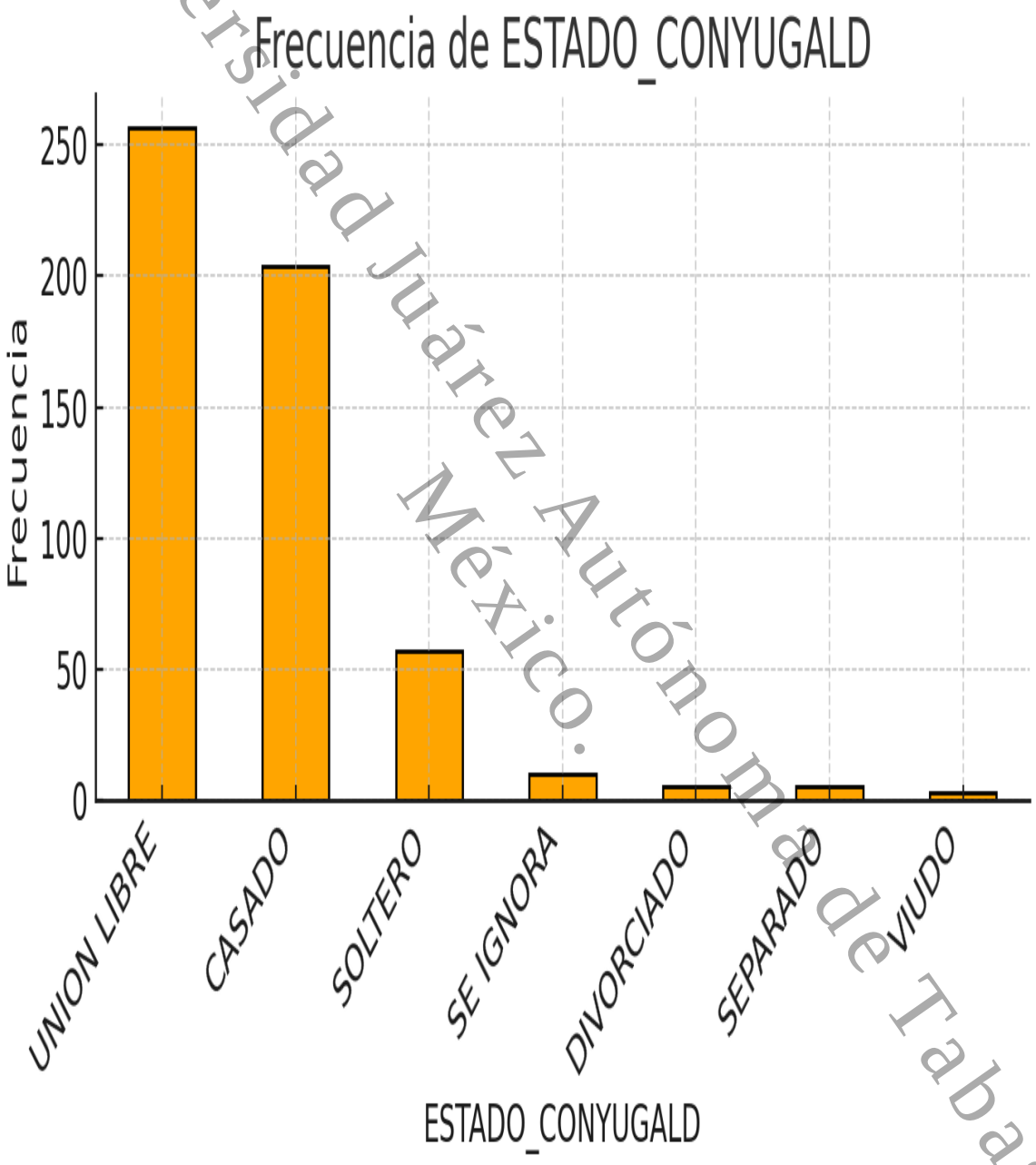
Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Figura A2.5 Distribución mensual de las defunciones maternas en Tabasco (2002–2022).



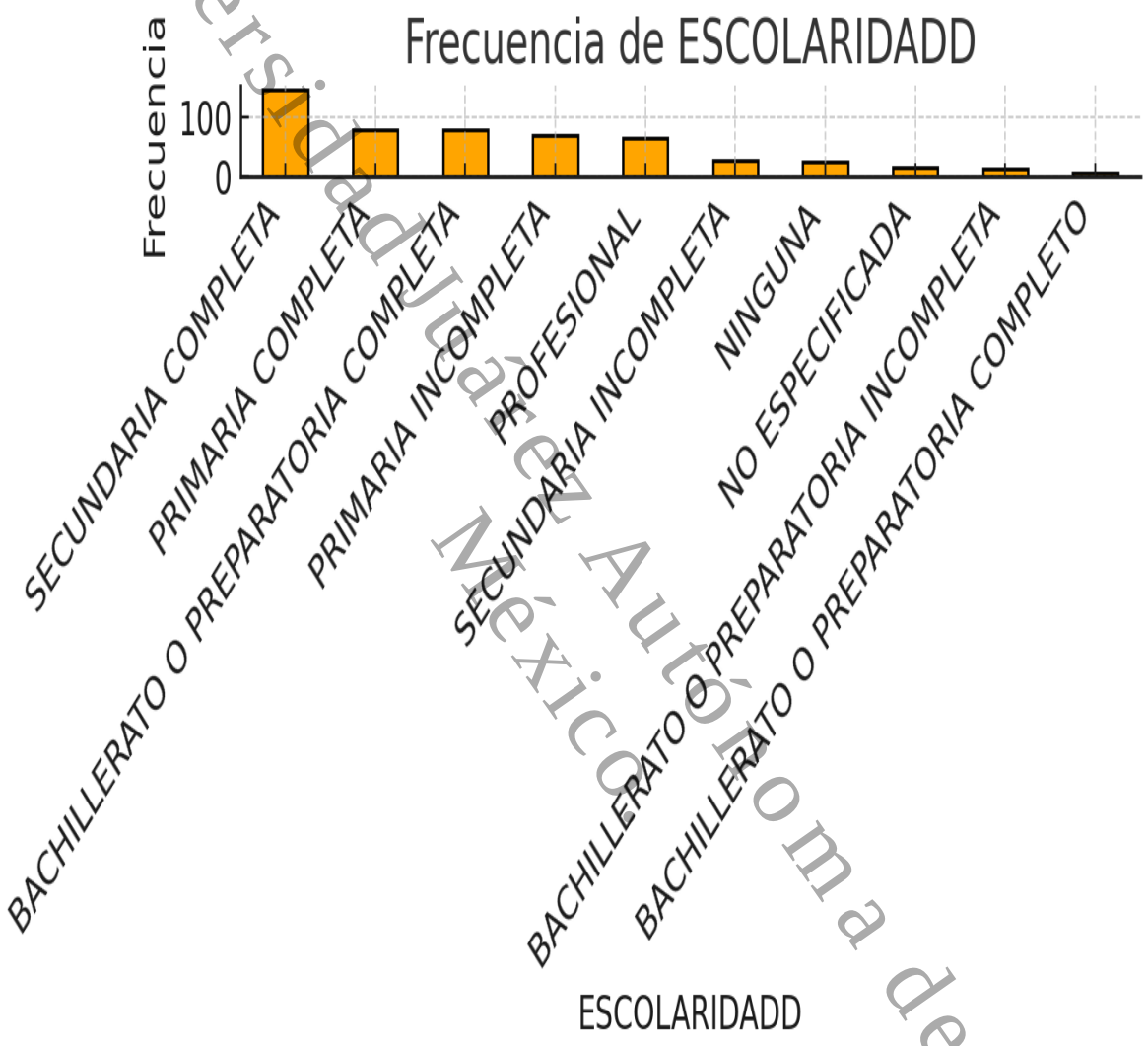
Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Figura A2.6 Frecuencia de estado conyugal de las mujeres fallecidas por causas maternas en Tabasco (2002–2022).



Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Figura A2.7 Frecuencia de escolaridad de las mujeres fallecidas por causas maternas en Tabasco (2002–2022).



Nota: Fuente: Elaboración propia con base en datos de la DGIS (2002–2022).

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la Tesis:	Descubrimiento del Conocimiento en Bases de Datos para la Prevención de la Mortalidad Materna
Autor:	L.C. Fredy López Meneses
ORCID:	https://orcid.org/0009-0005-3594-2086
Resumen:	El presente estudio analiza los registros abiertos de la Dirección General de Información en Salud (DGIS) correspondientes a las muertes maternas ocurridas en el estado de Tabasco entre 2002 y 2022. Se aplicó la metodología de Descubrimiento del Conocimiento en Bases de Datos (KDD) con un enfoque cuantitativo, exploratorio y descriptivo, empleando el algoritmo no supervisado Expectation–Maximization (EM) en la herramienta WEKA. El análisis permitió identificar tres perfiles diferenciados de riesgo de mortalidad materna: (1) mujeres jóvenes con escolaridad media y seguridad social, donde el principal factor es la calidad de la atención hospitalaria; (2) mujeres con baja o alta escolaridad, mayormente sin seguridad social, con muertes más frecuentes en fines de semana, asociadas a desigualdades de cobertura; y (3) un grupo crítico compuesto por adolescentes y adultas mayores, con baja escolaridad, sin seguridad social y residentes en zonas rurales dispersas. Estos hallazgos evidencian que la mortalidad materna responde a patrones sociodemográficos, territoriales y de acceso a servicios.
Palabras Clave:	Mortalidad materna, Minería de datos, KDD, EM, Tabasco.
Referencias Citadas:	En las páginas 54-57 se muestran las referencias.