# UNIVERSIDAD DE SONORA
# DIVISIÓN DE INGENIERÍA


POSGRADO EN INGENIERÍA INDUSTRIAL
**MAESTRÍA EN INGENIERÍA EN SISTEMAS Y TECNOLOGÍA**

REFINAMIENTO DE UN ALGORITMO DE MINERÍA DE DATOS PARA MEJORAR LA ATENCIÓN DE MEDICINA PREVENTIVA EN PACIENTES DIABÉTICOS

# T E S I S

PRESENTADA POR

**OMAR FERNANDO GARCÍA MORA**

Desarrollada para cumplir con uno de los
requerimientos parciales para obtener
el grado de Maestro en Ingeniería

DIRECTOR DE TESIS
DR. FEDERICO CIRETT GALÁN

CODIRECTOR
DR. MARIO BARCELO VALEZUELA

HERMOSILLO, SONORA, MÉXICO. DICIEMBRE 2020

# Universidad de Sonora

Repositorio Institucional UNISON


"El saber de mis hijos hará mi grandeza"

Hermosillo, Sonora a 15 de julio de 2020

**OMAR FERNANDO GARCÍA MORA**

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado vigente, otorgamos a usted nuestra aprobación de la fase escrita del examen de grado, como requisito parcial para la obtención del Grado de Maestro en Ingeniería: Ingeniería en Sistemas y Tecnología.

Por tal motivo este jurado extiende su autorización para que se proceda a la impresión final del documento de tesis: **REFINAMIENTO DE UN ALGORITMO DE MINERÍA DE DATOS PARA MEJORAR LA ATENCIÓN DE MEDICINA PREVENTIVA EN PACIENTES DIABÉTICOS** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

Dr. Federico Miguel Cirett Galán
Director de tesis y Presidente del jurado

Dr. Mario Barceló Valenzuela
Codirector y Vocal del Jurado

Dra. Raquel Torres Peralta
Secretaria del Jurado

Dr. René Francisco Navarro Hernández
Vocal del Jurado

c.c.p. Archivo

UNIVERSIDAD DE
MONTEMORELOS

DIRECCIÓN DE
POSGRADO E INVESTIGACIÓN

Ave. Libertad #1300 Pte.
Barrio matamoros,
Montemorelos, N.L.
México. C.P. 67510

(826) 263-0900
ext. 1510,1511,1512 y 1513

Montemorelos, Nuevo León, México, a 17 de noviembre de 2020

**OMAR FERNANDO GARCÍA MORA**

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado de la Universidad de Sonora, otorgo a usted mi aprobación de la fase escrita del examen profesional, como requisito parcial para la obtención del Grado de Maestro en Ingeniería: Ingeniería en Sistemas y Tecnología.

Por tal motivo, como sinodal externo y vocal del jurado, extiendo mi autorización para que se proceda a la impresión final del documento de tesis: **REFINAMIENTO DE UN ALGORITMO DE MINERÍA DE DATOS PARA MEJORAR LA ATENCIÓN DE MEDICINA PREVENTIVA EN PACIENTES DIABÉTICOS** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

DR. RAMÓN ANDRÉS DÍAZ VALLADARES
UNIVERSIDAD DE MONTEMORELOS
Sinodal Externo y Vocal del Jurado

DIRECCIÓN DE POSGRADO E INVESTIGACIÓN

# RESUMEN

El presente trabajo de investigación busca mejorar los resultados de un algoritmo de minería de datos dedicado a la atención de pacientes diabéticos por parte del departamento de medicina preventiva en un instituto de salud pública. El modelo que actualmente está en funcionamiento utiliza tres variables que se obtienen de la base de dato de citas a médico general. Con este trabajo se pretende explotar otras características que ya existen en la base de datos con el objetivo agregarlas a las que ya se consideran para describir mejor a los pacientes y poder brindar información más precisa.

La estrategia seguida se basa en la adaptación de la metodología CRISP-MED-DM de Niaksu (2015), la cual consta de seis etapas: 1) análisis del problema, 2) análisis de datos, 3) preparación de datos, 4) modelado, 5) evaluación y 6) despliegue. Dentro de la segunda etapa, análisis de datos, se hace uso de la metodología de Buczak et al. (2012) para asignar importancia a las variables de la base de datos según el objetivo que busca la investigación, en este caso para determinar enfermedades relacionadas a la diabetes.

Con la aplicación de esta estrategia se logra presentar un comparativo cuantitativo con métricas que permiten determinar qué tan homogéneos son los clústers de los algoritmos comparados, así como describir de acuerdo con las características seleccionadas a los grupos resultantes.

Al haber trabajado con una base de datos generada a partir de citas con médico general, varias de las enfermedades relacionadas no tuvieron la frecuencia necesaria para impactar en el modelo. Sin embargo, se agregó en la etapa del despliegue un informe que incluyen estos padecimientos para generar conocimiento de su comportamiento en pacientes diabéticos, así como en personas sin el diagnóstico de esta enfermedad.

# ABSTRACT

This research work seeks to improve the output of a data mining algorithm that supports the care of diabetic patients by the department of preventive medicine in a public health institute. The program that is currently in operation uses three variables that are obtained from a general medical appointment database. This work aims to find other characteristic in the database to add them to those already considered to better describe patients to provide more accurate information.

The strategy is based on the adaptation of the CRISP-MED-DM methodology developed by Niaksu (2015), which consists of six stages: 1) problem understanding, 2) data understanding, 3) data preparation, 4) modelling, 5) evaluation and 6) deployment.

In the second stage, data analysis, different sources are consulted to assign importance to the variables of the database according to the objective sought by the research, as indicated by the Buczak et al. methodology (2012). In this case to determine diseases related to diabetes mellitus.

With the application of this strategy, it is possible to perform a quantitative comparison with metrics that allow determining how homogeneous the clusters of the compared algorithms are, as well as describing the resulting groups according to the characteristics selected.

Having worked with a general medical appointment database, several of the related diseases were not frequent enough to impact the algorithm. However, a report that includes these conditions was added at the deployment state to generate knowledge of their behavior in diabetic patients, as well as in people without the diagnosis of diabetes mellitus.

# AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa de Fortalecimiento de la Calidad Educativa (PFCE) por su apoyo económico brindado en mi estudio de posgrado.

A la Universidad de Sonora y a departamento de Ingeniería Industrial, a los profesores que en cada materia aportaron algo a este trabajo, a mis compañeros de maestría y, en especial, a mi director de tesis, el Dr. Federico Cirett y a la Dra. Raquel Torres, por haberme guiado aportando experiencia y conocimiento en este proyecto.

A mi familia y a mi novia, por el apoyo que siempre me han brindado.

# ÍNDICE GENERAL

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# 1. INTRODUCCIÓN

En la última década las técnicas de minería de datos se han convertido en un factor importante en el sector de la salud para descubrir patrones que se pueden utilizar en el diagnóstico y la toma de decisiones para la atención del paciente.

El presente capítulo plantea la descripción de la institución y, en específico, de la problemática para cual se presenta una estrategia para dar solución a su situación, además, se plantean los objetivos y la hipótesis relacionada a esta investigación.

## 1.1. Presentación

El proyecto se desarrolló en el Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado de Sonora (ISSSTESON), una institución de salud pública que presta servicios de seguridad social a un aproximado de 180,000 derechohabientes. De esta cantidad, aproximadamente el 44.00% reciben atención en la coordinación Hermosillo.

La institución cuenta con una base de datos con el registro de consultas a médico general, la cual se ha ido recabando desde hace varios años. En este proyecto se propone explotar estos registros para ampliar el conocimiento del estado de salud de los derechohabientes y, en específico, optimizar el modelo actual de medicina preventiva en pacientes diabéticos y/o pacientes con riesgo de padecer diabetes.

La metodología utilizada para el programa actual trabaja con el algoritmo K-medias, esta decisión se fundamenta en que demanda poco tiempo para poder ser implementado en comparación con otros algoritmos, por lo que no se estudiaron resultados con otros algoritmos, pudiendo existir otro más eficiente respecto a tiempo y con mejor precisión. Entre 2014 y 2017 se detectaron un total de 2,748 casos totales; 1,401 en 2014, 564 en 2015, 462 en 2016 y 321 en 2017 (Sanez, 2018).

El algoritmo que actualmente está en uso tiene un amplio rango de mejora en la forma en que describe a la población, es decir, las variables que considera para realizar las

predicciones. Dicho modelo basa sus decisiones en edad, sexo y herencia, o sea, familiares con padecimientos de obesidad y diabetes. Sin embargo, la mejora del algoritmo es necesaria ya que hay otros factores que pueden ser de utilidad para hacer el modelo más robusto, esto con el objetivo de describir de mejor manera a los diferentes segmentos de población y mejorar la predicción. Lo anterior agregando nuevas variables derivadas del análisis de la información que se encuentra en la base de datos de la institución, como lo son las enfermedades relacionadas (ER) antes y después de padecer diabetes.

Según el modelo actual, la edad es un factor que influye en sufrir padecimientos de diabetes, siendo las personas mayores quienes tienen más riesgo de padecer esta enfermedad. Por lo cual, pacientes de edad avanzada que no cuentan con padres y madres afiliados, impide hacer un análisis de la herencia de estos.

Al momento, el algoritmo trabaja con cinco grupos distintos, de los cuales dos con un rango de 20 años o más entre la edad mínima y máxima de los integrantes que los componen; lo cual entra en conflicto con el objetivo de la estrategia, es decir, la información que se les hace llegar a personas de estos grupos no es totalmente personalizada.

La forma en la que están compuestos los diferentes grupos en cuanto a integrantes muestra un desequilibrio que puede estar ocasionando menor precisión en la predicción, además de sesgo en el proceso de tomas de decisiones. Esto, debido a que existe una diferencia de 535 integrantes entre el grupo con mayor cantidad de integrantes comparado con el grupo minoritario.

## 1.2. Planteamiento del problema

El algoritmo actual basa sus predicciones en edad, sexo y herencia del paciente, descartando otros datos para generar predicciones más certeras. Esto causa que exista conocimiento en la BD que no se está considerando para describir los segmentos de la población. Además, la manera en que están conformados dichos

segmentos presenta un desequilibrio de clases que pueden afectar las predicciones y generando un sesgo en el proceso de toma decisión.

## 1.3. Objetivo general

Refinar la precisión de la salida del algoritmo perteneciente al programa de medicina preventiva para dar seguimiento a pacientes con un bajo, mediano y alto riesgo de padecer diabetes mediante la inclusión de nuevas variables y el rediseño de agrupamientos.

## 1.4. Objetivos específicos

- Analizar las características de los pacientes con diabetes derivado de las citas con médico general.
- Revisar y seleccionar técnicas de minería de datos para el procesamiento de información.
- Afinar el algoritmo incluyendo nuevas variables y rediseño agrupamientos.
- Evaluar el algoritmo propuesto con nuevas variables respecto al algoritmo previo, validando que la distancia promedio de los elementos de un grupo a su centroide sea menor respecto al algoritmo anterior.

## 1.5. Hipótesis

El refinamiento del algoritmo mediante técnicas de minería de datos a través de la consideración de nuevas variables y la mejora en el agrupamiento de pacientes con riesgo de padecer diabetes permitirá generar resultados más precisos para el área de medicina preventiva.

## 1.6. Alcances y delimitaciones

El proyecto se desarrollará en el programa de medicina preventiva, específicamente en el algoritmo desarrollado para detección de pacientes con riesgo alto, medio y bajo de padecer diabetes. Se trabajará con datos de citas de médico general realizados

entre 2014 y 2017, esto en derechohabientes que pertenecen a la coordinación de Hermosillo.

## 1.7. Justificación

El proyecto tendrá un impacto positivo en la calidad de vida de la población, ya que está enfocado en la diabetes mellitus, la segunda mayor causa de muerte a nivel nacional y tercera a nivel estatal; Sonora contribuye con el 1.52% del total de defunciones por dicha enfermedad en el país (INEGI, 2018).

Además, repercute en la detección de pacientes que ya sufren de la enfermedad, pero no han sido detectados, el cual corresponde al 50% de las personas que padecen de dicho padecimiento, según la Federación Internacional de Diabetes (FID) (2019).

Al estar enfocado en medicina preventiva el impacto de este proyecto se puede reflejar en ahorros para la institución ya que se puede llegar a evitar consultas, tratamientos y medicamentos en derechohabientes.

# 2. MARCO DE REFERENCIA

En esta sección se presenta los fundamentos teóricos del problema a tratar, mismos que sirven como sustento de la metodología propuesta en la sección 3. Primero se presenta información sobre salud pública y medicina preventiva, seguido por una explicación amplia de la diabetes mellitus y su impacto en el mundo. Además, se exponen los temas de gestión del conocimiento (GC), descubrimiento de conocimiento en bases de datos y minería de datos (MD) para destacar la diferencia entre ellos. También, se incluyen temas de minería de datos, sus fundamentos, los retos de ésta en el sector salud, clasificación de algoritmos y técnicas, donde se explica específicamente el algoritmo k-medias. Por último, se incluyen estudios donde se ha aplicado la MD en el sector salud.

## 2.1. Salud Pública y Medicina Preventiva

Según la Asociación Americana de Salud Pública (2019) el objetivo de toda institución dedicada a la salud pública es promover y proteger la salud de las personas y las comunidades donde viven, aprenden, trabajan y juegan, por lo tanto, deben establecer estándares de seguridad que ayuden a rastrear brotes de enfermedades, prevenir lesiones y alertar sobre por qué algunas personas son más propensas de sufrir problemas de salud que otras. Un papel bien desempeñado de la salud pública ahorra dinero, mejor la calidad de vida y ayuda a reducir el sufrimiento humano.

En adición, Piédrola (2015) menciona que la salud pública se define en términos de sus objetivos: disminuir la incidencia de enfermedades, así como mantener y promover la salud de la sociedad a través de esfuerzos organizados de la comunidad. Uno de estos esfuerzos es el que aporta el sistema de salud en su doble vertiente, preventiva y asistencial.

Normalmente, se define la medicina como el arte y la ciencia de curar y prevenir enfermedades, es decir, tiene dos vertientes, la preventiva, que abarca actuaciones e indicaciones médicas dirigidas específicamente a la prevención de la enfermedad, y la

curativa, que se refiere a las actuaciones médicas dirigidas a la curación de la enfermedad (Salleras, 1985).

La medicina preventiva se considera menos amplia que la salud pública, su objetivo es la prevención de las enfermedades mediante intervenciones y consejos médicos que ayuden en la defensa y mantenimiento de la salud poblacional (Piédrola, 2015).

García Pérez y García Bertrand (2012) clasifican la medicina preventiva en cuatro diferentes tipos:

- **Prevención primaria:** Es el conjunto de actividades sanitarias que se ejecutan tanto por la comunidad o los gobiernos como por el personal de salud antes de que aparezca una determinada enfermedad.
- **Prevención secundaria:** Se refiere a las actividades que se realizan ante la presencia de los demostrados factores de riesgo de las enfermedades.
- **Prevención terciaria:** Es la que actúa sobre el enfermo, logra la curación o lo mantiene controlado para evitar complicaciones.
- **Prevención cuaternaria:** Incluye aspectos de rehabilitación integral necesaria y el conjunto de actividades sanitarias que ayudan a reducir o evitar las consecuencias de las intervenciones innecesarias o excesivas del sistema sanitario.

Los organismos de salud pública se enfrentan un desafío a nivel mundial en padecimientos como la diabetes y otras enfermedades no transmisibles. El gobierno, el sector de salud pública y las instituciones mismas, desempeñan un papel fundamental en la medicina preventiva de tipo primaria, por lo que deben crear sinergias para realizar incursiones sustanciales en la prevención de este tipo de enfermedades (Bergman *et al*., 2012).

## 2.2. Diabetes Mellitus

Según la Organización Mundial de la Salud (OMS) (2019) la diabetes mellitus es una enfermedad crónica que se caracteriza por altos niveles de glucosa en la sangre

resultado de que el páncreas no produce insulina suficiente o porque el organismo no la utiliza eficazmente. La insulina es una hormona permite que la glucosa de los alimentos se convierta en energía.

La Asociación Americana de Diabetes (AAD) (2019) clasifica la diabetes en dos distintos tipos, descritos a continuación:

- **Diabetes tipo 1:** Se presenta cuando el cuerpo no produce insulina. Alrededor del 5% de las personas con diabetes tienen este tipo y es más común el diagnóstico a niños y adultos jóvenes.
- **Diabetes tipo 2:** Es el tipo más común de diabetes, se presenta cuando el cuerpo no produce suficiente insulina o las células no hacen uso de la insulina.

Por su parte, la Federación Internacional de la Diabetes (FID) (2019) propone un tipo de diabetes extra, diabetes gestacional. Este tipo de diabetes consiste en niveles altos de glucosa en la sangre durante el embarazo y se asocia con complicaciones tanto para la madre como para el niño. Generalmente desaparece después del embarazo, pero las mujeres afectadas y sus hijos tienen un mayor riesgo de desarrollar diabetes tipo 2 más adelante en la vida.

Los pacientes con diabetes mellitus tienen una mayor incidencia de enfermedad en varios órganos y tejidos internos. Las enfermedades microvasculares y macrovasculares crónicas tienen mayor influencia en el pronóstico a largo plazo de pacientes con diabetes tipo 2 que las complicaciones agudas. Investigar las asociaciones de estas complicaciones con enfermedades comórbidas mediante el uso de datos de diagnóstico del paciente es útil para predecir su incidencia y, por lo tanto, para tratar de manera más efectiva a los pacientes con diabetes (Kim et al., 2012).

### 2.2.1. Diabetes en el mundo y en México

La diabetes es una de las mayores emergencias sanitarias mundiales del siglo XXI. Se encuentra entre las diez principales causas de muerte a nivel mundial. Se calcula que alrededor de 425 millones de personas en todo el mundo padecen de diabetes,

equivalente al 8.8% de los adultos entre 20 y 79 años. El número aumenta a 451 millones si el rango de edad se considera de 18 a 99 años. Si las tendencias continúan, para el año 2045, 693 millones de personas entre 18 y 99 o 629 millones de personas de 20 a 79 años, tendrán diabetes (Federación Internacional de Diabetes, 2017).

Uno de los principales problemas que presenta esta enfermedad es que alrededor de 212.4 millones de personas, o el 50% del total de quienes padecen diabetes no son diagnosticadas. Dicho grupo utiliza más los servicios médicos en comparación con las personas sin diabetes y, por lo tanto, incurren en mayores gastos sanitarios (Federación Internacional de Diabetes, 2017).

México es el 5to país con mayor número de personas con diabetes en el mundo, con 12 millones, y se estima que en 2045 ocupará el 4to lugar con 21.8 millones de personas con este padecimiento. Se estima que 4.5 millones personas no están diagnosticadas, equivalente al 37.5% del total (Federación Internacional de Diabetes, 2017).

Según el Instituto Nacional de Estadística, Geografía e Historia (INEGI), en México, la mortalidad por diabetes mellitus se ha incrementado constantemente desde 1998 hasta 2018, llegando hasta las 101,257 defunciones, y se posicionó como la causa número dos de mortalidad a nivel nacional (INEGI, 2018).

## 2.3. Sistemas de información en el Sector Salud

La atención médica y la investigación en el sector salud se han visto afectadas por la era de la información de distintas maneras, entre las que destaca la generación de conocimiento que contribuye significativamente a la mejora de la salud mediante el uso de las tecnologías de la información y comunicación (TICs). De acuerdo con la Organización Mundial de la Salud, las TICs para la salud son reconocidas como una de las áreas de salud de más rápido crecimiento (Organización Mundial de la Salud, 2017)

Los Sistemas de Información de Salud o Health Information Systems (HIS) en inglés, proporcionan los fundamentos para la toma de decisiones y tiene cuatro funciones claves: generación de datos, compilación, análisis y síntesis, y comunicación y uso. Los HIS recopilan datos del sector de la salud y otros sectores relevantes, analiza los datos y garantiza su calidad general, relevancia y oportunidad, y convierte los datos en información para la toma de decisiones. Además de ser esencial para el monitoreo y la evaluación, el sistema de información también sirve para fines más amplios, brindando una capacidad de alerta temprana, apoyando la gestión de pacientes y centros de salud (Organización Mundial de la Salud, 2008).

## 2.4. Gestión del conocimiento

El conocimiento es considerado como uno de los activos más importante para las organizaciones, empresas e individuos (Ward y Joe, 2002). El crecimiento de cualquier entidad se basa en la calidad de conocimiento disponible para la toma de decisiones, por lo tanto, las organizaciones deben crear, almacenar y compartir eficientemente el conocimiento (Zhan, 2008).

La gestión del conocimiento (GC) es una amplia combinación de estrategias, herramientas y técnicas que tiene como objetivo transferir el conocimiento desde donde se genera hasta el lugar donde se va a aplicar. La GC incluye todas aquellas tecnologías que surgen de la necesidad de procesar, analizar y aprovechar la información para convertirla en conocimiento. Dentro de la variedad de campos en lo que se basa la GC se encuentran las tecnologías de información, donde surgen el descubrimiento de conocimiento y la minería de datos como herramientas para la creación de conocimiento (Dalkir, 2005). Young (2010) lista cinco pasos claves del proceso de GC: identificar, crear, almacenar, compartir y aplicar conocimiento. Los pasos de identificación y creación son sinónimos del descubrimiento de conocimiento, y forman la base de los procesos productivos de GC en cualquier organización. Lo más importante en los procesos de descubrimiento de conocimiento es la creación de sistemas de conocimiento que pueden ser respaldados por el uso de tecnologías de

minería de datos para el descubrimiento de relaciones entre datos explícitos (Yi et al., 2008). En los segmentos 2.4 y 2.5 se clarifican los conceptos de Minería de Datos (MD) y el de Descubrimiento de Conocimiento en Bases de Datos o Knowledge Discovery in Databases en inglés (KDD).

## 2.5. Descubrimiento de conocimiento en bases de datos

El KDD datos es una intersección de varias disciplinas como estadísticas, bases de datos e inteligencia artificial (Seifert, 2004). Fayyad, Piatetsky-Shapiro y Smyth (1996) define el KDD como el proceso no trivial que busca patrones en datos que sean válidos, novedosos, potencialmente útiles y comprensibles. Por su parte (Luo, 2008), descompone el proceso de KDD los siguientes pasos que se muestran en la figura 2.1, los cuales corresponden a: selección de datos, limpieza de datos (preprocesar), transformación de datos, búsqueda de patrones (minería de datos), interpretación y evaluación.

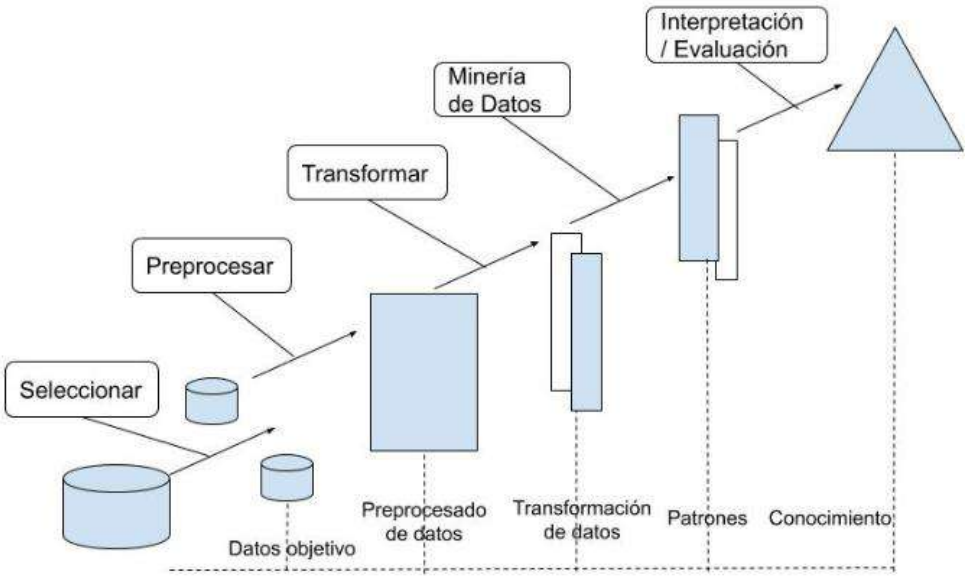



***Figura 2.1.*** *Proceso de KDD, adaptado de Luo (2008).*

Luo (2008) menciona que la búsqueda de patrones que se realiza a través de la MD es la parte central del KDD, por lo que es común el uso de ambas definiciones indistintamente del contexto.

## 2.6. Minería de Datos

Los orígenes de la Minería de Datos (MD) se remontan a partir de los años 60, cuando se basaba simplemente en el procesamiento de archivos (Durairaj y Ranjani, 2013). El crecimiento explosivo de las bases de datos ha creado la necesidad de desarrollar tecnologías que utilicen la información y el conocimiento de manera inteligente. Por lo tanto, la MD se ha convertido en un área de investigación cada vez más importante y es uno de los componentes principales en el proceso de KDD (Mahindrakar y Hanumanthappa, 2013).

Milley (2000) define la MD como el proceso de selección de datos y modelos de exploración y construcción que utilizan vastos almacenes de datos para descubrir patrones previamente desconocidos. En adición, Bhramaramba et al., (2011) mencionan que la MD es la búsqueda de relaciones y patrones globales que existen en grandes bases de datos pero que están ocultos entre la gran cantidad de información, como la relación entre los datos del paciente y su diagnóstico médico. Lamont (2012) destaca que las personas ejercen una función fundamental en el desarrollo de algoritmos, ya que éstos requieren de verificación de resultados y una optimización continua para mejorar su precisión y velocidad.

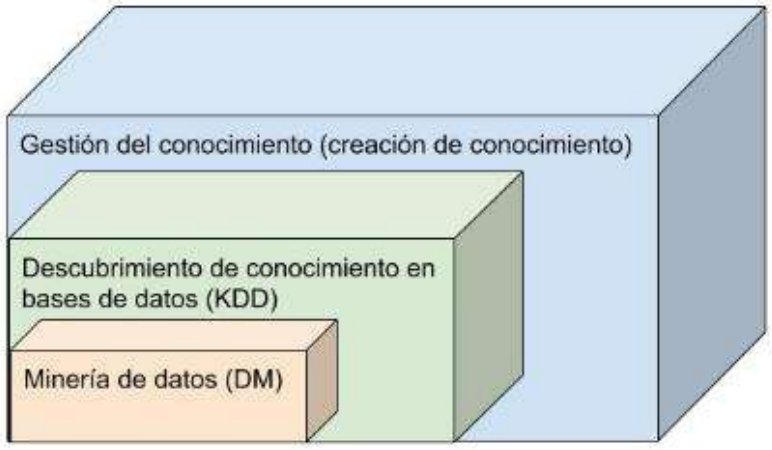



***Figura 2.2.*** *Minería de datos comparado con KDD y Gestión del Conocimiento.*

Además del término MD, existen otros términos más o menos equivalentes utilizados en la literatura, como aprendizaje automático, análisis predictivo y KDD (Iavindrasana, et al., 2009). Sin embargo, se puede concluir que la MD pertenece al KDD, que a su vez, es una herramienta de la fase de creación de conocimiento, como se muestra en la figura 2.2.

### 2.6.1. Minería de datos en el sector salud

La industria de la salud está evolucionando a un ritmo rápido, por lo que día a día se generan cantidades masivas de datos, incluidos registros médicos electrónicos, informes administrativos y otros resultados de evaluación comparativa, sin embargo, en muchas ocasiones los datos están siendo subutilizados (Wickramasinghe, Sharma y Gupta, 2008). Los algoritmos de minería de datos aplicados en la industria de la salud juegan un papel importante en la predicción y el diagnóstico de las enfermedades (Durairaj y Ranjani, 2013).

Derivado de la MD en la industria de la atención médica han surgido algunos sistemas confiables de detección temprana y otros sistemas relacionados con la atención médica a partir de los datos clínicos y de diagnóstico (Jothi, Rahisd y Husain, 2015).

De acuerdo con Durairaj y Ranjani (2013), la aplicación de la MD en el sector salud se puede clasificar en los siguientes grupos:

- **Efectividad del tratamiento**: Se pueden desarrollar aplicaciones de MD para evaluar la efectividad de los tratamientos médicos para comparar y contrastar causas, síntomas y tratamientos.
- **Gestión de la atención médica:** Se utiliza para identificar y rastrear mejor los estados de enfermedades crónicas y los pacientes de alto riesgo, diseñar intervenciones apropiadas y reducir el número de ingresos hospitalarios.
- **Fraude y abuso:** Utilizados identificar patrones de reclamos inusuales o anormales por parte de médicos, clínicas u otros.

- **Industria farmacéutica**: La tecnología se está utilizando para ayudar a las empresas farmacéuticas a desarrollar nuevos productos y servicios.

### 2.6.2. Desafíos de la minería de datos en el sector salud

Diversos autores como Kumar, Govardhan y Srinivas (2014), Bellazi y Zupan (2008), Cios y Moore (2002), entre otros, han destacado que la aplicación de la MD en el sector de la medicina se enfrenta a varias barreras como la comunicación tecnológica, proyectos interdisciplinarios, ética y protección de datos de pacientes. Los desafíos de la MD en el sector salud se describen en la tabla 2.1.

| Desafío | Descripción |
|---|---|
| Variedad de formatos | Los datos médicos por lo general se encuentran en varios tipos de formatos como archivos de video e imagen, archivos de texto, entre otros. Por lo tanto, se requieren preprocesamientos de datos adicionales, actividades de extracción de características o técnicas de MD no estándar para tratar estos tipos de datos. |
| Datos heterogéneos | El análisis de datos de varias especialidades médicas plantea otro tipo de desafíos. En medicina, el mismo concepto semánticamente puede tener múltiples nombres y diferentes identificadores en diferentes sistemas de códigos. Antes de aplicar algoritmos MD, los datos deben integrarse y unificarse semánticamente. En los casos, cuando los sistemas de información usan clasificadores biomédicos estándar, nomenclaturas y ontologías, la tarea de interoperabilidad semántica es definir una ontología común. Los especialistas en MD e informática médica tienen que crear métodos de transformación de datos para garantizar un correcto mapeo de datos semántico. |
| Privacidad de datos | El uso de la información clínica de pacientes es respaldado por leyes que protegen la privacidad de dichos datos, por lo que el hacer uso de ellos con fines de investigación es complicado. Este problema se puede resolver mediante técnicas de despersonalización automática de datos. Los datos utilizados para la investigación no deben incluir el nombre del paciente u otros atributos de identificación. |
| Calidad de los datos | La calidad de las variables disponibles se pueden ver afectados por mediciones inexactas, errores humanos o de equipo. Es por esto por lo que es importante considerar muestras grandes de datos y emplear técnicas de preprocesamiento para identificar y eliminar datos atípicos. |

***Tabla 2.1.*** *Desafíos de los proyectos de MD en el sector salud.*

### 2.6.3. Clasificación de algoritmos de minería de datos

Un algoritmo es un procedimiento paso a paso para realizar un cálculo. La arquitectura de algoritmo se expresa como una lista finita de instrucciones bien definidas para computar una función. En general, son utilizados para el cálculo, el procesamiento de datos y el razonamiento automatizado (Liao, Chu y Hsiao, 2012).

Los algoritmos de MD se clasifican en dos categorías: modelo descriptivo, o aprendizaje no supervisado, y modelo predictivo, o aprendizaje supervisado (Sondwale, 2015).

- **Algoritmos supervisados (predictivos):** Estos deducen reglas de predicción a partir de datos de entrenamiento y aplica las reglas a datos no predichos o no clasificados. En el aprendizaje supervisado, hay dos tipos de tareas de aprendizaje: clasificación y regresión. Los modelos de clasificación intentan predecir distintas clases, mientras que los modelos de regresión predicen valores numéricos (Dunham, 2003). Algunas de las técnicas más comunes son árboles de decisión, aprendizaje de reglas y aprendizaje basado en instancias, tales como k-vecinos más cercanos, algoritmos genéticos, redes neuronales artificiales y máquinas de vectores de soporte (Kavakiotis, et al., 2017).
- **Algoritmos no supervisados (descriptivos):** Agrupan los datos al medir la similitud entre objetos, o registros, y descubre patrones o relaciones desconocidos en los datos para que los usuarios puedan comprender fácilmente una gran cantidad de datos. Este tipo de modelos incluye técnicas como agrupamiento, asociación, resumen y descubrimiento de secuencias (Dunham, 2003).

### 2.6.4. Técnicas de minería de datos en el sector salud

Según Patil (2015), las técnicas de MD como asociación, clasificación y agrupamiento son utilizadas por organizaciones de atención médica para aumentar su capacidad

para construir conclusiones apropiadas con respecto a la salud del paciente a partir de datos y cifras sin procesar.

La asociación tiene un gran impacto en la industria del cuidado de la salud para descubrir las relaciones entre las enfermedades, el estado de la salud humana y los síntomas de la enfermedad (Patel y Patel , 2016).

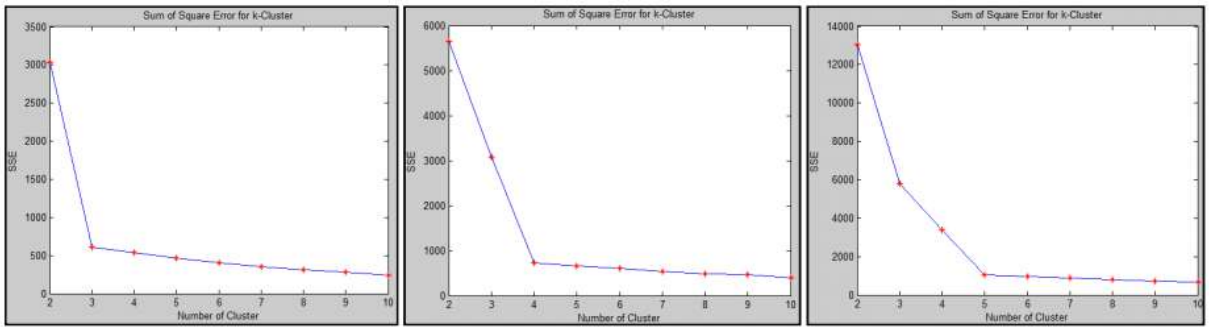
La técnica de clasificación comprende de dos pasos: 1) Entrenamiento y 2) Pruebas. El entrenamiento construye un modelo de clasificación, el cual consiste en reglas de clasificación, mediante el análisis de datos de entrenamiento que contienen etiquetas de clase. El segundo paso, la prueba, examina un clasificador, utilizando datos de prueba, para determinar la precisión o capacidad para clasificar registros desconocidos para la predicción. La precisión de los modelos de clasificación depende del grado en que las reglas de clasificación sean verdaderas, lo que se estima mediante datos de prueba (Yoo et al., 2012). Cataloluk y Kesler (2012) utilizaron esta técnica para analizar enfermedades de la piel mediante el uso de clasificador K-NN ponderado.

Por su parte, el agrupamiento divide los datos en función de las similitudes que tiene; los algoritmos de agrupación descubren colecciones de datos de manera que los objetos en la misma agrupación son más parecidos entre sí que otros grupos (Sharmila y Vethamanickam, 2015). Estos se dividen en agrupamiento duro, que son aquellos en los que cada dato solo pertenece a un grupo y agrupamiento suave, donde cada objeto puede estar categorizado a distintos clústers según su grado de pertenencia (Gupta y Bora, 2014).

El algoritmo de Agrupación Difusa C-medias o Fuzzy C-means en inglés (FCM) trabaja con un agrupamiento suave, se basa en que cada punto comparte con cada grupo una función, llamada función de pertenencia, cuyos valores están entre 0 y 1, por lo que un punto puede integrar en un porcentaje a mas de un clúster (Gupta, Shivhare y Sharmae, 2015).

De acuerdo con Haraty, Dimishkieh y Masud (2015), el algoritmo de agrupamiento K-medias es uno de los métodos de agrupación de datos más utilizados, el cual tiene como función la división de un conjunto de datos de “n” observaciones en “k” grupos, en donde cada dato observado corresponde al grupo “k” cuyo centroide es más cercano. Veloso et al., (2014) aplicaron técnicas de agrupamiento a una base de datos para predecir los reingresos de pacientes en medicina intensiva. Los algoritmos utilizados en el método de cuantificación vectorial fueron k-medias, k-mediods y x-medias. En adición, Belciug et al., (2009) utilizaron esta técnica con el agrupamiento jerárquico aglomerativo para agrupar a pacientes según su duración de estancia en el hospital para proporcionar una mejor utilización de los recursos del hospital y brindar mejores servicios a los pacientes.

Los algoritmos de agrupación pueden ser validados a través dos herramientas. La primera de ellas es el método del codo, el cual calcula la suma de errores al cuadrado (inercia) para cada valor de “k” estudiado, con la finalidad de elegir valor de “k” cuyo punto en la gráfica representa la forma del codo en un brazo, como se muestra en la figura 2.3 De esta forma, se determina el número apropiado de clústers minimizando la inercia, es decir, disminuyendo la distancia entre los miembros de un mismo clúster. La segunda herramienta es el análisis de silueta, el cual se utiliza para determinar el grado de separación entre los clústers, a partir del cual se busca seleccionar el valor de “k” donde los clústers están ampliamente separados entre ellos (Thinsungnoen et al., 2015).




(a) k=3 (b) k=4 (c) k=5

***Figura 2.3.*** *Método del codo para selección del valor de "k", (Thinsungnoen et al., 2015).*

El método del codo es fácil de implementar observando el gráfico del valor k ideal con la posición en el codo junto con la inercia. El mejor resultado del grupo k será la base para la agrupación a realizar. Cuanto menor sea el valor de la inercia y el gráfico del codo disminuya, mejores serán los resultados del clúster (Syakur et al., 2018).

### 2.6.5. Algoritmo K-medias

K-medias es uno de los algoritmos de aprendizaje no supervisados más simples que se enfoquen en la técnica de agrupamiento. El procedimiento sigue una manera sencilla y fácil de clasificar un conjunto de datos dado a través de un cierto número de grupos (k grupos) fijados previamente, como se muestra en la figura 2.4. La idea principal es definir k centroides, uno para cada grupo. Estos centroides deben colocarse de manera astuta debido a ubicaciones diferentes causan resultados diferentes. Por lo tanto, la mejor opción es colocarlos lo más lejos posible el uno del otro. (Rajagaru y Prabhakar, 2017).

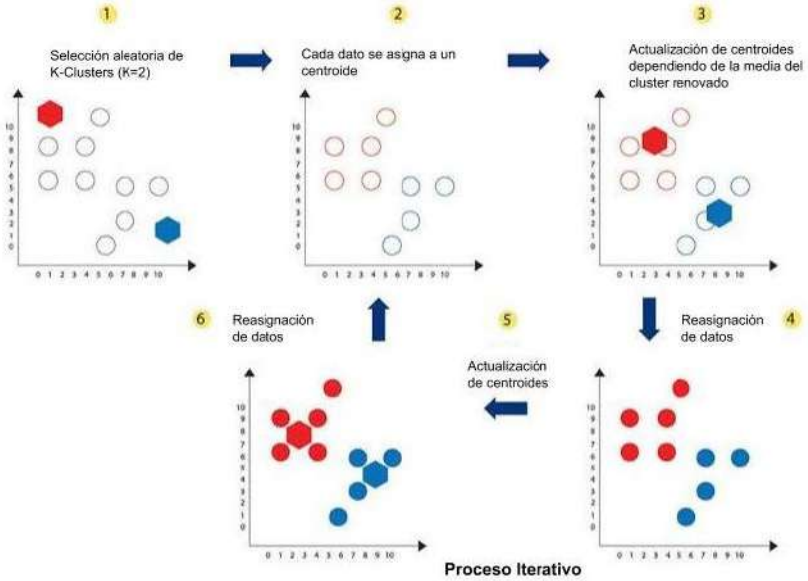



***Figura 2.4.*** *Proceso de algoritmo K-medias.*

Dicho algoritmo es una técnica que se basa en el centro de las agrupaciones creadas, esto a menudo está representado por el promedio del clúster. La agrupación mide la similitud del grupo iterando la distancia de medición entre cada objeto y el centro de cada grupo utilizando la medición de distancia euclidiana.

Según Hartono et al. ( 2018) k-medias es un algoritmo iterativo que puede describirse mediante los siguientes pasos.

a) Elegir arbitrariamente k objetos como los centros de agrupación inicial (centroides).
b) Calcular la distancia de cada objeto al centroide.
c) Clasificar los objetos en función de la distancia mínima a un centroide.
d) Actualizar el agrupamiento de acuerdo con el valor medio de los objetos para cada clúster.
e) Se repiten los pasos b), c) y d) hasta que el algoritmo logre la convergencia y los objetos no cambien de clúster.

## 2.7. Estudios previos

Existen varios estudios relacionado al refinamiento de algoritmos, como el de Hartono et al. (2018), donde se propone un enfoque para optimizar un algoritmo de agrupamiento K-medias para resolver problemáticas de desequilibrio de clases, situación en la que varias instancias en al menos una clase son mucho más altas que las instancias de números de otras clases. Esta situación crea un problema de clase mayoritaria versus minoritaria, en específico en el algoritmo a mejorar, la atención se centra en la clase mayoritaria por lo que puede llegar a afectar la precisión de la predicción. El objetivo de esta investigación fue determinar las coordenadas del centroide de un grupo en un proceso de agrupación de K-medias y analizar su efecto en el desequilibrio de clases. En general, la metodología utilizada consistió en dos etapas: la etapa de preparación y la etapa de agrupamiento. En la etapa de preparación, el conjunto de datos es entrenado por un algoritmo de red neuronal artificial para generar k centros de conglomerados iniciales (centroides). Por su parte, la etapa de agrupamiento se ejecutó de manera normal el algoritmo k-medias, ingresando los resultados de la etapa de preparación para asignar los centroides iniciales.

La propuesta de los autores fue probada en dos diferentes bases de datos: “Balanced Scale” con 625 instancias y “Abalone” con 4177 instancias, ambas con tres diferentes grupos, Clase B, Clase L y Clase R para la primera y Sexo F, Sexo I, y Sexo M para la segunda. En el algoritmo normal de K-medias, ambos conjuntos de datos presentaron un problema de desequilibrio en sus grupos, con hasta 136 instancias de diferencia en “Balanced Scale” y 2458 en “Abalone”. En consecuencia de estos problemas, el número de errores tiende a ser alto. El promedio de error en “Balanced Scale” es de 385.5 (61.68%) y en “Abalone” es de 2546.6 (60.96%).

Como resultados del algoritmo K-medias optimizado lograron reducir el desequilibrio de 136 a 51 en la base de datos “Balanced Scala”, y de 2458 a 1075 en “Abalone”. Con esto también se logró reducir el número de errores. Para “Balanced Scale”, el número de errores se redujo a 289 o alrededor del 46,24%, mientras que el resultado de agrupación del conjunto de datos de “Abalone” redujo el número de errores a 2166 o alrededor del 51,86%.

Derivado de la investigación se puede llegar a dos conclusiones. Primero, que la red neuronal artificial puede ayudar a determinar el centroide de la agrupación de K-medias, lo que se ve reflejado en la disminución del número de errores. En segundo lugar, se confirma que el modelo de optimización del algoritmo de agrupación de K-medias utilizando la red neuronal artificial también puede manejar el problema de desequilibrio de clase para resolver problemas de precisión de predicción.

En otro de los estudios, Haraty, Dimishkieh y Masud (2015) proponen un algoritmo de agrupamiento de k-medias mejorado para el descubrimiento de patrones en datos de atención médica. Los autores destacan como unos de los principales problemas del algoritmo la necesidad de realizar cálculos en repetidas ocasiones sobre todo el conjunto de datos en cada ciclo, y la necesidad de muchos de estos ciclos antes de enfocarse en un resultado de calidad. En consecuencia de esto, los autores mencionan que esto hace que su utilización sea extremadamente costosa, especialmente para conjuntos de datos de disco local sustancialmente enormes, por lo que proponen una

metodología para el algoritmo con menos cantidad de ciclos de cálculos, que a su vez pueda manejar los mismos resultados convergentes que el algoritmo K-medias.

Dicho algoritmo propuesto lo llaman G-medias, cuya esencia es facilitar el cálculo de los centroides iniciales para evitar la elección de los mismos de forma arbitraria como lo ejecuta el algoritmo K-medias, ya que los autores asumen que el algoritmo original necesita mejora con la selección aleatoria incial de la matriz de centroides. Por lo tanto, el paso inicial es calcular todos los elementos existentes que tienen el mayor grado en el espacio; a partir de ahí se puede tener una configuración inicial de cómo deberían verse los agrupamientos. En la segunda ejecución, se eliminan todos los centroides que están en un solo grupo y se selecciona $k$ grupos con los resultados más altos de la función de similitud que se tomarán como los centroides de grupo reales. Una vez hecho esto, se itera en el resto de los elementos de datos para ver si los centroides cambian. Esto se realiza exactamente como el medio $k$ original con las funciones de distancia y similitud.

La propuesta fue puesta a prueba en una base de datos para comparar los algoritmos K-medias y G-medias. Se analizo de acuerdo a la entropía del algoritmo, la presición de los resutados (Valor F), el coeficiente de varianza y el tiempo de ejecución. En cuanto a entropía, en el algorito K-medias, cuando el tamaño de los datos aumenta, la entropía tambien lo hace, por lo que disminuye la calidad de los clústers. Por su parte, en G-medias los resultados son consistentes, es decir, la calidad de los clúster no disminuye significativamente al aumentar la cantidad de datos. En la presición de los resultados, G-medias muestra mejor comportamiento y es más estable que el algoritmo K-medias. En el coeficiente de varianza los resultados muestran que cuando los conjuntos de datos son grandes la varianza en el algoritmo G-medias disminuye drasticamente.Por último, en el tiempo de ejecución los resultados indican que cuando se aumenta el conjunto de datos, el incremento del tiempo de ejecución para K-medias es casi constante, mientras que en G-medias el tiempo se vuelve al principio mucho más alto; A medida que crece el conjunto de datos, aumenta menos tiempo en cada ejecución. Esto debido al hecho de que cuenta con un tiempo constante para la

selección de los centroides iniciales, pero las iteraciones siguientes son mínimas, por lo tanto, se puede conluir que cuando aumenta el número de datos, G-medias convergerá más rápido que K-medias.

Por su parte, Silva-Cárcamo (2016) realizó un estudio para identificar las comorbilidades que presentan los pacientes con diabetes mellitus tipo 2 que asisten al Instituto Nacional del Diabético en Tegucigalpa, Honduras. El autor destaca que la diabetes es un reto para las instituciones de salud, ya que es una enfermedad que sigue en crecimiento a nivel mundial y existe una limitación de recursos para implementar estrategias que puedan mejorar la situación de este padecimiento. Además, resalta que los pacientes con diabetes mellitus tipo 2 tienen una mayor incidencia de enfermedad en varios órganos y tejidos internos. Las enfermedades microvasculares y macrovasculares crónicas tienen una mayor influencia en el pronóstico a largo plazo de pacientes con diabetes que las complicaciones agudas. Por lo tanto, propone que investigar las asociaciones de estas complicaciones con enfermedades comórbidas mediante el uso de datos de diagnóstico del paciente es útil para predecir su incidencia y, en consecuencia, para tratar de manera más efectiva a los pacientes con diabetes.

La metodología utilizada consitió en la recolección de  datos de los expedientes de pacientes, dicho datos constaban de tres distintas secciones: datos del paciente, características sociodemográficas y datos clínicos de la enfermedad y de enfermedades acompañantes. Se incluyeron todos los pacientes que cuplieran con las características de padecer diabetes mellitus tipo 2 y que presentaran comorbilidades.

Se revisaron un total de 382 expedientes, se encontró que el 43.98% presentó sobrepeso y dentro el rango de edad más crítico fue entre 60-70 años representando el 24.61%. También se hizo un análisis del tiempo transcurrido desde que fue diagnósticada la enfermedad, en relación a esto, el 30.97% presenta la enfermedad desde hace 5-10 años. El autor destaca la importancia del tiempo de diagnóstico ya que a mayor tiempo de evolución de la enfermedad comienzan a presentarse

afecciones en otrs sistemas del cuerpo, por lo que aumentan las comorbilidades. La comorbilidad que más se encontró en los pacientes fue hipertensión arterial, con 245 personas, equivalente al 64.14% de la población estudiada, seguida de neuropatía diabética y dislipidemia con 26.96% y 15.97% respectivamente.

El autor concluye en la necesidad de comprender los datos y sus patrones para lograr un mejor apego al tratamiento de los pacientes que asistan a consulta, teniendo en cuenta las comorbilidades encontradas, teniendo asi la capacidad de ofrecer mejores tratamiento y anticipandose para preveer ciertas de enfermedades relacionadas (ER).

Además, existen trabajos realizados con la misma base de datos que se utilizará para este proyecto, como el de Sanez (2018), quien diseñó una metodología para aprovechar las bases de datos electrónicas derivada de consultas médicas para mejorar las estrategias de medicina preventiva enfocadas en diabetes y obesidad. La metodología desarrollada la componen 5 etapas: análisis inicial, segmentación de la población, uso de minería de datos, obtención de reportes y difusión del contenido. Dentro de la etapa del uso de minería de datos se utilizó un algoritmo de agrupamiento K-medias que trabaja con variables de edad, sexo y herencia. Dentro de los resultados obtenidos en obesidad, se identificó que las mujeres son más propensas a padecer esta enfermedad en un 14%, en adición, se identificaron dos grupos de edad críticos (de 10 a 18 años y de 30 a 50 años) donde se incrementa el diagnóstico de obesidad. En cuanto a la diabetes, el autor destaca que las mujeres presentan alrededor de 30% más diagnósticos que los hombres y resalta que el padecimiento se intensifica a medida que aumenta la edad, en específico, a partir de los 35 años.

El trabajo desarrollado Llanes (2018), quien utilizó análisis de georreferenciación y MD para desarrollar una metodología para detectar brotes epidemiológicos, también hizo uso de la misma base de datos que se utilizó para este trabajo.

# 3. ESTRATEGIA

En esta sección se presenta la propuesta de estrategia que incluye el método a seguir para resolver la problemática expuesta en el capítulo 1. El planteamiento de dicho problema permite establecer que se trata de una investigación de tipo cuantitativa con un alcance descriptivo, cuya característica, según Hernández et al. (2014), es que son estudios que buscan especificar propiedades y características importantes de cualquier fenómeno que se analice. En este caso en particular, se trata de especificar como son y como se manifiestan las carácterísticas y perfiles de los pacientes para mejorar la forma en que se describen.

La estrategia presentada toma como referencias los trabajos de Chapman et al. (2000), Niaksu (2015) y Buczak et al. (2012) los cuales sirven como base para el proceso a seguir para resolver el tema de estudio. Niaksu (2015) propone una extensión del modelo Cross Industry Standard Process for Data Mining (CRISP-DM), el cual es un proceso estándar para el desarrollo de proyectos de MD. En dicha propuesta, se parte de la situación de que no existe un modelo o metodología de proceso bien definido que aborde los problemas y las limitaciones de la medicina y la atención médica, por lo que se pasa de tener una estrategia general a contar con un modelo donde se abordan los desafíos específicos de la MD en el sector salud, como lo son la variedad de formatos de los datos, la privacidad de datos del paciente y la calidad e integridad de los datos clínicos.

En la figura 3.1, se muestra la estrategia basada en la adaptación de la modelo CRISP-DM desarrollada por Niaksu (2015) para el desarrollo de proyectos de MD en el sector salud.

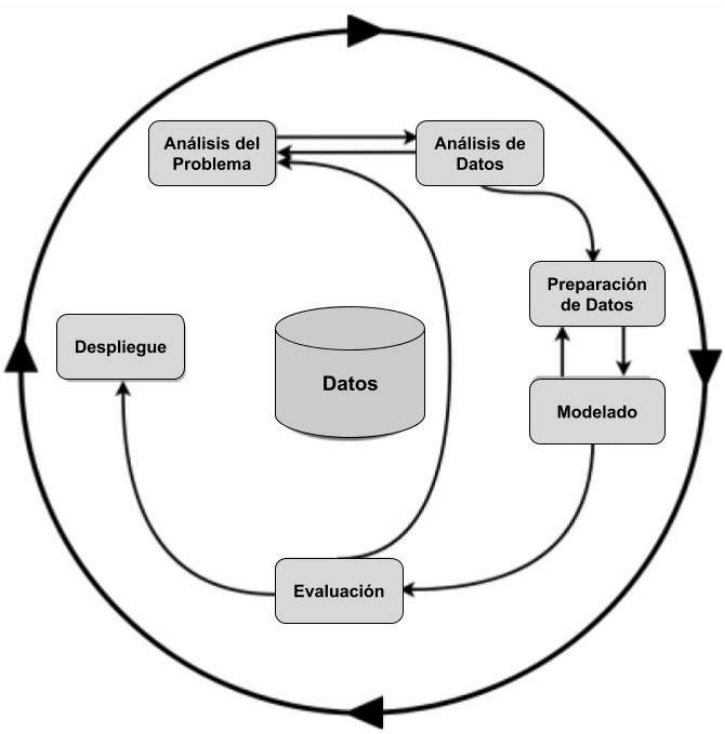



***Figura 3.1.*** *Etapas de la estrategia propuesta, adaptado de Niaksu (2015).*

La estrategia está compuesta de 6 etapas, los cuales son: Análisis del problema, Análisis de datos, Preparación de datos, Modelado, Evaluación y Despliegue; estas, a su vez, contienen actividades que son necesarias de llevar a cabo. A continuación, se describe cada una de las etapas.

## 3.1. Análisis del problema

La fase inicial tiene como objetivo obtener una comprensión clara de los objetivos deseados desde una perspectiva empresarial. En consecuencia, el objetivo de negocio debe convertirse en una definición de problema desde una perspectiva de MD.

Los pasos propios de esta fase son la determinar objetivos de negocio, evaluación de la problemática y determinar los objetivos de MD.

### 3.1.1. Determinar objetivos de negocio

En este paso se debe especificar el problema que da origen a la ejecución del proyecto, además de los objetivos perseguidos. Las tareas principales son:

1. **Conocimiento previo**: Se debe recolectar la información disponible sobre el entorno organizacional donde se desarrolla el proyecto, con el fin de establecer los objetivos e identificar los recursos disponibles. Dicho análisis se hace respecto a la organización y el problema.
2. **Objetivos de negocio**: Se debe describir detalladamente cada uno de los objetivos, por lo que es necesario describir el problema de MD, detallar los aspectos del problema y los beneficios que se esperan obtener por cada problema establecido.
3. **Criterios de éxito**: Cada uno de los objetivos planteados debe contar con un criterio para juzgar el grado de éxito obtenido.

### 3.1.2. Evaluación de la situación

En el segundo paso es necesario analizar profundamente todas las variables que puedan tener un impacto en el desarrollo del proceso de MD. El análisis se divide en:

1. **Recursos disponibles**: Estos se deben enlistar en cuanto a datos, hardware, software y personal.
    a. **Datos**: Se debe describir su origen, grado de fiabilidad, el modo de acceso, el tipo de datos y la cantidad.
    b. **Hardware**: Se enumera todo el software del que se dispone para cada una de las fases del proceso.
    c. **Software**: Se realiza una descripción de las herramientas de software. Se analizan los pros y contras que presentan cada una de las opciones, se pueden considerar el precio de la licencia del software, la experiencia previa del personal, así como las características que facilitan el desarrollo del proyecto según su naturaleza, como podría ser el trabajo en la nube si se va a trabajar de manera remota. La información se puede resumir en una matriz como la que se muestra en la tabla 3.1, para posteriormente comparar y elegir el software más conveniente para el proyecto.

| Nombre | Características | Lenguaje de programación | Sistema operativo | Precio / Licencia |
|---|---|---|---|---|
| Software 1 | | | | |
| Software n | | | | |

***Tabla 3.1.*** *Formato de comparativo para elección de software de MD.*

2. **Requerimientos, supuestos y restricciones**: Se deben enlistar todos aquellos relacionados tanto a la panificación del proyecto como a los datos y recursos disponibles.
    a. **Requerimientos**: Se decretan los requerimientos del proyecto, se consideran tanto los temporales como aquellos que se relacionan con los resultados (comprensibilidad, precisión, capacidad de explotación, mantenibilidad, repetibilidad). También se consideran aspectos de seguridad, restricciones legales y privacidad.
    b. **Supuestos**: Se establecen todos los supuestos relativos a los datos (calidad necesaria, tipos de datos a utilizar, cantidad de datos), así como la forma en que deben ser presentados los resultados.
    c. **Restricciones**: Por posibilidades de acceso, tiempo, acceso de datos confidenciales, acceso al conocimiento de expertos, costos, etc.
3. **Terminología**: Se realiza un glosario con la terminología referente al proceso de negocio y al proceso de MD, se incluyen términos que se usarán de manera constante y que no son de conocimiento general para los involucrados en el proyecto.

| Término | Definición |
|---|---|
| Término 1 | Definición 1 |
| Término n | Definición n |

***Tabla 3.2****. Formato para realizar glosario don definición más importantes para el proyecto.*

### 3.1.3. Determinar objetivos de MD

Una vez que se han establecido los objetivos desde la perspectiva del negocio, se deben determinar los objetivos desde una perspectiva orienta al proceso de MD, es decir, describirlos con una perspectiva más técnica.

1. **Objetivos de MD**: Es necesario desglosar los objetivos previamente establecidos en perspectiva del proceso de MD. En adición, se describen las herramientas o técnicas que podrán ser utilizadas.
2. **Criterios de éxito de MD**: Se establecen los criterios que se utilizarán para evaluar el grado de alance de los resultados alcanzados en cada una de las tereas de MD efectuadas.

Los resultados de este paso con los objetivos y la forma de evaluarlos pueden establecerse en una matriz como la que se muestra en la tabla 3.2.

| Objetivos de MD | Métricas |
|---|---|
| Objetivo 1 | Métrica 1 |
| Objetivo 2 | Métrica 2 |
| Objetivo n | Métrica n |

***Tabla 3.3.*** *Ejemplo de matriz para establecer objetivos de MD y sus métricas.*

## 3.2. Análisis de datos

Esta fase comienza con la recopilación de datos iniciales y el acceso al conjunto de datos. Los problemas de calidad de los datos deben identificarse y se determinan las relaciones más evidentes para establecer las primeras hipótesis que puedan ser de interés para pasos posteriores.

Esta fase se compone de los cuatro diferentes pasos, que se describen a continuación.

### 3.2.1. Adquisición de los datos

El primer paso de la fase consta en la recopilación de los datos. Se debe de crear un conjunto inicial de datos con lo que trabajar. Este paso cuenta con las siguientes tareas:

1. **Planificar requerimientos**: Se estudia la información que se requiere para establecer las variables necesarias y si analiza si es posible la adquisición de estas.
2. **Criterios de selección**:

a. Determinar los criterios para la elección de las variables.
b. Escoger las tablas o ficheros de interés.
c. Establecer el periodo de tiempo de los datos.

### 3.2.2. Descripción de los datos

Se deben de describir las características más importantes de las variables en cuanto a cantidad de registros, tipo de datos y descriptores estadísticos, de forma que se pueda entender el comportamiento de los datos. Las tareas que se incluyen en este se describen a continuación:

1. **Análisis volumétrico de los datos**: Se identifican los datos y los métodos de captura.
2. **Tipos y valores de las variables**: Se específica el tipo de datos, rango de los valores, uso de estadística descriptiva, grado de consistencia de los datos y la importancia que tienen según los criterios de selección establecidos.
   La información se puede resumir en una matriz como la que se muestra en la tabla 3.3.

| Variable | Descriptor de la variable | Rango de los valores | Formato de la variable | Importancia |
|---|---|---|---|---|
| Var1 | | | | |
| Var2 | | | | |
| Var3 | | | | |

***Tabla 3.4.*** *Formato para descripción de variables.*

### 3.2.3.Exploración de los datos

Derivado de la fase anterior, se procede a efectuar un primer análisis de las características de los datos donde se buscan relaciones entre variables, tipos de distribuciones, agrupamientos, etc. Para esto se pueden utilizar distintas técnicas de visualización, análisis de correlación, técnicas estadísticas, entre otras técnicas.

### 3.2.4. Verificar la calidad de los datos

El último paso de esta fase consta de establecer la calidad de los datos con los que se cuentan para saber si son suficientemente buenos o es necesarios regresar a la fase previa.

Se debe analizar si los datos cuentan con errores, describen lo establecido, cubren todo el rango, cantidad de datos incompletos o inexistentes, etc.

1. Revisión de las variables.
    a. Representan la realidad y son consistentes.
    b. Campos vacíos y sus causas.
    c. Variables no necesarias.
2. Ruido e inconsistencia de datos.
    a. Determinar cantidad de datos redundantes.
    b. Detectar ruido, procedencia y las variables afectadas.

## 3.3. Preparación de datos

En esta fase se busca obtener una base de datos lista para la fase de modelado (fase 4). Cubre todas las actividades que se requieren para preparar previamente el conjunto de datos final. Las actividades de la fase de preparación de datos dependen en gran medida de las características y la calidad de los datos sin procesar originales. Algunas de las tareas características de la preparación de datos incluyen la elección de una tabla, proyecciones de atributos y registros, transformación de atributos, clasificación, normalización, eliminación de ruido y muestreo.

Las actividades se desempeñan en repetidas ocasiones hasta que se obtiene la base de datos apropiada para la fase posterior.

### 3.3.1. Selección de datos

El primer paso inicia con los resultados obtenidos en la fase anterior y de la descripción de los datos previamente realizados. A partir de esto, se efectúa la selección de las

variables más importantes, en dónde se debe tener en cuenta que se trate de variables que sean los más independiente entre sí, que describan en gran parte el sistema a analizar, que tengan una importancia individual destacada y que estén libres de ruido, es decir, que sea datos fiables.

### 3.3.2. Limpieza de datos

Este paso hace énfasis en tener variables lo más fiables posibles para ejecutar la fase de modelado. Se definen dos tareas para corregir el ruido y los espurios.

1. Ruido en los datos
    a. Corregir, ignorar o eliminar aquellos datos con ruido.
    b. Estudiar los posibles causantes que generan el ruido y establecer la manera de corregirlo.
    c. Utilizar técnicas de filtrado para obtener datos de mayor calidad.
2. Espurio en los datos
    a. Analizar por separado y determinar los causantes.
    b. Eliminar o ignorar los datos con esta característica.
    c. Completar los datos con técnicas estadísticas.

### 3.3.3. Generación de variables adicionales

Es posible que sea necesario crear nuevas variables a partir de las ya existentes, esto con el objetivo de poder ejecutar la fase de modelado, además de agilizar los estudios posteriores.

Algunas de las actividades de transformación de datos que pueden ser utilizadas son:

1. Estandarizar o normalizar variables.
2. Asignar pesos según la importancia de cada variable.
3. Cambiar la codificación de la variable.
4. Uso de transformadas.
5. Adición de nuevas variables a partir de otras.

## 3.4. Modelado

En esta fase se realiza la selección adecuada de técnicas de modelado, algoritmos o combinaciones de estos. Después, se deben elegir los valores óptimos de los parámetros del algoritmo. En general, para la misma tarea, existen diversos métodos de modelado disponibles. Algunos de los métodos tienen restricciones de calidad de datos o tipos de datos específicos. En consecuencia, esta fase es posible realizarla de forma iterativa hasta que se alcance el criterio de calidad del modelo elegido. La calidad del modelo se evalúa formalmente, para esto, se utilizan métricas comunes en MD y estadística: sensibilidad, precisión, especificidad y curva ROC.

Los pasos que conforman esta fase son selección de las técnicas de modelado, diseño del método de evaluación y la generación del modelo.

### 3.4.1. Selección de las técnicas de modelado

En este paso se deben elegir las técnicas a utilizar para el modelado, dicha elección debe hacerse en base al tipo del problema, los datos a utilizar, el tiempo del que se dispone para obtener el modelado, las herramientas de que se disponen, así como el conocimiento y experiencia que se tiene de la técnica.

Las técnicas que se seleccionen previamente deben ser calificadas y ordenadas en función de los puntos establecidos anteriormente.

### 3.4.2. Diseño del método de evaluación

Previo al inicio del proceso para construir el modelo, se debe definir la manera en que se validara el mismo. Para esto es necesario establecer:

1. Función que determine el error y umbral de calidad estimado: error cuadrático medio, indicador de error de la predicción, métricas de distancia intra-clúster e inter-clúster, etc.

2. Establecer el tipo de método de evaluación de los modelos a generar: división de dos grupos para entrenamiento y validación, validación cruzada, etc.
3. En caso de ser necesario, el tamaño de los grupos de validación y el tipo de entrenamiento.

### 3.4.3. Generación del modelo

Después de establecer el método de evaluación del modelo, se deben aplicar las técnicas de modelado a los datos preparados. Para esto es necesario describir:

1. Los parámetros utilizados, se debe incluir su importancia, influencia en el resultado del modelo y valores iniciales asignados.
2. Modelos resultantes, gráficas de entrenamiento y validación, resultados numéricos.
3. Descripción específica del modelo, de los parámetros, de la exactitud y sensibilidad, así como de la forma de implementarlo.

### 3.4.4. Evaluación del modelo

En este paso el modelo debe evaluarse para garantizar que cumpla con los criterios de éxito con enfoque en MD, además de analizar los criterios de prueba deseados. Es una evaluación completamente técnica basada en el resultado de las tareas de modelado.

Algunos de los criterios a tomar en cuenta son:

1. Grado de calidad de la predicción.
2. Métricas de evaluación interna y externa de agrupamientos.
3. Velocidad del proceso.
4. Evaluación del proceso con nuevos datos.
5. Influencia de cada uno de los parámetros en el modelo.
6. Obtener información sobre por qué cierta técnica de modelado y cierta configuración de parámetros conduce a buenos o malos resultados.

En caso de que los resultados no sean adecuados, se deben repetir todos los pasos hasta obtener una solución óptima.

Para comparar algoritmos, se pueden contrastar los resultados ejecutados la cantidad de algoritmos a evaluar en métricas como entropía, valor de f, coeficiente de varianza, tiempo de ejecución, entro otros índices. En la tabla 3.4 se muestra el formato para comparar dos algoritmos estableciendo las métricas a contrastar.

| Métrica | Algoritmo actual | Algoritmo propuesto |
|---|---|---|
| Distancia promedio inter-clúster | | |
| Distancia promedio intra-clúster | | |
| Índice 1 | | |
| Índice 2 | | |

***Tabla 3.5.*** *Formato para comparación de algoritmos*

## 3.5. Evaluación

Los pasos de evaluación previos trataron factores como la precisión y la generalidad del modelo. Esta fase evalúa el grado en que el modelo cumple con los objetivos de negocio y busca determinar si hay alguna razón por la cual este modelo es deficiente. La evaluación se debe realizar comparando los resultados con los criterios de evaluación definidos al inicio del proyecto. El resultado final de esta fase es la elección de si los resultados de MD se pueden usar en entornos prácticos.

### 3.5.1. Evaluación de los resultados

Una buena forma de definir los resultados totales de un proyecto de minería de datos es usar la ecuación: Resultados = Modelos + Hallazgos. Donde se considera hallazgo cualquier cosa (aparte del modelo) que es importante para cumplir con los objetivos de negocio o juegan un papel importante a nuevas preguntas.

Se deben resumir los resultados de la evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final de si el proyecto cumple con los objetivos de negocio establecidos inicialmente.

Algunas de las actividades para este paso son:

1. Verificar los resultados de MD con la base de conocimiento dada para ver si la información descubierta es nueva y útil.
2. Evaluar los resultados con respecto a los criterios de éxito de negocio, es decir, ¿ha logrado el proyecto los objetivos comerciales originales?
3. Crear una clasificación de resultados con respecto a los criterios de éxito de negocio.

## 3.6. Despliegue

El conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda usarlo. Dependiendo de los requisitos, esta fase puede ser tan simple como generar un informe o tan compleja como implementar un proceso de MD repetible en toda la organización o departamento.

Esta fase se compone de un informe final, el cual se describe a continuación.

### 3.6.1. Generación de reporte final

Al final del proyecto, dependiendo del plan de implementación, el reporte puede ser solo un resumen del proyecto y sus experiencias, si aún no se han documentado como una actividad en curso, o puede ser una presentación final e integral de los resultados de la minería de datos.

Las actividades para este paso son:

1. Una descripción detallada del problema original.
2. El proceso utilizado para realizar la MD.
3. Notas sobre cualquier desviación del plan original del proyecto.
4. Resumen de los resultados de la MD, tanto de modelos como hallazgos.
5. Recomendaciones para el trabajo adicional de MD, incluyendo hallazgos descubiertos durante la exploración y el modelado.

# 4. IMPLEMENTACIÓN

Esta sección se compone de cinco diferentes fases. La inicial describe el entorno y la situación actual del algoritmo, la segunda tiene como objetivo la familiarización con los datos e identificar las relaciones más evidentes entre las variables a trabajar, en la tercera fase se preparan los datos para posteriores actividades, en la penúltima fase se seleccionan y se aplican técnicas de MD y, por último, se comparan los algoritmos para evaluar los resultados, dichas conclusiones son expuestos en el capítulo 5.

## 4.1. Fase I. Análisis del problema

La primera fase de la estrategia se desglosa en tres pasos, el objetivo es comprender a fondo, desde perspectiva del instituto, lo que se quiere lograr y desde dónde parte. Además, se descubren factores importantes que pueden influir en el resultado final. A continuación, se describen cada uno de los pasos con sus respectivas actividades.

### 4.1.1. Determinar objetivos de negocio

En este paso se debe especificar el problema que da origen a la ejecución del proyecto, además de los objetivos perseguidos. Estos se hacen a través de las siguientes actividades:

- Conocimiento previo: El algoritmo actual trabaja con tres de las veintiocho variables disponibles de la base de datos que se genera de las citas a médico general, estás variables son edad, sexo y herencia. El departamento de medicina preventiva es quien toma decisiones derivado de los resultados del modelo; cuando se detectan los grupos de mayor vulnerabilidad y las características que comparten los afiliados, este se define como el grupo de pacientes que requiere mayor atención. Además, el conocimiento generado se hace llegar al mayor número de personas posibles a través de redes sociales y medio electrónicos de la institución.

- Objetivo de negocio: El departamento de medicina preventiva busca poder generar conocimiento más exacto, es decir, ser capaz de detectar con mayor precisión los grupos con mayor riesgo de padecer DM.
- Criterio de éxito: Detectar los pacientes con características similares que tengan un riesgo de padecer DM.

### 4.1.2. Evaluación de la situación

- Recursos disponibles
    - Humano: De parte de la institución se conforma un equipo de trabajo con las áreas de medicina preventiva, informática y estadística. Además, para cualquier parte del desarrollo del proyecto, hay dos médicos generales que asesoran temas de medicina.
    - Datos: Surgen de citas a médico general que se llevaron a cabo entre 2015 y 2017 en la ciudad de Hermosillo.
    - Software: Las opciones disponibles para elegir se comparan en la tabla 4.1

| **Nombre** | **Características** | **Lenguaje de programación** | **Sistema operativo** | **Precio / Licencia** |
|---|---|---|---|---|
| MATLAB | Cajas de herramientas para MD.<br>Gráficas para visualizar datos y herramientas para crear diagramas personalizados.<br>Gestión de archivos en la nube.<br>Permite la comunicación con programas en otros lenguajes. | MATLAB | GNU/Linux macOS Windows | Versión estudiante $ 49.00 USD.<br>$29.00 USD por cada toolbox adicional. |
| Google Colaboratory | Basado en Jupyter Notebook.<br>Librerías preinstaladas para ciencia de datos y posibilidad de instalar otras.<br>Trabaja con valores numéricos.<br>Servicio en la nube.<br>Conexión con Google Drive. | Python 3<br>Python 2 | Se ejecuta desde el navegador, por lo que se puede operar en cualquier sistema operativo. | Gratuito |
| R | Proporciona un amplio abanico de herramientas estadísticas.<br>Permite generar gráficos con alta calidad.<br>Cuenta con una enorme cantidad paquetes.<br>Facilidad para ejecutar operaciones de aprendizaje automático | R | macOS Windows | Gratuito |
| GNU Octave | Sintaxis orientada a las matemáticas con herramientas integradas de trazado y visualización.<br>Compatible con muchos scripts de Matlab.<br>Soporte en español. | Octave | GNU/Linux macOS Windows | Gratuito |

***Tabla 4.1.*** *Matriz para comparar distintos softwares para ejecutar proyectos de MD.*

Las cuatro opciones son de las principales variantes en el mercado para proyectos de análisis de datos, uno de los principales criterios para elegir fue el precio de la licencia, por lo que se descarta MATLAB, ya que es el único que no es gratuito. Además, otro factor importante es la experiencia en el lenguaje de programación, en este caso se tiene conocimiento previo de Python y R, por lo que se descarta Octave. Por último, debido a las características del proyecto en el que varias personas están involucradas, el servicio en la nube resulta una ventaja importante que facilita la comunicación y el intercambio de documentos. Debido a estos criterios se decidió trabajar con Google Colaboratory.

- Requerimientos, supuestos y restricciones

Debido a la naturaleza del proyecto, en el que se trata con información real de pacientes que asisten a la institución, se establecieron varias restricciones para respetar la privacidad de los datos de los derechohabientes. Por esto, en ningún momento se puede acceder a la identidad de la persona a la que pertenecen los registros, su edad o dirección exactas. Para esto, en la base de datos se cuenta con un identificador de derechohabiente llamado “hashAfilia”, un rango de edad y solamente la colonia para conocer la ubicación de la persona.

Para el manejo de los datos fue necesario la firma de un documento de privacidad por parte de los miembros del proyecto involucrados en el manejo y procesamiento de cualquiera de los registros que la institución proporcionó.

Otra restricción fue el rango de fechas de la base de datos, solo se cuenta con citas que se efectuaron en Hermosillo de enero del 2015 a marzo del 2017.

- Terminología: En la tabla 4.2 se presenta un glosario con las definiciones más importantes en el sector salud y de MD.

| Término | Definición |
|---|---|
| Algoritmo | Serie ordenada de instrucciones, pasos o procesos que llevan a la solución de un determinado problema. |
| Clúster | Es un método de MD para agrupar datos, dónde se busca que los datos que están dentro de un grupo sean los más similares entre sí. |
| Enfermedades Cardiovascular | Son un conjunto de trastornos del corazón y de los vasos sanguíneos. |
| Gastroparesia | Es una enfermedad que afecta el movimiento normal espontáneo de los músculos (motilidad) del estómago |
| Neuropatía | Es una consecuencia del daño a los nervios fuera del cerebro y la médula espinal (nervios periféricos), a menudo causa debilidad, entumecimiento y dolor, generalmente en las manos y los pies. |
| Nefropatía | Se refiere al daño, enfermedad o patología del riñón. |
| Artropatía | Es un término que denomina a cualquier enfermedad de las articulaciones. |

***Tabla 4.2**. Glosario con definiciones de términos claves del ámbito del sector salud y de MD para la ejecución del proyecto.*

### 4.1.3. Determinar objetivos de MD

En el primer paso de esta fase se estableció el objetivo desde la perspectiva de la institución médica, en este paso se convierte dicha meta en aspectos más técnicos. Los objetivos desde la perspectiva de MD se describen en la tabla 4.3.

| Objetivos de MD | Métricas |
|---|---|
| Agrupar pacientes con características similares a factores de riesgo de la DM | Distancia inter-clúster |
| | Distancia intra-clúster |
| | Inercia |

***Tabla 4.3.** Objetivos desde la perspectiva de MD y sus respectivas métricas para evaluar su alcance.*

Las métricas que se presentan en la tabla 4.3 serán evaluadas con el algoritmo actual y el algoritmo propuesto que se deriva de esta investigación. Dicha evaluación se desarrolla en la Fase IV.

## 4.2. Fase II. Análisis de datos

Esta fase consiste en obtener los datos que sean necesarios de todas las fuentes disponibles. Es importante integrar las diferentes bases de datos con las variables más útiles para facilitar su manejo para obtener como resultado un conjunto inicial de datos con lo que trabajar. A continuación, se desglosan los pasos que componen esta fase.

### 4.2.1. Adquisición de los datos

Los datos fueron proporcionados por la institución en dos diferentes tablas, la principal con registros de pacientes que acudieron a citas con médico general y otra con registros de pacientes diabéticos. El periodo al que pertenecen los datos fue establecido por la institución de acuerdo con el rango de fechas de las citas de los registros de la base de datos principal.

La tabla con los registros de citas a médico general cuenta con veintiocho atributos distintos, los cuales de describen en la sección 4.2.2. Por su parte, la tabla con los registros de pacientes diabéticos solo cuenta con dos atributos, hashAfilia con el que se identifica el paciente y diabetes, que es una variable binaria para identificar los pacientes con diagnóstico de DM.

### 4.2.2. Descripción de los datos

En el segundo paso de la fase el objetivo es conocer más los datos con las que se va a trabajar. En la tabla 4.4. se muestra un resumen de las variables, con la descripción, formato e importancia (baja, media y alta) que tienen para el proyecto; las variables con ponderación baja no representan utilidad para este proyecto para describir los pacientes diabéticos, ya que no fueron identificados en la literatura como indicadores de riesgo para la DM. Aquellas variables con ponderación media son de utilidad para analizar los datos en cuanto a tiempo, es decir fecha y citas distintas. Por último, las que fueron etiquetadas con importancia alta, son variables que tanto la FID (2019), como la AAD (2019) en su encuesta para conocer el riesgo de padecer DMT2 (Anexo

A), consideran rasgos importantes que afectan el padecimiento, además se considera la variable que identifica a los derechohabientes.

| Variable | Descriptor | Formato | Importancia |
|---|---|---|---|
| Gpo Edad | Grupo de edad del paciente | String | Alta |
| Sexo | Sexo del paciente | String | Alta |
| SubGpodia | Diagnóstico derivado de la cita | String | Alta |
| Tder | Tipo de derechohabiente | String | Alta |
| hashAfilia | Identificador de paciente | Numérico | Alta |
| Id | Identificador de cita | Numérico | Media |
| Fecha | Fecha de la cita | Fecha | Media |
| Área | Área que atendió la cita | String | Baja |
| Cant Inter | Cantidad interconsultas | Numérico | Baja |
| Cant Lab | Cantidad laboratorios pedidos | Numérico | Baja |
| Cant Ray | Cantidad rayosx pedidos | Numérico | Baja |
| Colonia | Colonia de residencia del paciente | String | Baja |
| Cant Med | Cantidad de medicamento recetado | Numérico | Baja |
| Cant Rece | Cantidad de recetas | Numérico | Baja |
| Citología | ¿Requiere citología? | String | Baja |
| Controlado | ¿Medicamento controlado? | String | Baja |
| Crónico | ¿Enfermedad crónica? | String | Baja |
| Cuidados | ¿Requiere cuidados? | String | Baja |
| Edocivil | Estado civil del paciente | String | Baja |
| Electro | ¿Requiere electrocardiograma? | String | Baja |
| Epidemi | ¿Es epidemia? | String | Baja |
| Especialista | Especialista que atendió la cita | String | Baja |
| ExamMama | ¿Requiere examen de mama? | String | Baja |
| Rayosx | ¿Requiere rayos x? | String | Baja |
| Receta | ¿Se proporciono receta? | String | Baja |
| Tipo Med | Tipo de médico que atendió la cita | String | Baja |

***Tabla 4.4.*** *Descripción de variables de base de datos de citas a médico general.*

### 4.2.3. Exploración de los datos

En este paso se hizo uso de estadística descriptiva para conocer el comportamiento de las principales características de la base de datos y poder establecer las primeras hipótesis que alimentaran las siguientes fases.

Este paso se realiza con una muestra de la base de datos, dichos registros son citas a médico general que se realizaron de enero a marzo del 2017. Esto es debido a que fue solo en este periodo de tiempo en que se cuentan con la base de datos completa; para posteriores pasos como el modelado se obtuvieron los datos ya clasificados por las ER.

Se identificaron 1,333 personas con DM durante el período de estudio, 703 mujeres (52.74%) y 630 hombres (47.26%), dicha información se presenta en la figura 4.1.

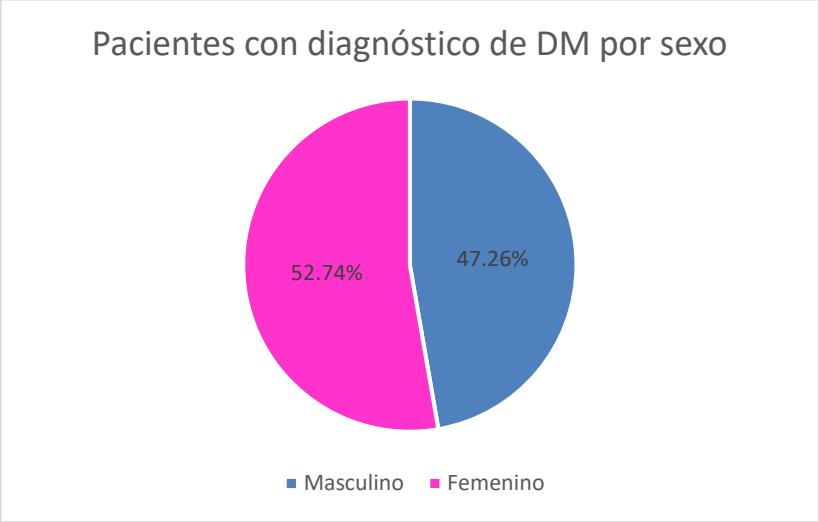



***Figura 4.1.*** *Pacientes con diagnóstico de DM separado por sexo en el periodo de enero a marzo del 2017.*

Los pacientes separados por sexo en la figura 4.1. es el punto de partida para el análisis que se desarrolla en este paso, es decir, todos aquellos pacientes con diagnóstico de DM.

En la tabla 4.5. se presenta la frecuencia de citas a médico general de los pacientes con diagnóstico de DM.

| Frecuencia de visita | Número de pacientes |
|---|---|
| 16 | 1 |
| 15 | 1 |
| 13 | 1 |
| 10 | 1 |
| 9 | 2 |
| 8 | 4 |
| 7 | 4 |
| 6 | 14 |
| 5 | 22 |
| 4 | 51 |
| 3 | 165 |
| 2 | 413 |
| 1 | 654 |

***Tabla 4.5.*** *Frecuencia de citas a médico general de pacientes con DM en el periodo de enero a marzo del 2017.*

En la tabla 4.5. se pueden identificar que existen 13 distintas frecuencias de visitas que van de 1 a 10, 13, 15 y 16 citas con médico general. En la figura 4.2. se resume la información de la tabla anterior.

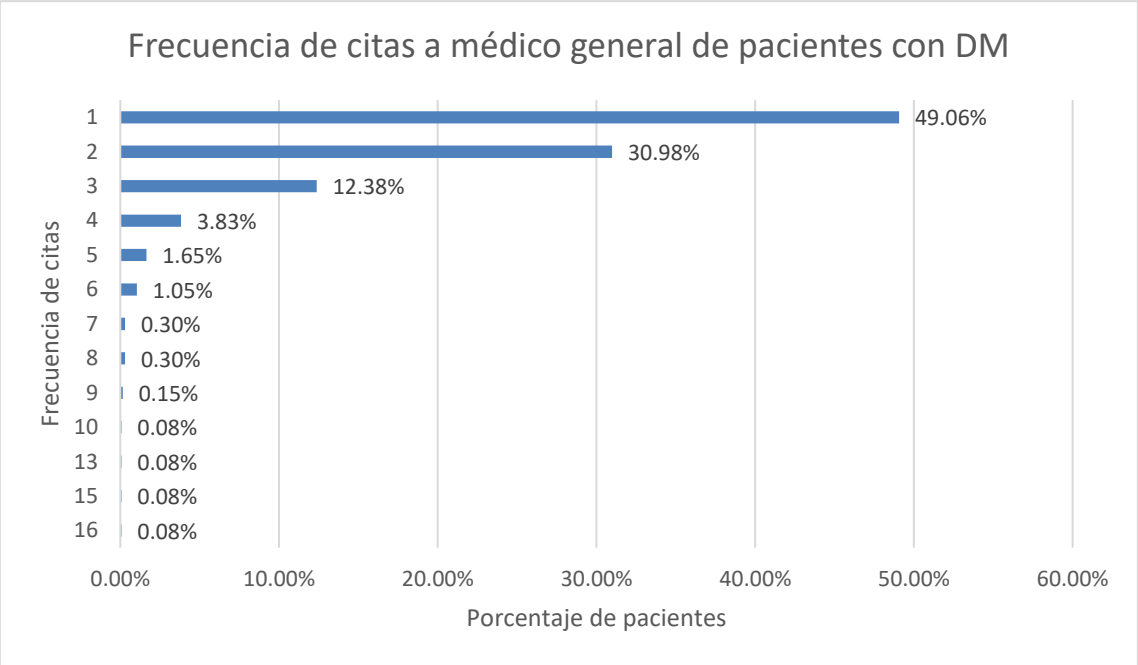



***Figura 4.2.*** *Frecuencia de citas a médico general de pacientes con DM en el periodo de enero a marzo del 2017.*

En la figura 4.2. se observa un comportamiento decreciente en número de pacientes a medida que aumentan el número de visitas. El 49.06% de los pacientes registran 1 sola cita, el 30.98% asistió a 2 citas y el 12.38% visitó en 3 ocasiones al médico general; entre estos 3 número de visitas representan el 92.42% de los pacientes. Las frecuencias de citas con valores de 10, 13, 15 y 16 solo registran un paciente cada uno, representando el 0.08% de los pacientes.

La tabla 4.6 contiene los diagnósticos derivados de citas con médico general de aquellos pacientes que tienen un diagnóstico de diabetes.

| Diagnóstico | Frecuencia | Porcentaje (%) |
|---|---|---|
| Enfermedades hipertensivas | 151 | 11.33% |
| Infecciones agudas de las vías respiratorias superiores | 145 | 10.88% |
| Enfermedades infecciosas intestinales | 27 | 2.03% |
| Síndromes del comportamiento asociados con alteraciones | 23 | 1.73% |
| Otras enfermedades del sistema urinario | 15 | 1.13% |
| Síntomas y signos generales | 12 | 0.90% |
| Trastornos metabólicos | 12 | 0.90% |
| Personas en contacto con los servicios de salud para investigación y exámenes | 11 | 0.83% |
| Síntomas y signos que involucran los sistemas circulatorio y respiratorio | 11 | 0.83% |
| Otros trastornos del oído | 9 | 0.68% |
| Personas en contacto con los servicios de salud por otras circunstancias | 9 | 0.68% |
| Micosis | 8 | 0.60% |
| Tumores (neoplasias) malignos | 8 | 0.60% |
| Enfermedades de los órganos genitales masculinos | 5 | 0.38% |
| Infecciones virales caracterizadas por lesiones de la piel y de las membranas mucosas | 5 | 0.38% |
| Otras enfermedades de los intestinos | 5 | 0.38% |

***Tabla 4.6.*** *Frecuencia de diagnósticos derivado de citas con médico general en pacientes con DM en el periodo de enero a marzo 2017. Se muestran enfermedades con frecuencia igual o mayor a 5 pacientes.*

La tabla 4.6. solo incluye aquellas enfermedades con frecuencia igual o mayor 5, con este criterio se identificaron dieciséis valores distintos. El diagnóstico más común es enfermedades hipertensivas con un 11.33%, seguido por infecciones agudas de las vías respiratorias superiores con 10.88% y enfermedades infecciosas intestinales con 2.03%. Cabe mencionar que el 84.25% de las citas de pacientes diabéticos no tiene registrado ningún dato en el diagnóstico.

En total, en la base de datos existen sesenta diferentes diagnósticos; con frecuencia de 4 (0.30%) hay seis enfermedades, con frecuencia de 3 (0.23%) hay tres diagnósticos, con 2 (0.15%) son diez enfermedades y con 1 solo registro (0.08%) existen veintitrés padecimientos.

### 4.2.4. Verificar la calidad de los datos

Este paso tiene como objetivo asegurar la calidad de los datos que se seleccionaron para el desarrollo del proyecto. Para esto se analizaron las distintas variables para verificar que sus valores contengan el tipo de dato correcto, estén dentro de rango y asegurar que no existan incongruencias.

La variable de sexo es un atributo excluyente cuyos posibles valores para cada hashAfilia (id paciente) son femenino o masculino. En la base de datos se identificaron 69 derechohabientes que tiene asignados ambos sexos, es decir, hay pacientes que en el registro de una cita está identificado como sexo femenino y en otra cita como sexo masculino o viceversa. De estos 69 casos, 3 son de pacientes con diagnóstico de DM, los hashAfilia de estos pacientes son: 3144481, 6105174 y 6159674.

## 4.3. Fase III. Preparación de datos

La fase 3 abarca aquellas actividades para construir el conjunto de datos final, es decir tener los datos listos que se incorporarán a las herramientas de modelado a partir de los datos sin procesar iniciales.

### 4.3.1. Selección de datos

Debido a que el proyecto se realiza con datos que fueron proporcionados por la institución, la elección de las variables fue determinada tomando en cuenta aquellas que conforman la base de datos de citas a médico general. Dicha base de datos cuenta con veintiocho variables, de las cuales el veinte (71.43%) no son importantes para el objetivo del proyecto, por lo que fue necesario determinar criterios para la elección de variables que afectan el resultado final, seleccionando aquellas con importancia media y alta asignada en la tabla 4.6. Debido a que el algoritmo anterior basa su predicción en sexo, edad y herencia, las variables de sexo, grupo de edad y tipo de derechohabiente fueron identificados como atributos importantes que fueron seleccionados. Además, los identificadores de cita y de pacientes también se agregaron. En adición, fueron elegidas las variables de fecha y diagnóstico (sub gpodia) para poder identificar las enfermedades relacionadas (ER) a la DM.

En resumen, en este paso se redujeron las variables de veintiocho a siete, seleccionando aquellas más importantes para el desarrollo de las siguientes fases y pasos.

| Variable | Descriptor de la variable | Formato de variable |
|---|---|---|
| id | Identificador de cita | Numérico |
| fecha | Fecha en que se llevó a cabo la cita | Fecha |
| GpoEdad | Grupo de edad al que corresponde el paciente | String |
| Sexo | Sexo del paciente | String |
| SubGpodia | Diagnóstico derivado de la cita | String |
| tder | Tipo de derechohabiente | String |
| hashAfilia | Identificador de paciente | Numérico |
| Diabetes | ¿Padece de diabetes? | Numérico |

***Tabla 4.7.*** *Descripción de variables seleccionadas, se incluye nombre, descriptor y formato de la variable.*

La variable diabetes fue agregada derivada de la tabla de diagnósticos de diabetes que también proporciono la institución, la relación se hizo a través de la variable clave hashAfilia para identificar aquellos pacientes con diagnóstico de DM.

Debido a que en el campo SubGpodia se cuentan con 133 diagnósticos distintos, se realizó una agrupación de padecimientos por tipo de enfermedades que se relacionan con la DM. Dicha asociación de padecimientos se realizó en base a información de la FID (2019), AAD (2019) y la Encuesta Nacional de Salud y Nutrición (2016), en la cual cuestionan a pacientes sobre otras enfermedades que han sufrido derivadas de la DM (Anexo B).

Los tipos de ER se agrupan en:

- Artropatía
- Padecimientos auditivos
- Padecimientos cardiovasculares
- Gastroparesia
- Nefropatía
- Neuropatía
- Obesidad
- Padecimientos oculares
- Padecimientos de piel

### 4.3.2. Limpieza de datos

En este paso se le dio continuidad a los 69 pacientes que se identificaron diferentes sexos en distintas citas, los cuales se explican en la sección 4.2.4. Los registros de dichos pacientes se eliminaron para análisis posteriores.

En adición, todos aquellos registros en los cuales la variable Sub Gpodia (diagnóstico) tiene el registro “sin dato”, fue excluida para generar los grupos de ER que se presenta en la sección 4.3.3. Estos registros representan el 84.25% de las citas de pacientes diabéticos.

### 4.3.3. Generación de nuevas variables

Debido a que actualmente se trabaja con el algoritmo k-medias, el cual tiene la característica que trabaja con variables numéricas, fue necesario realizar otro proceso para tratar propiamente las variables categóricas. Para esto, se creó un campo nuevo en la base de datos por cada uno de los posibles valores para una misma variable, es decir, se cambia el valor categórico por una variable binaria (“0” o “1”) en cada uno de los campos de cada registro.

Este proceso se realizó en las variables de sexo y edad. En el caso del atributo sexo, la cual es una variable categórica que tiene como posibles valores “femenino” o “masculino”, se crearon dos nuevas columnas con los nombres de los posibles valores, en el cual se registra con “1” en el campo al que pertenezca el sexo de cada paciente. En el caso de la edad se realizó el mismo proceso, pero en este caso se crearon 14 nuevas columnas, ya que son los distintos valores que puede contener la variable Gpo Edad.

Para la creación de los campos de ER, que se establecieron en el segmento 4.3.2, se siguió el mismo proceso de la creación de variables binarias. La diferencia en este caso fue que se agruparon varios diagnósticos para establecer a qué tipo de enfermedad relacionada pertenece. Este proceso se hizo con la asesoría de un médico general, al que se le cuestionó para cada uno de los valores distintos de los

diagnósticos a qué tipo de enfermedad relacionada pertenecía. En la tabla 4.8. se muestra el resultado de esta actividad, en la primera columna se describe el diagnóstico original que se introdujo al sistema y en la segunda columna se identifica el tipo de enfermedad relacionada al que pertenece el diagnóstico.

| Diagnóstico | Tipo de ER |
|---|---|
| Artropatías | Artropatía |
| Defectos De La Coagulación, Purpura Y Otras Afecciones Hemorrágicas | Cardiovascular |
| Dermatitis Y Eczema | Piel |
| Enfermedades De Las Arterias, De Las Arteriolas Y De Los Vasos Capilares | Cardiovascular |
| Enfermedades Del Esófago, Del Estómago Y Del Duodeno | Gastroparesia |
| Enfermedades Del Oído Interno | Auditiva |
| Enfermedades Del Oído Medio Y De La Mastoides | Auditiva |
| Enfermedades Hipertensivas | Cardiovascular |
| Enfermedades Infecciosas Intestinales | Gastroparesia |
| Glaucoma | Ocular |
| Infecciones De La Piel Y Del Tejido Subcutáneo | Piel |
| Infecciones Virales Caracterizadas Por Lesiones De La Piel Y De Las Membranas Mucosas | Piel |
| Insuficiencia Renal | Nefropatía |
| Litiasis Urinaria | Nefropatía |
| Micosis | Piel |
| Obesidad Y Otros Tipos De Hiperalimentación | Obesidad |
| Otras Enfermedades De Los Intestinos | Gastroparesia |
| Otras Enfermedades Del Sistema Digestivo | Gastroparesia |
| Otras Enfermedades Del Sistemas Urinario | Nefropatía |
| Otras Formas De Enfermedad Del Corazón | Cardiovascular |
| Otros Trastornos De Las Piel Y Del Tejido Subcutáneo | Piel |
| Otros Trastornos Del Oído | Auditiva |
| Otros Trastornos Del Sistema Nervioso | Neuropatía |
| Polineuropatías Y Otros Trastornos Del Sistema Nervioso Periférico | Neuropatía |
| Síndromes Del Comportamiento Asociados Con Alteraciones Fisiológicos Y Factores Físicos | Neuropatía |
| Síntomas Y Signos Que Involucran El Sistema Digestivo Y El Abdomen | Gastroparesia |
| Síntomas Y Signos Que Involucran El Sistema Urinario | Nefropatía |
| Síntomas Y Signos Que Involucran La Piel Y El Tejido Subcutáneo | Piel |
| Síntomas Y Signos Que Involucran Los Sistemas Circulatorio Y Respiratorio | Cardiovascular |
| Trastornos De La Conjuntiva | Ocular |
| Trastornos De La Coroides Y De La Retina | Ocular |
| Trastornos De La Esclerótica, Cornea, Iris Y Cuerpo Ciliar | Ocular |
| Trastornos De Las Feneras | Piel |
| Trastornos De Los Nervios, De Las Raíces De Los Plexos Nerviosos | Neuropatía |
| Trastornos Del Cristalino | Ocular |
| Trastornos Del Nervio Óptico Y De Las Vías Ópticas | Ocular |
| Trastornos Del Parpado, Aparato Lagrimal Y Órbita | Ocular |
| Trastornos Metabólicos | Gastroparesia |
| Urticaria Y Eritema | Piel |

***Tabla 4.8.*** *Relaciones de diagnósticos derivados de citas de médico general con tipo de enfermedades relacionadas.*

Cabe mencionar que en la base de datos se identificaron 133 diagnósticos diferentes, en la tabla 4.8. solo se consideran 39, aquellas que se relacionan a una de las 9 ER. El resto al no pertenecer a una enfermedad relacionada, se identifican como “sin relación”.

## 4.4. Modelado

La fase 4 incluye tareas como selección de técnicas de modelado, diseñar de método de evaluación, generación del modelo y la evaluación. Es una fase iterativa, por lo que puede regresar recursivamente a la fase de preparación de datos. Dentro de la misma, además, hay una iteración entre los pasos de generación del modelo y la evaluación de este.

### 4.4.1. Selección de técnicas de modelado

Debido a que el modelo actual trabaja con un algoritmo de agrupamiento k-medias, la optimización a presentar se efectúa en este tipo algoritmo. Además, con la intención de comparar dos algoritmos de agrupamiento distintos, se ejecutará El algoritmo de Agrupación Difusa C-medios o Fuzzy C-means en inglés (FCM).

A continuación en la tabla 4.9, se muestra un comparativo entre estos algoritmos con sus ventajas y desventajas.

| **Algoritmo** | **Tipo de agrupamiento** | **Ventajas** | **Desventajas** |
|---|---|---|---|
| K-medias (Abu Abbas, 2008) | • Duro | • Fácil de implementar.<br>• Puede producir grupos más ajustados que el agrupamiento jerárquico.<br>• Las instancias pueden cambiar de clúster cuando se vuelven a calcular los centroides. | • Elección arbitraria del número de clústers.<br>• Posición inicial de los centroides tiene impacto en los resultados. |
| FCM (Thomas y Nashipudimath, 2012) | • Suave | • Buen desempeño con datos superpuestos.<br>• Los datos pueden pertenecer a más de un clúster. | • Elección arbitraria del número de clústers.<br>• Dificultades para manejar datos atípicos.<br>• Problemas en el manejo de conjuntos de datos de alta dimensión. |

***Tabla 4.9.*** *Ventajas y desventajas de algoritmos k-medias y de agrupamiento jerárquico.*

Los modelos para comparar son el algoritmo actual (k-medias), así como los algoritmos k-medias y FCM con las nuevas variables propuestas.

### 4.4.2. Diseño de método de evaluación

Al tratarse de algoritmos de aprendizaje no supervisado, las métricas de evaluación deben enfocarse en calificar las separaciones de los datos, es decir, buscar en lo medida de lo posible que los miembros que pertenecen a la misma clase sean más similares que miembros de diferentes clases según alguna métrica de similitud.

A continuación, se enlistan y se describen las métricas a utilizar:

- Cohesión: Distancia promedio intra-clúster, es decir, entre todos los elementos de un clúster. Entre menor sea el valor, mejor son los agrupamientos.
- Separación: Distancia promedio inter-clúster, es decir, entre los centroides de los clústers. Entre mayor sea el valor, mejor son los agrupamientos.
- Inercia: La métrica evalúa la sumatoria de las distancias al cuadrado de cada dato al centroide de su grupo.
- Coeficiente de silueta: Es una métrica para evaluar la calidad de los clústers, cuyo objetivo es identificar cuál es el número óptimo de agrupamientos. Una puntuación más alta del coeficiente se relaciona con un modelo con grupos mejor definidos.
- Índice de Calinski-Harabasz: Se basa en la relación entre la varianza entre los grupos y la varianza dentro de los grupos, una puntuación más alta se relaciona con un modelo con grupos mejor definidos.
- Índice de Davies-Bouldin: Este índice significa la similitud promedio entre los grupos, donde la similitud es una medida que compara la distancia entre grupos con el tamaño de los grupos. Cero es el puntaje más bajo posible. Los valores más cercanos a cero indican una mejor partición.

### 4.4.3. Generación del modelo

Debido a que la elección del software fue Google Colaboratory, los archivos necesarios fueron almacenados en Google Drive en formato CSV. para posteriormente cargarlos a Colaboratory.

Antes de ejecutar el modelo como tal, se realizaron las tareas descritas en la sección 4.3 para preparar los datos según lo requiere el algoritmo a ejecutar. Se utilizaron librerías como pandas y numpy para la manipulación de datos, matplotlib para la generación de gráficos, scikit-learn tanto para ejecutar los algoritmos como para efectuar métricas de evaluación, entre otras librerías.

Como un paso anterior a generar el modelo con el algoritmo k-medias, se hizo el análisis del método del codo para la elección del número de clústers (valor de k) más adecuada al conjunto de datos (Thinsungnoen et al., 2015) (Anexo C), el cual resultó en k=5. Una vez obtenido el valor de k, se procedió a ejecutar el código para la obtener el modelo.

La ejecución del algoritmo propuesto con todas las variables de ER no arrojó un resultado mejor al actual en ninguna de las métricas seleccionadas para evaluar los modelos. En consecuencia, se realizó un Análisis de Componentes Principales, o Principal Component Analysis (PCA) en inglés (Anexo D), para disminuir el número de variables, ya sea que no aportan o son menos importantes para explicar el modelo. Como resultado del PCA se pasó de 32 a 23 variables, agregando solo la variable de enfermedades cardiovasculares, las cuales explican el 0.9949% de la varianza.

En adición a buscar mejorar el mismo tipo del modelo actual, es decir, k-medias, se decidió comparar el comportamiento de los datos en una técnica propuesta distinta, en específico el algoritmo de agrupamiento Fuzzy. C-means (FCM) (Anexo E). Para su ejecución se utilizó el paquete fuzzy-c-means; primero se evaluó con 5 grupos, sin embargo un clúster resultó vacío, al ejecutarlo con 4 grupos se presentó la misma

situación, por lo que al final se determinó trabajar con 3 grupos. Los resultados comparados contra el algoritmo propuesto de K-medias se muestran en la tabla 4.10.

Por último, se hizo un análisis estadístico de los clústers resultantes y se obtuvieron el valor de las métricas para posteriormente poder evaluar los distintos algoritmos.

### 4.4.4. Evaluación

En este paso se presentan los resultados que se obtuvieron con cada uno de los modelos que se realizaron. En la tabla 4.10 se resumen las métricas que se obtuvieron.

| Métrica | Algoritmo actual | Algoritmo propuesto (K-medias) | FCM |
|---|---|---|---|
| Clústers | 5 | 5 | 3 |
| Cohesión | 0.9377 | **0.7474** | 1.30800 |
| Separación | **1.7522** | 1.6942 | 0.2873 |
| Inercia | 59,623.21 | **57,030.22** | 61,148.32 |
| Coeficiente de silueta | 0.5007 | **0.5116** | 0.3974 |
| Índice de Calinski-Harabasz | **56,233.59** | 51,031.28 | 40,748.45 |
| Índice de Davies-Bouldin | 0.8931 | **0.8887** | 1.1312 |

***Tabla 4.10.*** *Comparativo de métricas entre algoritmo actual y propuestos K-medias y FCM. A los algoritmos propuestos se agrega la variable de enfermedades cardiovasculares. En negrita se resalta el mejor resultado de cada métrica.*

.

El algoritmo propuesto que se muestra en la tabla 4.10 tiene mejores resultados en 4 de las 6 métricas comparativas, esto derivado del añadido de la variable de enfermedades cardiovasculares y descartar el componente de asegurado. El agregado de dicha característica es la que mejor resultado da en cuanto a las distancias inter-clúster, intra-clúster e inercia. A pesar de que los grupos se encuentran más cercanos entre ellos (métrica de separación), cada grupo presenta mayor cohesión y en total se obtiene menor inercia. La elección de la variable agregada fue resultado del PCA (Anexo D), además de evaluar el modelo con las distintas ER, estos resultados se muestran en la tabla 4.11.

| Métrica | Propuesta (cardiovascular) | Propuesta (cardiovascular, gastroparesia) | Propuesta (cardiovascular, gastroparesia, neuropatía) | Propuesta (cardiovascular, gastroparesia, nefropatía) | Propuesta (cardiovascular, gastroparesia, nefropatía, neuropatía) | Propuesta (cardiovascular, gastroparesia, nefropatía, neuropatía, auditiva, piel) | Propuesta (todas) |
|---|---|---|---|---|---|---|---|
| Clústers (k) | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Cohesión (Intracluster) | **0.7474** | 0.7560 | 0.7574 | 0.7583 | 0.7597 | 0.7636 | 0.7645 |
| Separación (Interclúster) | 1.6943 | 1.6943 | 1.6943 | 1.6943 | 1.6943 | 1.6943 | 1.6943 |
| Inercia | **57,030.22** | 57,875.16 | 57,172.97 | 58,035.03 | 58,602.84 | 59,080.00 | 59,227.80 |
| Coeficiente de Silueta | 0.5116 | 0.4739 | 0.4882 | 0.4840 | 0.4795 | 0.5007 | **0.5252** |
| Índice de Calinski-Harabasz | 49,075.45 | 48,359.02 | **69,299.37** | 50,771.40 | 49,232.49 | 65,797.40 | 53,513.92 |
| Índice de Davies-Bouldin | 0.8887 | 0.8966 | 0.8449 | 0.9301 | 0.9140 | 0.8347 | **0.8342** |

***Tabla 4.11.*** *Resultados de algoritmos propuestos con ER agregadas como variables. Ninguna propuesta considera el componente asegurado. En negritas se resalta el mejor resultado de cada métrica.*

Una vez obtenido los valores de las distintas métricas para las propuestas establecidas en la tabla 4.11, se eligió la que mejor resultados arrojó de acuerdo con uno de los objetivos específicos del proyecto, es decir, la distancia intra-cluster (cohesión), distancia inter-cluster (separación) e inercia. La propuesta con enfermedades cardiovasculares fue el modelo que mostró mejor rendimiento, siendo mejor en dos de las tres métricas que afectan directamente los objetivos, por lo que es la que se seleccionó para el comparativo que se muestra en la tabla 4.10.

A pesar de que la propuesta que incluye todas las enfermedades también tiene 2 métricas con el mejor rendimiento entre los modelos comparados en la tabla 4.11, la forma en que se forman los clústers es igual a la propuesta que solo incluye las enfermedades cardiovasculares. Al estar las métricas de la propuesta elegida más apegada a los objetivos específicos del proyecto, se optó por esta.

El agregado de solo 1 de las 9 ER como variables al modelo, no excluye que los demás padecimientos no puedan aportar mejoría al algoritmo en un futuro, muchas de las enfermedades que no se incluyeron son atendidas por médicos especialistas, por lo que no ve su impacto real al estar trabajando con una base de datos de citas a médico general. A continuación, en la figura 4.3, se muestra un conteo de los casos de las distintas ER en la base de datos.

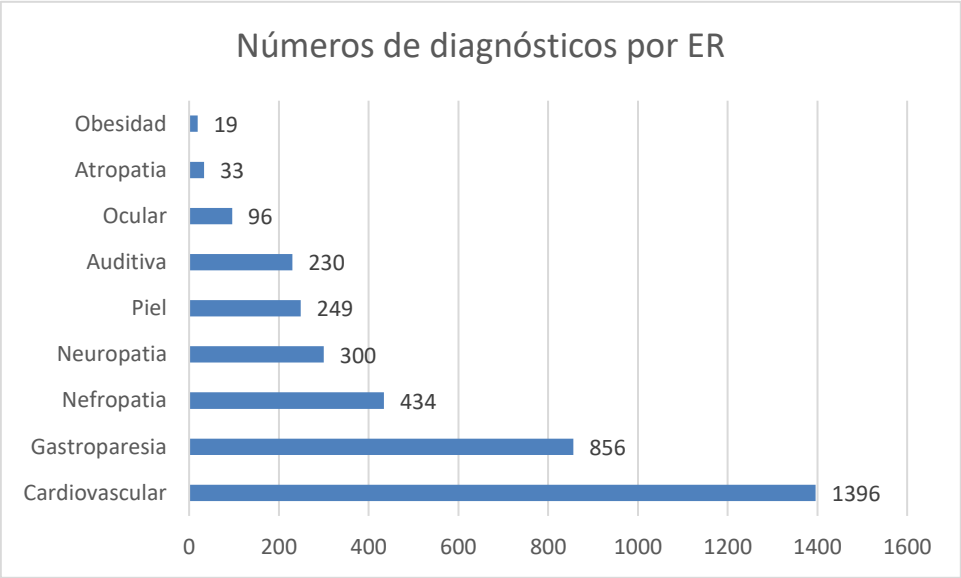



***Figura 4.3**. Frecuencia de diagnósticos de ER en pacientes con cita en médico general.*

Como se puede observar en la figura 4.3, las enfermedades cardiovasculares representan la que más diagnósticos tiene, con el 38.64%, por lo que existen suficientes datos para poder describir a la población respecto a esta característica. Enfermedades como la obesidad y artropatía, que presentan las ocurrencias más bajas, tienen respaldo en la bibliografía médica que son de las comorbilidades más comunes en pacientes con DM, sin embargo, al ser padecimientos que son tratados con médicos especialistas, no se refleja en estos datos que son derivados de citas con médico general.

## 4.5. Evaluación

La fase 5 evalúa los resultados más allá de los resultados técnicos derivado del comparativo de los modelos con el rendimiento de las distintas métricas, como se muestra en la tabla 4.10. Esta valoración incluye los hallazgos encontrados al ejecutar y obtener los resultados del modelo.

### 4.5.1. Evaluación de los resultados

Los hallazgos que son representativos para los objetivos del proyecto se centran, específicamente, en la nueva forma en que el modelo describe los distintos grupos, es

decir, como se componen los clústers, para posteriormente hacer un comparativo entre los grupos del algoritmo actual y el propuesto.

Los grupos del modelo obtenido se describen a continuación:

- Grupo 1: Se conforma por el 32.25% de la población estudiada, el 99.25% con mujeres, el 97.24% registradas como esposas. El 4.68% padece DM y el 2.93% alguna enfermedad cardiovascular. El 83.44% de los pacientes tiene de 35 años en adelante, siendo el grupo de edad con más frecuencia entre 40 y 44 años con el 16.56%.
- Grupo 2: Está compuesto por 21.52% de la población estudiada, en su totalidad de sexo masculino, el 97.17% registrados como esposos y el 1.70% como padres. El 6.90% padece DM y el 4.45% alguna enfermedad cardiovascular. El 86.85% de los pacientes tiene de 35 años en adelante, siendo el grupo de edad con más frecuencia de 65 años en adelante con el 16.93%.
- Grupo 3: Está compuesto por 16.56% de la población estudiada, en su totalidad de sexo masculino y registradas como hijos. El 0.07% padece DM y el 0.48% alguna enfermedad cardiovascular. El 88.66% va de los 0 a 9 años y de 15 a 19 años, el grupo de edad más representativo es de 5 a 9 años con el 33.17%.
- Grupo 4: Se conforma por el 23.04% de la población estudiada, el 100% se conforma por mujeres registradas como hijas. El 0.07% padece DM, el 0.35% alguna enfermedad cardiovascular. El 88.90% de los pacientes tiene entre 0 y 19 años, siendo el grupo de edad con más frecuencia el de 10 a 14 años con el 26.38%.
- Grupo 5: Se conforma por el 6.63% de la población estudiada, en su totalidad de sexo masculino y registrados como hijos. Ninguno padece DM, mientras que el 0.36% sufre alguna enfermedad cardiovascular. El 100% de los pacientes tiene entre 10 y 14 años.

En la tabla 4.12 se presenta un comparativo de los resultados obtenidos en los diferentes clústers con el algoritmo actual y el algoritmo propuesto. En esta se

muestran la proporción de población que representa cada grupo, la proporción de pacientes diagnosticados con diabetes, la edad promedio y el porcentaje según el género de los pacientes que integran cada clúster. Al ser las enfermedades cardiovasculares una nueva variable agregada al modelo sugerido, esta no se considera en la comparación.

| Clúster | Algoritmo Actual | | | | Algoritmo Propuesto | | | |
|---|---|---|---|---|---|---|---|---|
| | % Población | % Diabetes | Edad Media | Género % Fem / % Masc | % Población | % Diabetes | Edad Media | Género % Fem / % Masc |
| 1 | 19.09% | 6.06% | 50 | 98.74% / 1.26% | 32.25% | 4.68% | 47 | 99.25% / 0.75% |
| 2 | 20.91% | 6.90% | 49 | 0.00% / 100% | 21.52% | 6.90% | 50 | 0.00% / 100% |
| 3 | 23.80% | 0.23% | 12 | 0.00% / 100% | 16.56% | 0.07% | 11 | 0.00% / 100% |
| 4 | 23.04% | 0.07% | 11 | 100% / 0.00% | 23.04% | 0.07% | 11 | 100% / 0.00% |
| 5 | 13.16% | 2.68% | 43 | 100% / 0.00% | 6.63% | 0.00% | 12 | 0.00% / 100% |

***Tabla 4.12.*** *Comparativo de resultados entre algoritmo actual y propuesto. El "% Población" representa la proporción del total pacientes del clúster, "% Diabetes" representa la proporción de diagnosticados con DM del total de pacientes del clúster, "Edad Media" es el promedio de edad de los integrantes del clúster, "Genero %Fem / % Masc" es el porcentaje de pacientes de sexo femenino seguido del porcentaje de pacientes masculino del total de integrantes del clúster.*

Existe una diferencia entre los algoritmos en la forma en la que se componen los clústers en cuanto a la edad. En el modelo actual existen 3 grupos cuya edad media está por encima de los 40 años, en específico los grupos 1, 2 y 5; en contra parte, el algoritmo actual solo tiene dos grupos por con edad media por encima de 40 años, es decir, divide más su población en pacientes menores.

Uno de los principales cambios es la proporción de pacientes con diagnóstico de DM. En al algoritmo actual existen tres clústers cuya proporción está por arriba del 2.00%, el grupo 1 con 6.06%, el grupo 2 con 6.90% y el grupo 5 con 2.68%, por su parte, en el algoritmo propuesto solo hay dos clústers que superan el 2.00%, el grupo 1 con 4.68% y el grupo 2 con el 6.90%. También se destaca que en el algoritmo propuesto el grupo 5 cuanta con una proporción de pacientes diabéticos del 0%, situación que no se presenta en el algoritmo actual en ninguno de los clústers.

En el modelo actual, se destacan los clústers 1 y 2 como los pacientes de mayor riesgo, con de 6.06% y 6.90% respectivamente, seguido del clúster 5 con el 2.68% con un riesgo medio, por último, los clústers 3 con 0.23% y el 0.07%, siendo los de riesgo más bajo. En contraste, el algoritmo propuesto se divide en dos grandes grupos; de riesgo alto, compuesto por los grupos 1 y 2 con 4.68% y 6.90% respectivamente, y de riesgo bajo por los clústers 3 con 0.07%, el 4 con el 0.07% y el 5 con el 0.00%.

## 4.6. Despliegue

En esta fase se debe organizar el conocimiento adquirido gracias al proceso de MD y presentarlo de una manera que sea utilizable en el contexto del instituto. Es importante clarificar que el departamento de medicina preventiva cuenta con una estrategia de la que es parte el algoritmo actual, por lo que el modelo propuesto lo sustituirá sin modificar el proceso de ejecución o de toma de decisiones. Por lo tanto, esta fase se enfoca en la generación de un reporte final.

### 4.6.1. Generación del reporte final

Se incluye un resumen de lo generado a lo largo del proyecto, incluyendo experiencias obtenidas en el desarrollo de este.

- Descripción detallada del problema original: La institución utiliza un algoritmo que tiene un rango de mejora en la forma en que describe a la población para realizar predicciones; dicho modelo basa sus decisiones en edad, sexo y herencia.

- El proceso utilizado para realizar la MD: La estrategia utilizada fue CRISP-DM, el cual es un proceso estándar para el desarrollo de proyectos de MD enfocado en los problemas y las limitaciones de este tipo de proyectos en el sector salud. Los pasos de la estrategia son análisis del problema, análisis de datos, preparación de datos, modelado, evaluación y despliegue.

- Notas sobre cualquier desviación del plan original del proyecto: Unos de los requerimientos solicitados por parte del instituto fue hacer un análisis de

comorbilidad de pacientes diabéticos con artropatía, así como considerar su incorporación al modelo. Sin embargo, aunque se agregó en las ER analizadas, no fue de significancia para el modelo, debido a que este tipo de padecimientos los atiende un médico especialista y los datos con los que se trabajó son solo citas a médico general.

- Resumen de los resultados de la MD, tanto de modelos como hallazgos: Los resultados obtenidos de comprar el algoritmo actual y el propuesto se muestran em la tabla 4.13.

| Métrica | Algoritmo actual | Algoritmo propuesto (K-medias) |
|---|---|---|
| Clústers | 5 | 5 |
| Cohesión | 0.9377 | **0.7473** |
| Separación | **1.7522** | 1.6942 |
| Inercia | 59,623.21 | **57,024.58** |

***Tabla 4.13.*** *Comparativo entre algoritmo actual y propuesto. En negritas se resalta el mejor resultado por métrica*

El algoritmo propuesto es mejor en cohesión (distancia intra-clúster) e inercia. A pesar de que el modelo actual se comporta mejor en separación (distancia inter-clúster), se concluye que el algoritmo propuesto es preferible por comportarse mejor en 2 de las 3 variables que se compararon.

En adición, se presentan resultados obtenidos a partir del análisis de las ER en pacientes diabéticos.

En la tabla 4.14 se desglosan los padecimientos identificados en pacientes con DM y comorbilidad de 1 ER. Entre las enfermedades más comunes se encuentran padecimientos cardiovasculares, gastroparesia, nefropatía y piel.

| ER | Frecuencia | Porcentaje (%) |
|---|---|---|
| Cardiovascular | 120 | 56.87% |
| Gastroparesia | 31 | 14.69% |
| Nefropatía | 15 | 7.11% |
| Piel | 15 | 7.11% |
| Neuropatía | 14 | 6.64% |
| Ocular | 7 | 3.32% |
| Auditiva | 6 | 2.84% |
| Obesidad | 2 | 0.95% |
| Artropatía | 1 | 0.47% |
| Total | 211 | 100.00% |

***Tabla 4.14.*** *Distribución de pacientes con diagnóstico de DM y 1 ER.*

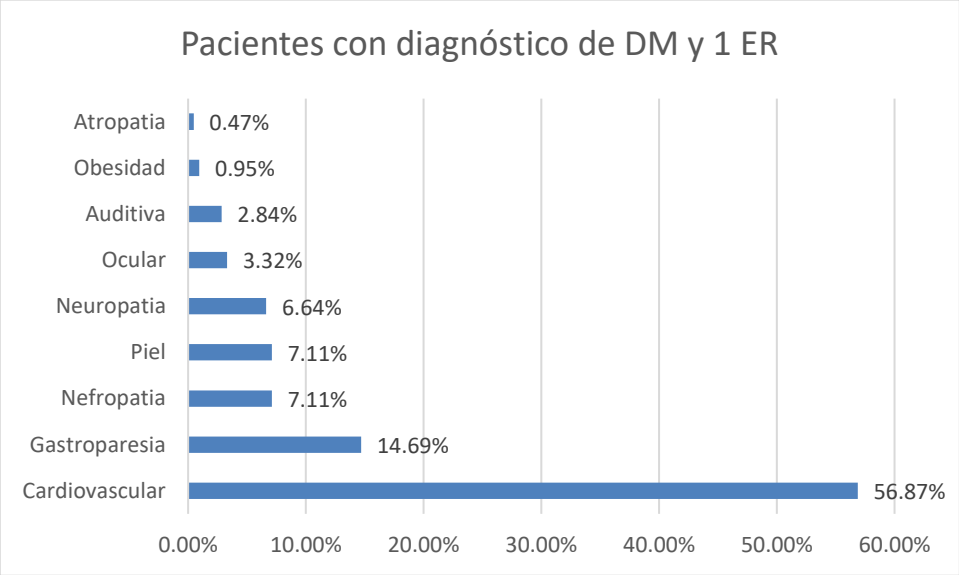



***Figura 4.4.*** *Distribución de pacientes con diagnóstico de DM y 1 ER*

En la figura 4.4 se muestra que las 9 ER se presentan en pacientes con 1 comorbilidad. Además, más de la mitad (56.87%) presenta un padecimiento cardiovascular. Por último, las principales cuatro enfermedades, es decir, cardiovasculares, gastroparesia, nefropatía y piel representan el 85.78% de los pacientes.

En la tabla 4.15 se desglosan los padecimientos identificados en pacientes con DM y comorbilidad con 2 ER.

| ER | Frecuencia | Porcentaje (%) |
|---|---|---|
| Cardiovascular y Gastroparesia | 4 | 33.33% |
| Cardiovascular y Piel | 1 | 8.33% |
| Cardiovascular y Neuropatía | 3 | 25.00% |
| Cardiovascular y Artropatía | 3 | 25.00% |
| Gastroparesia y Neuropatía | 1 | 8.33% |
| Total | 12 | 100.00% |

***Tabla 4.15.*** *Distribución de pacientes con diagnóstico de DM y comorbilidad con 2 ER.*

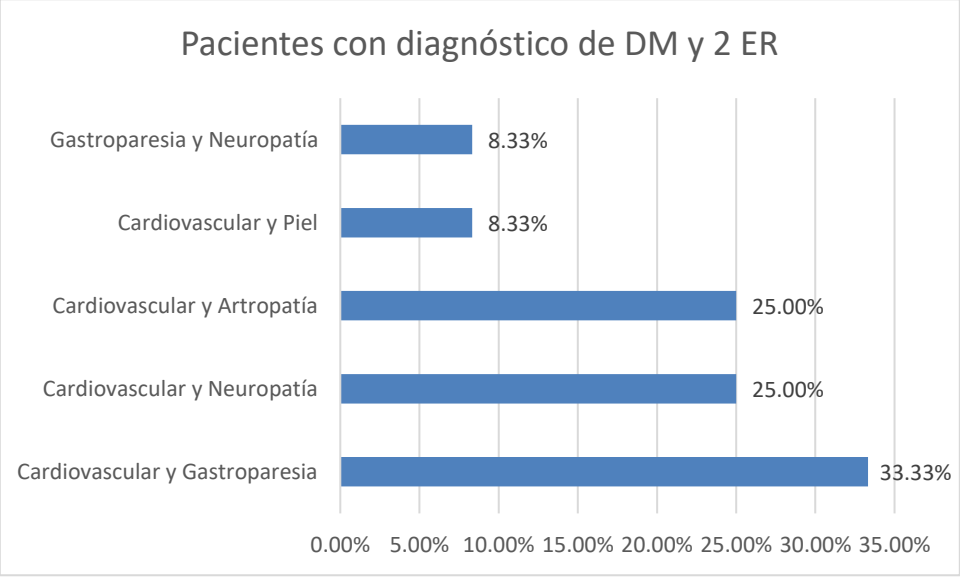



***Figura 4.5.*** *Distribución de pacientes con diagnóstico de DM y comorbilidad con 2 ER.*

En la figura 4.5 se muestra que solo se presentan 5 de las 9 enfermedades analizadas en 5 distintas parejas de enfermedades. Además, en el 91.67% de los casos, al menos una de las enfermedades es un padecimiento cardiovascular. Por último, las principales tres parejas, es decir, cardiovascular y gastroparesia, cardiovascular y neuropatía, así como cardiovascular y artropatía, representan el 83.33%.

En adición, se presentan resultados obtenidos a partir del análisis de las ER en pacientes sin diabéticos.

En la tabla 4.16 se desglosan los padecimientos identificados en pacientes sin diagnóstico de DM y 1 ER. Entre las enfermedades más comunes se encuentran padecimientos cardiovasculares, gastroparesia, nefropatía y neuropatía.

| ER | Frecuencia | Porcentaje (%) |
|---|---|---|
| Cardiovascular | 1169 | 37.97% |
| Gastroparesia | 740 | 24.03% |
| Nefropatía | 387 | 12.57% |
| Neuropatía | 234 | 7.60% |
| Piel | 211 | 6.85% |
| Auditiva | 210 | 6.82% |
| Ocular | 85 | 2.76% |
| Artropatía | 26 | 0.84% |
| Obesidad | 17 | 0.55% |
| Total | 3079 | 100.00% |

***Tabla 4.16.*** *Distribución de pacientes sin diagnóstico de DM y 1 ER.*

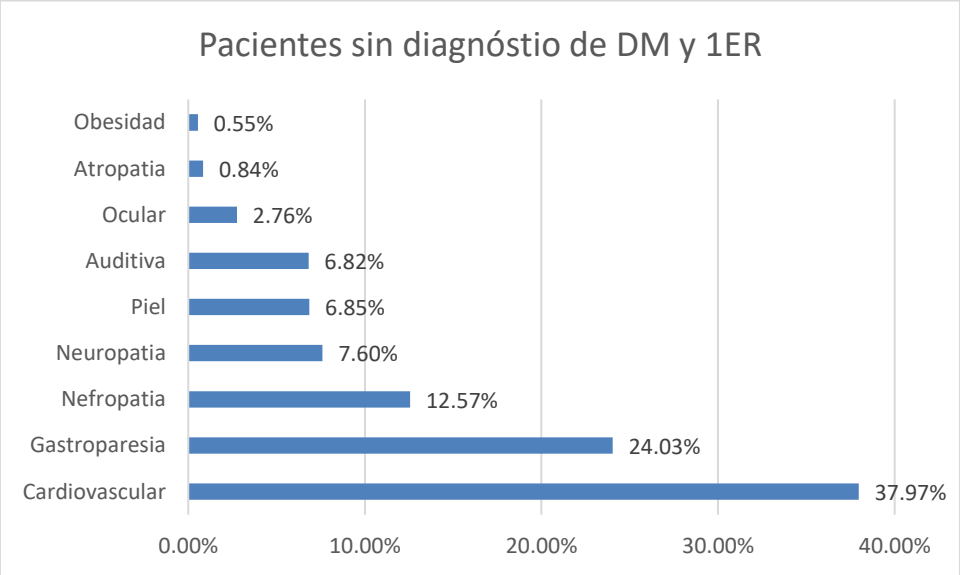



***Figura 4.6.*** *Distribución de comorbilidades en pacientes sin diagnóstico de DM y 1 ER.*

En la figura 4.6 las enfermedades cardiovasculares representan más de un tercio de los casos con el 37.97%. Los primeras tres padecimientos son las mismas que en pacientes con diagnóstico de DM y 1 ER, es decir, cardiovasculares, gastroparesia y nefropatía. Sin embargo, en este caso, neuropatía es el cuarto diagnóstico más frecuente, hasta esta enfermedad representan el 82.17%.

En la tabla 4.17, se desglosan los padecimientos identificados en pacientes sin DM y 2 ER.

| ER | Frecuencia | Porcentaje (%) |
|---|---|---|
| Cardiovascular y Gastroparesia | 42 | 29.58% |
| Cardiovascular y Neuropatía | 31 | 21.83% |
| Cardiovascular y Nefropatía | 10 | 7.04% |
| Gastroparesia y Nefropatía | 9 | 6.34% |
| Gastroparesia y Neuropatía | 8 | 5.63% |
| Gastroparesia y Piel | 8 | 5.63% |
| Gastroparesia y Auditiva | 6 | 4.23% |
| Cardiovascular y Piel | 5 | 3.52% |
| Cardiovascular y Auditiva | 4 | 2.82% |
| Neuropatía y Nefropatía | 4 | 2.82% |
| Nefropatía y Piel | 4 | 2.82% |
| Gastroparesia y Artropatía | 2 | 1.41% |
| Gastroparesia y Ocular | 2 | 1.41% |
| Neuropatía y Auditiva | 2 | 1.41% |
| Cardiovascular y Artropatía | 1 | 0.70% |
| Cardiovascular y Ocular | 1 | 0.70% |
| Neuropatía y Piel | 1 | 0.70% |
| Nefropatía y Auditiva | 1 | 0.70% |
| Piel y Ocular | 1 | 0.70% |
| Total | 142 | 100.00% |

***Tabla 4.17.*** *Distribución de pacientes sin diagnóstico de DM y 2 ER.*

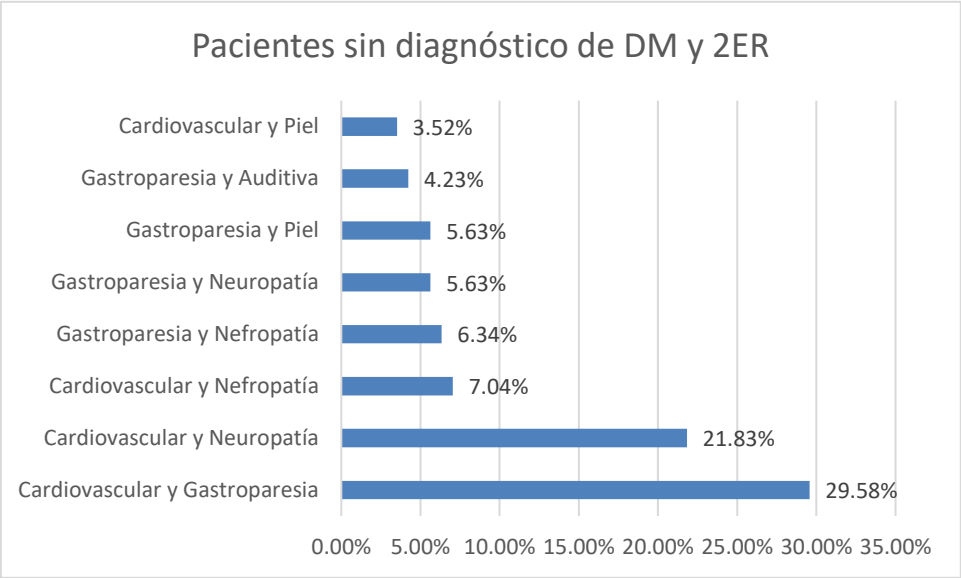



***Figura 4.7.*** *Distribución de pacientes sin diagnóstico de DM y 2 ER. Se muestran solo las parejas con frecuencia igual o mayor a 5.*

En total se presentan 19 parejas distintas de padecimientos, entre las siete parejas con más frecuencia representan el 80.28%.

En la figura 4.7 se muestra que las enfermedades cardiovasculares se presentan en las tres parejas con más frecuencias, entra estas representan el 61.97%. Por su parte, gastroparesia se presenta en cinco diferentes comorbilidades, entra estas representan el 51.41%.

En adición, en pacientes sin diagnósticos de DM se encontraron casos que presentan 3 ER, estos se muestran en la tabla 4.18.

| ER | Frecuencia | Porcentaje (%) |
|---|---|---|
| Cardiovascular, Nefropatía y Auditiva | 1 | 20.00% |
| Cardiovascular, Gastroparesia y Piel | 1 | 20.00% |
| Gastroparesia, Neuropatía y Nefropatía | 1 | 20.00% |
| Gastroparesia, Nefropatía y Piel | 1 | 20.00% |
| Neuropatía, Nefropatía y Piel | 1 | 20.00% |
| Total | 5 | 100.00% |

***Tabla 4.18.*** *Distribución de pacientes sin diagnóstico de DM y 3 ER.*

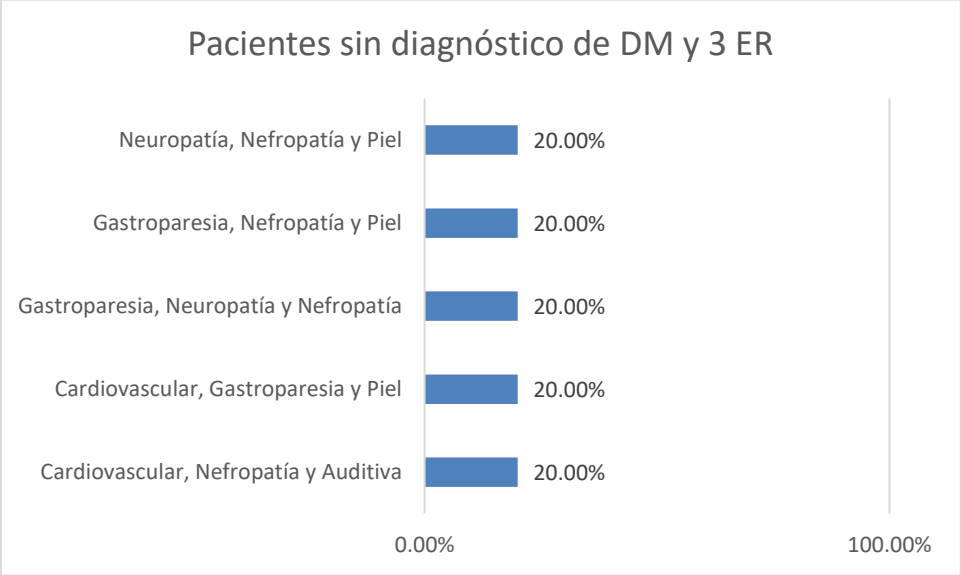



***Figura 4.8.*** *Distribución de pacientes sin diagnóstico de DM y 3 ER.*

Como se muestra en la figura 4.8, solo se presentan 5 casos de pacientes sin diagnóstico de DM y 3 ER, todos con componentes distintos en sus comorbilidades. A pesar del bajo número, tienen un gran impacto, ya que ni en pacientes con DM se presentan hasta 3 ER, estas cinco personas pertenecen a un grupo de alto riesgo de padecer DM.

A continuación, se presenta en la tabla 4.16 un resumen de las comorbilidades por separado en pacientes con diagnóstico de DM, se presenta el número de casos totales, así como su desglose en porcentaje por sexo.

| ER | Frecuencia | Porcentaje (%) | % Mujeres | % Hombres |
|---|---|---|---|---|
| Cardiovascular | 131 | 6.48% | 50.38% | 49.62% |
| Gastroparesia | 36 | 1.78% | 58.33% | 41.67% |
| Neuropatía | 18 | 0.89% | 50.00% | 50.00% |
| Nefropatía | 15 | 0.74% | 46.67% | 53.33% |
| Auditiva | 6 | 0.30% | 50.00% | 50.00% |
| Piel | 16 | 0.79% | 62.50% | 37.50% |
| Artropatía | 4 | 0.20% | 25.00% | 75.00% |
| Ocular | 7 | 0.35% | 85.71% | 14.29% |
| Obesidad | 2 | 0.10% | 50.00% | 50.00% |
| Total | 235 | 11.62% | 52.77% | 47.23% |

***Tabla 4.19.*** *Distribución de ER en pacientes con diagnóstico de DM y desglose por sexo.*

Los padecimientos cardiovasculares son los que más se presentan en pacientes diabéticos con 6.48%. Además, se puede concluir que los padecimientos cardiovasculares, gastroparesia, de piel y ocular se presentan en mayor proporción en pacientes de sexo femenino. Por su parte, los padecimientos como nefropatía y artropatía se presentan mayormente en hombres. Por último, los padecimientos de neuropatía y obesidad se reparten un 50% entre ambos sexos.

- Recomendaciones para el trabajo adicional de MD: Para poder tener un panorama completo del resto de ER que se analizaron y no fueron incluidas como variables en el algoritmo propuesto (artropatía, padecimientos auditivos, padecimientos cardiovasculares, gastroparesia, nefropatía, neuropatía, obesidad, padecimientos oculares y padecimientos de piel) se necesitaría tener una sola base de datos con citas de médico general y de especialistas, esto con el objetivo de ver el impacto real que tienen en la DM.

# 5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

En el presente trabajo se utilizó como base unas de las principales metodologías para proyectos de MD para aplicar una estrategia que permitiera aprovechar y generar nuevo conocimiento de la base de datos de médico general para su uso en el departamento de medicina preventiva.

## 5.1. Conclusiones

En el presente estudio se presentó el refinamiento de los resultados de un algoritmo K-medias para la atención a pacientes diabéticos, la propuesta para mejorar el modelo fue incluir enfermedades relacionadas (ER) a la diabetes mellitus (DM). Se analizaron 9 diferentes padecimientos: cardiovasculares, gastroparesia, neuropatía, nefropatía, artropatía, obesidad, padecimientos auditivos, oculares y de la piel. Tras analizar el algoritmo incluyendo las ER, el mejor resultado se presentó incluyendo solo las enfermedades cardiovasculares; con esta propuesta se logró mejor resultado en 4 de las 6 métricas comparativas.

Dentro de los objetivos específicos se estableció el analizar las características de los pacientes con DM. Dichas características que definieron el refinamiento del algoritmo fueron establecidas a través de las ER. El 10.44% de los pacientes padecen DM y comorbilidad con una ER, mientras que el 0.59% padece comorbilidad con dos ER. Dentro del grupo de pacientes con una ER los problemas cardiovasculares es el padecimiento que más se presenta con un 56.87%. Las principales 4 enfermedades, es decir, cardiovasculares, gastroparesia, nefropatía y problemas de piel representan el 85.78% de los pacientes. En el grupo de pacientes con dos ER el par de padecimientos más comunes que suceden concurrentemente son cardiovascular y gastroparesia con 33.33%, seguido de cardiovascular y neuropatía con un 25.00%.

Otro aspecto importante que se asocia a las ER es el sexo. Los padecimientos cardiovasculares, gastroparesia, así como problemas oculares y de piel, se presentan en mayor proporción en pacientes de sexo femenino; se destaca que el 85.71% de los casos de enfermedades oculares se presentan en mujeres. Por su parte, nefropatía y artropatía se presentan mayor proporción en pacientes de sexo masculino. Los padecimientos de neuropatía, obesidad y problemas auditivos se comportan de manera similar en ambos sexos, repartido los casos en un 50%.

La aplicación de la MD en el sector salud se ha vuelto un elemento necesario para poder brindar un mejor servicio médico a los derechohabientes, por consecuencia, la mejora de un algoritmo de MD se vuelve una parte esencial en la mejora continua de los procesos de medicina preventiva. En este caso, al refinar el algoritmo e incluir nuevas variables, permitió crear mejores conclusiones con la misma información que se recolecta día a día en las consultas médicas, pero que se dejaban fuera en el análisis que realiza el algoritmo actual.

Dicha mejora en el proceso del departamento de medicina preventiva es en esencia el incrementar la obtención de conocimiento con los mismos datos. Dicho conocimiento tiene un impacto directo en el área para ayudar en la detección precoz de la DM, así como determinar sus causas e identificar grupos de riesgo.

En adición, el agregado de variables permite describir de mejor manera a los pacientes, por lo que, con ayuda de la estrategia de medicina preventiva y haciendo uso de la aplicación y redes sociales del instituto, la información que se le hace llegar a cada uno de los derechohabientes es más exacta; esto no solo es de utilidad para generar dichos mensajes, sino que también el recibir mensajes con información más específica puede llegar a generar mejor respuesta en las personas.

Por último, los derechohabientes también se ven favorecidos, ya que se logra mejorar la prevención y detección de una de las principales enfermedades de causa de muerte a nivel nacional y estatal, como lo es la DM. Además de fomentar la prevención, el programa de medicina preventiva también se vuelve capaz de atacar uno de los

principales problemas dicho padecimiento, el cual es que la mitad de las personas que padecen esta enfermedad no son diagnosticadas.

## 5.2. Recomendaciones

Los puntos que a continuación se explican surgieron a partir de esta investigación en la institución de salud:

- Se recomienda tener el historial general de los pacientes en una sola base de datos, ya que, al contar solamente con datos de citas a médico general, hubo padecimientos que no tuvieron significancia para el modelo, pero existe sustento en la bibliografía que enfermedades como la artritis y la obesidad son comorbilidades presentes en pacientes diabéticos.
- Poder incluir información de pacientes sobre índice de masa corporal, estatura, peso, así como niveles de glucosa en sangre, ya que estos componentes podrían convertir las salidas del modelo en resultados más precisos.
- Se sugiere que, para un mayor impacto del proyecto, la información generada para el departamento de medicina preventiva sea utilizada para fortalecer la estrategia del departamento, es decir, que a partir del modelo se segmente la población a la que se envía mensajes a través de la aplicación y redes sociales, sin esperar a que el paciente acuda a cita al instituto para generar un aviso preventivo.
- Se aconseja mejorar el sistema de información en diferentes aspectos para disminuir el preprocesamiento de datos. Entre los cambios que se recomiendan es que se vuelva obligatorio el llenado de todos los atributos, ya que un campo tan importante como el diagnostico se encuentra vacío en más del 50% de los registros en la base de datos que se utilizó para el proyecto. En adición, crear estándares para los nombres de las enfermedades para evitar que un solo padecimiento sea registrado de distintas formas y, por último, impedir que una sola persona tenga ambos sexos en registros separados, de ser así crear alguna alerta para corregir el dato al momento.

- Debido a que el personal médico son los encargados de alimentar el sistema, sería favorable capacitar a los doctores para explicar la importancia del llenado correcto dónde se pueden explicar o llegar acuerdo con los estándares de los padecimientos, esto con el fin mejorar la calidad de datos que eventualmente se puedan convertir en información y conocimiento.

## 5.3. Trabajos futuros

Una vez identificadas las enfermedades que se relacionan a los pacientes con diagnóstico de DM, un paso posterior sería identificar los padecimientos temporales considerando la cronología, es decir, el patrón secuencial de aparición de las enfermedades.

Los resultados de esta propuesta de estudio futuro pueden dar mejor idea de los mecanismos que conducen a padecer diabetes y mejores formas de detectar, tratar y controlar desde los primeros indicios que se presentan con los padecimientos de las enfermedades relacionadas.

# 6. REFERENCIAS

Abu Abbas, O., 2008. Comparisons Between Data Clustering Algorithms. Int. Arab J. Inf. Technol, Volumen 5, pp. 320-325.

Asosiación Americana de Diabetes, 2019. Diabetes Disponible en: www.diabetes.org/diabetes. [Accedido: 01 Octubre 2019].

Asosiación Americana de Salud Pública, 2019. What is public health. Disponible en: www.apha.org/what-is-public-health. [Accedido: 19 Septiembre 2019].

Belciug, S., 2009. Patients length of stay grouping using the hierarchical clustering algorithm. Annals of University of Craiova, Math. Comp. Sci. Ser., 36(2), pp. 79-84.

Bellazzi, R. y Zupan, B., 2008. Predictive data mining in clinical medicine: Current issues. international journal of medical informatics, 77(2), pp. 81-97.

Bergman, M. et al., 2012. Diabetes prevention: global health policy and perspectives from the ground. Diabetes Management, Volumen 2, pp. 309-321.

Bhramaramba, R., Allam, A. R., Kumar, V. V. y Sridhar, G. R., 2011. Application of data mining techniques on diabetes related proteins. International Journal of Diabetes in Developing Countries, 31(1), pp. 22-25.

Buczak, A. et. al., 2012. A data-driven epidemiological prediction method. BMC Medical Informatics and Decision Making, Volumen 12, pp. 124-144.

Cataloluk, H. y Kesler, M., 2012. A diagnostic software tool for skin diseases with basic and wighted K-NN. Bilecik, s.n., pp. 1-4.

Chapman, P. et al., 2000. CRISP-DM 1.0: Step-by-step data mining guide.

Cios, K. y Moore, G., 2002. Uniqueness of medical data mining. Artificial Intelligence in Medicine, Volumen 26, pp. 1-24.

Dalkir, K., 2005. Knowledge Management in Theory and Practice. Third ed. Burlington, MA: Butterworth-Heinemann.

Dunham, M. H., 2003. Data mining introductory and advanced topics. 1st ed. New Jersey: Pearson Education.

Durairaj, M. y Ranjani, V., 2013. Data Mining Applications in Healthcare: A Study. International Journal of Scientific & Technology Research, 2(10), pp. 29-35.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. Artificial Intelligence Magazine, 17(3).

Federación Internacional de Diabetes, 2017. IDF Diabetes Atlas, Bruselas. Disponible en: www.idf.org/aboutdiabetes/what-is-diabetes.html. [Accedido: 01 Octubre 2019].

Federación Internacional de Diabetes, 2019. International Diabetes Federation. Disponible en: www.idf.org/aboutdiabetes/what-is-diabetes.html. [Accedido: 01 Octubre 2019].

García Perez, A. A. y García Bertrand, F., 2012. La medicina preventiva en la atención primaria de salud. Revista Habanera de Ciencias Médicas, 11(2), pp. 308-316.

Gupta, A. K. y Bora, D. J., 2014. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. International Journal of Computer Trends and Technology (IJCTT) , 10(2), pp. 108-113.

Gupta, A., Shivhare, H. y Sharma, S., 2015. Recommender system using fuzzy c-means clustering and genetic algorithm based weighted similarity measure. International Conference on Computer, pp. 1-8.

Haraty, R., Dimishkieh, M. y Masud, M., 2015. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. International Journal of Distributed Sensor Networks, 11(6), pp. 1-11.

Hartono, Sitompul, O. S., Tulus y Nababan, E. B., 2018. Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem. IOP Conference Series: Materials Science and Engineering, 288(1).

Hernández Samperi, R., Fernández Collado, C. y Baptista Lucio, P., 2014. Metodología de la investigación. 6ta edición ed. México: Mc Graw Hill Education.

Iavindrasana, J. et al., 2009. Clinical Data Mining: a Review. IMIA Yearbook of Medical Informatics, pp. 121-133.

INEGI, 2018. Características de las defunciones registrasdas en México durante 2017. Disponible en: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2018/EstSociodemo/DEFUNCIONES2017.pdf. [Accedido: 10 Septiembre 2019].

Jothi, N., Rashid, N. A. y Husain, W., 2015. Data Mining in Healthcare – A Review. Procedia Computer Science, Volumen 72, pp. 306-313.

Kavakiotis, I. et al., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, Volumen 15, pp. 104-116.

Kim, H. S., Shin, A. M., Kim, M. K. y Kim, Y. N., 2012. Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining. Korean J Intern Med, 27(2), pp. 197-202.

Kumar, C. S., Govardhan, A. y Srinivas, S., 2014. Data Mining Issues and Challenges in Healthcare Domian. International Journal of Engineering Research & Technology, 3(1).

Lamont, J., 2012. Big data has big implications for knowledge management. KM World Magazine, Abril.21(4).

Liao, S. H., Chu, P. H. y Hsiao, P. Y., 2012. Data mining techniques and applications - A decade review from 2000 to 2011. Expert Systems with Applications, 39(12), pp. 11303-11311.

Llanes, M., 2018. Una metodología para la detección de brotes epidemiológicos de rápida propagación utilizando minería de datos y análisis de georreferenciación. Universidad de Sonora.

Luo, Q., 2008. Advancing Knowledge Discovery and Data Mining. Australia, Chinese Society for Electrical Engineering - IEEE.

Mahindrakar, P. y Hanumanthappa, M., 2013. Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges. Int. Journal of Engineering Research and Applications, 3(6), pp. 937-941.

Milley, A., 2000. Healthcare and data mining. Health Management Technology, 21(8), pp. 44-47.

Niaksu, O., 2015. CRISP Data Mining Methodology Extension for Medical Domain. Baltic J. Modern Computing, 3(2), pp. 92-109.

Organización Mundial de la Salud, 2008. Health Metrics Network Framework and Standards for Country Health Information Systems. Disponible en: https://www.who.int/healthinfo/statistics/toolkit_hss/EN_PDF_Toolkit_HSS_InformationSystems.pdf. [Accedido: 03 Agosto 2020].

Organización Mundial de la Salud, 2017. Global Observatory for eHealth. Disponible en: https://www.who.int/goe/en/. [Accedido: 03 Agosto 2020].

Organización Mundial de la Salud, 2019. Diabettes Mellitus. Disponible en: www.who.int/topics/diabetes_mellitus/es/. [Accedido: 24 Septiembre 2019].

Patel, S. y Patel , H., 2016. Survey of Data Mining Techniques used in Healthcare Domain. International Journal of Information Sciences and Techniques, 6(1), pp. 9-10.

Patil, S. L., 2015. Survey of Data Mining Techniques in Healthcare. International Research Journal of Innovative Engineering, 1(9), pp. 1-3.

Piédrola, G., 2015. Medicina preventiva y salud pública. 12a ed. Barcelona: ELSEVIER.

Rajagaru, H. y Prabhakar, S. K., 2017. KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. A Detailed Analysis. First ed. Hamburg: Anchor Academic Publishing.

Sagar, H. K. y Sharma, V., 2014. Error Evaluation on K- Means and Hierarchical Clustering with Effect of Distance Functions for Iris Dataset. International Journal of Computer Applications, 86(16).

Salleras, L., 1985. Educación sanitaria: principios, métodos y aplicaciones. Primera ed. Madrid: Ediciones Diaz de Santos.

Sanez, E., 2018. Minería de datos para una estrategia de medicina preventiva más robusta en una institución de salud pública del estado de Sonora. Uiversidad de Sonora.

Seifert, J., 2004. Data Mining: An Overview. s.l., Congressional Research Service.The Library of Congress.

Sharmila, K. y Vethamanickam, S. A., 2015. Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform. International Journal of Emerging Technology and Advanced Engineering, 5(1), pp. 567-571.

Silva-Cárcamo, H., 2016. Comorbilidades en los Pacientes con Diabetes Mellitus Tipo 2 del Instituto Nacional del Diabético. iMedPub Journals, Volumen 12.

Sondwale, P., 2015. Overview of Predictive and Descriptive Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 5(4), pp. 262-265.

Syakur, M., Khotimah, B., Rochman, E. y Dwi Satoto, B., 2018. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conference Series: Materials Science and Engineering, Volumen 336.

Thinsungnoen, T. et al., 2015. The Clustering Validity with Silhouette and Sum of Squared Errors. International Conference on Industrial Application Engineering, pp. 44-51.

Thomas, B. y Nashipudimath, M., 2012. Comparative Analysis Of Fuzzy Clustering Algorithms In Data Mining. nternational J ournal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) , 1(7).

Veloso, R. et al., 2014. A Clustering Approach for Predicting Readmissions in Intensive Medicine. Procedia Technology, Volumen 16, pp. 1307-1316.

Ward, J. y Joe, P., 2002. Strategic Planning for Information Systems. Third ed. UK: Wiley.

Wickramasinghe, S., Sharma, S. K. y Gupta, J. N. D., 2008. Knowledge Management in Healthcare. Medical Informatics: Concepts, Methodologies, Tools, and Applications, pp. 186-197.

Yi, P., Gang, K., Yong, S. y Zhengxin, C., 2008. A Descriptive Framework for the Field of Data Mining and Knowledge Discovery. International Journal of Information Technology and Decision Making, 7(1), pp. 639-682.

Yoo, I. et al., 2012. Data mining in healthcare and biomedicine: A survey of the literature. Journal of Medical Systems, 36(4), pp. 2431-2448.

Young, R., 2010. Knowledge Management Tools and Techniques Manual. UK: Asian Productivity Organisation.

Zhan, J., 2008. Privacy-preserving collaborative data mining. IEEE Computational Intelligence Magazine, 3(2), pp. 31-41.

# 7. ANEXOS

## 7.1. Anexo A

A continuación se muestra la encuesta que tiene disponible la AAD para identificar el riesgo de padecer diabetes.


# ¿Su madre, padre, hermana o hermano tiene diabetes?

¿Por qué importa esto?

Los antecedentes familiares de diabetes podrían contribuir a su riesgo de desarrollar diabetes tipo 2.

NO SÍ

# ¿Alguna vez le han diagnosticado hipertensión?

¿Por qué importa esto?

Tener hipertensión contribuye a su riesgo general de desarrollar diabetes tipo 2.

NO SÍ

# ¿Es físicamente activo?

¿Por qué importa esto?

Ser inactivo puede aumentar su riesgo de desarrollar diabetes tipo 2.

NO SÍ

# ¿Qué raza o grupo étnico lo describe mejor?

¿Por qué importa esto?

Las personas de ciertos grupos raciales y étnicos tienen más probabilidades que otras de desarrollar diabetes tipo 2.

| | |
|---|---|
| BLANCO | ASIÁTICO |
| HISPANO O LATINO | INDÍGENA AMERICANO O NATIVO DE ALASKA |
| NEGRO O AFROAMERICANO | NATIVO DE HAWAII U OTRA ISLA DEL PACÍFICO |
| OTRO | PREFIERO NO DECIRLO |

# ¿Cuál es su altura y peso?

¿Por qué importa esto?

La combinación de su peso y altura nos permite conocer su índice de masa corporal o IMC. Las personas con un IMC más alto tienen un mayor riesgo de desarrollar diabetes tipo 2.

Altura ▾ Peso ▾

## 7.2. Anexo B

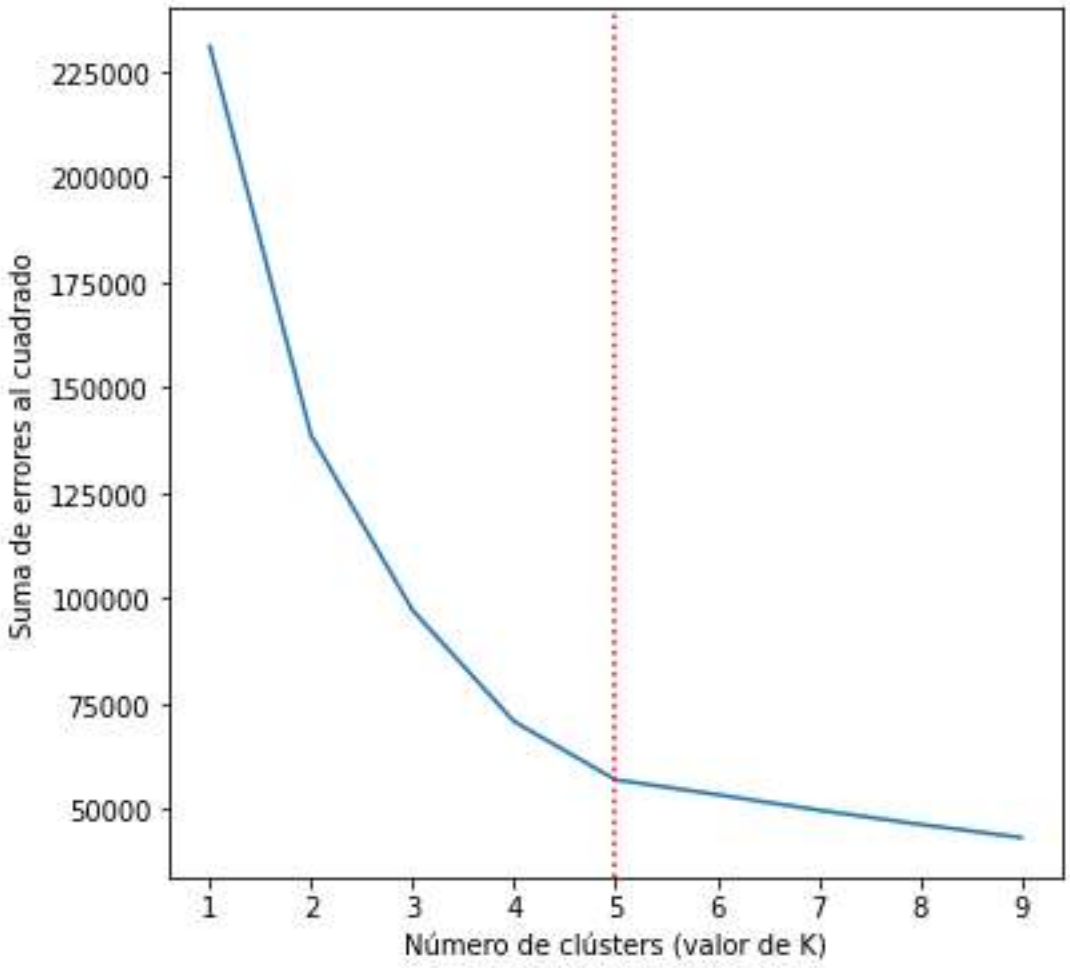

**ENSANUT-MC 2016**

**DOCUMENTACION DE LA BASE DE DATOS**

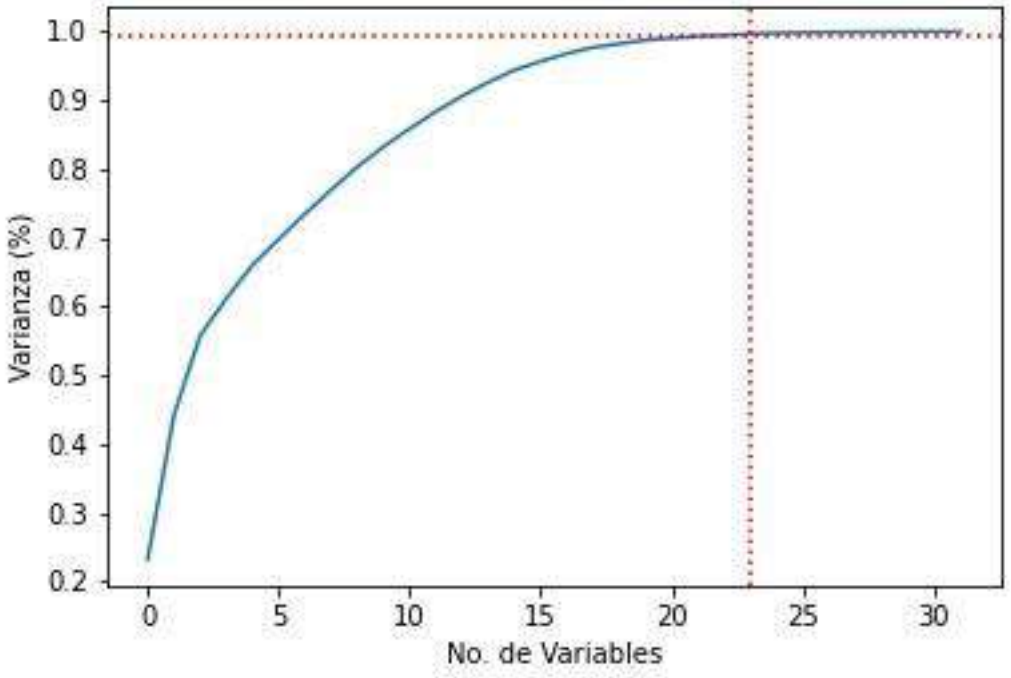
**Información de la Base: adultos_cronicas**

| Variable | Descriptor de variable | Ancho de columna | Formato de variable |
|---|---|---|---|
| a313a | ¿debido a la diabetes: ulceras en piernas o pies tarden sanar más de 4 semanas | 8 | NUMERICO |
| a313b | ¿debido a la diabetes le han amputado alguna parte del cuerpo? | 8 | NUMERICO |
| a313c | ¿debido a la diabetes le ha disminuido su visión? | 8 | NUMERICO |
| a313d | ¿debido a la diabetes ha sufrido daño en la retina? | 8 | NUMERICO |
| a313e | ¿debido a la diabetes ha perdido la vista? | 8 | NUMERICO |
| a313f | ¿debido a la diabetes le han hecho diálisis? | 8 | NUMERICO |
| a313g | ¿debido a la diabetes ha sufrido de un infarto del miocardio? | 8 | NUMERICO |
| a313h | ¿debido a la diabetes ha tenido infarto cerebral? | 8 | NUMERICO |
| a313i | ¿debido a la diabetes: ardor, dolor o pérdida de sensibilidad en planta de pies? | 8 | NUMERICO |
| a313r | ¿debido a la diabetes ha requerido cirugía ocular? | 8 | NUMERICO |
| a313s | ¿debido a la diabetes ha sufrido arritmia? | 8 | NUMERICO |
| a401 | ¿algún médico le ha dicho que tiene la presión alta o hipertensión? | 8 | NUMERICO |
| a605a | ¿le ha dicho el médico que usted tiene (o tuvo) nfección de vías urinarias en má | 8 | NUMERICO |
| a605b | ¿le ha dicho el médico que usted tiene (o tuvo) cálculos renales? | 8 | NUMERICO |
| a605c | ¿le ha dicho el médico que usted tiene (o tuvo) insuficiencia renal? | 8 | NUMERICO |
| a701a | ¿su padre tiene o tuvo diabetes o azúcar alta en la sangre? | 8 | NUMERICO |
| a703a | ¿su padre tiene o tuvo hipertensión o presión alta? | 8 | NUMERICO |
| a705a | ¿su padre tuvo un infarto? | 8 | NUMERICO |
| a701b | ¿su madre tiene o tuvo diabetes o azúcar alta en la sangre? | 8 | NUMERICO |
| a703b | ¿su madre tiene o tuvo hipertensión o presión alta? | 8 | NUMERICO |
| a705b | ¿su madre tuvo un infarto? | 8 | NUMERICO |
| edad | Edad | 8 | NUMERICO |
| sexo | sexo | 8 | NUMERICO |

## 7.3. Anexo C


225000
200000
175000
150000
125000
100000
75000
50000
Suma de errores al cuadrado
1
2
3
4
5
6
7
8
9
Número de clústers (valor de K)

## 7.4. Anexo D




## 7.5. Anexo E

| Clúster | Algoritmo FCM | | | |
|---|---|---|---|---|
| | % Población | % Diabetes | Edad Media | Género % Fem / % Masc |
| 1 | 23.04% | 0.07% | 11 | 100% / 0.00% |
| 2 | 32.25% | 4.68% | 47 | 99.25% / 0.75% |
| 3 | 44.71% | 3.35% | 30 | 0.00% / 100% |