UNIVERSIDAD DE SONORA
DIVISIÓN DE INGENIERÍA

POSGRADO EN INGENIERÍA INDUSTRIAL
MAESTRÍA EN INGENIERÍA EN SISTEMAS Y TECNOLOGÍA

MINERÍA DE DATOS PARA UNA ESTRATEGIA DE MEDICINA PREVENTIVA MÁS ROBUSTA EN UNA INSTITUCIÓN DE SALUD PÚBLICA DEL ESTADO DE SONORA

T E S I S

PRESENTADA POR

EZEQUIEL ALONSO SANEZ MORENO

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

DIRECTOR DE TESIS
DRA. RAQUEL TORRES PERALTA

CODIRECTOR
DR. MARIO BARCELÓ VALENZUELA

HERMOSILLO, SONORA, MÉXICO. OCTUBRE 2018

Universidad de Sonora

Repositorio Institucional UNISON

"El saber de mis hijos hará mi grandeza"

Hermosillo, Sonora a 3 de octubre de 2018

EZEQUIEL ALONSO SANEZ MORENO

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.

Por tal motivo este jurado extiende su autorización para que se proceda a la impresión final del documento de tesis: **MINERÍA DE DATOS PARA UNA ESTRATEGIA DE MEDICINA PREVENTIVA MÁS ROBUSTA EN UNA INSTITUCIÓN DE SALUD PÚBLICA DEL ESTADO DE SONORA** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

Dra. Raquel Torres Peralta
Directora de Tesis y Presidente del Jurado

Dr. Mario Barcelo Valenzuela
Codirector y Vocal del Jurado

Dr. Federico Miguel Cirett Galán
Secretario del Jurado

Dr. René Francisco Navarro Hernández
Vocal del Jurado

c.c.p. Archivo

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Ciències de la Computació

Barcelona, España, a 29 de junio de 2018.

EZEQUIEL ALONSO SANEZ MORENO

Con fundamento en el articulo 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 Ingenieria.

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: **MINERÍA DE DATOS PARA UNA ESTRATEGIA DE MEDICINA PREVENTIVA MÁS ROBUSTA EN UNA INSTITUCIÓN DE SALUD PÚBLICA DEL ESTADO DE SONORA** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE,

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Ciències de la Computació

Dr. MIQUEL SÀNCHEZ-MARRÈ FiEMSs
UNIVERSITAT POLITÉCNICA DE CATALUNYA · BARCELONATECH
Sinodal Externo y Vocal del Jurado

Edifici OMEGA, Campus Nord
Jordi Girona, 1-3
E-08034 Barcelona

Dr. Miquel Sànchez-Marrè FiEMSs
Associate Professor
Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC)
Knowledge Engineering & Machine Learning Group (KEMLG)
Computer Science Dept. (CS)
Universitat Politècnica de Catalunya - BarcelonaTech

Campus Nord - Building OMEGA, Office 203
Jordi Girona 1-3
08034 Barcelona
Phone: + 34 93 4137841 / Fax: + 34 93 4137833
miquel@cs.upc.edu
http://www.cs.upc.edu/~miquel

RESUMEN

Con el presente trabajo se busca fortalecer las estrategias de medicina preventiva que una institución de salud pública del estado de Sonora desarrolla y divulga entre su población ya que no estaban teniendo el impacto deseado al carecer de personalización para el afiliado, evitando una atención oportuna de diversas enfermedades, en específico Obesidad y Diabetes.

Para contrarrestar la problemática se optó por un enfoque tecnológico al detectar que la institución contaba con registros electrónicos de consultas médicas del periodo 2014-2017, maneja una App institucional y tiene presencia en redes sociales. Con esto en mente surgió la idea de desarrollar una metodología de cinco etapas, 1: visualizar la situación inicial, 2: segmentar a la población, 3: utilizar minería de datos, 4: obtener conocimiento y proporcionarlo a la institución, 5: difundir el conocimiento y evaluar.

Tras la implementación de la metodología se comenzaron a apreciar las tendencias que la obesidad y la diabetes han estado teniendo en los afiliados de la institución a través del tiempo, sobre todo, algunas de las características que los diagnosticados con estas enfermedades comparten entre sí, mismas que pudieron ser utilizadas para determinar grupos de enfoque específicos. También se obtuvo información hasta ahora desconocida, como los tipos de familias afiliadas y sus integrantes, la mayor proporción de hombres diagnosticados con diabetes aún y cuando otros estudios señalan que las mujeres son más propensas, entre otros. Con esto se logró alcanzar el objetivo de generar conocimiento que pueda ser utilizado en el soporte a la toma de decisiones, a la vez que se creó material informativo para su difusión en medios electrónicos.

Aun y cuando se encontraron dificultades para la implementación de la metodología al ser una de las primeras aproximaciones de esta institución con la integración de tecnologías y minería de datos, los resultados obtenidos irán incrementando y mejorando al ir perfeccionando y retroalimentando los procesos sugeridos.

ABSTRACT

The present work seeks to strengthen the preventive medicine strategies that a public health institution of the state of Sonora develops and disseminates among its population since they were not having the desired impact due to lack of personalization for the member, avoiding timely attention of diverse diseases, specifically Obesity and Diabetes.

To counteract the problem, a technological approach was chosen to detect that the institution had electronic records of medical consultations for the 2014-2017 period, manages an institutional App and has a presence in social media. With this in mind, came the idea of developing a five-step methodology, 1: visualize the initial situation, 2: segment the population, 3: use data mining, 4: obtain knowledge and provide it to the institution, 5: disseminate the knowledge and evaluate.

After the implementation of the methodology began to appreciate the trends that obesity and diabetes have been having in the affiliates of the institution over time, especially some of the characteristics that those diagnosed with these diseases share with each other, same that could be used to determine specific focus groups. Information was also obtained until now unknown, such as the types of affiliated families and their members, the highest proportion of men diagnosed with diabetes even when other studies indicate that women are more prone, among others. With this, it was possible to achieve the objective of generating knowledge that can be used to support decision making, while creating informative material for its dissemination in social media.

Even when the methodologies are similar to those of integration with the integration of technologies and data mining, the results are enlarged and improved by refining and feeding back the suggested processes.

AGRADECIMIENTOS

Agradezco a mi directora de tesis, Dra. Raquel Torres Peralta, por el apoyo brindado a lo largo de esta investigación, su conocimiento sobre los temas relacionados con mi tesis fue fundamental para guiarme a lo largo del camino. También a mis codirectores, los Doctores Federico Cirett Galán y Mario Barceló Valenzuela, y al Dr. Alonso Pérez Soltero, quienes con sus constantes consejos y apoyo ayudaron a que esta fuese una investigación de provecho y bien llevada.

Agradezco a mi familia, amigos, compañeros y maestros que siempre estuvieron cuando las cosas iban bien, pero sobre todo cuando no se veía el final del trayecto, con esa palabra de aliento para seguir, sin ellos hubiese sido un camino bastante complicado.

Al departamento de Ingeniería Industrial y a la misma Universidad de Sonora por tener el orgullo de pertenecer a sus filas y considerarla mi casa durante estos años en que formé parte de ella, lo aprendido y vivido en ella formará parte de la persona que soy y seré en el futuro.

También agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa de Fortalecimiento de la Calidad Educativa (PFCE) por su apoyo económico, sin el cual no habría sido posible el cursar este posgrado y obtener el grado de Maestro en Ingeniería en Sistemas y Tecnología.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

El actual incremento en el número de personas afectadas por padecimientos no transmisibles prevenibles que está afectando al mundo entero y que tiene a México en los primeros planos en cuanto a enfermedades como obesidad y diabetes, ha obligado a las Instituciones de Salud Pública (ISP), principales responsables del cuidado y fomento de la salud en el país, a enfocar sus esfuerzos en contrarrestar los efectos negativos que se están sintiendo en el presente, así como prever los posibles escenarios en un futuro cercano para generar las condiciones que permitan una acción efectiva ante este gran reto, que no solo afecta la salud de los individuos, sino también perjudica a la economía y sociedad en general. Para hacer frente a dicho reto, las ISP concuerdan en la planeación y aplicación de Campañas de Medicina Preventiva, mismas que deben de estar debidamente fundamentadas y sustentadas en conocimiento obtenido de experiencias pasadas o en la generación de este mediante análisis de registros, tales como, historiales médicos de pacientes, bitácoras epidemiológicas, entre otros. El objetivo es tener un alcance masivo, que permita llegar al mayor número de personas en el menor tiempo posible para hacer las recomendaciones pertinentes que mejor se adapten a sus características físicas, fisiológicas y demográficas, que ayuden a prevenir enfermedades o mitigar sus efectos.

Es importante que se plantee el contexto y entorno en el cual el presente proyecto se realizó, esto con el fin de denotar las características y objetivos clave que permitirán un mejor entendimiento de la problemática, el impacto que se tendrá y la justificación de la realización de este.

1.1. Presentación

El presente proyecto se desarrollará en una Institución de Salud Pública del Estado de Sonora. Presta sus servicios a un aproximado de 182,000 personas. Debido a la creciente demanda de servicios, a la actual reducción en presupuestos y medidas de

austeridad, se está apostando a la predicción y prevención de padecimientos. Para lograrlo esta Institución cuenta con un Área de Medicina Preventiva enfocada en enfermedades como obesidad, diabetes, entre otras.

Algunos de los programas de medicina preventiva se centran en el envío de recomendaciones, alertas sobre brotes de enfermedades y riesgos epidemiológicos que pudieran estar afectando a un sector en particular, esto mediante mensajes SMS y la difusión en distintos medios. Se tiene una aplicación para celular (App), y una "Tarjeta Inteligente", con dos códigos QR (Quick Response), uno para el historial clínico, y el otro código para dar recomendaciones a través de la App.

Pese a que se cuenta con la aplicación antes mencionada, el uso de los mensajes SMS y la tarjeta inteligente, las campañas de medicina preventiva carecen del impacto deseado, pues al tratarse del envío masivo de información sin que exista algún tipo de filtro que discierna entre los tipos de afiliados, se difunde material informativo que no es de su interés o que no representa un beneficio para ellos.

La institución cuenta con grandes bases de datos digitales con expedientes médicos recolectados por más de una década, por desgracia, al no explotar estos recursos, la generación de información de calidad no es la mujor, incurriendo en gastos de atención, hospitalización y medicamentos para tratar enfermedades prevenibles.

Aunque los beneficios del análisis de datos son conocidos, se carece de un método efectivo que permita la extracción de conocimiento que ayude a direccionar de manera efectiva y sustentada las estrategias de medicina preventiva, de modo que el material gráfico y contenido informativo (artículos, posts, tweets o mensajes de texto preventivos) sean dirigidos a los distintos segmentos de la población de acuerdo al nivel de riesgo y características demográficas.

1.2. Planteamiento del problema

La Institución de Salud actualmente realiza programas de medicina preventiva que no se están enfocando y personalizando debidamente según las características físicas y demográficas de sus afiliados, aunado a la carencia de herramientas para el análisis

de registros históricos que ayuden a dicha personalización, generación, difusión de conocimiento y contenido informativo, reduciendo la efectividad de estos programas y provocando un menor impacto al deseado por esta Institución.

Al no contar con una segmentación adecuada de afiliados que ya padecen enfermedades como obesidad y diabetes, o que están en riesgo de padecerlas, las acciones tomadas no presentan los resultados esperados, evitando beneficios tanto para la Institución como para la salud de su población y sociedad en general.

1.3. Objetivo general

Diseñar una metodología para incrementar la generación de conocimiento y contenido informativo sobre obesidad y diabetes mediante el análisis de registros de consultas médicas de pacientes de una institución de salud utilizando herramientas de minería de datos para detectar grupos a los cuales enfocar campañas de medicina preventiva y realizar informes que sirvan de base a estrategias más dirigidas y personalizadas.

1.4. Objetivos específicos

- Analizar las características de los registros para un prefiltrado, enfocándose en Obesidad y Diabetes, así como un análisis inicial de la situación.
- Evaluar técnicas de análisis y procesamiento de datos para el diseño de una metodología que logre la detección de grupos que:
    - Compartan características específicas para determinar un posible riesgo de padecer obesidad o diabetes.
    - Padezcan ya alguna de estas enfermedades y clasificarlos de acuerdo con su edad, sexo, tipo de afiliado, entre otros aspectos.
- Implementar la metodología propuesta para la generación de conocimiento.
- Generar reportes para el área de medicina preventiva con el fin de personalizar los mensajes y contenido informativo para sus afiliados.

- Determinar el alcance y beneficios obtenidos tras la implementación, así como también los productos generados (publicaciones, contenido para redes sociales, y demás).

1.5. Hipótesis

La creación de una metodología de análisis de registros de consultas médicas utilizando herramientas de minería y procesamiento de datos incrementará la generación de conocimiento y permitirá conocer la situación actual de la población derechohabiente de una institución de salud en cuanto a obesidad y diabetes, además de ayudar a la detección de grupos entre los pacientes para una difusión segmentada y personalizada de material informativo como parte de una estrategia de medicina preventiva más robusta y sustentada.

1.6. Alcances y delimitaciones

Se busca como objetivo principal el enfocarse a aspectos de medicina preventiva, relacionada con obesidad y diabetes, para ello se estará trabajando con información proveniente del Departamento de Estadística de la Institución, y sólo en la ciudad de Hermosillo, Sonora. En todo momento se cuidará el manejo de los datos personales, debido a que se estará trabajando con bases de datos de afiliados reales. El campo de acción de este proyecto es el de la prevención, no se pretende realizar aportaciones sobre el cómo tratar un determinado tipo de enfermedad cuya competencia corresponde a los profesionales de la salud.

Este proyecto forma parte de uno mayor que abarca varios aspectos y oportunidades dentro de la misma Institución, haciendo necesario la intervención de un grupo multidisciplinario, por lo que al realizar este análisis de datos se proporcionará información y conocimiento que ayudarán a conocer la situación actual de los afiliados, sirviendo de base para una estrategia de medicina preventiva dirigida, segmentada y personalizada. Por esto mismo se espera que se puedan observar y medir los resultados que se obtengan de este proyecto durante los próximos 5 años. Los

resultados de este proyecto se medirán con estadísticas, así como los cambios en los procedimientos en la atención a los afiliados producto de las recomendaciones surgidas del análisis, mismas que se entregarán a modo de informe y de viva voz en reuniones con el personal de la institución.

1.7. Justificación

Actualmente en esta Institución no se cuenta con una metodología que permita el análisis de registros y la generación de conocimiento con el cual se sustente y respalde la toma de decisiones para la planeación y lanzamiento de campañas de medicina preventiva que se adapte a las características de segmentos específicos de afiliados. Tomando en cuenta el gran número de afiliados a los que esta Institución presta sus servicios (aproximadamente a un 9.5% de la población de Hermosillo), conocer las tendencias de salud de un cierto grupo de individuos es clave para una correcta generación de estrategias de medicina preventiva, mismas que podrán ser aplicadas de forma masiva tanto a afiliados fuera del grupo analizado como a la población en general a través de sus redes sociales, pues estas son abiertas a todo público.

A su vez, la Institución tiene la intensión y el interés para realizar un análisis de sus registros de consultas médicas, por lo que este proyecto estará respaldado y apoyado por personal de esta.

2. MARCO DE REFERENCIA

En esta sección se describen los términos de salud pública y algunos conceptos del análisis de datos, a su vez, se da sustento a la metodología propuesta. En la sección de Anexos se amplían algunos de estos puntos.

2.1. Salud Pública y sus principales conceptos asociados

Es de suma importancia comenzar con una explicación del significado de Salud Pública, así como de las instituciones encargadas de acercarla a quienes la necesitan, pues es en estas donde se concentra el mayor número de personas que requieren recibir algún tipo de atención médica, consulta, orientación o recomendaciones sobre el cuidado de su salud, entre muchos otros servicios en México y el mundo. Tal es el caso del esfuerzo realizado por la actual administración federal, que desde años atrás ha lanzado una estrategia global contra la obesidad basada en 3 pilares: Salud Pública, Atención Médica, Regulación Sanitaria y Política Fiscal, como se observa, dicha estrategia es protagonizada por las medidas de Salud Pública y las Instituciones encargadas de suministrar atención médica (Córdova-Villalobos 2016). Comúnmente los servicios prestados buscan combatir enfermedades y evitar que estas crezcan y se conviertan en epidemias que afecten a la población (anexo 7.1).

Por otra parte, independientemente del gran impacto social que se tiene al trabajar con este tipo de instituciones, también es en una de estas donde se implementa la metodología que esta investigación propone para el aprovechamiento de los registros disponibles mediante herramientas de Minería de Datos (MD) para la generación de conocimiento y contenido informativo útil en la toma de decisiones en cuanto a la creación, puesta en marcha y difusión de campañas de prevención personalizadas para enfermedades específicas.

2.1.1. Salud Pública

Uno de los mayores exponentes en cuanto a Salud Pública y Medicina Preventiva fue Gonzalo Piedrola Gil, en su libro se mencionan varias definiciones de salud pública

provenientes de diversos autores a través del tiempo, todas ellas comparten el tener objetivos como reducir, minimizar y prevenir los efectos de una enfermedad, y también está implícito en dichas definiciones, el cómo conseguir estos objetivos, mediante esfuerzos conjuntos, sociales o de las instituciones (Piedrola et al. 2014), de las definiciones consultadas se adaptó la siguiente:

"La salud pública es tanto un arte como una ciencia que se encarga de prevenir, curar o mitigar enfermedades, promover y mejorar condiciones de sanidad e higiene de la población en general mediante los esfuerzos organizados de individuos, grupos, instituciones y gobierno con el fin de prolongar la vida gozando de un estado de salud integral".

En los trabajos de Cascón (2008), Álvarez-Alva y Kuri-Morales (2012) se puntualiza que los objetivos de la salud pública son prevenir, prolongar la vida, fomentar y conservar la salud y eficiencia de la comunidad; lo destacable en sus enfoques es que no se centra en el individuo, sino que abarca a la comunidad en general, y por otra parte, dejan ver que se trata de una ciencia amplia, dinámica y de suma importancia en la lucha contra los problemas ocasionados por las enfermedades. Estos autores también señalan que otra de las responsabilidades de esta ciencia es el proporcionar las herramientas que permitan a los profesionales de la salud un diagnóstico temprano, medidas y tratamiento preventivos contra enfermedades.

La salud pública confiere sus medios y recursos, con el fin de alcanzar los objetivos antes citados, a la realización de campañas para la promoción de la salud y buenos hábitos entre las personas, en este sentido Longo et al. (2012) hablan de cómo las campañas de medicina preventiva informan oportuna y precisamente sobre riesgos, promueven modificaciones en el estilo de vida de las personas mediante el cambio de aspectos cotidianos, lo que trae como resultado reducir la aparición, transmisión y prevalencia de ciertas enfermedades. Existen instituciones en México que se encargan de brindar estos servicios a la población en general (anexo 7.2).

2.2. Medicina Preventiva, principal arma de la Salud Pública contra las enfermedades y su propagación.

Es inevitable hablar de salud pública sin hablar de medicina preventiva, de hecho estas dos disciplinas suelen ser impartidas conjuntamente en muchas de las instituciones de educación superior encargadas de la formación de profesionales de la salud, como en el caso de España (Piédrola Gil et al. 2014), por otra parte para Zhao et al. (2017) si se mejora la educación en medicina preventiva los procesos de salud pública se verán beneficiados, pues los objetivos de ambas suelen ser los mismos (Álvarez y Kuri 2012), por ello es indispensable definir qué es y cuáles son las funciones de esta ciencia colaboradora de la salud pública.

No existe un único término para definir a la medicina preventiva, la siguiente adaptación fue extraída de Piedrola Gil et al. 2014, Cascón 2008, Borja Aburto 2013 y Piédrola Gil 2001:

"Es la ciencia y el arte que busca evitar o disminuir el riesgo de contraer una enfermedad mediante la educación, prestación de servicios médicos y rehabilitación, con el fin de prolongar la vida y promover la salud al interceptar las enfermedades en cualquier fase de su evolución a través de múltiples programas preventivos".

Existe la opinión de que la medicina preventiva es la única especialidad que equilibra tanto el cuidado clínico individual como la prevención basada en la población, esto significa que se puede partir de la información de la sociedad en general para identificar factores de riesgo y realizar acciones que impacten en lo individual y sean acordes a cada persona, ya que se tiene confianza en que la medicina preventiva personalizada representa una alternativa prometedora a los enfoques de salud pública cotidianos pues sus esfuerzos van dirigidos tanto a personas enfermas como sanas promoviendo y mejorando la salud; pero también al diagnosticar y tratar oportunamente a un enfermo, y al realizar actividades de rehabilitación, evitando así complicaciones o secuelas del padecimiento tratando de evitar la invalidez o la incapacidad de los individuos, readaptándolos a su medio social (Jani et al. 2015; Clarke 2010; Álvarez

Alva y Kuri Morales 2012). Existen diversos niveles en que la medicina preventiva puede enfocar sus acciones. En la Tabla se presentan los tres niveles de la medicina preventiva.

PREVENCIÓN	OBJETIVOS	ACTIVIDADES	NIVELES DE PREVENCIÓN
PRIMARIA	Promover la educación. Evitar la aparición de enfermedades.	Educación para la Salud. Prevención específica. Detección.	Promoción de salud. Prevención específica.
SECUNDARIA	Evitar que las enfermedades progresen. Limitar la Invalidez.	Diagnóstico oportuno y tratamiento adecuado.	Diagnóstico temprano y tratamiento inmediato. Limitación de invalidez.
TERCIARIA	Rehabilitar al enfermo.	Rehabilitación.	Rehabilitación.

Tabla 2.1. *Niveles de Medicina Preventiva (Leavell y Clark, 1958 en Álvarez y Kuri, 2012).*

Algunas de las enfermedades no transmisibles en las que más se enfoca la medicina preventiva son el Cáncer, la Diabetes Mellitus, el Sobrepeso y Obesidad en conjunto con sus padecimientos cardiovasculares asociados, las cuales representan las principales causas de muerte en el mundo (OMS: Organización Mundial de la Salud, 2017b). Algunas de las acciones para evitar la aparición y efectos negativos de estas enfermedades son la realización de campañas preventivas que promuevan los buenos hábitos y medidas paliativas. Arredondo y De Icaza (2011) hablan de como las instituciones de salud pública han comenzado a reforzar sus campañas preventivas para combatir los efectos negativos de las actuales epidemias más importantes en el mundo (diabetes y obesidad), a la vez que se evitan altos costos, en su mayoría asociados a las complicaciones subyacentes a estos padecimientos.

Cuando el concepto y los alcances están definidos, se debe de encontrar la forma de comenzar con estas campañas preventivas, para lo cual Lemus et al. (2008) en su libro presentan dos principales estrategias para llevar a cabo un diagnóstico preventivo mediante el análisis de datos recabados de manera específica para este propósito o de manera indirecta cuando el paciente acude por otros motivos a consulta médica. El

término para la primer estrategia es “Catastro” y consta de invitar a personas del público en general a someterse a algún tipo de prueba con el objetivo de discriminar entre aquellos con mayor y menor probabilidad de contraer alguna enfermedad, para posteriormente inducir a la persona a que busque un diagnóstico definitivo; la segunda estrategia denominada “Hallazgo de Caso” se centra en el hecho de que la mayoría de personas visitan al médico algunas veces al año, por lo que no se necesita de voluntarios, y haciendo uso de datos sobre síntomas no relacionados se puede realizar el diagnóstico precoz de alguna enfermedad.

2.3. Las grandes epidemias del siglo XXI: Impacto en la Salud Pública

Si se desean generar mejores mecanismos para la prevención y eficientizar los ya existentes en la Instituciones de Salud Pública, es necesario determinar cuáles enfermedades son las que están causando mayores estragos en todo el sistema y afectan a la gran mayoría de los individuos, usualmente un número muy reducido de padecimientos son los que generan mayores estragos en la población, pudiendo aplicar el Principio de Pareto donde un 80% de las afectaciones son causadas por el 20% o menos de las fuentes de problemas (enfermedades en este caso) (Kumar, 2015). Una vez que se han detectado la o las enfermedades objetivo, se deben de definir con el fin de encontrar las variables que permitan hacer uso de las técnicas de análisis necesarias para la generación de conocimiento y mejoramiento de estos mecanismos en pro de la salud, lo que trae beneficios tanto para la sociedad en general como para las Instituciones de Salud.

Tomando en cuenta lo anterior, la gravedad del actual aumento de los casos de Diabetes, Sobrepeso y Obesidad ha llevado a las autoridades de salud en todo el mundo a tomar medidas y acciones contra ellas por los problemas físicos, sociales y económicos que conllevan junto con los padecimientos e incapacidades asociados y derivados de las mismas. Autores como Córdova (2016) proponen a la obesidad como la verdadera epidemia del siglo XXI por su rápido crecimiento y afectaciones a nivel mundial. La diabetes es la principal causa de padecimiento cardiovascular, ceguera,

fallo renal y amputación de las extremidades inferiores (Ponce y Kánter 2016), al igual que el sobrepeso y obesidad, afecta a una gran cantidad de personas en el mundo.

2.3.1. Diabetes

Según datos de la Organización Mundial de la Salud (OMS) actualizados a Noviembre de 2016, la prevalencia de la diabetes ha aumentado en los países con ingresos medios y bajos; en 2014 aproximadamente 422 millones de personas la padecían en el mundo y pasó del 4.7% en 1980 al 8.5% del total de adultos (mayores de 18 años) en 2014, para el 2015 el 7% de la población mundial la padecía (Panam et al. 2015). La OMS define a la diabetes como "una enfermedad crónica que aparece cuando el páncreas no produce insulina suficiente o cuando el organismo no utiliza eficazmente la insulina que produce". La insulina es una hormona que regula el azúcar en la sangre. El efecto de la diabetes no controlada es la hiperglucemia (aumento del azúcar en la sangre), que con el tiempo daña gravemente órganos y sistemas, especialmente los nervios y los vasos sanguíneos (OMS: Organización Mundial de la Salud, 2016). Existen dos principales tipos de diabetes y la OMS los define como:

- **Diabetes de tipo 1**

También llamada insulinodependiente, juvenil o de inicio en la infancia, se caracteriza por una producción deficiente de insulina y requiere la administración diaria de esta hormona. Se desconoce aún la causa de la diabetes de tipo 1 y no se puede prevenir con el conocimiento actual (anexo 7.3.1).

- **Diabetes de tipo 2**

También llamada no insulinodependiente o de inicio en la edad adulta, se debe a una utilización ineficaz de la insulina. Este tipo representa la mayoría de los casos mundiales y se debe en gran medida al peso corporal excesivo y a la inactividad física (anexo 7.3.2).

2.3.2. Sobrepeso y Obesidad

En datos actualizados a junio del 2016, la OMS ha estimado que en 2014 en el mundo 1,900 millones de personas mayores de 18 años padecían sobrepeso, y de ellos 600 millones eran obesos, esto representa del total de adultos a nivel global 39% con sobrepeso y 13% con obesidad (un 11% de los hombres y un 15% de las mujeres), así mismo establece las condiciones para determinar si una persona padece alguno de estos padecimientos, como se observa en los siguientes apartados. Por otra parte, en el mismo año se estimó que 41 millones de niños menores de 5 años tenían sobrepeso u obesidad. La OMS (2016b) y Yepez et al. (2007) definen el sobrepeso y obesidad como "una acumulación anormal o excesiva de grasa en el cuerpo en relación con la talla que puede ser perjudicial para la salud"; el IMC es un indicador simple que ayuda a detectar estos padecimientos en adultos. En el anexo 7.4 se muestra a detalle las distintas clasificaciones según edad y sexo, tanto para sobrepeso como para obesidad.

Las principales afectaciones que estos padecimientos han causado en México van desde el ámbito personal, hasta el productivo, económico y social, en el anexo 7.4.2 se describen algunos de ellos.

2.4. Sistemas de Información en las Instituciones de Salud Pública, fuentes de "Big Data"

Como cualquier otro sistema que utiliza tecnologías de la información y la comunicación (computadoras, software, telecomunicaciones), un Sistema de Información de Salud o Health Information Systems (HIS) en inglés, es un conjunto de componentes tecnológicos y de comunicación que trabajan juntos para la captura, almacenaje, procesamiento, recuperación y diseminación de información, sólo que con la gran diferencia que dicha información es referente a la condición de pacientes (citas médicas, análisis de laboratorio, seguimiento, entre otros), así como cambios y tendencias en el panorama de la salud (Kenny et al. 2017; Muhaise y Kareeyo 2017).

Los HIS se han implementado en las instituciones de salud porque permiten un mejor control y mayor calidad en la información que se tiene y con la que se puede trabajar,

esto ayuda a vigilar en tiempo real o históricamente lo que está pasando con un determinado paciente o toda una población, logrando así brindar un mejor soporte a la toma de decisiones y acciones clínicas precisas (anexo 7.6).

2.5. Análisis estadístico de Datos

Como en toda investigación, independientemente de si los datos provienen de encuestas, mediciones de campo o de algún sistema informático, como un HIS, lo primero que se suele hacer es conocer la historia y la situación actual del fenómeno de estudio, claro si es que este lo permite. En el caso de los registros médicos almacenados en los HIS, esto ayudará a conocer los antecedentes y el comportamiento a través del tiempo de individuos o una población completa, además de poder describir clara y precisamente los aspectos relevantes ocurridos, se podrán pronosticar las tendencias y comportamientos futuros con algún grado de incertidumbre, ayudando a la prevención. Para esto la estadística cuenta con un amplio repertorio de herramientas aplicables (anexo 7.7).

2.6. Descubrimiento de conocimiento en bases de datos

El Descubrimiento de Conocimiento en Bases de Datos o Knowledge Discovery in Databases (KDD) en inglés, es todo un campo de estudio que abarca teorías, métodos y técnicas, que tratan de dar un sentido a los datos y extraer conocimiento utilizable de ellos. El KDD considerado un proceso de pasos múltiples (selección, preprocesado, transformación, minería de datos, interpretación/evaluación), donde el de mayor importancia, sin duda alguna, en todo el proceso es la minería de datos, independientemente de la herramientas que de ella se utilice (Kavakiotis et al. 2017).

"KDD es el proceso no trivial que busca darle sentido a los datos involucrando múltiples métodos y técnicas para la identificación de patrones válidos, novedosos, potencialmente útiles y entendibles, en grandes volúmenes de registros provenientes de bases de datos, teniendo como principal herramienta la utilización de algoritmos de minería de datos" (Fayyad, Piatetsky-shapiro et al. 1996; Dash et al. 2016; García-Peñalvo y Conde-González 2017) (anexo 7.8).

A grandes rasgos y tomando en cuenta la definición del KDD, la figura 2.1 muestra resumidamente el proceso que lo compone y el producto final que surge de su aplicación, es decir, el conocimiento proveniente de los patrones y tendencias encontrados mediante la minería de datos, útil en el apoyo a la toma de decisiones.

Figura 2.1. *Proceso básico del KDD, adaptado de García y López (2012).*

Se puede decir que el KDD está conformado por dos aspectos, el primero son los datos con que se pretende obtener conocimiento, sus patrones de comportamiento actual y futuro, así como las tendencias e interpretación de estos; y el segundo aspecto es el cómo se obtiene lo anterior a partir de simples datos, es decir, la aplicación de técnicas de preprocesamiento y minería de datos.

Los autores Fayyad, Piatetsky-Shapiro et al. (1996) y retomado por Ho (2017) incluyen en sus trabajos una lista detallada conteniendo 9 pasos que el proceso de KDD contempla en su metodología (anexo 7.8.1). Estos pasos van desde entender la importancia y relevancia que el conocimiento generado va a tener en la organización en donde vaya a ser utilizado, hasta finalmente consolidar el conocimiento obtenido haciéndolo llegar a quienes toman las decisiones e incorporándolo a la memoria organizacional para futuras consultas.

En la figura 2.2 se representa la incorporación de los 9 pasos que integran el proceso completo de KDD antes descrito, de manera detallada se observa cada uno de los subprocesos principales y las tareas intermedias a realiza entre cada uno de ellos, yendo desde los datos en bruto hasta la toma de decisiones.

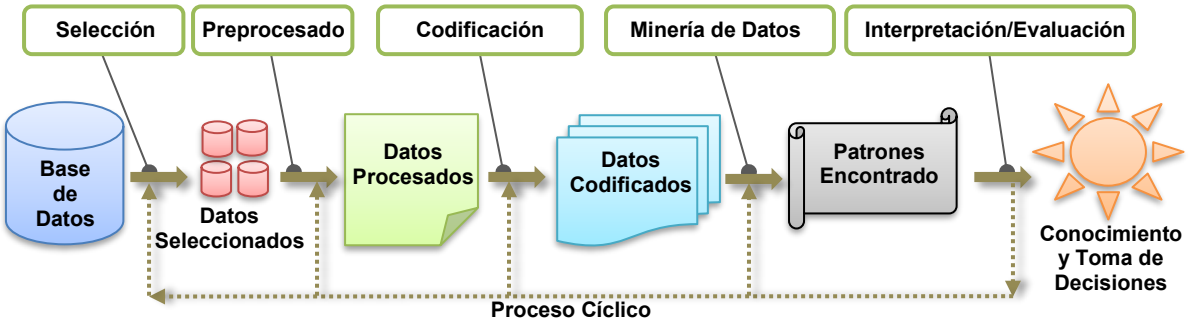


Figura 2.2. *Pasos que integran el proceso de KDD, adaptado de Dash et al. (2016) y Ltifi et al. (2016).*

Un aspecto a considerar durante el KDD, es que existe un proceso cíclico entre cada uno de los subprocesos o pasos, es decir, se puede volver a algún estado anterior de ser necesario una vez que los datos han sido seleccionados, esto con el fin de lograr el grado de calidad de datos o de confianza en el conocimiento descubierto que se espera. También hay que recordar que el paso central del proceso de KDD es la aplicación de técnicas de minería de datos, de ahí surge la importancia de conocer con mayor profundidad cuales son las características que la definen e integran, y poder así aplicarla de la mejor forma posible y obtener la mayor cantidad de conocimiento.

2.7. Minería de Datos

La minería de datos es uno de los métodos que integran un gran número de herramientas para los distintos pasos del análisis de datos, ya que integra aspectos de áreas como tecnologías de bases de datos, inteligencia artificial, aprendizaje máquina, redes neuronales, estadística, reconocimiento de patrones, sistemas de adquisición de conocimiento, recuperación de información, procesamiento computacional complejo y visualización de datos (Han y Kamber 2000).

"Minería de datos es un proceso que se basa en la aplicación de algoritmos específicos mediante software especializado para encontrar y describir patrones estructurados útiles y significativos en grandes volúmenes de datos con el fin de ayudar a explicar dichos datos, realizar predicciones y servir de apoyo en el proceso de toma de decisiones" (García-Peñalvo y Conde-González 2017; Aljumah y Siddiqui 2016; Witten et al. 2011; Fayyad, Piatetsky-shapiro et al. 1996) (anexo 7.9).

Fayyad, Piatetsky-Shapiro et al. (1996) establecen una clara diferencia entre KDD y minería de datos; KDD se refiere al proceso en general de descubrir conocimiento útil a partir de un conjunto de datos, mientras que minería de datos es un paso en particular en este proceso. Los pasos adicionales del proceso de KDD (selección, preparación y limpieza de datos) ayudan a que la utilización de las técnicas de minería de datos no se convierta en una actividad que conduzca fácilmente al descubrimiento de patrones sin sentido e inválidos.

Basándose en la anterior definición y el trabajo de Gheorghe y Petre (2014), se adaptó el siguiente diagrama que describe el proceso de minería de datos mediante el que se obtienen los patrones y estructuras ocultas en los grandes volúmenes de datos en bruto (figura 2.3).

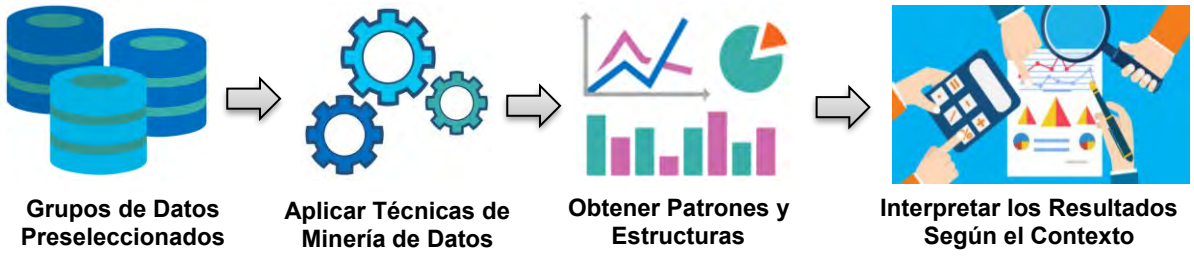


Figura 2.3. *Diagrama general del proceso de Minería de Datos.*

El proceso en sí parece sencillo, solo que la complejidad del mismo parte desde la obtención de los datos que serán procesados, la selección del tipo de algoritmo(s) que se empleará(n) dependiendo del tipo de datos y lo que se esté buscando obtener, y

hasta la detección de los patrones en las distintas representaciones gráficas y numéricas que el software arroje.

Independientemente de la complejidad del proceso de minería de datos, esta se puede emplear básicamente para realizar dos tipos de tareas (Han y Kamber 2000) que pueden ser del tipo:

- **Descriptivas**: Caracteriza las distintas propiedades de los datos dentro de los grupos seleccionados y las resume para su análisis de manera comprensible para el humano.
- **Predictivas**: Realiza inferencias basadas en las variables y datos actuales para realizar predicciones de valores futuros desconocidos de variables de interés.

En base al tipo de investigación se determinará el tipo de tarea que se realizará, y con esto el tipo de modelo y de algoritmo(s) de procesamiento que se seleccionará(n). Dentro de estas tareas existen subcategorías, por así llamarlas, de las distintas técnicas de minería de datos que pueden ser empleadas (Kaur et al. 2017; González 2013; Fayyad, Piatetsky-shapiro et al. 1996), algunas de ellas pueden ser de (anexo 7.9.1):

- ✓ **Clasificación**: Los elementos o datos son ubicados sistemáticamente dentro de un grupo o clase predefinida dependiendo del valor de sus atributos. Su objetivo es predecir la clase a la que un nuevo elemento pertenece a partir de aquellos que ya se encuentran clasificados. Las clases pueden ser tan simples o complejas como se desee, los atributos que deben de tener los elementos para pertenecer a una clase son ilimitados.
- ✓ **Agrupación (Clustering)**: Se emplea para la obtención de grupos naturales de los datos a partir de criterios, por lo general distancia o similitud entre ellos; no es necesario que exista una definición previa de los criterios para poder realizar el agrupamiento (aprendizaje máquina). Forma parte de las tareas descriptivas, y su objetivo es el de encontrar un número finito de categorías para, precisamente, describir a los datos dentro de cada grupo creado.

Si se busca ver desde una perspectiva concreta y resumida, independientemente del tipo de tarea y técnica utilizada, la minería de datos es el proceso mediante el cual se descubren estructuras interesantes en los datos (Roiger, 2017), entiéndase estructura como un conjunto de reglas, una red, ecuaciones, correlaciones, y más. Con lo anterior en consideración, es así como este proceso será interpretado para los fines del presente proyecto, pues se está buscando el descubrir las relaciones y estructuras que yacen ocultas entre los registros analizados para el apoyo a la toma de decisiones y prevención de riesgos en temas de salud.

2.7.1. Minería de Datos en la salud

La salud es una de las áreas de estudio que mayor interés causan entre la comunidad científica y de investigación, debido a la importancia que tiene para el cuidado y bienestar de la sociedad, por lo que no es de extrañar que constantemente se estén adoptando nuevas tecnologías y herramientas diseñadas para otros campos que ayuden al cumplimiento de sus objetivos de prevenir, curar o mitigar enfermedades. En este sentido, la minería de datos ha sido adoptada por las instituciones de salud como una valiosa herramienta que permite, como Oswal y Shah (2017) señalan, obtener una serie de beneficios como los que ahora se presentan:

- Beneficia a todas las distintas partes involucradas en el préstamo de servicios de salud, instituciones de salud, aseguradoras, laboratorios, y por supuesto a los pacientes que reciben estos servicios.
- Los pacientes podrán recibir mejores y más accesibles servicios de salud, mediante la identificación y seguimiento de aquellos con alto riesgo o enfermedades crónicas, interviniendo de manera oportuna con lo que necesiten. Esto también reduce el número de hospitalizaciones.
- Las instituciones de salud pueden utilizar minería y análisis de datos para encontrar las mejores prácticas y los tratamientos más efectivos en una determinada población. Estas herramientas comparan síntomas, causas, tratamientos, efectos negativos y luego analizan qué acción será la más efectiva para un grupo.

- Los registros de consultas médicas contienen toda la información relevante de los pacientes, así como la de las partes involucradas en el préstamo de servicios de la salud. El almacenamiento de este tipo de datos se está incrementando rápidamente, por lo que la minería permitirá conocer a fondo su comportamiento y aprovecharlo en favor de la salud.

A estos beneficios se pueden agregar las capacidades predictivas y preventiva de las técnicas de minería de datos, pues se pueden extraer reglas para la generación de modelos con capacidades predictivas, proveyendo de nuevas maneras de explorar y entender los datos (Chaurasia y Pal 2014), evitando que los profesionales de la salud tengan que basarse solamente en su criterio para el diagnóstico, e inclusive detectar a aquellos pacientes con posibilidad de desarrollar alguna enfermedad aun y cuando los síntomas no son evidentes. Para lograrlo cada una de estas técnicas agrupan a determinados algoritmos que cuentan con características particulares y funcionan mejor en distintas circunstancias y con ciertos tipos de datos, por lo que hay que conocer aquellos de mayor relevancia y uso en aspectos relacionados con salud para seleccionar el adecuado para los fines mencionados.

2.7.2. Algoritmos de minería de datos

La minería de datos en la salud busca resolver problemas que afectan a la población en todo el mundo al mejorar el diagnóstico y tratamiento de enfermedades, Chaurasia y Pal (2014) también establecen que en las investigaciones recientes, la minería de datos se utiliza para el diagnóstico de enfermedades como tuberculosis, diabetes, cáncer, enfermedades cardíacas, entre otras. Entre las técnicas que más se utilizan para el diagnóstico sobresalen aquellas que involucran algoritmos como:

1. K Vecinos Próximos (K-NN).
2. Redes Neuronales.
3. Clasificadores Bayesianos.
4. Agrupación (Clustering).
5. Árboles de decisión.

6. Algoritmos Genéticos.

Para definir cuáles técnicas son las apropiadas para la investigación relacionada con la salud, es necesario definir las características de algunos de los algoritmos de minería de datos en particular, con especial énfasis en aquellos que son más utilizados en temas de la salud. Pero primero, aunque se ha estado mencionando constantemente la palabra algoritmo, es bueno definir su significado para evitar confusiones y dudas, este es:

"Un conjunto de operaciones o reglas definidas y estructuradas que representan un modelo cuyo fin es la solución para un determinado problema".

Por lo que un algoritmo de minería de datos es aquella secuencia de instrucciones o reglas que se ejecutarán por el software especializado para descubrir los patrones y estructuras dentro de los grupos de datos seleccionados. Existen cientos de algoritmos de este tipo, por ello, la selección de aquel o aquellos deberá de ser cuidadosa. Se debe de recordar que no existe un "mejor" o "peor" algoritmo, todo depende del tipo de problema que se investigue y los datos disponibles, con esto en cuenta es que se debe de seleccionar aquel que entregue los resultados más exactos y confiables (Cáceres, 2016). Para este caso en particular, se pueden utilizar algoritmos que han demostrado ser eficaces en el análisis de datos en la salud (anexo 7.9.2).

2.7.3. Tipos de algoritmos y los más utilizados para el diagnóstico y predicción de enfermedades

Sin importar el área de aplicación de los algoritmos de minería de datos, existen dos principales tipos en que estos se clasifican, los Supervisados o Predictivos y los No Supervisados o de Descubrimiento de Conocimiento, Moreno-García et al. (2001) los definen de la siguiente manera:

- **Algoritmos Supervisados o Predictivos**: Son aquellos que buscan predecir el valor de un atributo de un conjunto de datos cuando ya se conoce el valor de otros atributos. A partir de datos cuyas etiquetas (clasificadoras o agrupadoras. Ejemplo: alto riesgo, enfermo, sano) son conocidas se induce una relación entre estos y

aquellos cuya etiqueta se desconoce, de esta forma se busca realizar la predicción de los valores de las etiquetas que se desconocen. A esto se le conoce como aprendizaje supervisado y consta de dos etapas: Entrenamiento (se crea un modelo con un subconjunto de datos con etiquetas conocidas) y Prueba (se prueba el modelo sobre el resto de los datos que no pertenecen al subconjunto de entrenamiento).

- **Algoritmos No Supervisados o de Descubrimiento de Conocimiento**: Este tipo de algoritmo es utilizado cuando no se cuenta con un conjunto de datos lo suficientemente "maduros" para utilizar una solución predictiva, pues se carece del suficiente conocimiento de los atributos como para determinar qué etiquetas pueden identificar correctamente a los integrantes del grupo de datos. Los algoritmos no supervisados descubren patrones, tendencias y estructuras en los datos de manera autónoma. El descubrimiento de esta información sirve para llevar a cabo acciones de soporte a la toma de decisiones y la obtención de beneficios.

Uno de los más utilizados en salud es:

- **K-Medias**

Es considerado uno de los algoritmos más simples, se utiliza para resolver el problema de agrupación, divide a los elementos del grupo de datos en un número preespecificado de clúster (grupos) sin que existe alguna jerarquización en ello. Se aplica cuando se busca agrupar por similitud y ayuda a tener una comprensión tanto cualitativa como cuantitativa de grandes cantidades de datos. Realiza una serie de iteraciones dividiendo el grupo de datos (n) en un número (K) de clusters, está basado en la minimización de la distancia, agrupando aquellos datos cuyas medias sean las más cercanas las unas a las otras (Cáceres, 2016).

Este algoritmo busca asegurar la menor suma de distancias cuadradas de todos los puntos (datos) y el clúster central, el proceso es el siguiente (Ray y Turi 1999):

- ✓ Definir el K número de clusters iniciales.
- ✓ Realiza una serie de n-iteraciones para distribuir la totalidad de elementos del grupo de datos (n) en los K clusters.

- ✓ Cuando un nuevo dato es agregado a un clúster, se vuelve a calcular el valor medio (centro) para asegurar que la suma cuadrada de distancias al centro sea la mínima.
- ✓ Si todos los elementos han sido asignados a un clúster, se termina el proceso, de lo contrario se regresa al paso 2.

2.7.4. Preprocesamiento y optimización de datos de entrada

Independientemente del tipo de técnica o algoritmo que se vaya a utilizar para el análisis de los registros, es necesario, al igual que cuando se emplea estadística, llevar a cabo un preprocesamiento de los datos de entrada, Moreno-García et al. (2001) en su investigación establecen que el preprocesamiento de datos consume aproximadamente el 60% del esfuerzo total, mientras que la parte de minería en sí, sólo representa un 10%, de ahí la importancia de esta etapa.

La idea es evitar que se introduzcan “impurezas” que puedan afectar el correcto funcionamiento de la técnica de minería de datos, causar una ejecución lenta de los algoritmos (tomando en cuenta que a mayor cantidad de registros, mayor será el tiempo y capacidad de procesamiento requeridos), o simplemente hacer que los resultados obtenidos pierdan objetividad y precisión (Mansingh et al. 2017).

Se debe de considerar que el tipo y contenido de los registros dependerá del enfoque de investigación y del objetivo de esta, así que los criterios para llevar a cabo el filtrado dependerán de estos factores, de ahí el descartar o mantener registros. Swapna et al. (2016) proponen una serie de pasos para la limpieza de los registros con el fin de mejorar la calidad de estos, comienzan describiendo los distintos tipos de datos “sucios” que suelen encontrarse en los registros, en la tabla 2.2 se enumeran algunas de las impurezas que los registros pueden contener y que es deseable eliminar antes de aplicar minería de datos por las razones antes expuestas.

Tipo de Dato "Sucio"	Descripción
Sin Valor	Se presentan cuando uno de los campos del registro no fue llenado y ha quedado un espacio en blanco.
Ruido	Son datos que no se relacionan de ninguna manera con el resto, pertenecen a otro contexto.
Valores Tontos	Son aquellos valores que son ingresados a los registros sólo con el fin de cumplir con campos de carácter obligatorio (abc, xyz, aaaa, qwqsas, etc.).
Datos Crípticos	Datos que se encuentran codificados.
Datos Contradictorios	Un ejemplo común es que la fecha de nacimiento no coincida con la edad de la persona.
Datos sin Identificador Único	Cuando el identificador único aparece más de una vez en distintos registros.
Problemas de Integración de Datos	Cuando existen datos cuyas características o codificación no coinciden con el resto de los datos en los registros, por lo que no pueden ser procesados por la herramienta o software seleccionado

Tabla 2.2. *Descripción de los datos que se busca filtrar previo al procesamiento.*

Además de estos datos "sucios" existen muchos más, en este caso en particular, los ejemplos de datos de la tabla 2.2 suelen encontrarse más que nada en registros electrónicos, el ejemplo concreto de Valores Tontos es comúnmente visto cuando se están capturando documentos con campos sin llenar en un formulario en algún sistema (HIS, PHIS), pero dichos campos son de llenado obligatorio, así que sólo se ingresan caracteres al azar.

Centrando un poco el tema de filtrado y optimización de datos, existe toda un área del KDD que se encarga justamente de realizar la preparación, filtrado y limpieza de datos, suele llamársele Preprocesamiento de Datos. Este procesado previo se vuelve de mayor importancia cuando los datos provienen de la vida real, es decir, de actividades cotidianas donde intervienen múltiples actores y manipulan (agregan, eliminan, modifican) los registros, donde lo más probable es que se presenten inconsistencias y se cometan errores, tales como registros duplicados o contradictorios, valores fuera de rango, campos sin llenar, entre otras. También se deberá de tomar en cuenta el tipo de codificación o de archivo de origen de los datos (.csv, .xml, .sql) para determinar la forma de tratamiento de los datos (Rahm y Do 2000; Németh y Michalconok 2016), por lo general se trabaja con lenguajes de consultas a bases de datos como SQL

(Structured Query Language) o MySQL en su versión gratuita. Por último se deberán de normalizar los datos, lo que significa ajustar los registros para que puedan ser procesados por el software o la herramienta de minería de datos que haya sido seleccionada, ya que por lo general estos sólo trabajan con datos de tipo numérico, y si se están utilizando datos que representan características cualitativas (tipo de falla, enfermedad, entre otras), se deberá de crear un índice que sustituya estas características por valores numéricos (Mansingh et al. 2017).

Como ya se mencionó, el objetivo del filtrado será el asegurar la calidad de los datos de entrada para su posterior procesado mediante minería de datos. Con este filtrado se busca reducir el tiempo de procesado una vez que los datos se encuentren cargados en el software seleccionado, agilizando las interacciones o cálculos de los algoritmos. En la figura 2.4 se describen los pasos generales que se pueden seguir al momento de realizar una limpieza o filtrado de datos, el aspecto relevante de este proceso de filtrado es que se enfoca en la preparación de registros provenientes de consultas médicas, por lo que se ajusta al enfoque de esta investigación.

Figura 2.4. *Preparación de Datos, previo a la aplicación de Minería de Datos en la Salud (adaptado de Varlamis et al. (2017)).*

1. Recopilar Datos: El origen y formato en que se encuentren los datos puede variar, provenir de distintas bases de datos o sistemas gestores, codificados en archivos de diferente extensión (.xls, .csv). Si estos datos pueden ser integrados en un mismo archivo desde el comienzo se facilitará el filtrado.

2. Identificación de Pacientes: Las bases de datos de los HIS contienen múltiples registros de un mismo paciente por diversas causas, al menos uno por cada visita a la institución de salud, por lo que se requiere seleccionar sólo aquellos registros que sean relevantes para la investigación y asociarlos a un identificador único para cada afiliado.
3. Eliminar Redundancia: Se deben eliminar todos aquellos registros que se encuentre más de una vez en la muestra seleccionada. Se deberá de tener especial cuidado de no eliminar aquellos registros que provengan del seguimiento del mismo afiliado y que no representan duplicidad.
4. Estandarización de Datos de Entrada: Suele ocurrir que los datos sean capturados sin seguir un formato predeterminado o alguna regla de estandarización de entrada, por ejemplo, cuando se captura ubicaciones o nombres de lugares se puede cometer el error de nombrar de distintas maneras al mismo sitio (Hermosillo, Ciudad de Hermosillo, Hermosillo Sonora, entre otras) lo que ocasiona un aumento en los agrupadores de datos cuando en realidad deberían de ser menos agrupadores con mayor número de integrantes.
5. Codificación de Datos: En esta parte del proceso se deberán de exportar los datos a los formatos correspondientes según la herramienta de minería que se va a utilizar.

Si bien esta serie de pasos no representa todo lo que se puede hacer para asegurar la calidad de los registros de entrada, retoma los aspectos clave que todo proceso de preprocesamiento de datos debe de considerar mínimamente antes de aplicar alguna técnica de minería de datos, ayudando al desempeño del software que ejecutará el algoritmo de procesado, reduciendo el tiempo de cada iteración e incrementando la probabilidad de que los resultados obtenidos sean precisos, de calidad y útiles.

2.7.5. Software para Minería de Datos

Según la asociación mundial de ingenieros “Institute of Electrical and Electronics Engineers” (IEEE), el software es el conjunto de los programas de cómputo, procedimientos, reglas, documentación y datos asociados que forman parte de las

operaciones de un sistema de computación. Es importante elegir el software correcto que sea capaz de soportar los algoritmos, que cumpla con los requerimientos y capacidades con las que se cuentan para la investigación; algunos ejemplos son: WEKA, MATLAB Y ORANGE (Anexo 7.10).

2.8. Utilización de Sistemas de Información Geográfica para la visualización de enfermedades

Otra de las herramientas tecnológicas, además de la minería de datos, que ha tenido un gran crecimiento en su utilización en el área de la salud en todo el mundo durante los últimos años, son los Sistemas de Información Geográfica o Geographical Information System (GIS) en inglés, en gran parte debido a la facilidad de contar con acceso a software que permita el monitoreo de casi cualquier zona en el mundo, y por supuesto, a que aportan información sobre la visualización, tendencias y comportamiento que un fenómeno puede tener en un futuro cercano. Si bien los GIS se basan casi por completo en la utilización de computadoras para el procesamiento y visualización de los registros que se desean estudiar, la idea base del mapeo de casos de enfermedades para su estudio y análisis surgió junto con la epidemiología moderna de la mano del médico inglés John Snow, quien en 1854 generó la hipótesis de que la epidemia de cólera que azotaba a la zona de Soho (Londres), era causado por el suministro de agua contaminada y no debido al aire con polución, y para comprobar su hipótesis, Snow utilizó mapas para demostrar la fuerte correlación entre las muertes por cólera registradas y las fuentes de agua contaminada en Soho (Loslier, 2016).

Según los trabajos consultados (Arab 2011; Boonchieng et al. 2014; Torio 2015; Shaw y McGuire 2017), los GIS en la salud pueden ser definidos como:

"Es un sistema computacional para recolectar, editar, integrar y analizar datos espacialmente referenciados o geográficos, incluyendo atributos del entorno e infraestructura, con el fin de representar y visualizar las conexiones dinámicas entre individuos, su salud, cambios físicos y sociales en su entorno, así como la interacción

compleja en una amplia variedad de contextos, mejorando la eficacia de los servicios de salud y medicina preventiva".

En la actualidad, aunado a las medidas convencionales para estimar el grado de propagación y tendencias de una enfermedad (censos, estudios de campo, entrevistas, entre otras), los GIS están ayudan a visualizar el panorama completo del comportamiento y distribución de una epidemia mediante la administración, representación, visualización, análisis, y comunicación de registros médicos en forma de tablas, gráficas o mapas (Arab, 2011), inclusive en tiempo real, lo que aumenta la efectividad de las acciones que se implementan, ya que se está enfocando el tipo de recurso específico en los sectores que los necesitan.

Como ya se definió previamente, las enfermedades que han sido consideradas para la realización de este estudio, presentan un comportamiento epidémico, lo que significa que deben de ser estudiada tanto su distribución (geográficas), así como los determinantes de la salud (factores favorables) y la enfermedad (factores adversos) en las poblaciones humanas, con el fin de contribuir a mejorar su salud (Arias-Valencia, 2017), por ello es que herramientas tecnológicas como los GIS están causando gran impacto en el alcance y disminución del tiempo de respuesta en las acciones que se realizan en contra de las enfermedades al automatizar pronósticos y prever escenarios futuros (propagación, número de afectaciones, regiones de interés, entre otras). La figura 2.5 representa un ejemplo de la visualización del número de registros de casos de diabetes en un sector de una ciudad, cada marcador representa una zona que concentra a un cierto número de personas que la padecen.

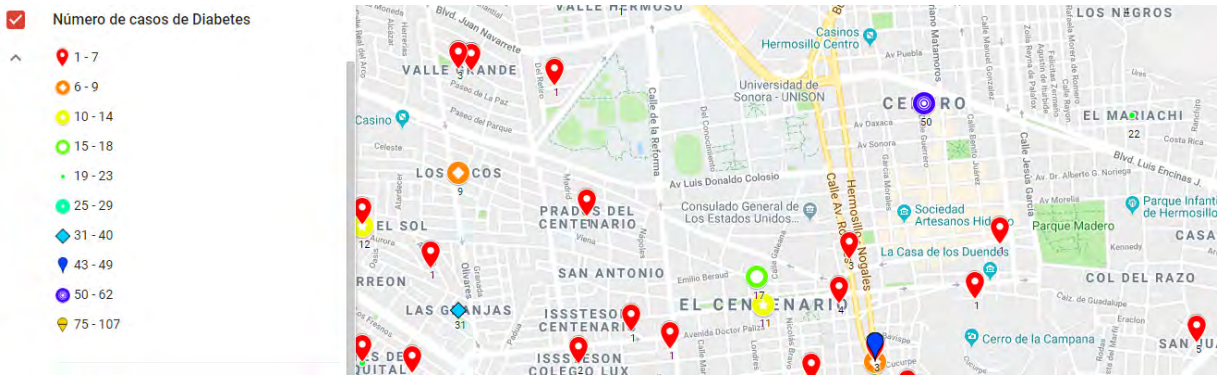


Figura 2.5. *Visualización del número de casos de Diabetes diagnosticados por zonas.*

La herramienta utilizada para crear la anterior figura fue “MyMaps” de Google, la cual permite de manera gratuita, cargar y visualizar registros geográficos, así como agruparlos según se necesite, medir distancias, trabajar colaborativamente, entre otros, por lo que es una buena opción para iniciar; por otra parte, tiene limitación en cuanto lo que se puede hacer con los datos, pues carece de aplicaciones que otros softwares de GIS poseen. Y aunque esta herramienta no sea sumamente compleja, permite observar que existen diversos puntos en la ciudad donde el número de casos se concentran, con lo que se puede trabajar para definir estrategias de acción; como Murad (2007) menciona, uno de los beneficios de los GIS cuando se trabaja con datos médicos es que incrementa la habilidad de mostrar más que lo que un atributo de los datos permite en una sola vista, en otras palabras, permite observar patrones en los datos de manera gráfica.

Tomando en cuenta la cantidad de casos de cierta enfermedad que se hayan o estén presentando en una determinada zona, el impacto de las acciones tomadas puede ser el esperado o superado, y habrán mayores posibilidades de que los beneficios lleguen a aquellos que realmente los necesitan, tanto atención como contenido informativo acorde a sus padecimientos (Shaw y McGuire 2017) (anexo 7.11).

2.9. Difusión de contenido preventivo y conocimiento en la salud vía medios electrónicos y redes sociales

Después de utilizar distintas herramientas como Minería de Datos y GIS, y la respectiva interpretación de los resultados, se necesita difundir el contenido generado para que pueda ser utilizado en favor de la mejora de la salud de los afiliados. Pero el objetivo es llegar a aquellos para quienes fue diseñado el contenido informativo en particular, de lo contrario los esfuerzos invertidos difícilmente tendrán los resultados esperados. Para lograr llegar al mayor público afiliado objetivo de una manera rápida, cómoda y atractiva, pero sin perder de vista la personalización, se tiene la posibilidad de utilizar herramientas tecnológicas para este fin, en particular la utilización de redes sociales y aplicaciones móviles propias de las instituciones de salud que trabajan en coordinación con sus HIS, que por sus características cumplen con los requerimientos para mejorar el alcance y difusión del material preventivo y recomendaciones.

2.9.1. Utilización de redes sociales y aplicaciones móviles en campañas de salud

El término en inglés “Social Media” será adaptado a Red Social (RS), el cual es comúnmente utilizado para referirse a aquellas plataformas electrónicas (Facebook, Twitter, WhatsApp, Instagram, entre otras) a través de las cuales las personas interactúan diariamente desde cualquier parte a cualquier hora del día, compartiendo y discutiendo puntos de vista sobre temas diversos, entre ellos la salud (Paul et al. 2016).

En 2013 el 90% de adultos jóvenes menores de 30 años en USA eran usuarios de alguna RS, y uno de cada cuatro adolescentes contaba con un teléfono inteligente para acceder a estas (Wong et al. 2014). La salud y el cuidado de esta no podía faltar como uno de los principales tópicos en estas plataformas. El incremento de la importancia de las RS utilizadas por personas de cualquier nivel social en todo el mundo está provocando que la forma en que las instituciones de salud pública realizan sus intervenciones e interacciones con estas deban de ser rediseñadas y reinterpretadas para comenzar a visualizarlas desde una óptica que incluya a estos

medios de comunicación (Valente et al. 2015); según Mano (2014) este reenfoque se debe a que la utilización de RS por parte de las personas, permite a las instituciones de salud mejorar su perspectiva sobre los problemas médicos que los aquejan y ayuda a eliminar las fuentes de preocupación que las personas pudieran estar percibiendo, mediante el monitoreo de este esfuerzo social y de compartición de información que surge de la simple interacción de las personas (anexo 7.12).

Si el objetivo es llegar a un público que va de los niños a adultos jóvenes en su mayoría, es indiscutible que el uso de RS y aplicaciones móviles son un muy buen primer paso para lograr tener la atención que lo que se desea transmitir necesita, por ejemplo, la edad media del billón de usuarios de Facebook son 22 años y el 30% de usuarios de internet de 18-24 años utiliza Twitter, lo que las convierte en escaparates perfectos para este público, sin olvidar que cada vez se suman más usuarios de todos los grupos de edad. Las RS son una de las actividades en línea más populares y la mayoría de usuarios las utiliza diariamente permitiendo que penetren en los grupos sociales sin importar el nivel educativo, económico o acceso a servicios de salud (Hswen et al. 2013).

George et al. (2013) presentan algunas de las características que la utilización medios electrónicos en la salud involucran, estas pueden ser tanto positivas como negativas y para determinarlo deben de existir métricas para su evaluación (anexo 7.12.1). algunas otras de las consideraciones para el uso de redes sociales se presentan en el anexo 7.12.2.

2.10. Estudios previos

Es importante conocer las investigaciones y trabajos que se han realizado en el campo de estudio de este proyecto o que tienen una similitud con el mismo, los siguientes son algunos ejemplos de estudios previos que involucran la utilización de técnicas de minería de datos orientadas al área de la prevención y salud.

- Kaur et al. (2017) comienzan describiendo los tipos y características que la diabetes presenta, para después señalar que actualmente existen grandes volúmenes de

datos en hospitales e instituciones de salud que pueden ser aprovechados para la temprana detección de esta enfermedad, lo que reduciría el número de muertes por esta causa, para esto proponen la utilización de minería de datos, en específico técnicas de agrupamiento (clustering), regresión y árboles de decisión.

Utilizan un ejemplo sencillo para describir la metodología que proponen, lo hacen utilizando un árbol de decisión para determinar, en base a una serie de variables, si es recomendable salir a jugar o no, pues lo que están buscando una vez que la metodología se emplee en el análisis de datos referente a individuos con posibilidad de padecer diabetes se obtengan sólo dos resultados, si este padecerá o no diabetes y detectarlos de forma temprana.

- Al-hagery et al. (2015) tuvieron por objetivo el construir tres modelos independientes, uno para cada tipo de hepatitis (A, B, C), donde cada modelo cubriera todas las posibles condiciones de cada tipo virus que la causa para ser capaces de predecir la condición precisa del paciente, a la vez que se provee de conocimiento valioso a los médicos para predecir y diagnosticar de manera temprana la enfermedad.

  Utilizando minería de datos, en específico algoritmos de agrupamiento (K-Means) y redes neuronales, en su metodología basada en los pasos del proceso de KDD ajustado a sus necesidades, lograron analizar 31,574 registros reales con 15 atributos cada uno, obteniendo tres distintos modelos para el diagnóstico y prevención de los diferentes tipos de hepatitis. Lo porcentajes de error al momento de realizar las clasificaciones (susceptibles, antes infectados, no infectados, infectados, crónicos) de los registros arrojaron errores menores al 0.3%, debido a la correcta asignación del número de grupos (clusters) y elección de las técnicas de minería de datos adecuadas.

  Los autores agregan que con el modelo generado se pueden crear herramientas web para que sea utilizado directamente por los médicos o pacientes en un futuro para el diagnóstico y predicción de hepatitis en cualquier momento y lugar.
- Koh y Tan (2005) abordaron en su trabajo la utilización de técnicas de minería de datos desde una perspectiva integral para su uso en el área del cuidado de la salud,

dejando ver las cualidades que esta tiene en la toma de decisiones necesarias para alcanzar metas en las instituciones de salud, no solo en el presente sino en un futuro, convirtiendo a la minería de datos en una herramienta que gana popularidad, y que también se está volviendo esencial. Concuerdan con que la minería de datos permite encontrar patrones previamente desconocidos en vastos volúmenes de datos, haciendo énfasis en que las transacciones (registros de citas médicas, diagnósticos, expedientes, análisis clínicos) y sus relaciones dentro de una institución de salud son demasiado complejas y voluminosas como para poder ser analizadas mediante métodos tradicionales. Posteriormente proponen una secuencia para la aplicación de técnicas de minería de datos, comenzando con la obtención de una fuente de datos, los cuales deberán de pasar por una selección previa para obtener una muestra que será transformada (codificada) antes de su procesamiento con alguna de las técnicas de análisis, vía software especializado, posteriormente lo resultante de este procesamiento deberá de ser evaluado, para finalmente aplicar el conocimiento obtenido en beneficio de la institución y sus pacientes.

Como resultado de esta secuencia propuesta, los autores muestran un caso de estudio detallado donde la aplicación de técnicas de minería de datos ayuda al diagnóstico de diabetes. Partiendo de cómo ciertas variables pueden incidir en el comienzo de esta enfermedad, se pretende identificar a aquellos individuos con un elevado nivel de riesgo de padecer diabetes para que reciban una notificación oportuna y puedan tomar acciones oportunas para contrarrestarla. Las variables seleccionadas fueron la edad y el IMC de los individuos, y se decidió que la técnica de árbol de decisiones era la más apropiada en este caso. Concluyen con que la aplicación más común e importante de la minería de datos involucra modelos predictivos, es decir, se utiliza para la proyección de escenarios futuros y prevención, por lo que las instituciones de salud pueden lanzar campañas preventivas donde se enseñe a las personas a cuidar su IMC y asesorar o realizar diagnósticos oportunos.

Por otraparte, observando los resultados obtenidos y oportunidades que se muestran en estos estudios previos, se debe de recordar que existen una serie de retos cuando se analizan registros médicos, Lee y Yoon, (2017) hablan de los problemas lógicos como la calidad de los datos, inconsistencias en los registros, validación, problemas de análisis, entre otros, e inclusive problemas legales (asociados al carácter de la información con que se trabaja).

En particular en México se pueden observar otros tipos de retos que están comenzando a ser resueltos en pro de un mejor uso e integración de tecnologías en la salud. Algunos ejemplos son la falta de un registro correcto de la información por parte de quienes la capturan; la ausencia de un estandar para un expediente médico que pueda ser utilizado por cualquier institución de salud independientemente de la entidad federativa o si esta es de carácter público o privado; el no contar con un sistemas de información para la captura y manejo de pacientes en las instituciones, sobre todo en aquellas de primer nivel (clínicas, consultorios rurales, etre otros), y por otra parte la incompatibilidad, pues hay quienes desarrollan sus propios sistemas con características que otras no consideraron.

Pese a las anteriores carencias y al hecho de que en la gran mayoría de instituciones aún no se aplican técnicas de minería de datos ni inteligencia artificial para explotar la información, existe la posibilidad de comenzar con la implementación de este tipo de herramientas de tal forma que sin la necesidad de grandes requerimientos de recursos por parte de las instituciones de salud se vean beneficiadas, lo que es un reto, pues primeramente se debe de romper con el ecepticismo que la inclusión de nuevas tecnologías en los procesos cotidianos siempre trae consigo. Una vez superado este primer paso, lo siguiente es que se comiencen a proporcionar resultados y hacer llegar el conocimiento generado a quienes toman decisiones para su aplicación, de lo contrario si se queda sólo como un estudio interno o académico el objetivo de mejorar la salud y las acciones de prevención no se cumplirá.

3. METODOLOGÍA

En este capítulo se describe la metodología seguida en esta investigación del tipo Cuantitativa, que tendrá un alcance Descriptivo.

La metodología propuesta ha sido planeada para abordar problemáticas en el sector salud del Estado de Sonora, en particular, se hace uso de la estrategia de “Hallazgo de Caso”, donde sin la necesidad de voluntarios para la obtención de datos, se utilizan los registros de consultas médicas de afiliados para realizar descubrimientos relacionados con Diabetes, Sobrepeso u Obesidad (Lemus et al. 2008); pero a la vez la metodología se diseñó tomando en cuenta su posible utilización en distintos tipos de entornos organizacionales, ajustando y adecuando las etapas que la componen.

Retomando el proceso del KDD, su paso de mayor importancia es utilizar minería de datos (Roiger, 2017), por ello la presente metodología la utiliza como su base para encontrar patrones y estructuras que el conjunto de datos oculta; también se integran otras herramientas para el análisis y preparación de datos. La selección del algoritmo K-Means (Cáceres, 2016) fue con base a sus características y requerimientos que se ajustan mejor al tipo y cantidad de variables con las que se trabajará, y también, aprovecha mejor los recursos (hardware y software) con los que se cuentan, además de resultar menos compleja su programación en comparación con otros algoritmos.

Basándose en lo antes mencionado y con el objetivo de generar conocimiento para el campo de la salud, en específico el área de medicina preventiva, se propone una metodología de cinco etapas como se muestra en la figura 3.1. Cada una de las etapas se subdividen en aspectos particulares para profundizar y ampliar lo que en cada una de ellas se debe de realizar.

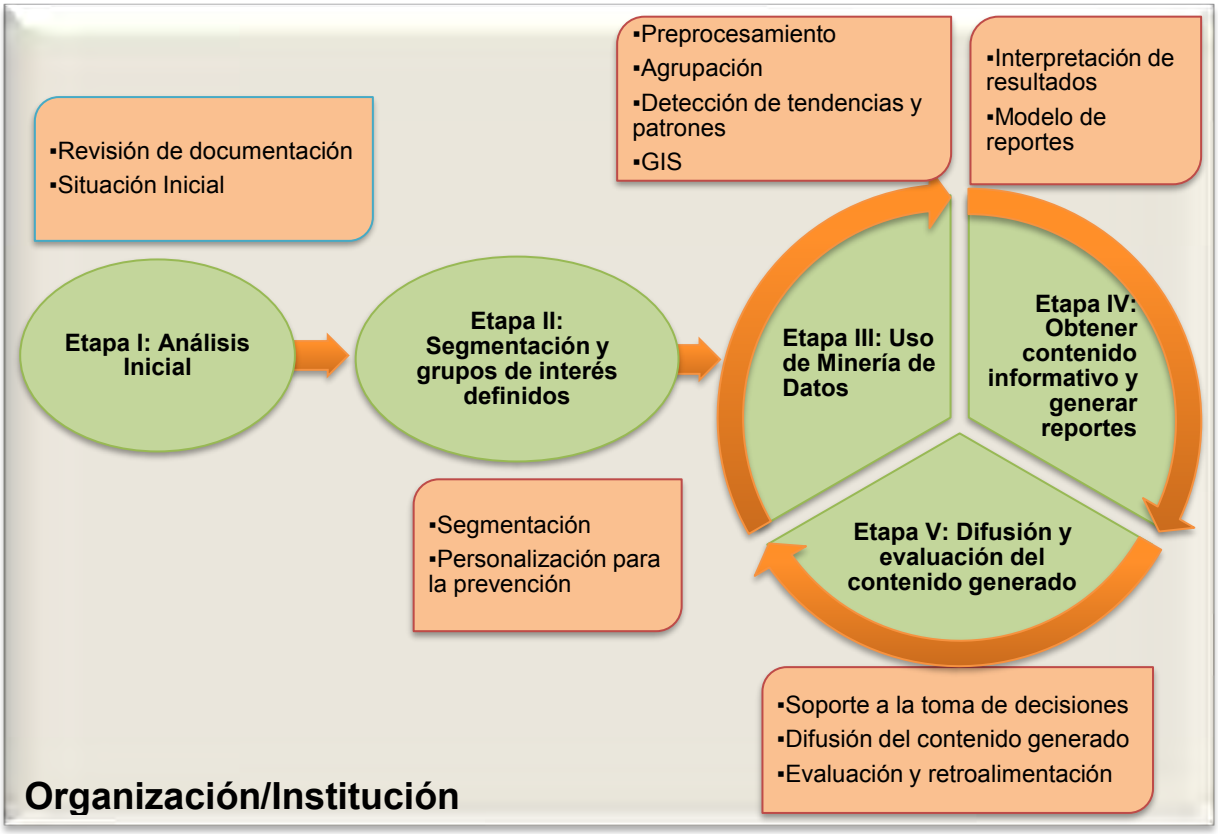


Figura 3.1. *Modelo de la metodología propuesta para la generación de conocimiento para estrategia de medicina preventiva.*

Ahora se describirán a detalle cada uno de los elementos que conforman a las 5 etapas de la metodología propuesta para que, además de poder tener un mejor entendimiento, esta pueda ser replicada en un momento dado en algún otro contexto organizacional.

3.1. Etapa I. Análisis inicial

Se deberá realizar una cuidadosa revisión de los registros a los que se tenga acceso, acatar las normativas y criterios que rigen a la institución y procesos que en esta se realizan, también considerar tendencias actuales en el campo de estudio, así pues, una revisión de documentación y un análisis de la situación inicial de la institución y sus registros es necesaria para un mejor procesamiento de estos últimos mediante las herramientas de minería de datos que se vayan a seleccionar.

El resultado de esta etapa es el análisis descriptivo de la situación actual de las enfermedades seleccionadas para su estudio en la población derechohabiente.

3.1.1. Revisión de documentación y publicaciones oficiales

Se requiere una revisión de documentación y estadísticas oficiales, ya que esto ayudará en la segmentación, detección y creación de los distintos grupos de interés, permitirá observar las tendencias y comportamiento de la o las poblaciones de estudio a través del tiempo, entre algunos aspectos más. Se podrá revisar documentación internacional de organizaciones públicas y privadas que se enfoquen en el área de interés (salud y enfermedades objetivo), también las diversas normativas nacionales vigentes sobre la prevención, diagnóstico, publicación y distribución de contenido informativo sobre enfermedades. Un ejemplo de fuentes oficiales que pueden ser consultadas son la Organización Mundial de la Salud (OMS), el Instituto Nacional de Estadística y Geografía (INEGI), el Banco Mundial, la Norma Oficial Mexicana (NOM-MX), entre muchas más, siempre y cuando se pueda verificar la integridad y veracidad de la información que de ahí se consulte o utilice.

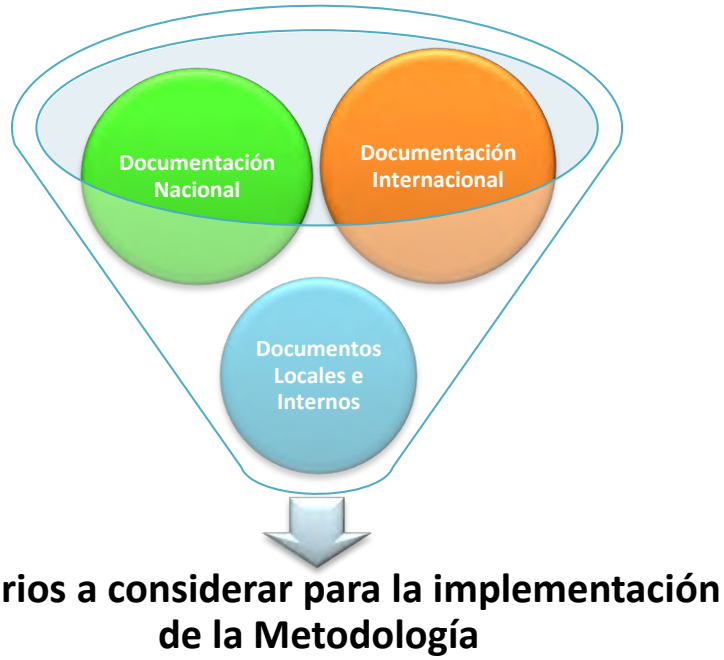


Figura 3.2. *Objetivo de la consulta de documentación y publicaciones oficiales.*

Tal y como se aprecia en la figura 3.2, el objetivo de consultar múltiples fuentes internacionales, nacionales, y por supuesto, las regulaciones y documentación local

que rigen a cada organización, es definir los criterios que se deberán de considerar a lo largo de toda la implementación de la metodología. Esto abarcará desde la difusión y ética de trabajo, hasta aspectos técnicos propios de la investigación como la definición de variables de estudio utilizadas para la definición y agrupamiento de las poblaciones de estudio.

3.1.2. Análisis descriptivo según los registros de datos, situación actual

Para el análisis inicial es necesario contar con acceso a la base de datos de la institución o a un conjunto de registros proporcionados por esta. De ser necesario para esta etapa se pueden seguir algunos de los pasos descritos en el apartado 3.3.1 *"Preprocesamiento de datos"* de este documento. Téngase siempre presente que en esta etapa se podrán identificar algunas de las variables compartidas por los registros, mismas que pueden ser utilizadas en los pasos posteriores de la metodología.

Para el análisis descriptivo se recomiendan los siguientes pasos:

1. Realizar una estandarización de los campos en los registros para asegurar la integridad de la base de datos.
2. Para el análisis de registros y presentación de los resultados, la estadística descriptiva (tablas, gráficas, histogramas, medidas de tendencia central) ayudan a un mejor entendimiento del denominado "Estado actual en la Institución".
3. Antes de aplicar alguna técnica estadística (tendencia central, predictivas) se deberá de comprobar si los datos con que se trabaja presentan una distribución normal, en especial si se trabaja con medidas de tendencia central, para que las observaciones o hallazgos que se realicen puedan ser aplicados o representen a una gran parte de la población de donde se obtuvo la muestra de datos.
4. Realizar un análisis georreferenciado para encontrar zonas o sectores de riesgo en la ciudad.
5. Presentación de resultados mediante reportes.

3.2. Etapa II. Segmentación y grupos de interés definidos

La segmentación en grupos permite la personalización de las recomendaciones preventivas y del contenido informativo que se genere, es por esto que dichos grupos deben de ser definidos con bases sólidas, basándose tanto en la documentación y literatura, y en el análisis de la población local, así asegurar que quien pertenece a cada uno de los grupos va a recibir una mejor información en función de sus características y padecimientos.

3.2.1. Segmentación

La parte de segmentación conlleva el determinar cuáles son las distintas características que la población afiliada y así separarlos en segmentos en donde compartan las mismas características generales, esta segmentación es el paso previo necesario para la determinación de los grupos de enfoque que se describirán en el siguiente apartado (3.2.2.). La figura 3.3 muestra los criterios generales que se podrán tomar en cuenta para la segmentación de los afiliados de la institución de salud según la literatura consultada.

Figura 3.3. *Criterios para Segmentación General.*

Como se describió en el marco de referencia, la salud pública y la medicina preventiva ofrecen sus servicios a las personas de manera integral, es decir, se encargan de ellas antes, durante y después de que padezcan alguna enfermedad, por lo que es importante determinar primeramente cual es el estado de salud, para después definir sus características y necesidades particulares según se requiera.

Dependiendo del riesgo, antecedentes o evolución que un afiliado tenga, es que las medidas deberán de ser tomadas para el cuidado de su salud. Existen otros criterios que van a permitir segmentar a la población en distintos conjuntos, por ejemplo, la OMS cuenta con estándares que le permiten determinar si una persona tiene bajo peso, pero normal, sobrepeso u obesidad; la NOM-MX establece una serie de categorías en las que una persona que padece o es propensa a padecer diabetes puede ser incluida; estas características particulares serán distintas según el interés de la investigación. En este sentido, Ríos-Julián et al. (2017) en su investigación sobre la viabilidad del desarrollo de una herramienta que les permitiera diagnosticar a niños con sobrepeso a través de la creación de segmentos de aquellos con tendencia a padecerlo basándose en variables antropométricas, deja ver que la segmentación es necesaria para enfocar esfuerzos en aquellos que más lo necesiten, a la vez que se utilizan criterios para la agrupación, en este caso variables antropométricas.

En resumen, la segmentación busca encontrar los criterios para separar a la población y brindarle mejor información y recomendaciones, algunos de estos criterios provienen de las variables que se detectarán en esta y la etapa anterior, al ser características propias de la población, el conocimiento generado va a influir de manera distinta en cada segmento, pero con la seguridad de que está sustentado en la propia población.

3.2.2. Personalización mediante grupos de enfoque definidos

Una vez segmentada la población afiliada, continúa el proceso de definición de grupos de interés para la personalización, esto se realiza mediante la definición de grupos de enfoque, dichos grupos permitirán un nivel mayor de especificación o acercamiento. Si bien el hecho de conocer el estado de salud de una persona permite saber qué tipo de información necesita, esta será de mayor ayuda y más específica si está adecuada al tipo de persona, estilo de vida y hábitos. Algunos de los grandes criterios para la personalización se muestran en la figura 3.4, estos pueden ser tan específicos como se necesite, hasta el grado de incluir el nivel de estudios, nivel socioeconómico, historial familiar de enfermedades, ubicación geográfica, entre otros, que una vez

combinados, permiten un grado de personalización mayor del contenido informativo y recomendaciones que se le hagan a las personas.

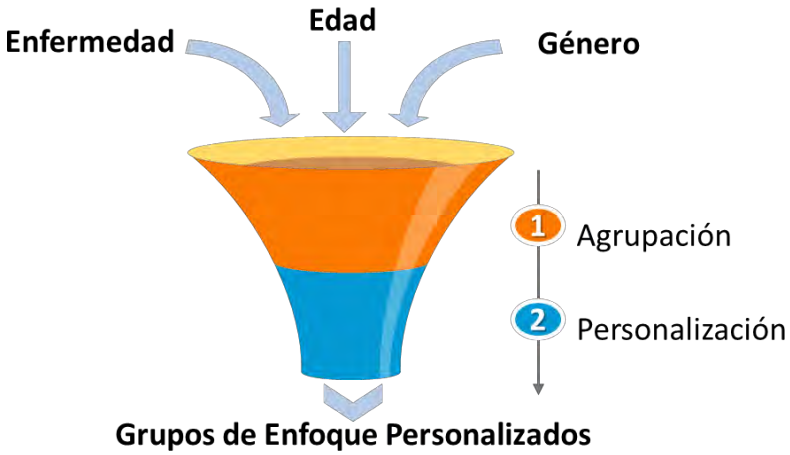


Figura 3.4. *Principales criterios para la personalización de acciones para afiliados.*

Los principales grupos de enfoque estarán determinados por la edad de la persona y su sexo, cada grupo contendrá sólo a aquellos afiliados que cumplan con estos criterios específicos, por lo que la información estará mejor dirigida, siendo clave en la realización de campañas de prevención y cuidado de salud. Los criterios que se seleccionen para la personalización dependerán de la enfermedad, por ejemplo, en el caso del sobrepeso la tabla 3.1 muestra algunos en específico, y un posible grupo de enfoque: *"género Femenino, de entre 20 y 25 años, con un IMC mayor a 24 kg/m^2 y estudiante"*, distinto a: *"mujer con posible sobrepeso".*

Variable	Descripción
Género	Masculino/Femenino
Edad	Años
Altura	En centímetros
Peso	En kilogramos
IMC	En $^{kg}/_{m^2}$
Circunferencia de cintura	En centímetros
Circunferencia de brazo	En centímetros
Espesor del tejido subcutáneo abdominal	En centímetros

Tabla 3.1. *Variables propuestas para el diagnóstico de obesidad en niños, adaptada de Ríos-Julián et al. (2017).*

Chawla y Davis (2013) investigaron sobre el impacto que la personalización tiene en el cuidado de la salud, establecen que el cuidado médico está dejando atrás el modelo donde las decisiones se toman basándose sólo en la experiencia y exámenes clínicos, para pasar al modelo que se centra en el paciente. En este modelo el paciente juega un rol activo en su propio cuidado y recibe atención de acuerdo a sus necesidades individuales, características y preferencias, mejorando la prevención y manejo de enfermedades. La forma en que este modelo enfocado a la personalización trabaja es a través de la fusión de los datos médicos del individuo y los de la población en general para determinar similitudes y diferencias, con lo que se crea un perfil de riesgo personal para recomendaciones de mayor utilidad.

3.3. Etapa III. Uso de Minería de Datos

El objetivo de esta etapa es encontrar, mediante las herramientas de minería de datos y de GIS, los patrones y tendencias ocultos en los datos procesados para su posterior interpretación y utilización en el soporte a la toma de decisiones.

3.3.1. Preprocesamiento de datos

Como en todo proceso, la calidad de los resultados finales dependerá de las entradas con que este sea alimentado, (Rodríguez Funes, 2008). La figura 3.5 muestra los tres pasos utilizados para el Preprocesamiento y filtrado de datos para el análisis de la situación actual en la institución de salud pública.

Figura 3.5. *Preprocesamiento y filtrado de datos.*

1. **Definición del Conjunto de Datos.** Las bases de datos de los HIS o sistemas de registro de consultas médicas albergan miles de registros sobre aspectos que van

desde diagnósticos, historial, análisis clínicos, recetas médicas, entre muchos otros; es importante tener en claro qué es lo que se está buscando para la correcta consulta y selección del conjunto de datos con el que se pretende trabajar. En este caso se pueden definir criterios como tipo de enfermedad, historial clínico, diagnóstico de enfermedades, entre otros, esto dependerá del objetivo del estudio que se vaya a realizar con estos registros.

2. **Asegurar Anonimato y Confidencialidad de los Datos.** Al estar trabajando con datos de afiliados reales de instituciones de salud pública se deben de tomar las medidas pertinentes para que los registros no puedan ser utilizados para fines distintos a los de la investigación y asegurar la anonimidad de las personas. También se debe de considerar el hecho de que los sistemas de información de estas instituciones trabajan continuamente, por lo que cualquier alteración o modificaciones en ellos puede ocasionar graves problemas o pérdida de información, así que se debe de crear una copia de los datos que contenga sólo la información no confidencial del paciente (no incluir: nombre, dirección, número de afiliado) y de la institución para su trabajo fuera del sistema. Es importante tomar las medidas necesarias para asegurar que el acceso a la información se dará solamente por el personal autorizado (contraseñas en equipos locales y remotos, encriptación de datos).
3. **Codificación y Estandarización.** Con el fin de un mejor manejo y procesado de los datos, se deben de eliminar todas aquellas “impurezas” mediante una tamización, es decir, un filtrado que dé como resultado sólo aquellos registros cuyos campos cumplen rigurosamente con los criterios que se necesitan, y siempre evitando la redundancia, tanto de duplicidad de registros completos como de campos en particular que no van a aportan información útil y harán lento el análisis. La estandarización se tiene que implementar en todos los registros que sean filtrados. Por último, asegurarse de que se utiliza tanto el léxico como sintaxis adecuadas (idioma y caracteres estandarizados), y cumplir con el tipo de codificación requerido por el software especializado en análisis de datos que se vaya a utilizar después en la aplicación de herramientas de minería de datos o el

de GIS (algunos software no admite palabras con acentos o caracteres especiales, puede que sólo trabajen con valores numéricos, otros admiten datos en forma de texto plano o en formato .csv).

Se debe de considerar que los registros estarán continuamente siendo actualizados o agregando nuevos por lo que este preprocesado es constante, una vez realizado el filtrado, se podrá repetir este último paso (paso 3) tantas veces como sea necesario hasta alcanzar el nivel de calidad requerido, o dependiendo del tipo de algoritmo que se elija para el procesamiento la codificación podrá variar. El preprocesamiento generará un conjunto de datos menor que con el que se inició pero con mayor calidad y relevancia de datos, lo que permite mejorar la eficiencia de la fase de procesado (Mansingh et al. 2017).

3.3.2. Selección y aplicación de algoritmo(s) de minería de datos

Al contar con una gran variedad de opciones en cuanto a algoritmos para el procesamiento de los datos filtrados, es importante seleccionar aquel que se ajuste mejor a los requerimientos del problema a tratar, llámese capacidad de procesamiento (equipo computacional y software con el que se cuente), facilidad de implementación (requerimiento de conocimiento de programación avanzada o no), tiempos de ejecución, tipos de datos que se procesarán, cantidad de datos por procesar, entre otras características que posee toda investigación cuando se aplican técnicas de minería de datos.

Selección de algoritmos

La figura 3.6 muestra el proceso de selección que se puede seguir para encontrar el algoritmo correcto y aplicarlo, este puede ser repetido tantas veces como sea necesario hasta encontrar el algoritmo que obtenga mejores resultados, pues hay que recordar que no existe un “mejor” o “peor” algoritmo.

Figura 3.6. *Proceso de selección de algoritmos de minería de datos.*

Las tareas de minería de datos pueden ser Descriptivas o Predictivas; algunas de las técnicas son de Clasificación, Agrupación, Asociación, Correlación, Regresión, entre otras; los tipos de algoritmos se clasifican en Supervisados o Predictivos, y No Supervisados o de Descubrimiento de Conocimiento; existen cientos de algoritmos de minería de datos, por ejemplo, K-NN, Árboles de Decisión, Algoritmos Bayesianos, entre otros; por último, la aplicación y evaluación del algoritmo seleccionado.

Una vez que se hayan seleccionado varias opciones de algoritmos, se puede elaborar una matriz comparativa (tabla 3.2) para evaluar las característica, pros y contras de algunos de ellos.

Algoritmo	Aplicación/ Fuente	Ventajas	Desventajas
Algoritmo 1	Predicción de:/Autor	▪ Es no supervisado. ▪ Rápida ejecución. ▪ Requiere pocos recursos de hardware.	◦ Desventaja 1 ◦ Desventaja… ◦ Desventaja n
Algoritmo…	Descripción de:/Autor	▪ Ventaja 1 ▪ Ventaja… ▪ Ventaja n	◦ Desventaja 1 ◦ Desventaja… ◦ Desventaja n
Algoritmo n	Predicción de:/Autor	▪ Ventaja 1 ▪ Ventaja… ▪ Ventaja n	◦ Requiere conocimientos de programación muy avanzados. ◦ Alto consumo de recursos.

Tabla 3.2. *Ejemplo de matriz para comparar algoritmos de minería de datos.*

Una vez seleccionado el algoritmo se puede proceder con lo siguiente.

Elección del software

Primeramente, hay que:

- Conocer las capacidades de procesamiento con que se cuenta (tipo de hardware disponible).
- Evaluar la disponibilidad de acceso a software, si se puede comprar una licencia o será gratuito.
- Saber si se necesitan conocimientos de lenguajes de programación en particular para su operación.
- Revisar si soporta el algoritmo de minería de datos que se ha seleccionado.
- Determinar qué tipo de resultados arroja y de qué manera lo hace, en forma de resúmenes numéricos, de forma tabular, gráficamente.

Aplicación de algoritmos

Dependiendo del tipo de algoritmo que se haya seleccionado, se deberán de seguir los pasos específicos que lo acompañan para una correcta aplicación. Después de que las iteraciones, ajustes y cambios en parámetros o entradas fueron realizados, se obtendrá una serie de resultados que, dependiendo del tipo de algoritmo y software utilizados, podrán ser de tipo numérico o gráfico, estos resultados deberán ser interpretados para su posterior evaluación.

Evaluación de algoritmos

Beltrán Martínez (2003) propone algunas características que se deben de considerar al momento de evaluar un algoritmo, sumada a otras más:

- **Precisión**: Debe de ser capaz de generar un modelo lo más preciso posible a la realidad de los datos, pero reconociendo que las diferencias entre herramientas se pueden deber al muestreo de datos o se pueden despreciar.

- **Explicación**: La herramienta debe de ser capaz de poder explicar los resultados obtenidos (patrones, comportamientos, estructuras) al usuario final de una manera clara, esto pensando en el soporte a la toma de decisiones.
- **Integración**: La herramienta debe de poder integrarse en el proceso real de la institución, es decir, tener la capacidad de trabajar con el tipo y flujo de datos e información que se manejan.
- **Generalización**: Tener la capacidad de poder calcular algún grado de error para determinar si los resultados provenientes del análisis de una muestra pueden ser aplicados a la población universo de donde se extrajo.

Para evitar subjetividad, se recomienda la utilización de técnicas matemáticas para determinar cuál de los algoritmos arroja mejores resultados. Las técnicas de minería de datos tienen asociadas a ellas una serie de mecanismos que buscan la validación de lo obtenido y el evitar errores, algunos ejemplos de estos son: estimación de errores, matrices de confusión, matrices de pérdida, curvas de esfuerzo y aprendizaje, análisis sensitivo de entradas, entre muchos otros (Moreno García et al. 2001), esto permite obtener resultados más completos y con un mejor grado de fiabilidad.

3.3.3. Ubicación geográfica de los sectores de interés (GIS)

La figura 3.7 muestra el proceso propuesto para la utilización de herramientas de GIS en el procesado de datos y la detección de los puntos de interés según los fines de la presente investigación, en general es muy similar al proceso de minería de datos, con la diferencia de que los resultados obtenidos son meramente gráficos (imágenes) en su mayoría y pueden ser interpretados con un mayor grado de facilidad.

Figura 3.7. *Proceso de ubicación de los sectores de interés para la investigación.*

Los datos se deben de filtrar y estandarizar asegurándose de que exista congruencia entre los distintos agrupadores (nombres de colonias, ciudades), convertir a un archivo o formato adecuado (por lo general .csv), elegir el software de GIS, cargar y validar que se hayan marcado correctamente los puntos, e interpretar los resultados.

3.3.4. Detección de tendencias y patrones en los datos

Una vez que el software de minería de datos fue aplicado, independientemente de el o los algoritmos seleccionados, en conjunto con la utilización de algún GIS en los datos preprocesados, se deberá de realizar una interpretación de los resultados obtenidos. Es recomendable contar con algún registro o tabla de valores (glosario) que nos indique el significado de aquellas variables cualitativas (nominales, ordinales) que fueron codificadas a valores numéricos reconocibles por el software de procesado de datos, ya que de esto dependerá la correcta “traducción” de los resultados numéricos a instrucciones operativas o patrones entendibles para su explicación y reporte a las áreas que utilizarán el conocimiento.

En esta parte del proceso es donde verdaderamente se determinará si la elección del software y algoritmo(s) de minería de datos fueron los correctos, pues si se obtuvieron resultados gráficos entendibles y con una complejidad de interpretación no muy elevada, se podrán hacer observaciones rápidamente y detectar si existen estructuras que describan el comportamiento de los datos, así como evaluar la necesidad de requerir un nuevo procesado de los estos.

3.4. Etapa IV. Obtener contenido informativo y generación de reportes

El objetivo de esta etapa es traducir los resultados numéricos obtenidos después de la aplicación de las distintas herramientas de minería de datos y tecnologías de análisis (estadística, GIS) a un nivel que pueda ser entendido y representado, mediante la utilización de gráficos e imágenes, para su posterior utilización por parte del personal o áreas interesadas, en este caso la de medicina preventiva de la institución. Si bien los datos tras su procesamiento se vuelven información, es necesario dar un significado dentro de un contexto a esta información para que pueda convertirse en conocimiento útil para la toma de decisiones (Dalkir y Liebowitz 2005), y es lo que se pretende lograr mediante la estandarización de reportes y su contenido.

3.4.1. Interpretación de resultados del procesamiento de datos

La parte de interpretación de resultados busca, principalmente, determinar si los resultados obtenidos después de la aplicación de minería de datos son tanto útiles como interesantes, por lo que, si lo descubierto en los datos es menor a lo esperado, se podrán utilizando nuevos parámetros para repetir el proceso. En el caso de que la interpretación deje al descubierto que lo obtenido no es ni útil o interesante en lo absoluto para la institución (esto queda a criterio del representante de la institución en el proyecto), el proceso deberá de repetirse desde la selección de los grupos de datos.

Roiger (2017) en su modelo de minería de datos contempla una etapa de interpretación y evaluación, y esta tiene por propósito determinar si lo obtenido puede ser aplicado a problemas o situaciones fuera de un ambiente de pruebas, es decir, utilizarlo en el entorno organizacional real. Si los resultados cumplen con lo esperado, el conocimiento adquirido debe de ser traducido a un lenguaje que los usuarios finales (encargados de áreas, estrategas, quien tome decisiones) entiendan y puedan utilizar. Algunas de las formas de interpretación pueden ser:

- ✓ **Análisis Estadístico**: Mediante la utilización de elementos estadísticos se puede realizar una interpretación de los resultados obtenidos.

- ✓ **Interpretación Gráfica**: Una de las mejores opciones para interpretar los resultados obtenidos es a través de la visualiza gráfica, de este modo suelen apreciarse mejor las características particulares que un grupo de datos tiene. El tipo de representación gráfica dependerá de las características de las variables y lo que estas representen (tiempo, frecuencia), así como su relación con otras.
- ✓ **Interpretación Georreferenciada:** Con la ayuda de mapas satelitales se podrán marcar sitios de interés o de relevancia para la investigación, permitiendo mostrar en una imagen el comportamiento de las variables, su agrupación, crecimiento, entre otros, con ello detectar regiones en riesgo o que requieran pronta atención.

Es importante que quien se encargue de la interpretación de resultados conozca el origen de los datos y del significado de las variables utilizadas, así como de los parámetros utilizados para alimentar al o los algoritmos utilizados, o en su caso, de alguna otra herramienta empleada para el procesamiento de datos.

3.4.2. Modelo de generación de reportes

La metodología propuesta busca la obtención de información, patrones, tendencias, predicciones, escenarios futuros y conocimiento en general y específico de enfermedades de interés, esto será utilizado para el soporte a la toma de decisiones y la elaboración tanto de campañas de medicina preventiva como de material electrónico o impreso, para ello se contará una serie de reportes periódicos para el Área o departamento de la institución de salud encargado de Medicina Preventiva. La figura 3.8 muestra un esquema de los distintos tipos de reportes contemplados a lo largo de la investigación, algunos de los cuales serán actualizados periódicamente a medida que los registros para el análisis sean actualizados y se reciba retroalimentación del contenido distribuido y de las campañas preventivas realizadas.

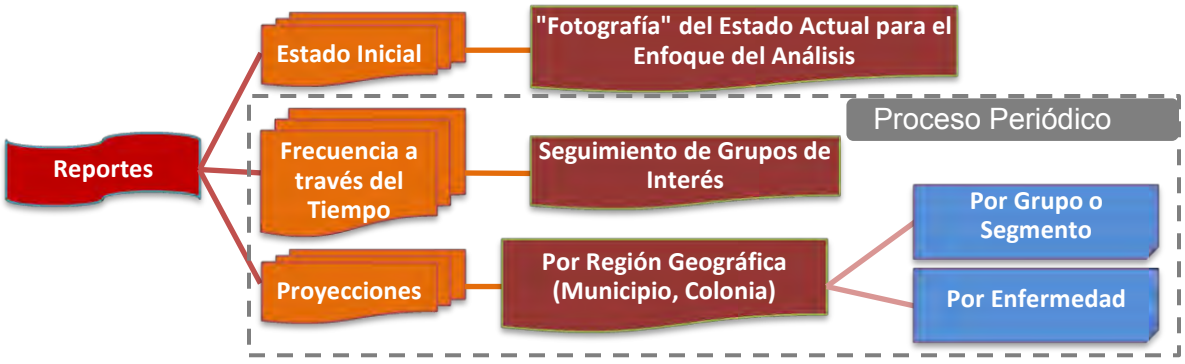


Figura 3.8. *Reportes propuestos según el tipo de información contenida y periodicidad.*

Si existe la posibilidad de hacer uso de medios electrónicos o del mismo HIS de la institución para hacerles llegar los reportes, estos pueden ser utilizados para asegurar una pronta comunicación y posible retroalimentación por parte de quienes reciban la información y conocimiento que los reportes contendrán, incluyendo gráficas, mapas y datos estadísticos que ilustren la situación.

➢ Estado Inicial

El reporte del Estado Inicial tiene como objetivo el presentar una fotografía de la situación en que se encuentran tanto las acciones que se están llevando a cabo en la Institución de Salud para la prevención y manejo de las enfermedades objetivo, así como de la evolución de las mismas enfermedades, es decir, qué se está haciendo y que tal se encuentra el panorama. De esto se desprenderán las acciones que se planteen seguir para el mejoramiento de las campañas preventivas y de seguimiento.

➢ Frecuencia a través del Tiempo

La clave de una estrategia de medicina preventiva está en los resultados sostenibles a mediano y largo plazo, no basta con tener beneficios en las etapas iniciales si se va a recaer en los mismos problemas con el paso del tiempo, por ello es clave una vez determinados los grupos de interés entre los afiliados y definido quiénes de ellos requieren de un seguimiento para vigilar la evolución de su/sus enfermedad/es o posibles riesgos de padecerlas, tener en los reportes de manera clara y simple de

observar el progreso con el paso del tiempo, ya sea mediante la utilización de tablas o gráficas donde se contengan las mediciones realizadas al paciente (IMC, Glucosa, Colesterol, Talla) en las consultas médicas posteriores al inicio del seguimiento. Se recomienda el uso de herramientas gráficas que faciliten la lectura e interpretación por parte del personal médico que trabajará con ellas, las tablas cargadas de cifras y números pueden ser complicadas y requieren de mayor tiempo para su entendimiento.

3.5. Etapa V. Difusión y evaluación del contenido generado

El objetivo de implementar esta metodología es el de descubrir y utilizar conocimiento que se encuentra oculto en los patrones y tendencias de los datos en registros de una institución de salud, esto con el fin de realizar campañas de medicina preventiva que tengan un mejor trasfondo y se sustenten en la realidad de la población afiliado a la que están enfocadas, en otras palabras, realizar campañas preventivas tomando en cuenta las características y necesidades particulares de los individuos que pertenecen a la institución de salud para buscar que las decisiones que se tomen tengan un mayor impacto, sean mejor percibidas y lleguen a más personas. El aspecto de la toma de decisiones se puede apreciar en el trabajo de García y López (2012), quienes establecen que la utilización de herramientas de análisis como la minería de datos tienen como objetivo el incorporar el conocimiento obtenido en algún proceso real, tomar decisiones a partir de los resultados, o simplemente registrar la información conseguida y suministrársela a quien tenga interés en ella.

3.5.1. Propuesta y ajuste del contenido informativo para su utilización en campañas preventivas a través de redes sociales

Para la difusión del material informativo se deben de contemplar los distintos medios de comunicación con los que se cuentan, el tipo de público al que se pretende llegar, así como los requerimientos técnicos que se deben de cumplir. Esto toma mayor relevancia cuando se van a utilizar medios electrónicos, pues se deberán de tomar en cuenta las actuales tendencias que se estén presentando en estos canales de contacto entre los afiliados o la sociedad en general y el contenido, información y recomendaciones que se quieran hacer llegar por parte de la institución.

En redes sociales se debe de ser rápido para transmitir un mensaje a la vez que se necesita llamar la atención del usuario, por lo que la elección correcta de palabras y la forma en que se estructure el mensaje, así como su presentación (texto, imagen, audio, video) deberán estar pensadas para un rápida transmisión y un impacto que despierte la curiosidad de quien lo observe, siempre brindando la opción de acceder a mayor información a través de alguna liga o link, así evitas el texto o duración de video excesivo, tal y como se observa en la figura 3.9, es una publicación de Clínica Mayo en Twitter, el texto es breve y da información precisa sobre lo que se quiere transmitir, acompañado de una imagen representativa y una liga (url) para mayor información sobre el tema.

Figura 3.9. *Ejemplo de contenido informativo para redes sociales de @ClinicaMayo (Sep. 2018).*

La simplicidad y contundencia es clave, los grupos de enfoque (madres, niños, adolescentes, mujeres, hombres) van a variar y así debe de cambiar el contenido para ajustarlo a, por ejemplo, el nivel de comprensión, la edad, el tipo de población, entre otras características específicas de cada grupo.

Por otra parte, se está volviendo común que las instituciones de salud utilicen aplicaciones móviles en conjunto en conjunto con sus HIS para recabar información y mantener un contacto con el afiliado, por lo que también se deben de considerar las características técnicas y formatos con los que el contenido informativo deberá de contar si se pretenden utilizar como medio de difusión. Lo mismo se deberá de contemplar para cada medio electrónico que sea seleccionado para hacer llegar el mensaje al usuario, es decir, revisar requerimientos (resolución, tipo de mensaje, codificación).

3.5.2. Evaluación y retroalimentación

La parte final de esta metodología está enfocada en la determinación del impacto que se ha obtenido tras la implementación, pero también sobre las recomendaciones y observaciones que se tengan y que realicen por parte de los involucrados para la adecuación o mejora de la misma.

- **Evaluación.**

Para determinar el impacto que se tenga con la implementación de la metodología es necesario evaluar de manera cuantitativa, cualitativa o utilizando ambas, lo que dependerá del tipo de acción realizada y el producto obtenido de esta, por ejemplo, el conocimiento que se generó o el alcance de la difusión en redes sociales, por mencionar algunas. La definición o elección de métricas es el aspecto primordial que se debe de considerar en esta etapa, de ello dependerá el cómo se realizan las mediciones y la determinación del grado de impacto.

Al estar trabajando con la generación de conocimiento, este puede ser utilizado de múltiples maneras, en el caso del entorno organizacional específico de la institución de salud seleccionada, este está pensando para su aprovechamiento en campañas preventivas, por lo que como se muestra en la figura 3.10, el primer paso para la evaluación es identificar los puntos o aspectos que deben de ser medidos, después definir las métricas que se deberán de utilizar para evaluar, realizar una medición de la situación actual de los aspectos seleccionados (antes de la implementación),

medirlos mismos aspectos después de la implementación, realizar una comparación para determinar el grado de impacto, y por último, presentar los resultados obtenidos.

Figura 3.10. *Proceso de evaluación de los resultados de la metodología propuesta.*

Independientemente de los valores numéricos que se evalúen, existen aspectos cualitativos que también pueden ser medidos para determinar si realmente se percibe una mejoría. Se puede evaluar el nivel de aceptación de los resultados, la calidad del conocimiento generado desde el punto de vista de las personas que toman las decisiones en la organización, la cantidad de aportaciones que se hayan realizado a las campañas de salud, entre muchos otros aspectos. Para conocer lo que es percibido por las personas, se puede realizar una serie de preguntas o un diálogo directo que posteriormente sean analizados y se transformen en valores medibles.

- **Retroalimentación.**

Considerando que las últimas tres etapas de la metodología representan un ciclo que se podrá volver a poner en marcha cada vez que se actualicen o nuevos datos sean agregados para su procesado, la retroalimentación o sugerencias podrán surgir en cualquiera de estas etapas, e inclusive desde los inicios de la implementación, según sea el caso.

Se deberán de considerar los cambios y sugerencias que sean recibidas para su integración en los procesos o cambios de los productos (reportes, contenido informativo, conocimiento, entre otros) que se obtendrán en la implementación. Se puede pensar en un proceso de mejora continua gracias a las características cíclicas de la metodología, y para incrementar la calidad a medida que se evalúan los resultados, los procesos que se desarrollan en cada una de las etapas, así como los resultados que se obtengan, pueden ser adaptados a las características de la organización donde se implemente. Una de las principales ventajas de la metodología aquí descrita es que el modelo puede ser alimentado periódicamente para mejorar sus resultados e incluso, detectar la evolución en el comportamiento de las enfermedades.

4. IMPLEMENTACIÓN

Los resultados obtenidos y las observaciones realizadas durante la aplicación de la metodología antes descrita, así como todo el proceso requerido, se encuentra documentado en esta sección.

4.1. Etapa I. Análisis inicial

Se revisó la documentación interna/externa disponible y se realizó el análisis descriptivo de los registros obtenidos para el estado actual de la institución en cuanto obesidad, diabetes y la difusión de contenido informativo en su población afiliada.

4.1.1. Revisión de documentación y publicaciones oficiales

Esta tarea busca definir criterios que ayuden a una correcta la segmentación de las personas tomando en cuenta regulaciones nacionales e internacionales, por ejemplo, basarse en tablas de desarrollo físico de la OMS, los niveles de IMC para la clasificación del sobrepeso y la obesidad de la NOM, los niveles de glucosa en la sangre en el caso de la diabetes, entre otros. Se consultó información proveniente de:

- Organización Mundial de la Salud (OMS).
- Norma Oficial Mexicana (NOM).
- Instituto Nacional de Estadística y Geografía (INEGI).
- Encuesta Nacional de Salud y Nutrición (ENSANUT).

Fue importante basarse en normas nacionales vigentes ya que existen múltiples métodos para el diagnóstico y segmentación de las personas que son propensas, están en riesgo de padecer o padecen algunas enfermedades, pues de otra manera se estaría trabajando fuera de los criterios que las instituciones de salud en el país se ven obligadas a seguir, volviendo este trabajo no aplicable por este motivo.

Criterios en Sobrepeso y Obesidad

En la reciente actualización de la “NOM-008-SSA3-2010” Para el tratamiento integral del sobrepeso y la obesidad (Secretaría de Salud Pública 2010), se igualaron los

valores de IMC con los de la OMS para categorizar a las personas en peso normal, sobrepeso u obesidad. Se define que una persona de acuerdo con su IMC podrá ser clasificada dentro de las siguientes categorías (**¡Error! No se encuentra el origen de la referencia.**4.1):

Índice de Masa Corporal ($^{kg}/_{m^2}$)	**Categoría**
IMC ≥ 18.5 y ≥ 24.9	Peso Recomendable
IMC ≥ 25 y ≥ 29.9	Sobrepeso
IMC ≥ 30 y ≥ 34.9	Obesidad I
IMC ≥ 35 y ≥ 39.9	Obesidad II
IMC ≥ 40	Obesidad III

Tabla 4.1. *Clasificación según el IMC de la persona. NOM-015-SSA2-2010 y ENSANUT 2016.*

Existe una serie de consideraciones que hay que tener en cuenta adicionalmente para la categorización de las personas, es la estatura y la edad, a continuación, se enumeran:

- Se diagnosticará Obesidad en las personas adultas de estatura baja (menor a la considerada como normal para su edad) cuando su IMC sea igual o mayor a 25 kg/m².
- Se diagnosticará Sobrepeso en las personas adultas de estatura baja cuando su IMC sea igual o mayor a 23 kg/m² y menor a 25 kg/m².

Cuando la persona tiene menos de 19 años, se utilizan las tablas de IMC para edad y sexo de la OMS.

- Se diagnosticará Obesidad cuando el IMC se encuentra por encima del 95% del grupo al que pertenece.
- Se diagnosticará Sobrepeso cuando el IMC se encuentra desde el 85% y por debajo del 95% del grupo al que pertenece.

Criterios en Diabetes

Las siguientes definiciones y criterios para diagnóstico de casos descartados, probables, confirmados y bajo control, así como los tipos 1 y 2 de diabetes, fueron extraídos de la "NOM-015-SSA2-2010" (Secretaría de Salud Pública 2010) y se explican con detalle a continuación.

- ✓ **Diabetes Tipo 1:** Es aquel tipo de diabetes en la que existe destrucción de células beta del páncreas, generalmente con deficiencia absoluta de insulina. Los pacientes pueden ser de cualquier edad, casi siempre delgados y suelen presentar comienzo abrupto de signos y síntomas con insulinopenia (deficiencia en la secreción de insulina) antes de los 30 años.
- ✓ **Diabetes Tipo 2:** Es aquel tipo de diabetes en la que se presenta resistencia a la insulina y en forma concomitante una deficiencia en su producción puede ser absoluta o relativa. Los pacientes suelen ser mayores de 30 años cuando se hace el diagnóstico, son obesos y presentan relativamente pocos síntomas clásicos.

Posibles clasificaciones para una persona con diabetes o propensa a ella:

- **Descartados.** Un caso de diabetes descartado se presenta cuando no se cumple con los criterios de caso probable o confirmado, o bien el paciente presenta síntomas o signos de alguna otra enfermedad.
- **Sospechosos.** Un caso de diabetes sospechoso se presenta cuando la persona con factores de riesgo comunes para enfermedades no transmisibles: edad (mayor de 20 años), antecedente heredofamiliar (padres y/o hermanos), sobrepeso u obesidad, circunferencia abdominal mayor de 80 cm en mujeres o 90 cm en hombres, hijo macrosómico (bebé con un tamaño mayor al promedio al nacer) en mujeres, hipertensión arterial.
- **Probables.** Un caso de diabetes probable se presenta cuando la persona a la que se le ha realizado alguno de los exámenes de detección (ver Confirmados),

presenta una glucemia capilar en ayuno > 100 mg/dl, o una glucemia capilar casual > 140 mg/dl.

- **Prediabetes.** Un caso de prediabetes se presenta cuando la persona cuenta con antecedente de padre o madre o ambos con estado metabólico intermedio entre el estado normal y la diabetes, o cuando el nivel de glucosa en ayuno es igual o mayor a 100 mg/dl y menor o igual de 125 mg/dl. Este término se aplica para diabetes Tipo 1 y 2.
- **Confirmados.** Un caso de diabetes confirmado se presenta cuando el paciente da positivo para alguno de los siguientes exámenes: niveles de glucemia plasmática (azúcar en la sangre) en ayuno de 126 mg/dl; una glucemia plasmática casual (muestra de sangre tomada al azar) de 200 mg/dl; o bien una glucemia de 200 mg/dl a las dos horas después de una carga oral de 75 g de glucosa anhidra disuelta en agua. Estas pruebas que se les realizan a los pacientes son criterios aceptados por el Sistema Nacional de Salud en México.
- **Controlado.** Un caso de diabetes controlado se presenta cuando el paciente bajo tratamiento presenta de manera regular, niveles de glucemia plasmática en ayuno de entre 70 y 130 mg/dl o de Hemoglobina Glucosilada (HbA1c) (mezcla de células sanguíneas y glucosa) por debajo de 7% (Martínez Mandujano et al. 2015).

En el caso del diagnóstico de diabetes en niños y jóvenes se deben de considerar una serie de criterios y rangos, algunos de los cuales son:

- ❖ Sobrepeso en niños (IMC > del percentil 85 para la edad y sexo, peso para la talla > del percentil 85, o peso mayor de 120% ideal para la talla)
- ❖ Historia de diabetes tipo 2 en el primero o segundo grado familiar.
- ❖ Signos y/o condiciones de resistencia a la insulina (acantosis nigricans, hipertensión arterial, dislipidemia, o síndrome de ovarios poliquísticos).
- ❖ En la mayoría de las personas jóvenes, el diagnóstico de la diabetes tipo 1 deberá hacerse sin dificultad y de manera urgente. Los síntomas de sed, ingesta de líquidos y micción excesivos deben inducir a la realización inmediata de pruebas anticuerpos anti-insulares.

Criterios Institucionales

Al tratarse de registros reales de afiliados y familiares que actualmente están o pertenecieron a esta institución de salud, la principal restricción impuesta fue la confidencialidad absoluta en el manejo de los archivos y registros que fueran proporcionados por la misma. En ningún momento se podría tener conocimiento de la identidad de la persona a quien pertenecieran los registros ni a su ubicación en específico (sólo su colonia de residencia).

Otra de las consideraciones fue que siempre al momento de publicar o difundir algún tipo de información o contenido referente a la forma en que la institución realiza sus campañas preventivas o recomendaciones a sus afiliados, debería de ser aprobada primeramente por parte del contacto dentro de la institución, mismo que canalizaría a quien correspondiera el asunto para validar si lo que se propone tendría o no el respaldo y aprobación necesarios para su difusión.

Para lo anterior 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 proporcionara.

4.1.2. Análisis descriptivo según los registros de datos, situación actual

De la consulta de fuentes oficiales se seleccionaron a la *edad, sexo y tipo de enfermedad* para realizar el análisis descriptivo, tomando en cuenta las fechas de diagnóstico y el tiempo (en años) de seguimiento que se les brinda a los afiliados.

- **Acceso y filtrado de registros**

Los registros sobre el diagnóstico y seguimiento de los casos de obesidad y diabetes fueron adquiridos del departamento de Estadística, perteneciente a la Subdirección Médica de la institución de salud pública. En total se tuvo acceso a 8,575 registros de afiliados con diabetes y 771 de aquellos con obesidad, mismos que representaban al periodo comprendido entre el año 2014 y mayo de 2017. Al tratarse de un primer

análisis sólo sobre aquellos que ya presenta alguna de las enfermedades, los datos podrán ser utilizados para el entrenamiento de algoritmos supervisados, si se selecciona uno de este tipo.

Los archivos que contenían los registros fueron proporcionados en formato .CSV para su manipulación mediante las diversas herramientas estadísticas, de graficación y análisis de datos. Siguiendo el proceso de preprocesamiento de esta metodología (sección 3.3.1) se desglosaron los 8,575 y los 771 registros de casos de diabetes y obesidad respectivamente, para su procesamiento (tabla 4.2).

Diabetes		**Obesidad**	
Total de Registros	8,575	Total de Registros	771
De Diagnóstico	2,748	De Diagnóstico	689
De Seguimiento	5,827	De Seguimiento	82
Georreferenciados	1,656	Georreferenciados	670

Tabla 4.2. *Preprocesamiento inicial de registros. Los datos de diagnóstico corresponden a la primera vez que el derechohabiente fue notificado con el padecimiento de alguna de las enfermedades. Los datos de seguimiento corresponden a visitas posteriores del derechohabiente una vez que fue diagnosticado. Los registros que contaban con ubicación precisa y específica se utilizaron para georreferenciación.*

La descripción detallada del preprocesamiento se muestra en los siguientes puntos:

❖ Obesidad

Partiendo de los 771 registros originales, 689 correspondían a los diagnósticos de obesidad (en cualquiera de sus variantes), y los 82 restantes son de las citas de seguimiento que se dieron a los afiliados.

❖ Diabetes

De los 8,575 registros, 2,748 representan a los diagnósticos de diabetes (en cualquiera de sus variantes), mientras que los 5,827 restantes corresponden a las citas de seguimiento que estos tuvieron a través del periodo de tiempo analizado.

➢ Resultados del análisis Descriptivo

Al igual que con el preprocesamiento de datos, se optó por realizar el mismo tratamiento con los registros de diabetes y obesidad, es decir, los dos padecimientos se procesaron y estudiaron de modo similar, siguiendo la misma técnica.

El primer paso es determinar el grado de representación de la muestra obtenida respecto a la población afiliada y la población en general.

Año	2014	2015	2016	2017
Población Hermosillo	857,223	870,096	882,716	895,100
Población Institución	85,689	85,371	84,099	84,761
% de Representación	10.00%	9.81%	9.53%	9.47%

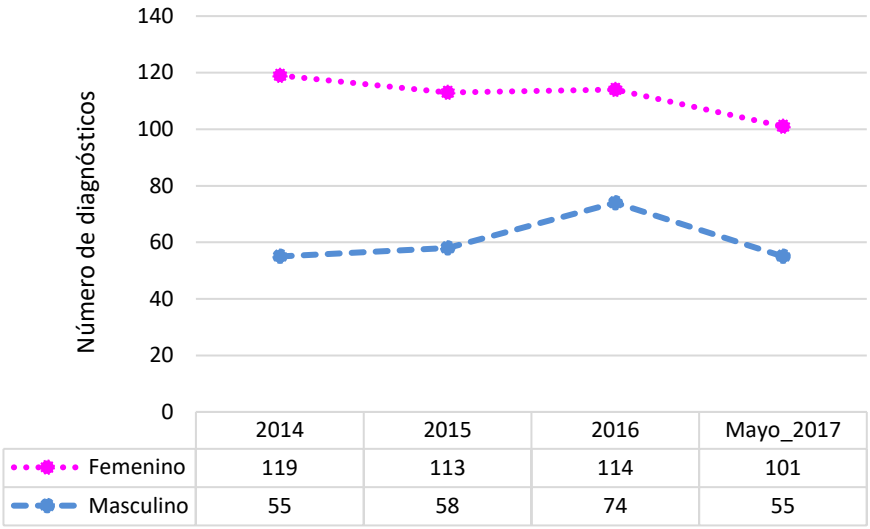
Tabla 4.3. *Porcentaje de la población de Hermosillo perteneciente a la institución de salud*.

La tabla 4.3 fue creada con información de la institución y el Consejo Nacional de Población. En esta se aprecia que la población institucional representa aproximadamente un 10% de los habitantes del municipio de Hermosillo, además se sabe que el factor socioeconómico también se encuentra representado en esta muestra (por la organización en la que trabajan, dato presente en los registros obtenidos) mismo que es considerado de importancia cuando se trata del diagnóstico de alguna de las enfermedades aquí analizadas, pues se ha demostrado que afectan en mayor medida a aquellos con menores ingresos.

➢ Análisis a través del tiempo

Diagnósticos totales y relativos por año y género:

❖ Obesidad.



	2014	2015	2016	Mayo_2017
Femenino	119	113	114	101
Masculino	55	58	74	55

Figura 4.1. *Total de diagnósticos de Obesidad por año y género. Para el año 2017 se contemplan las cifras hasta el mes de mayo.*

En la figura 4.1 se observa una tendencia ligeramente al alza en ambos géneros si consideramos que en 2017 sólo se habían contabilizado los primeros 5 meses y ya se está a menos de 15 diagnósticos de alcanzar al total del año anterior. El diagnóstico total en mujeres es en promedio un 45% más que en hombres, también es notable el bajo número de diagnósticos (689 en casi 4 años), situación que puede ser atribuida a la falta de un diagnóstico oportuno, por lo que se buscará corroborar esta tendencia mediante el análisis de los registros de estatura y peso.

En cuanto a los valores relativos, tomando en cuenta el total de diagnósticos por género y el total de afiliados del mismo año, la tendencia es parecida a la de los valores totales, pero se acortan las distancias (figura 4.2) entre hombres y mujeres.

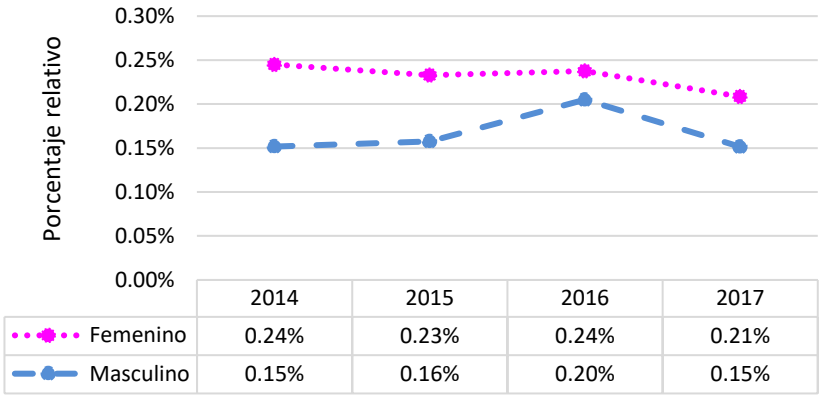


Figura 4.2. *Valores relativos del diagnóstico de Obesidad por año y género. Para el año 2017 se contemplan las cifras hasta el mes de mayo.*

En 2014, el 0.24% de las mujeres de la población general fueron diagnosticadas con obesidad, mientras que poco un 0.15% en el caso de los hombres.

Para el cálculo de los valores relativos se tomó el total de diagnósticos en mujeres en un año en particular y se dividió entre el total de afiliados del mismo sexo y del mismo año; el proceso fue igual para los hombres. Del total de diagnosticados con obesidad durante el periodo de análisis el 64.88% correspondieron a mujeres y el 35.12% a hombres.

❖ Diabetes.

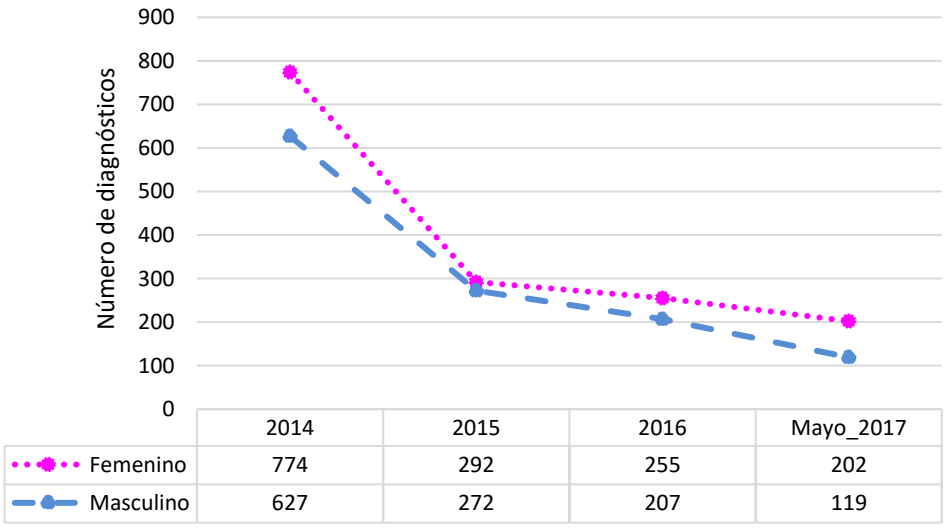
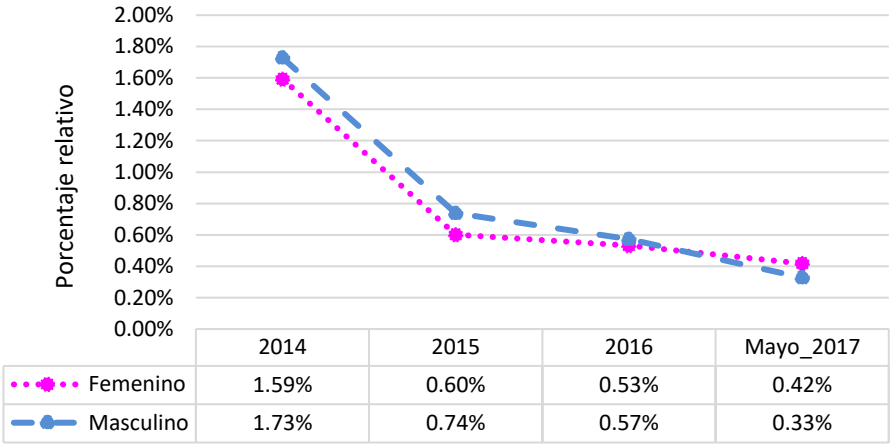


Figura 4.3. *Total de diagnóstico de Diabetes por año y género (2,748 casos totales).*

En la figura 4.3 se aprecia que el número de casos va considerablemente a la baja sobre todo en hombres, aunque sólo en los primeros 5 meses de 2017 se está por alcanzar al total de diagnósticos del año anterior en mujeres, lo que sugiere que se presentará un repunte. Aunque el número de mujeres es superior al de los hombres, una vez que se profundizó en el análisis calculando valores relativos en relación con el género se pudo observar que el número de diagnósticos es menor en mujeres que el presentado en hombres en 3 de los 4 años de registros analizados (figura 4.4).



	2014	2015	2016	Mayo_2017
Femenino	1.59%	0.60%	0.53%	0.42%
Masculino	1.73%	0.74%	0.57%	0.33%

Figura 4.4. *Valores relativos del diagnóstico de Diabetes por año y género.*

Se observa un aproximado del 3.24% de población con diabetes en general, además en el año 2017 las mujeres estaban presentando un mayor número de diagnósticos después de 3 años de que los hombres estuvieran liderando. Del total de diagnosticados con diabetes durante el periodo de análisis el 55.42% correspondieron a mujeres y el 44.58% a hombres, reduciendo la brecha entre géneros, a diferencia de lo observado con la obesidad donde es más marcada la separación. Es importante recalcar que considerando el mayor riesgo de las mujeres a padecer obesidad será un factor de riesgo que conlleve a un aumento en los casos de diabetes.

➢ Análisis por género y edad

Para conocer la situación de incidencia de diabetes en hombres y mujeres, se procedió a hacer un análisis por género, complementándolo con grupos de edad.

Género	2014	2015	2016	2017
Femenino	57.24%	56.86%	57.07%	57.15%
Masculino	42.76%	43.14%	42.93%	42.85%

Tabla 4.4. *Porcentajes de población por año y género en la institución de salud.*

Como se observa en la taba 4.4, es casi constante la proporción 57% mujeres y 43% hombres durante todos los años de estudio, la diferencia promedio es de 14% entre géneros. Se busca descartar que el mayor número de diagnósticos totales se deba a que existe más mujeres que hombres en la población afiliada. Según la bibliografía, con el simple hecho de ser mujer ya se está en mayor riesgo y se es más propensa a padecer obesidad y diabetes.

❖ Obesidad.

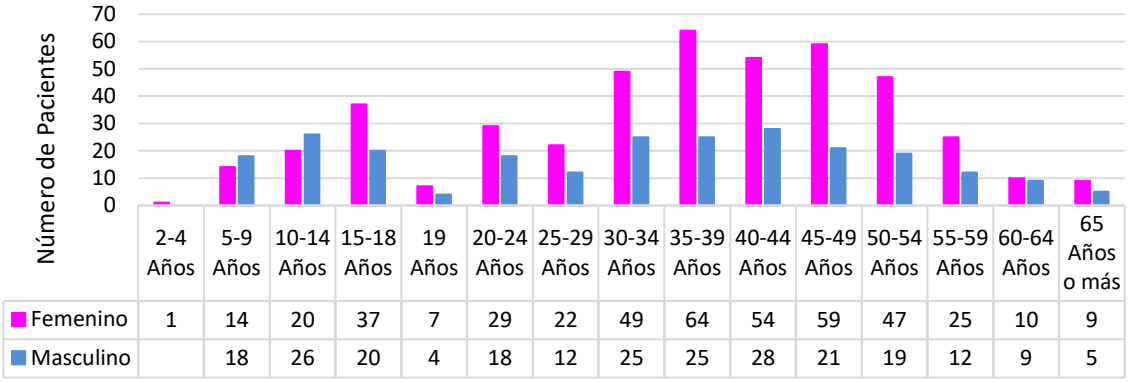
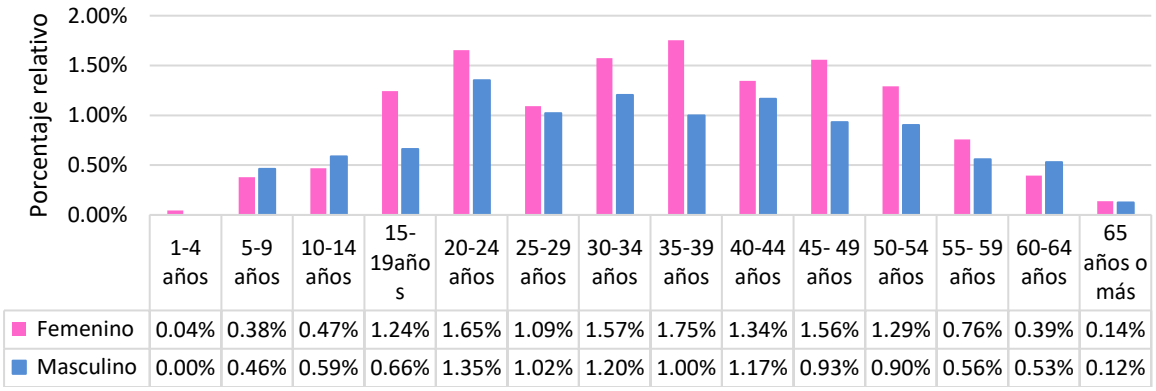


Figura 4.5. *Total de diagnóstico de Obesidad por género y grupo de edad.*

El total de diagnósticos muestra que existe un mayor número en mujeres diagnosticadas en 13 de los 15 grupos de edad definidos por la institución (figura 4.5), llegando a presentar hasta más del doble de diagnósticos que los hombres en ciertos grupos. Concretamente se observa que existen edades en donde se intensifica el diagnóstico, como en el caso de niños y jóvenes (10 a 18 años) y en adultos (30 a 54 años), lo que sugiere prestar especial atención a estos grupos.

Al calcular los valores relativos de los diagnósticos se siguió presentando la tendencia de mayor proporción de mujeres con obesidad, pero con una notable disminución de

la brecha entre géneros, pasando de un estimado de 2 diagnósticos en mujeres por cada diagnóstico en hombres (en ciertos grupos de edad) a sólo 1.4 diagnósticos en mujeres por cada diagnóstico en hombres como máximo (figura 4.6).



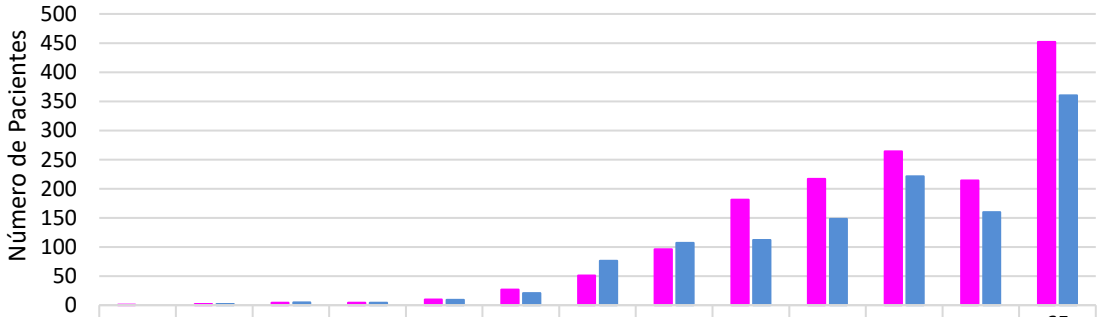
	1-4 años	5-9 años	10-14 años	15-19años	20-24 años	25-29 años	30-34 años	35-39 años	40-44 años	45- 49 años	50-54 años	55- 59 años	60-64 años	65 años o más
Femenino	0.04%	0.38%	0.47%	1.24%	1.65%	1.09%	1.57%	1.75%	1.34%	1.56%	1.29%	0.76%	0.39%	0.14%
Masculino	0.00%	0.46%	0.59%	0.66%	1.35%	1.02%	1.20%	1.00%	1.17%	0.93%	0.90%	0.56%	0.53%	0.12%

Figura 4.6. *Valores relativos del diagnóstico de Obesidad por género y grupo de edad.*

Los valores relativos muestran que el grueso de diagnósticos se presenta en jóvenes 15 a 24 y adultos de 30 a 54 años tendencia que no se observaba con los totales.

❖ Diabetes.

En cuanto a la diabetes, los totales muestran que existe alrededor de 20% más diagnósticos en mujeres que en hombres en algunos grupos de edad (figura 4.7).

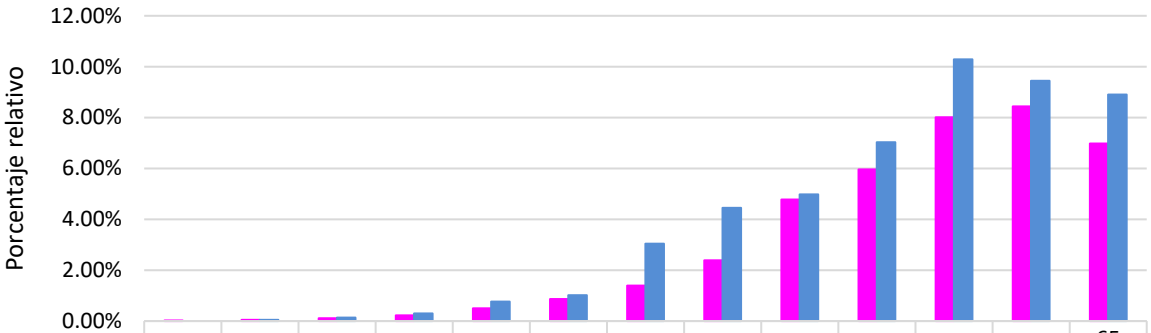


	5-9 Años	10-14 Años	15-18 Años	20-24 Años	25-29 Años	30-34 Años	35-39 Años	40-44 Años	45-49 Años	50-54 Años	55-59 Años	60-64 Años	65 Años o más
Femenino	1	2	4	4	10	27	51	96	181	217	264	214	452
Masculino		2	5	4	9	21	76	107	112	148	221	160	360

Figura 4.7. *Totales del diagnóstico de Diabetes por género y grupo de edad.*

A diferencia de la obesidad, el diagnóstico de diabetes se intensifica a medida que aumenta la edad, en el caso particular de la población de la institución, a partir de los 35 años comienza a incrementar el número de diagnósticos en ambos géneros y la tendencia es a aumentar, se sugiere vigilar los grupos de adultos y adultos mayores. Esta tendencia es ya conocida y propia del padecimiento.

Después de calcular los valores relativos se esperaba que existiera semejanza con el comportamiento presentado en obesidad y que las mujeres se mantuvieran con un mayor número de diagnósticos que los hombres, como lo sugieren otras investigaciones y los totales calculados, sólo que esta tendencia simplemente no se presentó en la población de la institución (figura 4.8). Solamente en los rangos de 5 a 24 años hubo mayor proporción de mujeres diagnosticadas con diabetes, aunque con valores casi imperceptibles.



	5-9 Años	10-14 Años	15-18 Años	20-24 Años	25-29 Años	30-34 Años	35-39 Años	40-44 Años	45-49 Años	50-54 Años	55-59 Años	60-64 Años	65 Años o más
Femenino	0.03%	0.05%	0.11%	0.23%	0.50%	0.87%	1.40%	2.39%	4.78%	5.96%	8.01%	8.44%	6.98%
Masculino	0.00%	0.05%	0.14%	0.30%	0.77%	1.01%	3.04%	4.46%	4.97%	7.03%	10.29%	9.44%	8.91%

Figura 4.8. *Valores relativos del diagnóstico de Diabetes por género y grupo de edad.*

Con este análisis de proporciones relativas se logró detectar que se está presentando una tendencia sumamente interesante en la población de la institución en Hermosillo, donde existen grupos de edad en que los hombres superan en aproximadamente un 30% al número de mujeres diagnosticadas con diabetes, y a partir de los 25 años esta tendencia es constante, aun y cuando existe un mayor número de mujeres en la población general.

➢ Principales variantes de las enfermedades

Ya que cada uno de los distintos tipos o variantes de las enfermedades requieren un tratamiento y definición de criterios para su diagnóstico diferentes, se necesitaba conocer en qué proporción se están presentando los distintos tipos. En el caso de la obesidad, aquella catalogada como “no especificada” es la que mayor número de casos ha presentado, seguida de la ocasionada por un exceso en la ingesta de calorías (tabla 4.5), este último permite enfocar esfuerzos preventivos en el control y recomendaciones sobre alimentación.

TIPOS DE OBESIDAD	Número de Casos	Valor Relativo
Obesidad debida a exceso de calorías	202	29.32%
Obesidad inducida por drogas	1	0.15%
Obesidad, no especificada	485	70.39%
Otros tipos de obesidad	1	0.15%

Tabla 4.5. *Tipos de obesidad y proporciones.*

TIPO DE DIABETES	Número de Casos	Valor Relativo
Diabetes mellitus no insulinodependiente, sin complicación (Tipo 2)	1,136	41.34%
Diabetes mellitus Tipo 1	58	2.11%
Otros Tipos	1,154	56.55%

Tabla 4.6. *Tipos de diabetes y proporción. Los “Otros Tipos” de diabetes más relevantes son aquellos asociados con desnutrición, diabetes insípida, los insulinodependientes (tipo 1 o 2) con complicaciones, por mencionar algunos.*

Lo destacable en los tipos de diabetes diagnosticados es que un 41.34% corresponden a aquellos afiliados que aún no presentan ningún tipo de complicación ni requieren de insulina para su control (tabla 4.6), este grupo debe de ser tomado en consideración para evitar que se evolucione y agrave la enfermedad. Por otra parte, el 56.55% también padecen diabetes tipo 2 sólo que ya cuentan con otro tipo de complicaciones, por lo que alrededor del 97% de los diagnósticos pudieron ser prevenidos o detectados antes de que tuvieran una mayor evolución, ya que sólo aquellos que padecen diabetes tipo 1 son por razones fuera de su control o prevención.

➢ Entorno laboral de los afiliados

En 40 de las organizaciones afiliadas a la institución de salud con algún tipo de obesidad, en 10 de estas se encuentra concentrado el 88.97%% o 613 de los 689 casos (tabla 4.7). El 75% de las organizaciones registraron entre 9 y 1 caso solamente.

Se sustituyeron los nombres de las organizaciones por fines de confidencialidad.

ORGANIZACIÓN	Diagnósticos (2014 - mayo 2017)	% Proporcional
A	203	0.80%
B	123	0.87%
C	105	0.85%
D	62	0.63%
E	33	0.50%
F	24	0.59%
G	22	1.06%
H	15	1.03%
I	15	0.94%
J	11	2.78%

Tabla 4.7. *Organizaciones afiliadas con mayor número promedio de diagnósticos de Obesidad. La proporción es un promedio del número de diagnósticos de todo el periodo de análisis entre el número de afiliados de cada organización.*

ORGANIZACIÓN	2014	2015	2016	Mayo 2017	Total
A	64	45	47	47	203
B	27	23	38	35	123
C	14	31	34	26	105
D	15	19	16	12	62
E	9	11	7	6	33
F	4	8	8	4	24
G	7	9	3	3	22
H	5	4	4	2	15
I	9	1	2	3	15
J	1	3	2	5	11

Tabla 4.8. *Organizaciones afiliadas con mayor número de diagnósticos de Obesidad por año.*

La columna “% Proporcional” de la tabla 4.7 se calculó considerando el número total de afiliados que la organización registró ante la institución de salud y dividiéndolo entre el número de casos de obesidad diagnosticados. En promedio casi un 1% de los miembros de las organizaciones padecen de obesidad, destacando el 2.78% de la organización “J”. En la tabla 4.8 se observan los valores totales registrados en las 10 organizaciones con mayor número de diagnósticos de obesidad, hay que puntualizar que lo más probable es que existan personas que aún no han sido diagnosticadas, por lo que la proporción se espera sea mayor.

En cuanto a diabetes, esta se presentó en un mayor número de organizaciones afiliadas (51) en comparación con la obesidad, pero con tendencias similares a las registradas en obesidad ya que las 10 las organizaciones volvieron a encabezar la lista de casos de diabetes sólo que con algunas variaciones en su orden (organizaciones H e I) juntas estas 10 agrupan al 93.12% con 2,748 casos (tabla 4.9).

Organización	Diagnósticos (2014 - mayo 2017)	% Proporcional
A	794	3.12%
B	641	4.54%
C	401	3.23%
D	300	3.05%
E	130	1.98%
F	93	2.30%
G	65	3.15%
I	58	2.94%
H	43	3.63%
J	34	8.59%

Tabla 4.9. *Organizaciones afiliadas con mayor número de diagnósticos de Diabetes. La proporción es un promedio del número de diagnósticos de todo el periodo de análisis entre el número de afiliados de cada organización.*

ORGANIZACIÓN	2014	2015	2016	Mayo de 2017	Total
A	401	154	150	89	794
B	380	109	112	40	641
C	178	77	57	89	401
D	159	68	43	30	300
E	66	32	16	16	130
F	34	14	30	15	93
G	36	14	9	6	65
I	21	15	2	5	58
H	30	16	7	5	43
J	23	7	4		34

Tabla 4.10. *Organizaciones afiliadas con mayor número de diagnósticos de Diabetes por año.*

En el 74.5% de estas organizaciones se presentaron 11 o menos casos de diabetes o un, por lo que se debe de vigilar a aquellas que encabezan la lista ya que son quienes concentra a mayor número de diagnosticados. Ahora la organización “J” es donde proporcionalmente al total de empleados que tuvo en el periodo analizado se presentó un mayor número de diagnósticos de diabetes, por otra parte, en la tabla 4.10 se observan los valores totales registrados por año en las 10 organizaciones con mayor número de diagnósticos de diabetes.

Hay que considerar que el total de afiliados que pertenece a una u otra organización es sumamente desigual, por esto se realizó el cálculo proporcional para poder tener una comparación con un mayor grado de objetividad y darle la importancia debida a cada organización y no descartar a aquellas con menor cantidad de miembros.

- **Difusión de contenido informativo en redes sociales**

Se realizó una serie de “recorridos” por las cuentas en redes sociales de la institución para determinar cómo y qué era lo que los afiliados y público general encontraban en estos medios que fuese relevante para el cuidado de su salud, también para determinar el grado de interacción que se tenía con las publicaciones (conteo de likes, reacciones, compartidos, comentarios).

La institución tiene presencia en Facebook con 8,691 amigos y en Twitter con 7,004 seguidores. Se decidió tomar las últimas 100 publicaciones en estos medios y clasificarlas según el fin de su contenido: Salud o No Salud (tabla 4.11).

Red Social	Salud	No Salud
Facebook	10%	90%
Twitter	9%	91%

Tabla 4.11. *Tipo de publicaciones en Facebook y Twitter institucional.*

Los resultados del tipo de publicaciones de la tabla 4.9 son casi idénticos ya que el mismo contenido por lo general se publica en ambos medios, resaltando la baja cantidad de material referente a temas de salud como eventos, medidas de prevención, alerta de riesgos o cualquier otro tipo de información relevante para el cuidado y bienestar del afiliado.

Posteriormente, se buscó profundizar en el contenido y tipo de interacción que se tuviera con publicaciones referentes a sobrepeso, obesidad y diabetes, en la tabla 4.12 se presentan 3 de estos ejemplos en conjunto con las cifras de interacción.

Medio	Likes	Compartido/ Retweet	Comentado	Contenido
Facebook	16	4	0	Recomendaciones para la activación física.
	23	25	1	Factores de riesgo y recomendaciones sobre diabetes.
	14	0	0	Enfermedades asociadas al sobrepeso, cómo evitarlas.
Twitter	29	34	0	Recomendaciones para la activación física.
	32	48	0	Factores de riesgo y recomendaciones sobre diabetes.
	29	43	0	Enfermedades asociadas al sobrepeso, cómo evitarlas.

Tabla 4.12. *Interacción con contenido sobre obesidad y diabetes.*

Las 3 publicaciones seleccionadas fueron publicadas en ambos medios por igual y en los mismos periodos de tiempo, destacando una muy baja interacción por parte de los afiliados y público en general. En Twitter es donde se observa una mayor respuesta para las publicaciones casi duplicando a las de Facebook.

En estos recorridos se destacó la gran importancia que se le da a la promoción de eventos y noticias mayormente relacionadas con temas administrativos, aunado a la baja proporción de publicaciones referentes a temas de salud y prevención, lo que de cierta manera podría explicar la baja interacción.

4.2. Etapa II. Segmentación y grupos de interés definidos

Basándose en los grupos detectados en el análisis descriptivo, como por ejemplo el género más propenso, las edades en donde se intensifica el número de diagnósticos, los tipos de enfermedad, entre otros, es que se trabajará en esta etapa.

4.2.1. Segmentación

La segmentación busca definir quiénes son propensos a padecer alguna enfermedad, quién ya está diagnosticado y quién está sano. Partiendo de lo anterior, la figura 4.9 muestra la primer gran segmentación que se realizará en la población afiliada.

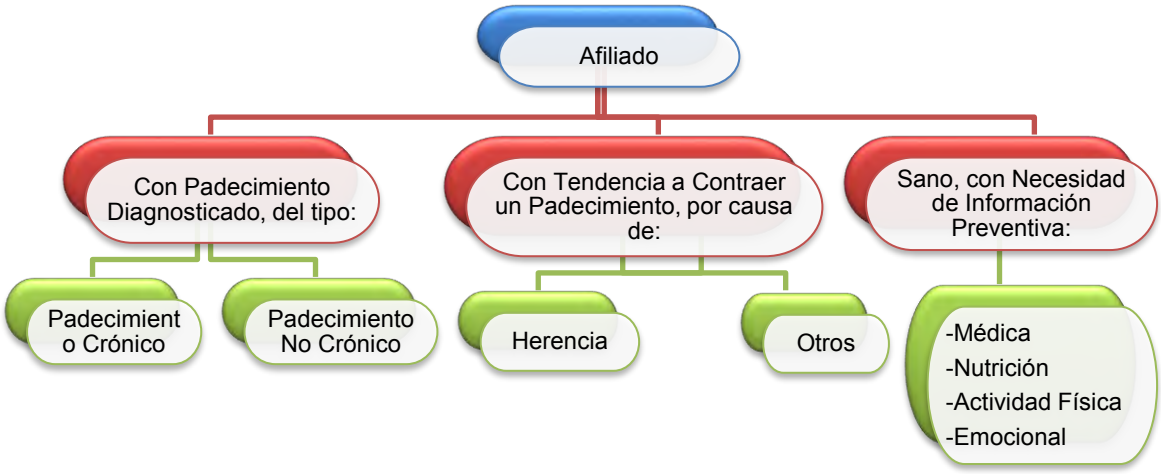


Figura 4.9. *Principales tipos de afiliados según su estado de salud.*

Como se presenta en la anterior figura, cada uno de los segmentos de la población requerirá de distinto tipo de información y recomendaciones, por lo que las características de cada grupo se deberán de tomar en cuenta, por ejemplo, dar recomendaciones sobre actividades que pueden realizar cotidianamente para mantenerse en forma a aquellos que están sanos; cómo evitar que los afiliados con tendencia a padecer obesidad o diabetes no lleguen a desarrollar ninguna de ellas, considerando su bajo, medio o alto grado de riesgo.

Analizando los registros familiares con antecedentes de obesidad y diabetes se podrá definir quiénes se encuentran en un bajo, medio o alto riesgo de sufrir sobrepeso, obesidad y diabetes por cuestión de herencia, mediante la aplicación del algoritmo de minería de datos, las variables y escalas establecidas en el análisis inicial.

4.2.2. Personalización mediante grupos de enfoque definidos

Se confirmó que el sexo es uno de los factores principales a considerar, por ello el primer gran determinante de la personalización es el género de la persona. Se debe de profundizar un poco más en el grado de especificidad si se quiere lograr una verdadera personalización del contenido informativo y recomendaciones que se difundan o conocimiento que se apliquen en las diversas campañas preventivas que se elaboren y pongan en marcha. Se crearon 5 distintos grupos de enfoque:

1. **Mujeres en general**: debido al número total de diagnósticos que estas han presentado en ambas enfermedades.
2. **Amas de casa**: considerando que son estas quienes se encargan de la alimentación de la familia, haciendo recomendaciones para que sean implementadas entre todos los miembros.
3. **Hombres adultos jóvenes:** a partir de esta edad en hombres se comienza a ver una tendencia al alza en el diagnóstico de sobrepeso y obesidad.
4. **Jóvenes**: en el caso de obesidad se está presentando con bastante regularidad entre adolescentes.

5. **Adultos mayores**: sobre todo en el caso de diabetes, se deberá de prestar especial cuidado a este grupo, pues se observó que a medida que aumenta la edad también lo hace el número de diagnósticos.

Es importante recalcar que, pese a la existencia de estos grupos definidos, se creará contenido para la sociedad en general, esto con el fin de lograr un mayor impacto y una mejora en las campañas preventivas.

Además del conocimiento que se espera obtener al aplicar minería de datos, se busca generar contenido informativo para su difusión en redes sociales, por lo que estos grupos de enfoque también servirán al momento del diseño y elaboración del contenido que se vaya a difundir, considerando características como, tipo de aplicación que utilizan para la consulta de redes sociales, tendencias de diseño y colores, entre otras.

Para encontrar segmentos de la población derechohabiente basados en la similitud de casos utilizaremos una técnica de minería de datos, en específico, un algoritmo de agrupamiento que automáticamente clasifique a los derechohabientes en distintas categorías de acuerdo a sus características demográficas y de salud, como explicamos a continuación.

4.3. Etapa III. Uso de Minería de Datos

Gracias al análisis previo se obtuvieron los principales grupos de enfoque para la generación de contenido informativo. Así como algunas de las clasificaciones a las que una persona puede ser sujeta según el grado de propensión a padecer o la evolución que presente, tanto para obesidad como para diabetes.

4.3.1. Preprocesamiento de datos para aplicación del algoritmo

Se contaba con acceso a tres distintas tablas de base de datos, una donde se encuentra la totalidad de afiliados pertenecientes a la institución de salud en la delegación de Hermosillo al momento de la consulta (febrero 2018), una más sólo con los diagnosticados con obesidad, en cualquiera de sus variantes, y por último una con los afiliados con algún tipo de diabetes, estas dos últimas con los registros desde el año 2014 y hasta mayo de 2017.

Se procedió a relacionar las dos tablas de diagnósticos con la del total de afiliados mediante sentencias SQL, con el fin de encontrar a todos aquellos miembros de una misma familia donde por lo menos alguno de ellos padeciera obesidad, diabetes o ambas, con lo que posteriormente se procedería a aplicar el algoritmo de minería seleccionado.

Una vez generada las consultas se obtuvieron una serie de resultados con todos aquellos registros que contaban con los criterios establecidos, llegando a la conclusión, como en el análisis descriptivo previo, de que aún y con las restricciones impuestas en las consultas se necesitaba estandarizar gran cantidad de registros y omitir columnas de las tablas originales que sólo sirven para control administrativo interno de la institución. En el caso del tipo de afiliado (hijo, esposa, madre...) se tenían 27 clasificaciones (anexo 7.13), de estas se redujo a sólo 6 para establecer la relación familiar entre los afiliados y encontrar indicios de herencia en las enfermedades.

Para poder definir el tipo de relación entre familiares registrados se hizo uso del campo “Familia” que agrupa a todos los miembros de una familia directa (padres, hijos, abuelos), pero de manera explícita la institución no registra el tipo de parentesco, por lo que en base al identificador familiar, el sexo y edad de los miembros de una familia se determinó de manera implícita si el rol que representaba la persona era el de un padre, madre, hijo, hija, esposo o esposa, dejando de lado términos como “jubilado”, “trabajador”, “incapacitado”, y demás, que no eran relevantes para la investigación. Por ejemplo, una familia con identificador “xxxx” se conforma por esposa, esposo, madre e hija, lo que significa que los esposos tienen una hija y la madre una nieta.

Cuando se comenzaron a agrupar a los afiliados por familias para determinar si existía herencia de enfermedades se tuvo que asegurar que los hijos estuvieran debidamente identificados por su género (hijo, hija) y no sólo como “HIJO”, ya que así es como lo maneja la institución.

Después del preprocesado general de los registros se obtuvieron dos tablas de familias de afiliados, una para obesidad y otra para diabetes, mismas que contenían las

siguientes columnas: Familia, Gpo. Edad, Edad, Sexo, Edad, si la persona estaba diagnosticada y el tipo de relación entre miembros de la familia directa, y por último si su padre o madre ha sido diagnosticado con la misma enfermedad (figura 4.10).

Colum Name	Data Type
Familia	int
Gpo Edad	int
Edad	int
Sexo	bit
Enfermedad	bit
Esposa	bit
Esposo	bit
Hija	bit
Hijo	bit
Madre	bit
Padre	bit
Madre con Diabetes	bit
Padre con Diabetes	bit

Figura 4.10. *Tabla "familias de afiliados con miembros diagnosticados". El valor binario 1 se utiliza para denotar si se trata de algún miembro de la familia, si padece la enfermedad o si cuenta con padres diagnosticados, mientras el valor 0 se presenta por default.*

Debido a que se aplicaría un algoritmo que trabaja sólo con variables numéricas se debió de realizar otro proceso para tratar propiamente las variables categóricas: Se crea un campo nuevo en la base de datos por cada uno de los posibles valores categóricos para una misma variable.

1. Se sustituye el valor categórico por una variable binaria (“0” o “1”) en cada uno de los campos de cada registro.

Por ejemplo, en el caso de “sexo”, una variable categórica con dos posibles valores: H y M, se suple por dos campos llamados Hombre y Mujer, cuyos valores pueden ser 1 o 0 de acuerdo con el sexo del afiliado. Este proceso permite al algoritmo de agrupamiento trabajar propiamente, pues clasifica las muestras de acuerdo con distancias euclidianas.

En los siguientes dos puntos se detalla el contenido de estas y su tamaño.

❖ Obesidad

De los 689 diagnósticos del periodo de análisis sólo se accedieron a 578, ya que 111 no contaban con registro vigente ante la institución al momento de realizar las consultas en las bases de datos, por lo que no se pudo obtener el número de identificación correspondiente que permitiera determinar si alguno de sus familiares estaba registrado o si padecía algún tipo de obesidad.

En total entre los diagnosticados y sus familiares se obtuvieron 1,610 registros de los cuales 1,032 pertenecen a los familiares y el resto a los diagnosticados, agrupados en 556 familias donde al menos uno de sus miembros padece obesidad, todas estas fueron codificadas para la aplicación del algoritmo (figura 4.11).

FAMILIA	EDAD	GPO EDAD	SEXO	OBESIDAD	ESPOSA	ESPOSO	HIJA	HIJO	MADRE	PADRE	MADRE CON OBESIDAD	PADRE CON OBESIDAD
1	73	13	1	1	1	0	0	0	0	0	0	0
1	101	13	0	0	0	1	0	0	0	0	0	0
2	57	11	1	1	0	0	0	0	1	0	0	0
3	44	8	1	0	1	0	0	0	0	0	0	0
3	47	9	0	1	0	1	0	0	0	0	0	0
3	13	2	0	0	0	0	0	1	0	0	0	1

Figura 4.11. *Ejemplo de 3 familias con algún miembro diagnosticado con obesidad.*

La familia 1 se compone de un matrimonio donde la mujer fue diagnosticada; la 2 sólo cuenta con un miembro; la 3 consta de un matrimonio con un hijo adolescente.

Diabetes

De los 2,748 diagnósticos de diabetes del periodo de análisis sólo se accedieron a 1,508 ya que 1,240 no contaban con registro vigente ante la institución al momento de realizar las consultas en las bases de datos, por lo que no se pudo obtener el número de identificación correspondiente que permitiera determinar si alguno de sus familiares estaba registrado o si padecía algún tipo de diabetes.

En total se contabilizaron 2,974 registros de diagnosticados con algún tipo de diabetes incluyendo a sus familiares directos, donde 1,736 registros pertenecen a los familiares

y el resto a los diagnosticados, todos agrupados en 1,376 familias diferentes. Tras la codificación se obtuvo una tabla como la mostrada en la figura 4.12.

FAMILIA	GPO EDAD	EDAD	SEXO	ENFERMEDAD	ESPOSA	ESPOSO	HIJA	HIJO	MADRE	PADRE	MADRE CON DIABETES	PADRE CON DIABETES
1	9	46	1	1	1	0	0	0	0	0	0	0
1	4	20	0	0	0	0	0	1	0	0	1	0
1	2	14	1	0	0	0	1	0	0	0	1	0
2	13	75	0	1	0	1	0	0	0	0	0	0
2	9	49	1	0	1	0	0	0	0	0	0	0
3	13	77	0	1	0	1	0	0	0	0	0	0
3	13	74	1	0	1	0	0	0	0	0	0	0

Figura 4.12. *Ejemplo de 3 familias con algún miembro diagnosticado con diabetes.*

La diferencia de diagnosticados entre el periodo de análisis y la consulta a la base de datos de la población general se puede deber a la rotación de afiliados.

4.3.2. Selección y aplicación de algoritmo(s) de minería de datos

Considerando el tipo de datos y las variables elegidas, se buscó el mejor algoritmo para la obtención de información útil, también se tomaron en cuenta la disponibilidad y conocimiento en el uso y programación, tanto del software como del algoritmo.

❖ Selección de algoritmos

Se requiere un algoritmo que permita describir el comportamiento de la muestra de forma automatizada y permita la exploración de los datos, mediante la generación de grupos que posteriormente puedan ser analizados con mayor profundidad por parte del investigador para describir a sus miembros y características. La tabla 4.13 presenta los algoritmos elegidos para su evaluación, tanto supervisados como no supervisados.

Algoritmo	Características	Ventajas	Desventajas
Redes Neuronales (Suca *et al.*, 2016)	No supervisado, predictivo, clasificación.	▪ Mejora su predicción con cada nueva corrida. ▪ Subdivide procesos complejos. ▪ Aprendizaje autónomo.	◦ Su funcionamiento es de caja negra. ◦ Difícil detectar cuando se ha cometido un error de programación.
K-NN (Shahi y Kaur 2013)	No supervisado, predictivo, agrupación.	▪ No paramétrico (no se necesita conocer la distribución de datos). ▪ De los más utilizados en reconocimiento de patrones.	◦ Suele ser propenso a datos atípicos. ◦ Tardado con abundantes datos.
K-Medias (García y Gómez 2006)	No supervisado, descriptivo, agrupación.	▪ Fácil aplicación. ▪ Representación gráfica. ▪ El de mayor uso para agrupar elementos. ▪ Puede trabajar con pocos datos. ▪ Se pueden añadir restricciones.	◦ Necesita sucesivas iteraciones para alcanzar el mejor resultado. ◦ Se necesita determinar "K" arbitrariamente. ◦ Suele ser propenso a datos atípicos.
Regresión (Beltrán 2003)	Supervisado, Predictivo.	▪ Genera una función matemática utilizable. ▪ Ampliamente utilizados para determinar probabilidades de un determinado diagnóstico.	◦ Calcula valores continuos, requiere utilizar otros algoritmos para valores discretos. ◦ Múltiples requisitos estadísticos al aplicar.

Tabla 4.13. *Tabla comparativa de algoritmos de minería de datos.*

Para la elección final del algoritmo se consideraron tres principales factores, el primero, como se mencionó, que haya demostrado ser de utilidad en el área de la salud, segundo, la naturaleza del problema y de datos, y tercero, que fuera un algoritmo de no muy alta complejidad de programación e interpretación debido al tiempo que se disponía para su implementación.

K-Means (K-Medias), desarrollado en 1967 por MacQueen J., es el algoritmo de agrupación que más se utiliza debido a que requiere menor cantidad de tiempo para su aplicación y es muy eficaz (García y Gómez 2006), las posibles desventajas pueden ser controladas con un buen preprocesamiento de datos y la correcta definición de variables, por lo que fue la primera elección para su aplicación.

❖ Elección del software

Para la elección del software se consideraron tres principales aspectos: accesibilidad, facilidad de uso y capacidad. Se optó por Google Colaboratory ® (figura 4.13), pues el lenguaje de programación con el que trabaja es de alto nivel y puede ser aprendido en un tiempo razonable.

Figura 4.13. *Logo de Google Colaboratory ®.*

Google lo describe como: “Un proyecto de investigación creado para divulgar contenido de investigación y formación sobre el aprendizaje automático. Es un entorno de Jupyter Notebook, el cual es una aplicación web de código abierto que permite crear y compartir documentos que contienen código vivo (ejecutable, editable), ecuaciones, visualizaciones y texto narrativo, no requiere configuración y se ejecuta completamente en la nube. Los cuadernos de Colaboratory se almacenan en Google Drive y pueden ser compartidos, además es un servicio gratuito”.

Las ventajas de este software son la portabilidad, ya que se ejecuta en la nube y puede ser compartido y accedido por cualquiera de los involucrados para trabajar de forma conjunta y simultánea, también evita la necesidad de contar con equipos de cómputo propios de alta gama pues utiliza el poder de procesamiento de Google, otra de las ventajas es su plataforma multilenguaje que soporta más de 40 lenguajes de programación (Python, R, Julia, Scala, entre otros).

Se tuvo que considerar que es un software para trabajo con valores numéricos, por lo que se debieron tomar ciertas consideraciones para denominar a aquellas variables de tipo cualitativo o alfabético y asignarles una codificación acorde, creando diccionarios para poder interpretar los resultados después del procesado.

❖ Aplicación de algoritmos

Se realizó un análisis exploratorio para encontrar patrones significativos o interesantes.

➢ Análisis exploratorio de datos

Se calcularon los valores totales y relativos de las tablas de familias obtenidas del preprocesado. Para esto sólo se consideraron a los miembros pertenecientes a las familias con a lo menos un diagnosticado y con vigencia ante la institución al momento de la consulta de registros, tomando la suma de todos los integrantes de todas las familias como el total para calcular valores relativos.

• Obesidad

Primeramente, se determinó el número de diagnosticados con obesidad en cualquiera de sus variantes y el tipo de miembro de familia que era, al igual que el porcentaje relativo tomando en cuenta el total de cada tipo de miembro de las familias con a lo menos un miembro diagnosticado (tabla 4.14).

	ESPOSA	ESPOSO	HIJA	HIJO	MADRE	PADRE
TOTAL	254	133	71	60	51	9
%	54.98%	43.18%	18.25%	15.35%	100.00%	100.00%

Tabla 4.14. *Valores totales y relativos de diagnosticados con obesidad.*

Según los registros, todas aquellas personas registradas como “Padres” (con una media de 43 años) y “Madres” (con una media de 52 años) fueron diagnosticadas con obesidad, seguidas en número por las “Esposas” y “Esposos”.

• Diabetes

Se separaron todos los derechohabientes de la institución según el tipo de integrante de familia que era, posteriormente se calculó la proporción dividiendo el total de diagnosticados “ESPOSA” con cualquier tipo de diabetes entre el total de afiliados del tipo “ESPOSA” de la institución, proceso que se repitió con los otros 5 tipos de integrante (Esposo, Hija, Hijo, Madre, Padre) (tabla 4.15).

	ESPOSA	ESPOSO	HIJA	HIJO	MADRE	PADRE
TOTAL	715	746	8	6	22	11
%	62.39%	71.46%	2.31%	1.65%	42.31%	47.83%

Tabla 4.15. *Valores totales y relativos de diagnosticados con diabetes tipo 1 y 2*.

Con este análisis exploratorio se logró determinar qué tipo de integrante de familia es el que se está viendo mayormente afectado tanto en el padecimiento de obesidad como de diabetes, así como la detección de tendencias y comportamientos interesantes sobre la relación familiar y la posible herencia en algunos de los casos, ahora se deberá de comprobar mediante el algoritmo K-Means.

➢ Aplicación del algoritmo K-Means

Los archivos con las tablas codificadas y en formato ".csv" fueron almacenadas en Google Drive (Drive) para poder cargarlas en el software seleccionado para la aplicación del algoritmo K-Means, teniendo en cuenta que una de las ventajas de utilizar Colaboratory es la modularidad del código, donde se puede ir avanzando paso a paso y observar los resultados de cada sentencia.

Lo primero que se realizó fue la binarización de aquellas variables categóricas que así lo necesitaran, cuidando siempre la integridad y calidad de los datos.

Posteriormente se importó la librería Science Kit de Python, en particular el algoritmo K-Means, y se definieron distintos números de clusters iniciales (valores de k para ver cuál era el comportamiento de los resultados y la densidad de los clusters). Cada vez que se ejecutaba el código con distintos "K" iniciales se revisaban los resultados para evaluarlos y ver las diferencias y similitudes, así como la densidad y la amplitud de los rangos de cada grupo, por lo que se decidió que el mejor valor para "K" sería de 5 ya que se ajustaba de mejor forma a los distintos grupos de enfoque definidos que también se obtuvieron en el análisis descriptivo.

Los resultados que se obtienen son una serie de valores para cada uno de los campos en particular. Por ejemplo, en el caso de los 5 clusters que se formaron en la figura

4.14, se tienen 5 renglones que representan a cada clúster con los valores medios de cada columna de las distintas variables.

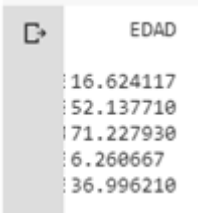
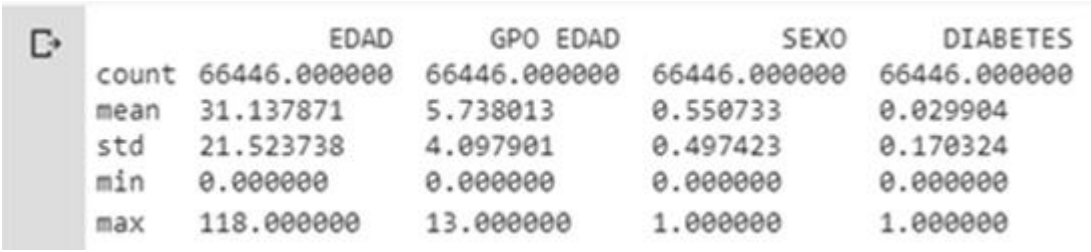


Figura 4.14. *Ejemplo del resultado de la aplicación de K-Means con 5 Clusters. En este campo en particular se observa la edad media de las personas que conforman cada uno de los 5 clusters.*

Si se desea visualizar con mayor detalle cada uno de los clusters se pueden saber la media, la desviación estándar, los máximos y mínimos, así como los cuartiles donde se distribuyen los elementos de cada clúster (Figura 4.15).



	EDAD	GPO EDAD	SEXO	DIABETES
count	66446.000000	66446.000000	66446.000000	66446.000000
mean	31.137871	5.738013	0.550733	0.029904
std	21.523738	4.097901	0.497423	0.170324
min	0.000000	0.000000	0.000000	0.000000
max	118.000000	13.000000	1.000000	1.000000

Figura 4.15. *Ejemplo del resultado detallado de un Clúster. Al seleccionar un cluster en particular se pueden obtener medidas de tendencia central y de distribución de los registros que lo conforman, esto por cada una de las variables representadas por las columnas de la tabla.*

Con un comando de código se probaron distintas combinaciones de codificación de las tablas de base de datos, así como de número de clusters iniciales.

Además de la información encontrada, se creó un manual para describir a detalle el proceso a seguir para poder aplicar el algoritmo de agrupación sobre futuras muestras de datos para dar continuidad al proyecto.

La ventaja de utilizar un algoritmo de aprendizaje no supervisado como K-Means es que se puede solicitar la agrupación de pacientes con un valor de k más alto, lo que dará como resultado grupos más compactos, pero con una mayor similitud entre los miembros.

4.3.3. Ubicación geográfica de los sectores de interés

Para la georreferenciación se utilizó “Google Fusion Tables” (figura 4.16) debido a la sencillez de utilización y las distintas formas de visualización que ofrece, además es un servicio gratuito de Google y es compatible con archivos en formato “.csv”, siendo este formato el de origen de los registros que se analizarían.

Figura 4.16. *Logo Google Fusion Tables ®.*

Basándose en los registros para georreferenciar que se mostró en la Etapa 1 (tabla 4.2), se obtuvo la siguiente información durante el preprocesado por enfermedad:

- **Obesidad**

De los 689 registros de diagnóstico, 17 no contaban con datos en el campo “colonia” o no era posible ubicarla para su geográficamente, dejando 672 registros para su mapeo. Con esto se encontraron 274 colonias que agrupan a los diagnosticados con obesidad en el Municipio de Hermosillo, en su mayoría sólo fue registrado un caso por colonia, por lo que no se puede hablar de “puntos críticos” al momento de realizar el análisis, el 96.30% de las colonias registraron 9 o menos casos. Las 10 colonias con mayor número de casos de obesidad registraron 141 o un 21.04% del total y en promedio tuvieron 14 diagnósticos (tabla 4.16).

COLONIA	CASOS	COLONIA	CASOS
Colonia Puerta Real	26	Colonia Las Lomas	11
Colonia Sahuaro	20	Colonia Villa De Seris	11
Colonia Altares	16	Colonia Villa Sonora	11
Colonia Palo Verde Sur	14	Colonia Balderrama	10
Colonia Solidaridad	12	Colonia Las Granjas	10

Tabla 4.16. *Principales colonias con alta incidencia de obesidad en Hermosillo.*

Para corroborar si la dispersión que se presentó en los datos también estaba presente geográficamente, es decir, que las colonias con mayor incidencia de diagnósticos de obesidad no se encontraban cerca unas de otras, se procedió a utilizar el software de georreferenciación obteniendo que la mayoría de los casos se ubican en las periferias, en especial en el norponiente y sur de la ciudad (anexo 7.14.1.).

Cabe aclarar que existe la posibilidad de que pacientes con obesidad no hayan sido formalmente diagnosticados. Esto puede resolverse con los datos de peso y talla que la institución ha estado recolectando en el último año.

- ## Diabetes

Considerando únicamente los registros de diagnóstico (2,748), 164 no contaban con una ubicación (nombre de la colonia) o no era posible ubicarla para su georreferenciación, dejando 2,584 registros para su mapeo. Después de este primer filtro, se procedió con la estandarización del nombre de las ubicaciones pues presentaban ciertas discrepancias que hacían que una misma colonia tuviera hasta tres distintas maneras para referirse a ella, con esto se encontraron 323 colonias con registros de personas con diabetes por parte de la institución en el Municipio de Hermosillo. Se observó que en 10 de las 323 colonias se registraba en conjunto el 23.45% del total de casos de diabetes diagnósticos (tabla 4.17), haciéndolas críticas para la implementación de medidas preventivas.

COLONIA	CASOS	COLONIA	CASOS
Colonia Ley 57	81	Colonia Jesus Garcia	58
Colonia Balderrama	80	Colonia Villa De Seris	51
Colonia Olivares	78	Colonia Palo Verde	46
Colonia Sahuaro	63	Colonia Jardines	45
Colonia Villa Sonora	60	Colonia San Benito	44

Tabla 4.17. *Principales colonias con alta incidencia de diabetes en Hermosillo.*

Posterior a esto se procedió a la ubicación de las colonias en el mapa de la ciudad para determinar qué sectores eran donde se estaban presentando esta situación de alta incidencia de diabetes (anexo 7.14.2.), un factor que se detectó es que se trata de algunas de las colonias con mayor antigüedad en la ciudad de la parte centro-norte, por ende, con más población adulta mayor, siendo este un factor clave para el padecimiento de la diabetes, también de la parte sur como Villa de Seris, Palo Verde y Cuauhtémoc.

Cabe aclarar que existe la posibilidad de que pacientes con obesidad o diabetes no hayan sido formalmente diagnosticados.

4.4. Etapa IV. Obtener contenido informativo y generación de reportes

Tras el procesado de los registros, se interpretaron los resultados obtenidos para su posterior utilización en contenido informativo, también para su entrega al personal del área de medicina preventiva mediante una serie de reportes.

4.4.1. Interpretación de resultados del procesamiento de datos

Tras la aplicación del algoritmo de minería de datos K-Means para los datos de diagnosticados y sus familias, se obtuvo una tabla con los 5 clusters resultantes para obesidad y otra para diabetes, mismas que se muestran en los siguientes 2 puntos.

- **Obesidad**

Lo primero que se observó fueron los rangos de edad de los integrantes de cada clúster, estos comenzaron a partir de los 7 años y hasta los 61 en promedio. En el

primer cluster la edad media fue de 17 años en su mayoría del grupo de edad de 13-27, en el segundo de 37 años en su mayoría de los rangos de 2843, en el tercero fue de 61 años de los rangos de 56-101, en el cuarto fue de 49 años del rango 44-55 y en el quinto clúster fue de 7 años de los rangos 1-12.

Con esto se podrá crear material y contenido informativo que se enfoque principalmente a adolescentes y niños, así como a adultos de mediana edad, sin descartar a los adultos mayores. El clúster con mayor densidad fue el número 1 con 416 integrantes.

Posteriormente se analizaron los resultados con otra de las variables principales, el género, considerando que los hombres fueron identificados como “0” y las mujeres como “1”, si los valores medios de la columna “SEXO” supera el 0.5 significa que predominan las mujeres en ese determinado clúster; y por último se observó el tipo de afiliado del diagnosticado (hijo, madre, etc.) y si sus padres presentaban la misma enfermedad, para ello se interpretaron las cifras obtenida en cada uno de los clusters.

Del análisis de género destaca que tres de los clusters se conforman ligeramente por un mayor número de mujeres que de hombres. En cuanto a la herencia, en el cluster 1 se tiene 44.71% de afiliados con madres diagnosticadas y con 17.06% de padres diagnosticados, al igual en el cluster 5 se aprecia que 48.92% de los niños tienen una madre diagnosticada y 23.65% cuentan con un padre diagnosticado con obesidad, ambos clusters conformados por menores de edad en promedio.

Cluster	Integrantes	% Diagnóstico	Edad media	Edad Min/Max	Género
1	416	23.55%	17	13/27	50.24% mujeres
2	340	57.94%	37	28/43	66.17% mujeres
3	150	50.66%	61	56/101	49.33% mujeres
4	332	49,69%	49	44/55	63.55% mujeres
5	372	11.29%	7	0/12	49.46% mujeres

Tabla 4.18. *Resumen de resultados de los clusters de obesidad. El “% Diagnóstico” representa la proporción de diagnosticados del total de integrantes del cluster, “Edad media” es el promedio de edad de los integrantes del cluster, “Género” es la proporción que predomina entre los integrantes de cada cluster según su sexo.*

En la tabla 4.18 se muestra el resumen de lo encontrado tras la aplicación del algoritmo en los diagnosticados con obesidad y sus familias, destacando que en los clusters 2 y 3 más de la mitad de los integrantes fueron diagnosticados.

- **Diabetes**

Para la diabetes se siguió el mismo procedimiento utilizado en la interpretación de los resultados de los clusters de obesidad. Lo primero fue observar los rangos de edad de los distintos clusters, estos comenzaron en los 13 años (mayores que en obesidad) y hasta los 79 años en promedio. Al observar los 5 clusters obtenidos, en el primero la edad media de los integrantes fue de 55 años, en el segundo de 13 años, en el tercero fue de 65 años, en el cuarto fue de 41 años y en el quinto clúster fue de 79 años, claramente las edades promedio corresponde a personas adultas o adultas mayores tal y como se espera por la correlación entre la edad y el diagnóstico de diabetes.

El material propuesto se puede enfocar en a adultos y adultos mayores pues es la población más afectada en la institución, poniendo especial cuidado en los niños ya que están comenzando a ser diagnosticados a una temprana edad. El clúster con mayor densidad fue el número 1 con 823 integrantes y 65.61% de ellos diagnosticados con diabetes y con una edad media de 55 años.

Del análisis de género pasó lo opuesto que con la obesidad, aquí tres de los clusters se conforman ligeramente por un mayor número de hombres que de mujeres. En cuanto a la herencia, se presentaron clusters que no se conformaban por menores de edad o hijos e hijas donde los afiliados tenían padres y madres con diabetes. El clúster 1 tiene 1.45% de afiliados con madres diagnosticadas, en el clúster 2 se aprecia que 45.16% de los niños tienen una madre diagnosticada y 50.07% cuentan con un padre diagnosticado con diabetes, el clúster 3 tiene un 0.01% con madres diagnosticadas y el clúster 4 0.05% también con madres con diabetes, de estos cuatro clusters sólo en uno se tenía referencia de padre con diabetes.

Cluster	Integrantes	% Diagnóstico	Edad media	Edad Min/Max	Género
1	823	65.61%	55	49/60	54.92% mujeres
2	713	2.10%	13	0/27	49.08% mujeres
3	640	67.34%	65	61/72	46.56% mujeres
4	510	64.70%	41	28/48	59.41% mujeres
5	288	66.66%	79	73/109	45.13% mujeres

Tabla 4.19. *Resumen de resultados de los clusters de diabetes. El "% Diagnóstico" representa la proporción de diagnosticados del total de integrantes del cluster, "Edad media" es el promedio de edad de los integrantes del cluster, "Género" es la proporción que predomina entre los integrantes de cada cluster según su sexo.*

En la tabla 4.19 se muestra el resumen de lo encontrado tras la aplicación del algoritmo en los diagnosticados con diabetes y sus familiares, destacando que en los clusters 1, 3, 4 y 5 alrededor del 65% fueron diagnosticados.

Para cada cluster se recomienda la división por género y diagnóstico, esto es, dividir los grupos por diabéticos y no diabéticos, hombres y mujeres.

Además de lo descrito en los dos puntos anteriores, se analizaron otras medidas de tendencia central, máximos/mínimos y los cuartiles en que se distribuían los clusters.

4.4.2. Modelo de generación de reportes

Se presentaron ante las autoridades de la institución una serie de reportes, donde de manera resumida y sumamente visual (utilizando gráficas), se encontraba lo más destacado del análisis descriptivo de registros, así como los resultados de la aplicación del algoritmo K-Means donde se detectaron los grupos de mayor vulnerabilidad y las características compartidas por aquellos afiliados con los que se debiera tener mayor atención, evitando un futuro diagnóstico de obesidad o diabetes.

Además, se profundizó en el tema de la diabetes, por petición de la institución, generando una publicación sobre el estado de esta enfermedad en Hermosillo, en específico en la población afiliada, con ello se espera una mayor difusión de este proyecto y de la actual situación que la diabetes está planteando en esta ciudad. Por razones de confidencialidad los reportes no pudieron ser publicados en este documento.

4.5. Etapa V. Difusión y evaluación del contenido generado

Con la finalidad de dar a conocer los hallazgos realizados sobre la población afiliada de la institución en la ciudad de Hermosillo, así como utilizar la información y conocimiento generado, se necesitaba de una manera práctica para hacer llegar este contenido a la mayor cantidad de personas posibles, independientemente del personal encargado de la toma de decisiones quienes fueron informados mediante reportes. La propuesta fue la utilización de redes sociales institucionales y medios electrónicos propios de esta, tomando en cuenta criterios de diseño y normativas para la publicación de recomendaciones sobre salud.

4.5.1. Propuesta y ajuste del contenido informativo para su utilización en campañas preventivas a través de redes sociales

Se tomaron en cuenta recomendaciones de diseño y aspectos legales para respaldar lo que se publique y hacerlo para beneficio de los afiliados y la institución al mejorar sus campañas de medicina preventiva.

- **Consideraciones Legales**

La Norma Oficial Mexicana NOM-008-SSA3-2010 (Secretaría de Salud Pública, 2010a) establece una serie de regulaciones cuando se realizará publicidad sobre el tratamiento y control de enfermedades, en particular de sobrepeso y obesidad.

Algunos de los criterios de la NOM-008-SSA3-2010 son:

- ✓ No anunciar la curación definitiva.
- ✓ No hacer referencia a tratamientos en los que no se distinga un tratamiento en particular.
- ✓ No promover la utilización de medicamentos secretos o fraccionados.
- ✓ No referirse a insumos o tratamientos que no estén respaldados científicamente en investigación clínica.
- ✓ No inducir la automedicación.

- ✓ La publicidad deberá estar orientada a inducir al paciente a que acuda con un médico, nutriólogo o psicólogo para que se determine la causa del problema y prescriba el tratamiento adecuado; cualquiera de los tres profesionales mencionados, podrán anunciarse y publicitarse en el tratamiento del sobrepeso y la obesidad, según su formación, materia y área de intervención.

En el caso de la diabetes, también se seguirán estas recomendaciones.

➢ Consideraciones de Diseño y Formato

El diseño del contenido para RS en la salud debe de hacerse pensando en que sólo se cuenta con un par de segundos para lograr llamar la atención de usuario, por lo que ser conciso y breve en extensión de palabras es necesario para que los mensajes y recomendaciones que se realicen tengan una mejor aceptación, sumándole que si son de carácter positivo y alientan a realizar alguna acción en lugar de sólo hacer notar que se tiene un problema.

El aspecto visual debe de desarrollarse según el tipo de contenido y el público a quien se quiere llegar, de hecho, si el mensaje puede transmitirse sin la necesidad de texto sería preferible, dejando un enlace a alguna fuente de mayor información; por otra parte, la contextualización del contenido puede necesitar cambios según la plataforma que se haya seleccionado (Puhl, Luedicke and Lee Peterson, 2013; Kornfield *et al.*, 2015; Park *et al.*, 2017; Ramanadhan *et al.*, 2017).

➢ Material propuesto para redes sociales

Tomando en cuenta los grupos que el análisis inicial planeó en la segmentación, se decidió diseñar cada una de las publicaciones que se realizaran con un código de color en particular para diferenciarlas. También se plantearon dos líneas de trabajo:

La primera con recomendaciones generales como recordatorios de chequeo o citas con el médico, opciones para una mejor alimentación, entre otras (figura 4.17 y 4.18).

Figura 4.17. *Recomendaciones para una mejor alimentación.*

La idea central de la figura 4.17 es la creación de un marco reutilizable en el cual sólo se deba de cambiar el texto y estarlo actualizando cotidianamente con recetas para los niños y jóvenes en edad escolar, en este mismo sentido se diseñó un marco parecido, pero para la publicación de recetas saludables orientado a las amas de casa.

Figura 4.18. *Recomendaciones generales para la prevención de diabetes.*

Se busca incrementar la conciencia sobre el padecimiento de la obesidad y diabetes en la comunidad, alentando siempre a que lleven un monitoreo y control adecuado si ya padecen alguna de estas enfermedades o cuidándose para evitarlas.

La segunda línea de trabajo busca difundir los hallazgos de mayor relevancia que el análisis de los registros de la institución en Hermosillo tuvo y separando por género con ilustraciones alusivas (figura 4.19 y 4.20), en los siguientes ejemplos se utilizaron estadísticas de la Encuesta Nacional de Salud sólo para fines ilustrativos.

Los colores institucionales fueron incorporados en el diseño del material de la segunda línea de trabajo para dar un mayor sentido de pertenencia y vincularlo con la población afiliada en particular, difundiéndolo entre sus allegados.

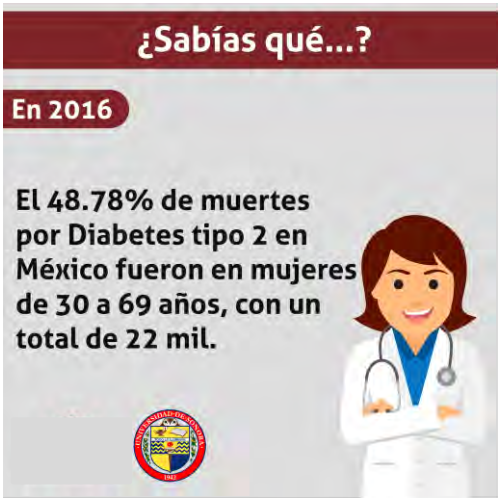


Figura 4.19. *Datos relevantes enfocados en el género femenino.*

Ambos diseños (figuras 4.19 y 4.20) pueden ser utilizados como marcos de trabajo para publicar el contenido informativo aprobado para su publicación sólo al cambiar el texto interno o agregar algún elemento extra ocasional, lo que se busca es la simplicidad para su actualización por el personal de comunicación.

Figura 4.20. *Datos relevantes enfocados en el género masculino.*

Las propuestas de material, ya aprobadas, están acompañadas de una descripción una vez publicadas en las redes sociales con una explicación y se alienta a visitar a su médico, en algunos casos acompañados de un link web para más información.

4.5.2. Evaluación y retroalimentación

Por otra parte, existen aspectos cuantitativos que pueden ser medidos mediante índices, porcentajes y estadísticas para evaluar el impacto que se tuvo después de la implementación, a la vez que contempla también cambios que no pueden ser medidos de esta manera, por ejemplo, cambios en los criterios para desarrollar un procedimiento, mejoras en procesos, entre otros, algunos aspectos cualitativos por evaluar son:

- Políticas para los procesos de atención y medicina preventiva en la institución. Por ejemplo, a raíz de la realización de este proyecto, se comenzó a medir el peso y estatura de forma obligatoria en cada consulta médica, antes se dejaba a criterio de la enfermera o al médico en turno.
- Mayor disposición de la institución para integrar nuevas tecnologías en el procesamiento de sus registros. Cada vez ven de mejor manera el hecho de que los registros con los que se cuentan pueden ayudar y mejorar la forma en que se

toman decisiones y se implementan nuevas campañas preventivas, entre otros aspectos.

- Se creó un manual para la utilización del código para la aplicación del algoritmo de minería de datos K-Means en Google Colaboratory, con el cual se podrá brindar capacitación al personal encargado del análisis de los registros en la institución. Dejándolo sumamente visual y con instrucciones cortas y exactas para evitar complicaciones.
- La institución ha lanzado una nueva aplicación (app) que incluye una sección de “Noticias” (figura 4.21), en esta sección es donde el material y contenido informativo generado por este proyecto será difundido, además de las RS.

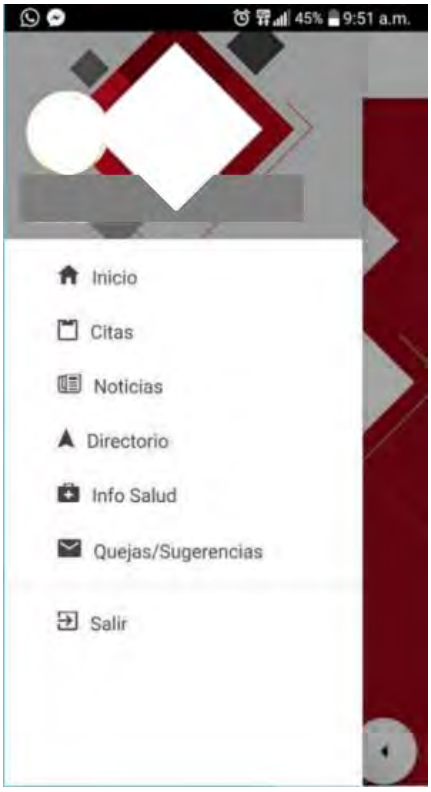


Figura 4.21. *Captura de pantalla de la aplicación móvil institucional.*

Un aspecto importante dentro de la evaluación es el conocimiento de la situación actual sobre las condiciones de la población de afiliados con respecto a diabetes y obesidad. La generación de reportes basados en el análisis exploratorio y el modelo de

agrupación (K-Means) arrojaron información hasta ahora desconocida por la institución y que podrá seguir siendo descubierta y actualizada mediante esta metodología. Dicha información será crucial para la toma de decisiones cuyo impacto podrá ser medible en los próximos 5 años con la continuación de este proyecto multidisciplinario.

A lo largo del desarrollo de proyecto se estuvieron recibiendo recomendaciones de parte de la institución, en específico se buscó dar mayor relevancia al tema de diabetes por su actual impacto en la población institucional y de México en general, siempre enfocándose en la obtención de información útil para la prevención.

5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

Tras la implementación de la metodología propuestas y de las experiencias obtenidas durante el proceso, se llegó a una serie de conclusiones y recomendaciones que buscan favorecer de la mejor manera posible los esfuerzos preventivos realizados en esta institución de salud.

5.1. Conclusiones

Recapitulando un poco, este proyecto forma parte de uno con mayor alcance que busca, con la ayuda de distintas disciplinas, aprovechar los recursos e información que una institución de salud pública del estado de Sonora posee y que esto se vea reflejado en mejoras en su proceso de toma de decisiones y en la atención de su población afiliada.

El objetivo principal de este trabajo fue el diseño de una metodología que aprovechara las bases de datos electrónicas (de consultas médicas) que la institución posee para mejorar en algún sentido sus estrategias de medicina preventiva enfocándose en obesidad y diabetes, lo cual se logró al desarrollar una metodología compuesta por 5 etapas donde se parte de un análisis inicial para una correcta segmentación de la población, posteriormente un procesado de registros con la ayuda de un algoritmo de minería de datos, finalizando con la generación de reportes para la institución y material con contenido informativo dirigido para su difusión masiva en redes sociales.

La ventaja de esta metodología es que puede ser aplicada para conocer el comportamiento de cualquier otra enfermedad, basta con realizar algunos ajustes en los parámetros de entrada en el algoritmo de minería de datos que se seleccione, a la vez que permite generar información de manera personalizada con sugerencias basadas en el análisis de los registros de la población local, por lo que se tendrá una mejor certeza que cuando se hace uso de información basada en la población de todo el país y fuentes internacionales donde existen distintos hábito alimenticios, factores

ambientales diferentes o simplemente la genética poblacional es otra, por lo que no se ve afectada de la misma manera que en otro lugar.

El trabajo presentó retos, al tratarse de las primeras aproximaciones con esta institución, en particular la propuesta y utilización de tecnologías de minería y análisis de datos, ya que se requiere asegurar la integridad y estandarización de la información que es capturada en sus sistemas, por lo que se tuvo que invertir aún más tiempo en el preprocesado de datos. Aunado a lo anterior, de cierta manera no se contemplaba utilizar los sistemas de consulta médica más allá de llevar un orden o administrar los recursos económicos que el atender a un afiliado conlleva, lo que está comenzando a cambiar gracias a la disposición y reciente apertura que la institución está teniendo al ver el potencial que proyectos como el que se ha realizado puede tener en beneficio de sus afiliados y por consecuencia en la forma en que esta enfrenta los nuevos retos que se presentan.

Con esta apertura mayor cantidad de conocimiento podrá ser generado y difundido, aunque se deben de seguir los lineamientos y protocolos internos para determinar qué puede ser compartido con la población o no, cómo y cuándo se hará, por lo que este proyecto seguirá en los próximos años, procurando siempre mejorar la toma de decisiones basándose en una eficiente forma de hacer llegar el conocimiento a quien lo requiera.

5.2. Recomendaciones

Los siguientes puntos surgen del trabajo realizado en la institución de salud, la experiencia obtenida y las oportunidades detectadas para su aprovechamiento:

- Se recomienda diversificar el contenido en redes sociales, enfocándose mayormente en temas relacionados con salud evitando temas de bajo interés para la población. También se debe alcanzar una cobertura mayor, sabiendo que aproximadamente 9 de cada 10 personas posee algún tipo de dispositivo para acceder a internet en México, se está desaprovechando este valioso medio, por ejemplo, si se considera a los 8,691 “amigos” registrados en la cuenta de Facebook

de la institución y suponiendo que todos ellos son afiliados, sólo se está llegando aproximadamente a un 10% de la población institucional total en Hermosillo. Las mejoras que se pueden realizar en esta área tienen un impacto verdadero en la forma en cómo las personas y la institución se comunican e interaccionan.

- Asegurarse de que los sistemas informáticos y software de registro de la institucionales cuenten con los candados o medidas de restricción que impidan la captura de información a criterio del personal en turno, ya que la calidad e integridad de esta no es la ideal, presenta un gran número de variaciones en campos donde se supone existe un estándar definido, también se aceptan campos vacíos tan importantes como el Diagnóstico, Dirección, y demás.
- De igual forma se recomienda brindar capacitación constante a las personas que se encargan de la captura de información, así como una correcta supervisión para asegurar que se esté realizando de manera satisfactoria esta tarea.
- Por último, llevar un registro familiar en donde esté plenamente identificado cuál es el rol de cada uno de los miembros, es decir, saber quién es hijo, hija, padre, madre, abuelo, u otro miembro que integre a la familia directa, ya que en los registros actuales sólo se categorizan para fines administrativos (pensionado, trabajador, incapacitado, entre otros) y no para el análisis en función de la salud y prevención de futuros riesgos donde influye la herencia.

Estas observaciones surgen debido a la excesiva cantidad de tiempo que se invirtió en esta investigación estandarizando y filtrando los registros para que se pudiese obtener información útil de estos, ya que actualmente están siendo capturados con información inexacta, incompleta, inválida o simplemente no se captura, evitando que sea aprovechada en su totalidad esta valiosa fuente de conocimiento para observar el potencial real y tangible en la institución, sin limitarse al entorno administrativo y de recursos económicos.

5.3. Trabajos futuros

La idea de utilizar la metodología propuesta sólo con obesidad y diabetes fue para probar su efectividad, el siguiente paso es utilizarla y adaptarla con todas aquellas enfermedades que puedan ser prevenidas y donde intervengan factores de riesgo como la herencia, el género o la edad. Para ello es importante que se capacite adecuadamente al personal que vaya a estar haciendo uso del equipamiento y del software especializado para el análisis de datos.

También se hay que recordar que uno de los objetivos era el de detectar a aquellos afiliados que se encontraran en un alto, medio o bajo riesgo de padecer alguna de las enfermedades objetivo, lo que no pudo ser alcanzado debido a la falta de información antropométrica y de química sanguínea de los afiliados diagnosticados, por lo que en trabajos futuros se espera poder concretar este importante objetivo.

Por otra parte, la mejora en la presencia en redes sociales es un punto importante para la institución. Un análisis completo sobre el proceso de elaboración, aprobación, publicación y seguimiento de campañas informativas y publicitarias dedicadas exclusivamente a temas de salud, tanto prevención como control y cuidados especiales o alertas epidemiológicas, donde se obtengan métricas para determinar si el mensaje llegó a quien se esperaba y fue de ayuda. El análisis debe llevar a una serie de sugerencias para el área encargada del manejo de redes sociales, así como de una propuesta de mejora con los pasos detallados a seguir para incrementar, primeramente, el número de personas (amigos, seguidores, entre otros), después la cantidad de interacción con las publicaciones, y por último la retroalimentación que se reciba.

5.4. Propuestas de estadía en el extranjero

Como parte de una estancia en la Universidad Politécnica de Cataluña durante el mes de junio de 2018, surgieron las siguientes propuestas para continuar trabajando en proyectos en la institución de salud pública.

Primeramente, una vez que se tuviera acceso a registros con mayor cantidad de atributos (estatura, peso, IMC, niveles de glucosa, y demás) se podrían utilizar otras herramientas de minería de datos como reglas de clasificación, para las cuales se debe de contar con información previa suficiente para “entrenar” al algoritmo seleccionado y poder así clasificar a los nuevos registros cada vez con un mejor porcentaje de acierto en su fiabilidad, esto ayudará a la mejor segmentación de la población institucional.

Otra de las propuestas fue la de utilizar algoritmos predictivos para poder detectar a aquellos que se encuentren en riesgo de padecer tanto obesidad como diabetes, y hacer proyecciones sobre la cantidad y tipo de personas que se verán afectadas en años por venir, inclusive determinar aquellos puntos emergentes de diagnósticos en la ciudad para monitorearlos con mayor detenimiento. También utilizar este tipo de algoritmo para hacer una comparación y ver si la herencia está influyendo en el diagnóstico de nuevos casos. Lo siguiente fue consultar una serie de trabajos realizados obteniendo las siguientes propuestas:

- Existen trabajos que se han realizado referentes a la utilización de técnicas de minería de datos es esta universidad, en específico el realizado por Gibert et al. (2016), que consiste en el desarrollo de una encuesta detallada que puede ser aplicada para mejorar el preprocesamiento de datos, detectar datos anormales, codificar, eliminar campos faltantes, así como una revisión de otras técnicas que se pueden utilizar para este fin, y observando que en la institución de salud existen inconvenientes con la calidad de los datos iniciales, este trabajo puede ser adaptado al tipo de datos de las consultas médicas y asegurar así la integridad y calidad de los datos de entrada, además se hacen alrededor de 28 recomendaciones finales que ayudarán a un correcto preprocesado de datos.
- Otro aspecto que se consideró fue que la metodología propuesta sería utilizada por personal de la institución que no cuenta con experiencia en el trabajo con algoritmos de minería, en específico con aquellos de agrupamiento (clustering), por lo que se propone una manera de cómo disminuir la brecha entre los resultados de

la aplicación del algoritmo y una efectiva toma de decisiones. Villanueva et al. (2013) en su trabajo propusieron una metodología que ayuda a definir el tipo de persona que deberá de encargarse de realizar la interpretación de los resultados obtenidos de manera gráfica, a la vez que proponen una manera de clasificar a individuos para una correcta asignación de dietas. Tomando lo anterior en consideración, se propone realizar una investigación entre los distintos miembros del personal de la institución de salud que utilizarán la metodología propuesta en esta tesis para que aquel con mayores capacidades sea quien realice la interpretación de los resultados para así generar el material y contenido informativo para difundir entre la población afiliada y los reportes a los jefes de departamento.

- También se propone desarrollar una prueba estadística para evaluar si la interpretación de los clusters obtenidos ha sido la apropiada, esto se vuelve especialmente importante cuando se está trabajando con algoritmos no supervisados e inclusive si se cuenta con la ayuda de expertos. La literatura establece que una de las formas para probar estadísticamente los clusters es mediante “índices”, los cuales evalúan desde un punto de vista estructural, pero la mejor manera es tomando en cuenta la significancia, es decir, darle es peso correcto a las variables correctas, fundamental cuando se cuenta con decenas de características descriptivas de los registros (Sevilla-Villanueva et al. 2014).
- La utilización del algoritmo G-Means como alternativa al utilizado en la metodología que se implementó en la institución de salud (K-Menas), permitirá evitaría o dejaría de lado la incertidumbre que se genera cuando se debe de seleccionar el número de clusters iniciales, pero con el mismo tipo de resultado para su interpretación post-procesado. Este algoritmo G-Menas comienza con una pequeña cantidad de centroides o “K´s” y va aumentando la cantidad de centros, en cada iteración del algoritmo cada centro inicial se divide en dos cuyos datos parecen no proceder de una distribución Gaussiana, en cada división de centroides se ejecuta K-Menas en todo el conjunto de datos y centros para ir refinando, se puede iniciar con k=1. La principal diferencia es que G-Menas toma decisiones en base a una prueba estadística sobre los datos de cada centroide para determinar si estos parecen ser

gaussianos, de ser así significa que deberán de estar agrupados en un mismo clúster, el proceso se repetirá “n” número de veces hasta encontrar el mejor número de grupos suponiendo una distribución gaussiana de los datos. Una de las ventajas de este algoritmo es que se auto prueba para evitar tomar malas decisiones sobre clúster con pocos datos (Hamerly y Elkan 2004; Oliveira et al. 2017).

- Por último, una alternativa para la elección de la técnica de minería de datos se aconseja para probar distintos tipos de algoritmos y elegir aquel que mejor se adapte a las necesidades de una investigación o datos en particular, tal y como lo proponen Gibert et al. (2006), con la ayuda de un mapa conceptual (figura 5.1) y las recomendaciones realizadas en su trabajo se tendrá una herramienta que podrá ser utilizada dentro de la institución de salud.

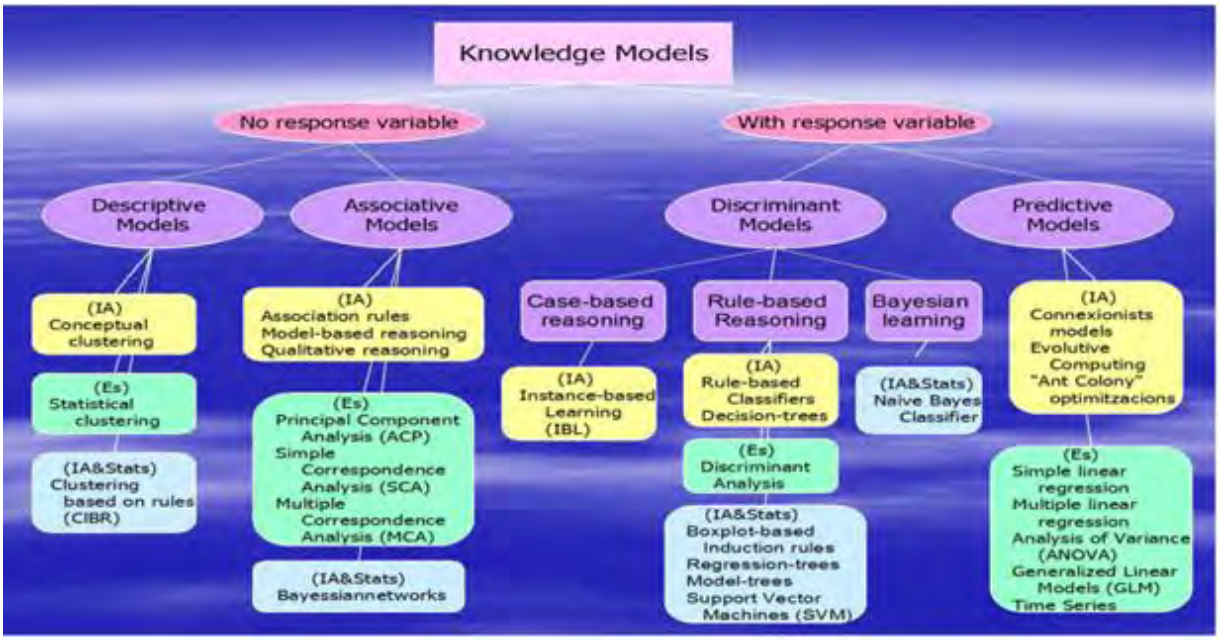


Figura 5.1. *Mapa conceptual de técnicas de minería de datos. De Gibert et al. (2006).*

Sin embargo, como ya se mencionó para la selección, la aplicación de la técnica de minería e interpretación de resultados siempre se deberá de considerar a aquella persona con mayor afinidad o conocimientos en el tema, aquella que se encuentre mayormente capacitada, para evitar lo más posible el error de tipo humano.

6. REFERENCIAS

70 Asamblea Mundial de la Salud, 2017. Informe de la Comisión para acabar con la obesidad infantil Informe de la Directora General. Organización Mundial de la Salud, 1, pp.1–42. Disponible en: http://apps.who.int/gb/ebwha/pdf_files/WHA69/A69_8-sp.pdf.

Abdar, M. et al., 2015. Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of Electrical and Computer Engineering (IJECE), 5(6), pp.1569–1576. Disponible en: http://iaesjournal.com/online/index.php/IJECE.

Agnelluti, C., 2014. Big Data: Seizing Opportunities, Preserving Values, Nova Science Publishers.

Aguilar Cordero, M.J. et al., 2014. Programas de actividad física para reducir sobrepeso y obesidad en niños y adolescentes; revisión sistemática. Nutr Hosp, 30(4), pp.727–740.

Al-hagery, M.A. et al., 2015. Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community. International Journal of Computer and Electronics Research, 4(6), pp.118–125.

Aljumah, A. y Siddiqui, M., 2016. Data Mining Perspective: Prognosis of Life Style on Hypertension and Diabetes. The International Arab Journal of Information Technology, (May), pp.93–99.

Álvarez Alva, R. y Kuri Morales, P.-A., 2012. Salud Pública y Medicina Preventiva 4a edición. C. A. Mendoza-Murillo, ed., Ciudad de México: El Manual Moderno.

Antonio, J. y Ávila, T., 2015. Brecha en los costos laborales debido a la obesidad de los trabajadores. Contaduría Universidad de Antioquia, 67, pp.21–44.

Arab, U., 2011. Geographie Information Systems (GIS) Application for Health: Case of Al Ain (UAE). Intimai of Geoinformatics., 7(1), pp.21–28.

Arias-Valencia, S., 2017. Epidemiology, health equity and social justice. Revista Facultad Nacional de Salud Pública, 35(2), pp.186–196. Disponible en: http://aprendeenlinea.udea.edu.co/revistas/index.php/fnsp/article/view/327006/20784990.

Arredondo, A. y De Icaza, E., 2011. Costos de la Diabetes en América Latina: Evidencias del Caso Mexicano. Value in Health, 14(5 SUPPL.), pp. S85–S88. Disponible en: http://dx.doi.org/10.1016/j.jval.2011.05.022.

Aschner, M.P. et al., 2016. Guía de práctica clínica para la prevención, diagnóstico, tratamiento y seguimiento de la diabetes mellitus tipo 2 en la población mayor de 18 años. Colombia Médica, 47(2), pp.109–131.

Asghar, O. et al., 2013. Diabetes, obesity and atrial fibrillation: Epidemiology, mechanisms and interventions. Journal of Atrial Fibrillation, 6(2), pp.47–55.

Bacardí-Gascón, M., Jones, E.G. y Jiménez-Cruz, A., 2013. Prevalence of obesity and abdominal obesity from four to 16 years old children living in the Mexico-USA border. Nutrición Hospitalaria, 28(2), pp.479–485.

Barrera, R. y Fernández, L., 2016. Programación metabólica fetal R. Perinatología y Reproducción Humana, 29(3), pp.99–105.

Baskaran, C. et al., 2015. A decade of temporal trends in overweight/obesity in youth with type 1 diabetes after the Diabetes Control and Complications Trial. Pediatric Diabetes, 16(4), pp.263–270.

Belle, A. et al., 2015. Big Data Analytics in Healthcare. BioMed Research International, 2015, pp.1–16.

Beltrán Martínez, B., 2003. Minería de datos, p.67. Disponible en: http://bbeltran.cs.buap.mx/NotasMD.pdf

Bhurosy, T. y Jeewon, R., 2014. Overweight and obesity epidemic in developing countries: a problem with diet, physical activity, or socioeconomic status? The Scientific World Journal, 2014, p.964236. Disponible en: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4212551&tool=pmcentrez&rendertype=abstract.

Bonvecchio, A. et al., 2014. Contribution of formative research to design an environmental program for obesity prevention in schools in Mexico City. Salud Publica de México, 56(2), pp. S139–S147.

Boonchieng, E. et al., 2014. Development of mHealth for Public Health Information Collection, with GIS, using Private Cloud: International Computer Science and Engineering Conference, pp.350–353.

Borja Aburto, V.H., 2013. Salud Pública 1a. B. F. Irene Paiz, ed., Ciudad de México: Editorial Alfil.

Brémond, P. et al., 2015. Evolution of Dengue Disease and Entomological Monitoring in Santa Cruz. PLOS ONE, pp.1–21.

Brusse, C. et al., 2014. Social Media and Mobile Apps for Health Promotion in Australian Indigenous Populations: Scoping Review. JOURNAL OF MEDICAL INTERNET RESEARCH, 16(12), p.20.

Cáceres, J.H., 2016. Clustering technique based on k- means algorithm for the identification of clusters of surgical patients. Universidad Santo Tomás, seccional Bucaramanga, pp.1–8.

Carr, S., Unwin, N. y Pless-Mulloli, T., 2007. An Introduction to Public Health and Epidemiology Segunda Ed., New York: Mc Graw Hill. Disponible en: http://www.lavoisier.fr/livre/notice.asp?id=O22WA3ARX3KOWA.

Cascón, M.M., 2008. Medicina Preventiva y Salud Pública y Comunitaria. pp.1–6.

CAUSES, 2010. Seguro Popular. Comisión Nacional de Protección Social en Salud, p.1. Disponible en: http://www.seguro-popular.gob.mx/index.php/conocenos/seguro-popular-1 [Accessed July 26, 2017].

Chaurasia, V. y Pal, S., 2014. Data Mining Approach to Detect Heart Diseases. International Journal of Advanced Computer Science and Information Technology, 2(4), pp.56–66.

Chawla, N. V. y Davis, D.A., 2013. Bringing big data to personalized healthcare: A patient-centered framework. Journal of General Internal Medicine, 28(SUPPL.3), pp.660–665.

Chen, H. et al., 2014. A review of data quality assessment methods for public health information systems. International Journal of Environmental Research and Public Health, 11(5), pp.5170–5207.

Chung, J.E., 2015. Computers in Human Behavior Antismoking campaign videos on YouTube and audience response: Application of social media assessment metrics. Computers in Human Behavior, 51, pp.114–121. Disponible en: http://dx.doi.org/10.1016/j.chb.2015.04.061.

Cisneros-González, N. et al., 2016. Índice De Amputaciones De Extremidades Inferiores En Pacientes Con Diabetes. Revista del Instituto Mexicano del Seguro

Social, 54(4), pp.472–479. Disponible en: papers3://publication/uuid/5EA39D01-7663-4CE3-A5BF-58375C36B207.

Clarke, J.L., 2010. Preventive medicine: A ready solution for a health care system in crisis. Population Health Management, 13(Suppl 2), pp. S3–S11. Disponible en: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc7&NEWS=N&AN=2010-21637-002.

Cordera, R. y Ziccardi, A., 2006. Las políticas sociales en México al fin del milenio Descentralización, Diseño y Gestión 1a edición., Ciudad de México: Coordinación de Humanidades UNAM. Disponible en: http://archivos.diputados.gob.mx/Centros_Estudio/Cesop/Comisiones/3_ssocial.htm# [Citar como] [Accedido Julio 27, 2017].

Córdova-Zamora, M., 1995. Estadística descriptiva e inferencial, Disponible en: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Estadistica+descriptiva+e+inferencial#2.

Córdova Villalobos, J.Á., 2016. La obesidad: la verdadera pandemia del siglo XXI. Cirugía y Cirujanos, 84(5), pp.351–355. Disponible en: http://dx.doi.org/10.1016/j.circir.2016.08.001.

Cruz Estrada, F.D.M. et al., 2017. Overweight or Obesity, Gender, and Age Influence on High School Students of the City of Toluca's Physical Fitness. BioMed Research International, 2017, p.11.

Dalkir, K. y Liebowitz, J., 2005. Knowledge Management in Theory and Practice,

Dash, P., Pattnaik, S. y Rath, B., 2016. Knowledge Discovery in Databases (KDD) as Tools for Developing Customer Relationship Management as External Uncertain Environment: A Case Study with Reference to State Bank of India. Indian Journal of Science and Technology, 9(4), p.11.

Dávila-Torres, J., González-Izquierdo, J. y Barrera-Cruz, A., 2015. Panorama de la obesidad en México. Revista Médica del Instituto Mexicano del Seguro Social, 53(2), pp.240–249.

Dawkins, R., 1982. The Extended Phenotype, Oxford: Oxford University Press.

Esfandiari, N. et al., 2014. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), pp.4434–4463. Disponible en: http://dx.doi.org/10.1016/j.eswa.2014.01.011.

Esper, R.J. y Machado, R.A., 2008. LA INVESTIGACIÓN EN MEDICINA: Elementos de Bioestadística Segunda., Buenos Aires: Prensa Médica Argentina.

Fayyad, U., Piatetsky-shapiro, G. y Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE, 17(3), pp.37–54.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P., 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. KDD-96 Proceedings, pp.82–88.

García-Peñalvo, F.J. y Conde-González, M.A., 2017. Statistical Implicative Analysis Approximation to KDD and Data Mining: A Systematic and Mapping Review in Knowledge Discovery Database Framework. Ninth International Conference on Advances in Databases, Knowledge, and Data Applications, pp.70–77.

García, H.J. y López, J.M.M., 2012. Técnicas De Análisis De Datos: Aplicaciones Prácticas Utilizando Microsoft Excel Y Weka, pp.5–43.

García, L.M. et al., 2015. The cost of polypharmacy in patients with type 2 diabetes mellitus. Revista Médica de Chile, 143, pp.606–611.

George, D.R., Rovniak, L.S. y Kraschnewski, J.L., 2013. NIH Public Access. Clin Obstet Gynecol, 76(October 2009), pp.211–220.

Gheorghe, M. y Petre, R., 2014. Integrating Data Mining Techniques into Telemedicine Systems. Informatica Economică, 18(1), pp.120–131.

Gibert, K., Sànchez-Marrè, M., y Codina, V. (2006). Elección de la técnica de minería de datos: Mapa conceptual de técnicas. In V Simposio de Teoría y Aplicaciones de Minería de Datos (p. 7).

Gibert, K., Sànchez–Marrè, M., y Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. AI Communications, 29, 627–663. https://doi.org/10.3233/AIC-160710

Gobierno-de-Perú, 2010. Pandemia De Influenza A H1n1, Definiciones Generales Fases Y Fase Actual De La Pandemia, pp.4–6. Disponible en: http://www.dge.gob.pe/influenza/flu/documentos técnicos/pandemia y fases de pandemia.pdf.

González, F., 2013. Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA). Universidad de Granada. Disponible en: http://masteres.ugr.es/moea/pages/tfm-1213/tfm_garciagonzalezfrancisco_1/!

Hamerly, G., y Elkan, C. (2004). Learning the k in k-means. Advances in Neural Information Processing …, 17, 1–8. https://doi.org/10.1.1.9.3574

Han, J. y Kamber, M., 2000. Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques. pp. 3–26.

Hauser, L.L.K.J.F., 2012. Harrison's Principles of Internal Medicine 18a edición. J. de León Fraga, ed., Ciudad de México: Mc Graw Hill.

Heavin, C., 2017. Health Information Systems – Opportunities and Challenges in a Global Health Ecosystem. Journal of the Midwest Association for Information Systems, 2017(2).

De Henauw, S. et al., 2007. Nutritional status and lifestyles of adolescents from a public health perspective. The HELENA Project - Healthy Lifestyle in Europe by Nutrition in Adolescence. Journal of Public Health, 15(3), pp.187–197.

Hernández, R., Fernández, C. y Baptista, P., 2014. Metodología de la investigación Sexta edición, Ciudad de México: Mc Graw Hill. Disponible en: https://mail.google.com/mail/u/1/#inbox/15a4cf4b02ab7f85?projector=1.

Hernández Ávila, M. et al., 2016. Encuesta Nacional de Salud y Nutrición de Medio Camino 2016. Instituto Nacional de Salud Pública, 2016, p.151. Disponible en: http://www.epidemiologia.salud.gob.mx/doctos/encuestas/resultados/ENSANUT.pdf.

Ho, A., 2017. Beyond the Dataset: Understanding Sociotechnical Aspects of the Knowledge Discovery Process Among Modern Data Professionals.

Hswen, Y. et al., 2013. Virtual Avatars, Gaming, and Social Media: Designing a Mobile Health App to Help Children Choose Healthier Food Options. Journal of mobile technology in medicine, 2(2), pp.8–14. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/25419244%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4239544.

Hu, H., Li, Z. y Dong, H., 2017. Graphical representation and similarity analysis of protein sequences based on fractal interpolation. TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 14(1), pp.182–192.

IMSS, 2016. Instituto Mexicano del Seguro Social. Conoce al IMSS, p.1. Disponible en: http://www.imss.gob.mx/conoce-al-imss [Accedido Julio 27, 2017].

INEGI, 2015. Porcentaje de población derechohabiente por institución de salud pública. INEGI Encuesta Intercensal, p.6. Disponible en: http://www.beta.inegi.org.mx/temas/derechohabiencia/ [Accedido Julio 11, 2017].

Jani, A.A., Trask, J. y Ali, A., 2015. Integrative Medicine in Preventive Medicine Education: Competency and Curriculum Development for Preventive Medicine and Other Specialty Residency Programs. American Journal of Preventive Medicine, 49(5), pp. S222–S229. Disponible en: http://dx.doi.org/10.1016/j.amepre.2015.08.019.

Joung, K.H., Jeong, J. y Ku, B.J., 2015. The Association between Type 2 Diabetes Mellitus and Women Cancer - PUBMED, 2015, pp.11–20.

Katchunga, P.B. et al., 2016. Obesity and diabetes mellitus association in rural community of Katana, South Kivu, in Eastern Democratic Republic of Congo: Bukavu Observ Cohort Study Results. BMC Endocrine Disorders, 16(1), p.60. Disponible en: http://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-016-0143-5.

Kaur, I. et al., 2017. Predictive Risk Modeling of Diabetes Using Data Mining. International Journal of Engineering Technology Science and Research, 4(4), pp.136–141.

Kavakiotis, I. et al., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, pp.104–116. Disponible en: http://dx.doi.org/10.1016/j.csbj.2016.12.005.

Kenny, G. et al., 2017. Trends, Findings, and Opportunities: An Archival Review of Health Information Systems Research in Nigeria. Journal of the Midwest Association for Information Systems, 2017(2), pp.73–92.

King, B.R. et al., 2012. A diabetes awareness campaign prevents diabetic ketoacidosis in children at their initial presentation with type 1 diabetes. Pediatric Diabetes, 13(8), pp.647–651.

Koh, H.C. y Tan, G., 2005. Data mining applications in healthcare. J Health Information Management, 19(2), pp.64–72. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/15869215.

Kornfield, R. et al., 2015. Televised obesity-prevention advertising across US media markets: exposure and content, 2010–2011. Public Health Nutrition, 18(6), pp.983–993. Disponible en: http://www.journals.cambridge.org/abstract_S1368980014001335.

Kumar, D.S., 2016. Headache; Primary Versus Secondary and Relationship with Age, Gender, Hypertension, Diabetes and Obesity in Lumhs Hyderabad. the Professional

Medical Journal, 23(2), pp.193–197. Disponible en: http://www.theprofesional.com/article/vol-23-no-02/prof-3051.pdf.

Kumar, P., 2015. Pareto Principle: The 80-20 Phenomenon, Jharkhand.

Kumar, S. et al., 2015. Effect of mobile reminders on screening yield during opportunistic screening for type 2 diabetes mellitus in a primary health care setting: A randomized trial. Preventive Medicine Reports, 2, pp.640–644. Disponible en: http://dx.doi.org/10.1016/j.pmedr.2015.08.008.

Lai, Y.-J., Chen, H.-C. y Chou, P., 2015. Gender Difference in the Interaction Effects of Diabetes and Hypertension on Stroke among the Elderly in the Shih-Pai Study, Taiwan. Plos One, 10(8), p. e0136634. Disponible en: http://dx.plos.org/10.1371/journal.pone.0136634.

Lanas, F. et al., 2016. Prevalence, Distributions and Determinants of Obesity and Central Obesity in the Southern Cone of America. PLOS ONE, 79, pp.1–13.

Lateef, F., 2016. Big Data: Applications in Healthcare and Medical Education. Education in Medicine Journal, 8(1), pp.85–89.

Lee, C. H., y Yoon, H. (2017). Medical big data: promise and challenges. Kidney Research and Clinical Practice, 36(1), pp.3–11.

Lempiäinen, J., 2017. Social Media Planning: Case company Forsman tea. Haaga-Helia.

Lemus, J.D., Oroz, V.A. y Lucioni, M.C., 2008. Epidemiología y Salud Comunitaria 1a. E. O. Mester, ed., Rosario - Argentina: Corpus Editorial y Distribuidora.

Levy, T.S. et al., 2015. Desnutrición y obesidad: doble carga en México. Revista Digital Universitaria, 16(5), pp.1–17. Disponible en: http://www.revista.unam.mx/vol.16/num5/art34/.

López Arellano, O. y Jarillo Soto, E., 2015. ¿Hacia dónde debe ir el sistema de salud en México? Una perspectiva desde el derecho a la salud. Medicina Social, 10(1), pp.1–13.

Losada Ruiz, C., 2014. Guía de Respuestas en Diabetes Colaboración Inter sociedades Andalucía. SEMERGEN, Revista Andaluza de atención Primaria, (2250–4410), pp.7–9. Disponible en: http://www.semergenandalucia.org/.

Loslier, L., 2016. Geographical Information Systems (GIS) from a Health Perspective. International Development Research Center, (September), pp.5–10.

Ltifi, H., Mohamed, E. Ben y Ayed, M., 2016. Interactive visual knowledge discovery from data-based temporal decision support system. Information Visualization, 15(1), pp.31–50.

Mano, R.S., 2014. Social media and online health services: A health empowerment perspective to online health information. Computers in Human Behavior, 39, pp.404–412. Disponible en: http://dx.doi.org/10.1016/j.chb.2014.07.032.

Mansingh, G. et al., 2017. Data preparation: Art or science? In International Conference on Data Science and Engineering, ICDSE 2016.

Martínez Mandujano, J.A. et al., 2015. La edad y sexo como factores condicionantes del control de enfermedad crónica en el primer nivel de atención: estudio retrospectivo. Cuidado Y Salud/Kawsayninchis, 2(2), p.213.

Mohamadali, N.A., Ab Aziz, N.F. y Mohd Zahari, N.A., 2017. A Novel Conceptual Framework of Health Information Systems (HIS) Sustainability, p.6.

Mohd Salleh, M.I., Abdullah, R. y Zakaria, N., 2017. Extending Health Information System Evaluation with an Importance-Performance Map Analysis. Advances in Health Management, pp.33–55.

Morales González, J.A., 2010. Obesidad Un Enfoque Multidisciplinario 1a edición., Pachuca, Hidalgo: Universidad Autónoma Del Estado De Hidalgo.

Moreno-Altamirano, L. et al., 2014a. Epidemiología y determinantes sociales asociados a la obesidad y la diabetes tipo 2 en México. Revista Médica del Hospital General de México, 77(3).

Moreno-Altamirano, L. et al., 2014b. Epidemiología y determinantes sociales asociados a la obesidad y la diabetes tipo 2 en México. Revista Médica del Hospital General de México, 77(3), pp.114–1. Disponible en: www.elsevier.es/hgmx.

Moreno García, M.N. et al., 2001. Aplicación De Técnicas De Minería De Datos En La Construcción Y Validación De Modelos Predictivos Y Asociativos A Partir De Especificaciones De Requisitos De Software. CEUR Workshop Proceedings, 84, p.14. Disponible en: http://ceur-ws.org/Vol-84/paper4.pdf.

Muhaise, H. y Kareeyo, M., 2017. Electronic Health Information Systems Critical Implementation Issues (E-HMIS): District Health Information Software Version.2 in the

Greater Bushenyi, Uganda. American Scientific Research Journal for Engineering, Technology, and Sciences, 34(1), pp.205–212.

Murad, A.A., 2007. Creating a GIS application for health services at Jeddah city. Computers in Biology and Medicine, 37, pp.879–889.

Napolitano, M. et al., 2013. Using Facebook and text messaging to deliver a weight loss program to college students. Obesity Journal, 21(1), pp.25–31. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/oby.20232/full.

Narang, B., Verma, P. y Kochar, P., 2016. Application based, advantageous K-means Clustering Algorithm in Data Mining: A Review. International Journal of Latest Trends in Engineering and Technology, 7(2), pp.121–126.

Németh, M. y Michalconok, G., 2016. Preparation and Cluster Analysis of Data from The Industrial Production Process for Failure Prediction. Materials Science and Technology, 24(39), pp.157–162.

Oliveira, G. V., Coutinho, F. P., Campello, R. J. G. B., y Naldi, M. C. (2017). Improving k-means through distributed scalable metaheuristics. Neurocomputing, 246, 45–57. https://doi.org/10.1016/j.neucom.2016.07.074

OMS: Organización Mundial de la Salud, 2016a. Diabetes OMS., p.6. Disponible en: http://www.who.int/mediacentre/factsheets/fs312/es/.

OMS: Organización Mundial de la Salud, 2017a. Diabetes OMS a. WHO Media centre, p.1. Disponible en: http://www.who.int/mediacentre/factsheets/fs312/es/ [Accedido agosto 27, 2017].

OMS: Organización Mundial de la Salud, 2017b. Las 10 principales causas de defunción en el mundo., p.3. Disponible en: http://www.who.int/mediacentre/factsheets/fs310/es/ [Accessed July 28, 2017].

OMS: Organización Mundial de la Salud, 2016b. Obesidad OMS. p.6. Disponible en: http://www.who.int/mediacentre/factsheets/fs311/es/.

Oswal, S. y Shah, G., 2017. A Study on Data Mining Techniques on Healthcare Issues and its uses and Application on Health Sector. International Journal of Engineering Science and Computing, 7(6), pp.13536–13538.

Palloni, A. et al., 2015. Adult obesity, disease and longevity in Mexico. Salud Publica de México, 57(1), pp.22–31.

Panam, R. et al., 2015. Aspectos epidemiológicos y genéticos de la diabetes mellitus en la población peruana. Revista Biomedicina | Medicina Familiar y Comunitaria BIOMEDICINA, 10(1), pp.20–33.

Park, B.K. et al., 2017. A Facebook-Based Obesity Prevention Program for Korean American Adolescents: Usability Evaluation. Journal of Pediatric Health Care, 31(1), pp.57–66.

Paul, M.J. et al., 2016. Social media mining for public health monitoring and surveillance. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 21, pp.468–79. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/26776210.

Piédrola Gil, G., 2001. La salud y sus determinantes. Concepto de Medicina Preventiva y Salud Pública. Revista Medicina Preventiva y Salud Pública.

Piédrola Gil, G. et al., 2014. Medicina Preventiva y Salud Pública 11a edición., MASSON.

Ponce Sernicharo, G. y Kánter Coronel, I. de R., 2016. Al día: las cifras hablan Día internacional de la mujer., (33), pp.1–18. Disponible en: http://www.senado.gob.mx/ibd/content/productos/ad/AD33.pdf.

Puhl, R., Luedicke, J. y Lee Peterson, J., 2013. Public reactions to obesity-related health campaigns: A randomized controlled trial. American Journal of Preventive Medicine, 45(1), pp.36–48. Disponible en: http://dx.doi.org/10.1016/j.amepre.2013.02.010.

Rahm E. y Do, H.H., 2000. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), pp.3–13. Disponible en: papers2://publication/uuid/30073F7F-1B7C-4496-ADA4-94FF4E6EE8F7.

Ramanadhan, S. et al., 2017. Graphic health warnings as activators of social networks: A field experiment among individuals of low socioeconomic position. Social Science & Medicine, 175, pp.219–227. Disponible en: http://linkinghub.elsevier.com/retrieve/pii/S0277953616307249.

Ray, S. y Turi, R.H., 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. Proceedings of the 4th international conference on advances in pattern recognition and digital techniques, pp.137–143.

Ríos-Julián, N. et al., 2017. Feasibility of a screening tool for obesity diagnosis in Mexican children from a vulnerable community of Me’ phaa ethnicity in the State of

Guerrero, Mexico. In Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges. pp. 20–25.

Rodríguez Funes, V., 2008. Artículo de Medicina Basada en la Evidencia. Rev Arch Col Med, 1(2), pp.64–68.

Roiger, R.J., 2017. Data Mining: A Tutorial-Based Primer Segunda., Boca Ratón, Florida: Taylor & Francis Group.

Russo, G.T. et al., 2016. Fracture Risk in Type 2 Diabetes: Current Perspectives and Gender Differences. International Journal of Endocrinology, 2016.

Sa'di, S. et al., 2015. Comparison of Data Mining Algorithms in the Diagnosis of Type Ii Diabetes. International Journal on Computational Science & Applications, 5(5), pp.1–12. Disponible en: http://www.wireilla.com/papers/ijcsa/V5N5/5515ijcsa01.pdf.

Secretaría de Gobernación, 2016. Anexo estadístico cuarto informe de gobierno. p.825. Disponible en: http://www.presidencia.gob.mx/cuartoinforme/.

Secretaría de Salud Pública, 2010a. NORMA Oficial Mexicana NOM-008-SSA3-2010, Para el tratamiento integral del sobrepeso y la obesidad. p.30. Disponible en: http://www.dof.gob.mx/nota_detalle.php?codigo=5154226&fecha=04/08/2010.

Secretaría de Salud Pública, 2010b. NORMA Oficial Mexicana NOM-015-SSA2-2010, Para la prevención, tratamiento y control de la diabetes mellitus. p.34. Disponible en: http://dof.gob.mx/nota_detalle.php?codigo=5168074&fecha=23/11/2010.

Sevilla-Villanueva, B., Gibert, K., y Sànchez-Marrè, M. (2014). The Role of Statistical Tests on Cluster Interpretation.

Shamah Levy, T., 2016. El sobrepeso y la obesidad: ¿son una situación irremediable? Boletín Médico del Hospital Infantil de México, 73(2), pp.65–66.

Shaw, N.T. y McGuire, S.K., 2017. Understanding the use of geographical information systems (GISs) in health informatics research: a review. Journal of Innovation in Health Informatics, 24(2), pp.228–233.

Simsek Gursoy, U.T., 2016. Defining Characteristics of Diabetic Patients by Using Data Mining Tools. Journal of Hospital & Medical Management, 2(2), pp.1–8.

Sobers-Grannum, N. et al., 2015. Female gender is a social determinant of diabetes in the Caribbean: A systematic review and meta-analysis. PLoS ONE, 10(5), pp.1–23.

Sosa-García, B.C. et al., 2017. Factores de riesgo metabólico y no metabólico en personas adultas de un centro de salud del Estado de México. Revista de Enfermería Instituto Mexicano del Seguro Social, 25(1), pp.29–35.

Stetson, B., Minges, K.E. y Richardson, C.R., 2017. New directions for diabetes prevention and management in behavioral medicine. Journal of Behavioral Medicine, 40(1), pp.127–144.

Stratebi, 2010. Comparativa de Algoritmos de Herramientas de Data Mining. p.26. Disponible en: http://www.stratebi.com/open-business-intelligence.

Suca, C. et al., 2016. Comparison of Classification Algorithms for Prediction of Cases of Childhood Obesity. ResearchGate, (April).

Swapna, S. et al., 2016. Data Cleaning for Data Quality. In 3rd International Conference on Computing for Sustainable Global Development. pp. 344–348.

Tarqui-Mamani, C. et al., 2017. Análisis de la tendencia del sobrepeso y obesidad en la población peruana. Revista Española de Nutrición Humana y Dietética, 21(2), p.137. Disponible en: http://renhyd.org/index.php/renhyd/article/view/312.

Tarro, L. et al., 2016. Obesity-promoting factors in Mexican children and adolescents: challenges and opportunities. Global Health Action, 9, p.13.

Thirumal, P.C. y Nagarajan, N., 2015. Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - a Case Study. Journal of Engineering and Applied Sciences, 10(1), pp.8–13.

Torio, C.M., 2015. The Role of the Geographic Information Systems Infrastructure in Childhood Obesity Prevention. American Journal of Preventive Medicine, 42(5), pp.513–515. Disponible en: http://dx.doi.org/10.1016/j.amepre.2012.02.003.

Turner, P., Kushniruk, A. y Nohr, C., 2017. Are We There Yet? Human Factors Knowledge and Health Information Technology – the Challenges of Implementation and Impact. IMIA Yearbook of Medical Informatics, pp.84–91.

Valente, T.W. et al., 2015. Social network analysis for program implementation. PloS one, 10(6), p. e0131712. Disponible en: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131712.

Varlamis, I. et al., 2017. Computer Methods and Programs in Biomedicine Application of data mining techniques and data analysis methods to measure cancer morbidity and

mortality data in a regional cancer registry: The case of the island of Crete, Greece. Computer Methods and Programs in Biomedicine, 145, pp.73–83.

Veale, H.J. et al., 2015. The use of social networking platforms for sexual health promotion: identifying key strategies for successful user engagement. BMC public health, 15, p.85. Disponible en: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4340797&tool=pmcentrez&rendertype=abstract.

Villanueva, B. S., Gibert, K., y Sànchez-marrè, M. (2013). Clustering and Interpretation on Real Nutritional Data. In Conferencia de la Asociación Española para la Inteligencia Artificial (pp. 1454–1463). Madrid, España. Disponible en: http://hdl.handle.net/2117/22903

Wang, Y. y Lim, H., 2012. The global childhood obesity epidemic and the association between socio-economic status and childhood obesity. International Review of Psychiatry, 24(3), pp.176–188. Disponible en: http://informahealthcare.com.elibrary.jcu.edu.au/doi/abs/10.3109/09540261.2012.688195%5Cnhttp://www.tandfonline.com/doi/full/10.3109/09540261.2012.688195.

Witten, I.H., Frank, E. y Hall, M. a, 2011. Data Mining: Practical Machine Learning Tools and Techniques, Disponible en: http://books.google.com/books?id=bDtLM8CODsQC&pgis=1.

Wong, C.A., Merchant, R.M. y Moreno, M.A., 2014. Using social media to engage adolescents and young adults with their health. Healthcare, 2(4), pp.220–224. Disponible en: http://dx.doi.org/10.1016/j.hjdsi.2014.10.005.

Yepez, R., Baldeon, M. y López, P., 2007. Obesidad, Quito: Sociedad Ecuatoriana de Ciencias de la Alimentación y Nutrición. Disponible en: http://secian.com/libros/1libro.pdf.

Zeeshan, A. et al., 2014. Obesity and Diabetes: An experience at a public sector tertiary care hospital. Pakistan Journal of Medical Sciences, 30(1), pp.81–85.

Zhang, P. et al., 2016. Prevalence of central obesity among adults with normal BMI and its association with metabolic diseases in northeast China. PLoS ONE, 11(7), pp.1–11.

Zhao, L. et al., 2017. Analysis of the Preventive Medicine Undergraduate Curriculum in China: The West China School of Public Heath Experience: A Case Study. Frontiers in Education, 2(July), pp.1–6.

7. ANEXOS

Aquí se muestran los elementos de apoyo utilizados para la implementación de la metodología propuesta, así como archivos extraídos de fuentes oficiales.

7.1. Enfermedad, epidemia y pandemia

La salud pública, como se verá en su definición, es una disciplina que se apoya de muchas otras para su correcto funcionamiento. Su principal objetivo es el combatir enfermedades, epidemias, pandemias, así como el proporcionar servicios preventivos y correctivos para el bienestar y salud de la población en general.

Estos conceptos, como enfermedad y epidemia, deben de ser conocidos para la comprensión general de la Salud pública. Morales González (2010) estructura una definición del significado de enfermedad desde un punto de vista integral, ya que abarca tanto los aspectos genéticos hasta llegar a las interacciones que un organismo tiene con el medio que lo rodea, esta está basada en la teoría genética y de sistemas, la cual dice que “las enfermedades son sistemas fenotípicos (cualquier característica o rasgo observable de un organismo, como su morfología, desarrollo, propiedades bioquímicas, fisiología, comportamiento y herencia (Dawkins 1982)) ocasionados por la interacción del genoma con el medio ambiente, considerado éste jerárquicamente desde el nivel citoplásmico hasta el social”. Esto significa que una enfermedad es todo aquel comportamiento o reacción que surja entre un organismo, su herencia y el medio que lo rodea, lo cual altere su estado habitual o de “bienestar” ocasionando alteraciones y mal funcionamiento en dicho organismo. También hay que considerar que un padecimiento es “la acción de sufrir una enfermedad”.

Si una enfermedad comienza a presentarse en múltiples individuos simultáneamente durante un mismo lapso de tiempo en una comunidad o región definida, es cuando se está ante una epidemia, para Brémond et al. (2015) una epidemia se presenta cuando un cierto grupo de personas en una determinada región y durante un periodo de tiempo similar padecen la misma enfermedad, influyendo también factores externos, del medio, y compartiendo características como sexo, edad, síntomas, ubicación, entre

otras. Por otra parte la epidemiología vista como ciencia, es definida como una disciplina científica responsable del estudio de la distribución y los determinantes de la salud y la enfermedad en poblaciones humanas, con el propósito de contribuir a mejorar la salud (Arias-Valencia, 2017), a la vez que juega un papel sumamente importante al momento de llevar a cabo acciones de salud pública pues la epidemiología es considerada como la principal fuente de teorías, métodos y técnicas (Lemus, Oroz and Lucioni, 2008) por su carácter de interacción directa con los pacientes, los factores y el medio cuando surge alguna enfermedad.

Cuando una epidemia comienza a extenderse hasta abarcar niveles mundiales, esta se convierte en una Pandemia, para que una enfermedad tome dicha denominación, esta debe de tener un alto grado de infección y un fácil traslado de un sector geográfico a otro, sin importar el grado de mortalidad o la letalidad de la enfermedad en cuestión (Gobierno-de-Perú, 2010). Este término es utilizado cuando se habla de Diabetes, Sobrepeso y Obesidad debido a que son padecimientos que están afectando a un gran número de personas en todo el mundo, como se presentará más adelante, aun y cuando no sean enfermedades transmisibles, por lo que se debe de tener especial cuidado al utilizar este término cuando se hable de este tipo de enfermedades.

7.2. Instituciones de Salud Pública en México

En México existen varias instituciones públicas y gubernamentales que se encargan de acercar y brindar los servicios de salud a sus afiliados y sociedad en general, por su importancia y la labor que desempeñan las principales son:

- **Instituto Mexicano del Seguro Social (IMSS):** Es “la institución con mayor presencia en la atención a la salud y en la protección social de los mexicanos combinando la investigación y la práctica médica, con la administración de los recursos para el retiro de sus asegurados. Hoy en día, más de la mitad de la población mexicana tiene algo que ver con el Instituto; hasta ahora es la más grande en su género en América Latina” (IMSS, 2016).

- **Seguro Popular:** Según la definición del Sistema de Protección Social en Salud, “es el mecanismo por el cual el Estado garantizará el acceso efectivo, oportuno, de calidad, sin desembolso al momento de su utilización y sin discriminación a los servicios médico-quirúrgicos, farmacéuticos y hospitalarios que satisfagan de manera integral las necesidades de salud” (CAUSES, 2010).
- **Instituto de Servicios y Seguridad Social de los Trabajadores del Estado (ISSSTE):** El ISSSTE atiende a los trabajadores al servicio del Estado, pensionados, jubilados y sus familiares afiliados; otorga servicios médicos, prestaciones económicas, sociales y culturales, de vivienda, tiendas y farmacias y servicios turísticos (Cordera y Ziccardi 2000; Lomelí 2006). Actualmente es la segunda institución con mayor número de afiliados en México.
- **Otras:** Por lo general en estas instituciones se encuentra afiliados trabajadores del estado que desempeñan labores especiales, como aquellos que pertenecen a Petróleos Mexicanos (PEMEX), a la Secretaría de la Defensa Nacional (SEDENA), a la Secretaría de Marina (SEMAR), entre otras.

Un ejemplo del alcance e impacto que estas instituciones tienen en el préstamo de servicios de salud en el país, se pueden hay observar sus cifras, para ello en la Tabla 7. se muestra la cantidad de personas inscritas a cada una de ellas, tomando en cuenta asegurados, familiares y pensionados; en la tabla 7.2 se tiene la cantidad de personas que reciben servicios de salud sin ser afiliados (Secretaría de Gobernación, 2016). Hay que señalar que se tiene un registro de aproximadamente 54,923,952 personas inscritas al Seguro Popular.

POBLACIÓN ASEGURADA	
IMSS	44,905,000
ISSSTE	8,953,000
PEMEX	643,000
SEDENA	720,000
SEMAR	303,000
Estatales	289,000
Total	55,813,000

Tabla 7.1. *Total de población afiliada por Institución de Salud.*

Las Instituciones con mayor impacto son el IMSS y el ISSSTE debido a que se ligan al empleo formal, es decir, sus afiliados son del sector productivo (López Arellano y Jarillo Soto 2015), gran número de afiliados, cuentan con la mayor infraestructura, mayor formación de profesionales de la salud (médicos, especialistas, enfermeras, entre otros) y presencia en las 32 entidades federativas del país.

POBLACIÓN NO ASEGURADA	
SECRETARÍA DE SALUD	34,214,000
IMSS-PROSPERA	11,960,000
UNIVERSITARIOS	29,000
TOTAL	46,203,000

Tabla 7.2. *Total de población No afiliada por Institución de Salud.*

Si la suma de las personas, aseguradas o no, es mayor al total de habitantes de México, se debe a que pudieron estar registradas en dos o más de estas Instituciones al momento de realizar los cálculos y levantamiento de los mismos (INEGI, 2015).

Combinadas ambas modalidades, asegurados y no asegurados, el préstamo de servicios de salud alcanza a más de cien millones de mexicanos, lo que demuestra el impacto antes mencionado de estas Instituciones para con la sociedad en general, sin olvidar que se siguen teniendo rezagos en materia de salud como lo mencionaba Borja Aburto (2013), pues en 2013 alrededor del 33% de la población no contaba con ningún tipo de afiliación que le asegurara servicios médicos, tendencia que se ha venido modificando en los últimos años gracias a la incorporación de estas personas a programas especiales de la SSA o directamente al denominado Seguro Popular.

Todas las instituciones mencionadas, así como las acciones que toman para la aplicación de medidas de salud pública se encuentran coordinadas y deben de pasar por las normativas y regulaciones que dicta la Secretaría de Salud en México, pues este organismo es quien se encarga de planear, dirigir, coordinar y evaluar todas aquellas acciones en beneficio de la salud en el país (Álvarez y Kuri 2012).

7.3 Diabetes y sus principales complicaciones

La tabla 7.3 presenta una serie de enfermedades asociadas al padecimiento de la diabetes.

Padecimiento Asociado	Descripción de lo encontrado en el estudio realizado	Referencia
Insuficiencia renal, Neuropatía, Retinopatía	En su estudio García y su equipo encontraron que estas 3 complicaciones de la diabetes prevalecían entre los miembros de su muestra de pacientes. (fallo renal, inflamación en nervios y retina)	(García et al. 2015)
Cáncer en la mujer	La diabetes es un importante factor de riesgo para varios tipos de cáncer. La resistencia a la insulina e inflamación crónica están fuertemente relacionados con el cáncer, además, los cambios en las hormonas producidas en los ovarios (incremento de estrógeno y andrógeno, baja de progesterona) son considerados potencialmente cancerígenos en las mujeres (pecho, endometrio y ovarios).	(Joung et al. 2015)
Incrementa riesgo de fracturas	En diversos estudios epidemiológicos se ha demostrado un aumento en el riesgo de fracturas en personas con osteoporosis y diabetes, es especial en mujeres	(Russo et al. 2016)
Amputaciones de miembros inferiores	Mencionan que, entre los distintos padecimientos asociados a la diabetes, se encuentran las amputaciones de miembros inferiores y pie diabético.	(Sobers-Grannum et al. 2015)
Infarto de miocardio, Derrame cerebral	El infarto de miocardio y accidentes cerebrovasculares como derrames son otras de las consecuencias de la diabetes, pues se ha demostrado fuerte correlación.	(OMS 2016a; Lai et al. 2015)

Tabla 7.3. *Principales padecimientos y complicaciones asociados a la Diabetes.*

Las anteriores son solo algunos de los padecimientos y enfermedades asociadas a la diabetes, el impacto en la salud por esta enfermedad y sus complicaciones provocaron solo en 2015 alrededor de 1.6 millones de muertes en el mundo (OMS 2017a).

7.3.1. Síntomas de la Diabetes Tipo 1

Consisten, entre otros, en excreción excesiva de orina (poliuria), sed (polidipsia), hambre constante (polifagia), pérdida de peso, trastornos visuales y cansancio (OMS: 2016a; Ponce Sernicharo y Kánter Coronel 2016). Estos síntomas pueden aparecer de forma súbita.

7.3.2. Síntomas de la Diabetes Tipo 2

Pueden ser similares a los de la diabetes de tipo 1, pero a menudo menos intensos. En consecuencia, la enfermedad puede diagnosticarse solo cuando ya tiene varios años de evolución y han aparecido complicaciones. Hasta hace poco, este tipo de diabetes sólo se observaba en adultos, pero en la actualidad también se está manifestando en niños (Wang y Lim 2012).

- **Prevención**

Se ha demostrado que medidas simples relacionadas con el estilo de vida son eficaces para prevenir la diabetes de tipo 2 o retrasar su aparición:

- ✓ Alcanzar y mantener un peso corporal saludable.
- ✓ Mantenerse activo físicamente: al menos 30 minutos de actividad regular de intensidad moderada la mayoría de los días de la semana; para controlar el peso puede ser necesaria una actividad más intensa. En múltiples estudios presentados en el trabajo de Aschner et al. (2016) se encontró que el cambio en el estilo de vida, dejar el sedentarismo y comenzar a realizar activación física intensiva acompañadas de pérdida de peso mediante un régimen de dieta, reducen significativamente los riesgos de complicaciones y muerte.
- ✓ Consumir una dieta saludable, higiene dietética para Losada Ruiz (2014), que evite el azúcar y las grasas saturadas.
- ✓ Evitar el consumo de tabaco, puesto que aumenta el riesgo de sufrir diabetes y enfermedades cardiovasculares.

La diabetes es una condición crónica-degenerativa que se encuentra entre las 10 principales causas de muerte en el mundo, donde la mayoría de los deseos de

personas con diabetes ocurre debido a complicaciones asociadas a esta enfermedad (OMS: Organización Mundial de la Salud 2017b; Carr et al. 2007).

7.4. Sobrepeso y Obesidad

El sobrepeso y obesidad pueden deberse a múltiples factores entre los cuales destacan:

➢ Causas

El sobrepeso y la obesidad es originado por una ingesta mayor de calorías en relación con las gastadas, esto se debe principalmente al consumo de alimentos con alto contenido calórico y grasa, si se le suma el sedentarismo y bajos niveles de actividad física debido a patrones de comportamiento adoptado de países desarrollados, cambios ambientales y sociales asociados al desarrollo, nuevos tipos de trabajos y medios de transportes, el riesgo es mayor; también el origen se puede deber a tendencias genéticas y hereditarias (Lanas et al. 2016; Zeeshan et al. 2014; De Henauw et al. 2007).

➢ Prevención

Para evitar el sobrepeso, la obesidad y los padecimientos asociados a estas (anexo 7.4.1), tanto el comportamiento del individuo, como organizaciones (escuela, trabajo) y de la sociedad en general, deben de colaborar en programas de activación física, inclusive fuera de las horas de enseñanza o trabajo (Aguilar Cordero et al. 2014; Shamah Levy 2016), también una alimentación más saludable es necesaria para prevenir y corregir tanto la obesidad como el sobrepeso (Wang y Lim 2012; Tarqui-Mamani et al. 2017; Bhurosy y Jeewon 2014). En lo individual, se debe de limitar la ingesta energética procedente de alimentos pobres en nutrientes y altos en grasa total y azúcares; aumentar el consumo de frutas y verduras, así como de legumbres, cereales integrales y frutos secos.

Existen diversas clasificaciones según la edad y género de la persona para clasificarla dentro de alguno de estos padecimientos.

➢ Sobrepeso y Obesidad en adultos

Normalmente la obesidad es definida en términos del Índice de Masa Corporal (IMC) que se calcula dividiendo los kilogramos de peso entre el cuadrado de la estatura en metros $\frac{kg}{m^2}$, pero existen otras mediciones e índices como la circunferencia abdominal, pruebas sanguíneas, entre otras para determinar si una persona padece sobrepeso u obesidad (Simsek Gursoy 2016; Zhang et al. 2016).

La OMS define el sobrepeso cuando se tiene un IMC igual o superior a 25; y obesidad cuando el IMC es igual o superior a 30. Esta medida se debe de ver como un valor aproximado porque puede no corresponderse con el mismo nivel de grosor en diferentes personas. La prueba de Impedancia Bioeléctrica es una forma bastante precisa para el diagnóstico de estos padecimientos tanto en adultos como en niños, esta consta de hacer pasar una corriente eléctrica de baja intensidad por el cuerpo y a mayor tejido graso mayor resistencia (Yepez et al. 2007; Morales González 2010).

➢ Sobrepeso y Obesidad en niños menores de 5 años

El sobrepeso es el peso para la estatura con más de dos desviaciones típicas por encima de la mediana establecida en los patrones de crecimiento infantil de la OMS; y la obesidad es el peso para la estatura con más de tres desviaciones típicas de la misma escala.

La totalidad de tablas con los patrones de crecimiento en niños se pueden encontrar en la siguiente dirección web:
https://www.aepap.org/sites/default/files/curvas_oms.pdf

➢ Sobrepeso y Obesidad en niños de 5 a 19 años

El sobrepeso es el IMC para la edad con más de una desviación típica por encima de la mediana establecida en los patrones de crecimiento infantil de la OMS, y la obesidad es mayor que dos desviaciones típicas por encima de la mediana establecida en la misma escala.

La totalidad de tablas con los patrones de crecimiento en niños se pueden encontrar en la siguiente dirección web:
https://www.aepap.org/sites/default/files/curvas_oms.pdf

La línea central, representa la media esperada de IMC en relación con la edad tanto para niños como para niñas mayores de 5 y hasta los 19 años.

7.4.1. Principales padecimientos asociados al Sobrepeso y Obesidad

Esta enfermedad se encuentra asociada a una gran cantidad de padecimientos, por ejemplo, el 80% de los casos de diabetes tipo 2 pueden ser atribuidos al sobrepeso y obesidad, en la tabla 7.4 se muestran más ejemplos.

Padecimiento Asociado	Descripción de lo encontrado en el estudio realizado	Referencia
Diabetes Mellitus Tipo 2 (DM2)	En innumerables estudios se ha demostrado la correlación entre diabetes y obesidad, una de las principales causas de la DM2 es la obesidad abdominal, factor que se aconseja mantener bajo control.	(Antonio y Ávila 2015) (Yepez et al. 2007) (Córdova 2016) (Zhang et al. 2016) (Lanas et al. 2016)
Enfermedades cardiovasculares	El alto peso y el alto nivel de grasa en el organismo conlleva a sufrir: ataque al miocardio, obstrucciones por colesterol, hipertensión, aterosclerosis, etc., estas son algunas de las enfermedades de este tipo asociadas a la obesidad y sobrepeso	(Zhang et al. 2016) (Bacardí-Gascón et al. 2013) (Baskaran et al. 2015)
Cáncer y trastornos del aparato locomotor	Favorece especialmente el padecer osteoartritis (enfermedad degenerativa de las articulaciones muy discapacitante); también padecer algunos cánceres como: endometrio, próstata, hígado, vesícula biliar, riñones y colon, también de mama y ovarios en mujeres.	(OMS 2016b)
Trastornos del sistema nervioso	La presencia de cefaleas (trastornos primarios dolorosos como migraña) se encuentran fuertemente correlacionados con la obesidad, género y edad.	(Kumar, 2016)

Tabla 7.4. *Principales padecimientos y complicaciones asociados al Sobrepeso y Obesidad.*

Independientemente de todos estos padecimientos, anualmente mueren como mínimo alrededor de 2.8 millones de personas en el mundo por el sobrepeso y obesidad (OMS 2017b). Esto se presentaba mayormente en países de altos ingresos, pero ahora es común en aquellos con medios y bajos ingresos.

7.4.2. Impacto de la Diabetes, Sobrepeso y Obesidad en México y las Instituciones de Salud

Para tener una idea de la gravedad de estas en enfermedades, basta con saber que México ocupa el primer lugar en el padecimiento de Obesidad en niños y el segundo en adultos en todo el mundo, destacando que 7 de cada 10 personas presentan sobrepeso u obesidad en el país (Cruz Estrada et al. 2017; Moreno-Altamirano et al. 2014a; Barrera y Fernández 2016; Lanas et al. 2016), por otra parte, la Diabetes Mellitus es considerada como uno de los mayores problemas de salud pública en el mundo, es la primera causa de muerte en México, ya que se asocia con otras enfermedades cardiovasculares, renales, cáncer, entre muchos otros más (Sosa-García et al. 2017; Katchunga et al. 2016; Martínez Mandujano et al. 2015). Por si esto no fuera suficiente, entre un 8 y 10% de las muertes prematuras en México son atribuidas a la obesidad, y se calcula que existe una disminución en la esperanza de vida de entre 6 y 20 años, dependiendo del rango de edad, por razones asociadas a esta enfermedad (Dávila-Torres et al. 2015; Córdova Villalobos 2016).

Por otra parte, cuando se habla de estas enfermedades es común pensar en automático en su impacto en la calidad de vida de las personas, pero se deben de tomar en cuenta otros factores como el social y el económico debido al constante incremento en los costos para el tratamiento y cuidado de enfermedades crónicas (anexo 7.5).

El impacto económico en las familias también es sustancial Moreno-Altamirano et al. (2014b) enfatizan el hecho de que por cada $100 pesos que se gastan en diabetes en México, $51 provienen de las familias, es decir, son recursos que no son suministrados por las instituciones de salud públicas o privadas a las que los que la padecen se encuentran afiliados, y esto sólo tomando en cuenta los costos directos. Algo similar

al Principio de Pareto ocurre con la distribución de los recursos destinados a tratar y combatir la diabetes; por cada $100 pesos gastados en diabetes, $53 se gastan en el 10% de la población, $33 en el 48% (asegurados) de la población y $15 pesos en el 42% restante de la población (no asegurados), esto implica que gran parte de los recursos se encuentran enfocados a un número muy pequeño de personas.

Todas las cifras y costos mencionados están relacionados con la diabetes, sobrepeso y obesidad, pero se debe de sumar la pérdida de productividad. En el país según el Instituto Mexicano de la Competitividad en total (sumando costos de tratamiento y pérdida de productividad) en el año 2015 estos costos ascendieron a 85,000 millones de pesos (Palloni et al. 2015; Shamah Levy 2016; Tarro et al. 2016).

Una gran causa de esta pérdida de productividad se debe a las amputaciones de miembros inferiores como consecuencia de ulceraciones por mal cuidado del paciente con diabetes, por citar un ejemplo. En México del 40 al 85% de las amputaciones de miembros inferiores se realizan en personas con diabetes, quienes son 15 veces más propensas, y este riesgo aumenta conforme aumenta la edad (Cisneros-González et al. 2016).

El llevar registro de este tipo de cifras y estadísticas por parte de las instituciones de salud siempre ha sido necesario para la toma de indicadores y conocer cómo se están prestando los servicios, aplicando tratamientos, diagnosticando enfermedades, contabilizando gastos, entre otros, por lo que se volvió común la utilización de sistemas de información computarizados que faciliten este proceso.

7.5. Ejemplos del Impacto de la Diabetes, Sobrepeso y Obesidad

Por citar un ejemplo, Clarke (2010) asegura que este tipo de enfermedades representan una amenaza a la economía nacional (refiriéndose a EUA), donde los costos por cuidado se incrementaron a una mayor tasa tan sólo en 2009 que en más de la mitad del siglo pasado, consumiendo aproximadamente el 60% del presupuesto para salud en ese año. Caso parecido es el de México, donde en el año 2016 se

asignaron a la estrategia de control de diabetes, obesidad y sobrepeso 589 millones 129 mil pesos (Ponce Sernicharo y Kánter Coronel 2016), independientemente de los recursos que las instituciones de salud destinan para la prevención y tratamiento de las mismas. Aunque estas cifras puedan parecer cuantiosas, el hecho es que los costos reales superan por mucho lo presupuestado, en 2008 el total de costos tan sólo por obesidad y sobrepeso para la Secretaría de Salud fue de 42,000 millones de pesos, y se espera que para 2017 este costo aumente a 151,000 millones; aunque económicamente el país puede solventar el incremento, demuestra que las estrategias tomadas no están presentando los resultados esperados (Dávila-Torres et al. 2015; Levy et al. 2015).

7.6. Sistemas de Información en las Instituciones de salud

También con el uso de los HIS se evitan ensayos clínicos, exámenes y tratamientos innecesarios, reduciendo costos, pues agrupan de manera integral lo financiero, administrativo y clínico en una institución de salud; se puede ver un HIS desde el punto de vista de la industria, donde la tecnología es utilizada con el propósito de lograr eficiencia en el trabajo y efectividad en la organización, a la vez que se proporciona el mejor servicio al cliente (Mohd Salleh et al. 2017; Heavin 2017; Mohamadali et al. 2017).

7.6.1. Big Data

Si se observa que en un país pueden existir miles de unidades médicas (hospitales, clínicas) a las que acuden diariamente otros miles o millones de personas a recibir algún tipo de atención, mismas cuyos registros quedan almacenados en los HIS, se puede estimar que la cantidad de datos que son recopilados en estos sistemas es inmensa e imposible de manejar sin el uso de tecnología. El término “Big Data” es utilizado cuando se habla de estos grandes volúmenes de registros que crecen y cambian constantemente a una tasa imposible de supervisar o analizar para el aprovechamiento de estos. No se cuenta con una definición única pero la mayoría se caracterizan por hablar de la creciente habilidad tecnológica para capturar, agregar y procesar datos con un gran volumen, velocidad y diversidad de tipos (texto, números,

imágenes, video, entre otros), atributos longitudinales (seguimiento en el tiempo), complejidad de relaciones, distribuidos en diversas bases de datos (repositorios que contienen los registros de un sistema), entre otras fuentes (Agnelluti, 2014).

En este mismo contexto han surgido los Sistemas de Información de Salud Pública o Public Health Information Sistems (PHIS) en inglés, los cuales tienen los mismos componentes y definición que los HIS, sólo que un PHIS tiene como misión el proporcionar una guía para la toma de decisiones en las prácticas de salud pública en las agencias de salud del gobierno al facilitar el intercambio y transmisión de datos entre estas. Prácticas tales como proveer alertas y advertencias tempranas sobre riesgos sanitarios y enfermedades, apoyar la administración de la salud pública, ayudar a estimar los estados y tendencias (Chen et al. 2014), y en especial estimular, promover y mejorar la investigación mediante el acceso y uso de datos provenientes de pacientes, son solo algunos de los beneficios de los PHIS (Mohd Salleh et al. 2017).

Una de las características de Big Data es el hecho de que los registros rebasan las capacidades comunes de las base de datos para su manejo, administración, almacenaje y análisis, por lo que se necesita constantemente de nuevos mecanismos tanto de hardware como de software (algoritmos, paquetes especializados) para un exitoso manejo y aprovechamiento de los datos (Lateef 2016; Belle et al. 2015).

Heavin (2017) utiliza el término “Big Health Data” (BHD) para referirse a los grandes cúmulos de los HIS, explica que de esta inmensidad de datos se pueden extraer nuevos conocimientos para informar de riesgos tanto en lo individual como en lo colectivo. Si a esto se le suma el hecho de que cotidianamente se utilizan aplicaciones médicas de las instituciones de salud en teléfonos inteligentes, intercambio de información de forma portátil, entre otros usos de tecnologías, se están recopilando, analizando e intercambian cantidades masivas de datos e información relacionados con la salud (Turner et al. 2017), y como menciona Lateef (2016) “Si los datos son lo suficientemente numerosos y la búsqueda suficientemente exhaustiva, los patrones de hecho pueden ser suficientemente convincentes”.

Tomando en cuenta todo lo anterior y en especial esto último, surge y se sustenta la importancia de contar tanto con las tecnologías, así como con las herramientas de análisis y procesamiento de datos que ayuden a aprovechar los registros y el conocimiento implícito en ellos y en las bases de datos de los HIS para, precisamente, utilizarlos en la investigación y generación de conocimiento que sirva en la promoción de cambios y mejoras en los individuos, a la vez que ayuden en la elaboración y modificación de políticas institucionales basándose en la toma de decisiones informadas y sustentadas en el conocimiento proveniente de la realidad de estas mismas y de sus pacientes.

7.7. Estadística

En su libro “Estadística descriptiva e inferencial” Córdova (1995) trata los aspectos de mayor relevancia en cuanto a estadística se refiere, su definición, tipos, elementos, entre otros aspectos, y es de este de donde se extrae la siguiente información.

➢ Definición y tipos de estadística

Primeramente, la Estadística en su definición más simple puede ser vista como:

“La ciencia que proporciona un conjunto de métodos, técnicas o procedimientos para Recopilar, Organizar (clasificar, agrupar), Presentar y Analizar datos con el fin de describirlos o de realizar generalizaciones válidas”.

Figura 7.1. *Tipos básicos de Estadística.*

La figura 7.1 muestra las dos principales ramas de la estadística, de las cuales se desprenden otras vertientes, para efecto de esta investigación sólo se abordarán las presentadas en dicha figura, pues suelen ser las de mayor utilización, junto con otras herramientas, para el análisis y proyección de datos.

- Estadística Descriptiva: es el conjunto de métodos estadísticos que se relacionan con el resumen y descripción de los datos, como tablas, gráficas, y el análisis mediante algunos cálculos, como las medidas de tendencia central o de localización (Media, mediana, moda, desviación, varianza, asimetría).
- Estadística Inferencial: es el conjunto de métodos con los que se hace la generalización o inferencia sobre una población utilizando una muestra. Existen diversas técnicas de inferencia como la Estimación Puntual, por Intervalos (de confianza, variabilidad de parámetro, límites, error de estimación) o Bayesiana (utilizada para saber qué tan probable es que una hipótesis sea cierta), donde se busca estimar los valores de algún parámetro poblacional a partir de los datos de una muestra.
- Probabilidad: cuando se utiliza la Inferencia estadística es necesario proporcionar una medida de confiabilidad (porcentaje de error, nivel de confianza, entre otros), ya que se puede llegar a conclusiones que no son ciertas en forma absoluta, y es la probabilidad quien agrupa las técnicas para brindar estas medidas de certeza.

Tanto la estadística descriptiva como la inferencial no pueden ser consideradas mutuamente excluyentes, pues para realizar una inferencia es necesario tener el conocimiento suficiente de los métodos descriptivos.

➢ Elementos de la estadística

Para poder utilizar métodos estadísticos se deben de conocer los conceptos básicos que se relacionan y son utilizados en esta, la tabla 7.5 muestra a continuación los de mayor significancia y uso, así como una pequeña descripción de estos para una mejor comprensión.

Elementos	Definición
Población	Conjunto de elementos que contienen una o más características observables de naturaleza cualitativa o cuantitativa que se pueden medir en ellos. Puede ser finita o infinita.
Muestra	Es una parte de la población seleccionada. Debe ser elegida de manera que sea representativa de la población.
Dato Observado	Es el resultado de medir alguna de las características observables de un elemento.
Parámetro	Es una medida descriptiva que resume una característica de la población, tal como la media (μ) o la varianza (σ^2), calculada a partir de los datos observados de toda la población.
Estimador	Es una medida descriptiva que resume una característica de la muestra, tal como la media ($\bar{x}$) o la varianza (s^2) calculada a partir de los datos observados de una muestra aleatoria.
Variable estadística	Es una característica definida en la población por la investigación estadística, que puede tomar cualquier valor (cualidades o números).

Tabla 7.5. *Principales elementos para el trabajo con estadística.*

➢ Organización y representación de datos

Una vez que los datos con los que se trabajará son recopilados, deben de ser presentados de forma resumida, de forma tal que se facilite su comprensión, análisis y posterior interpretación. En primera instancia los datos pueden ser presentados en forma de tablas numéricas y mediante gráficos, es importante siempre incluir un título representativo, la fuente de los datos y las unidades en que expresan los datos.

Las gráficas que más se utilizan para la representación de datos son las de barras (horizontales y verticales), las de sectores circulares, de puntos, de líneas, de batón, histogramas, polígonos y curvas de frecuencias, ojiva, entre muchos otros tipos, la selección y tipo de representación gráfica dependerá en gran medida del tipo de dato y escala utilizada, también de lo que se esté tratando de enfatizar, pues algunas resaltan aspectos que otras ni siquiera consideran o no muestran con detalle. En el caso de Hu et al. (2017) quienes en su investigación necesitaban mostrar los resultados de cálculos y mediciones de muestras, utilizaron diversos tipos de gráficos de barras, líneas, y tablas para ello, facilitando la lectura y transmisión de información.

Usualmente estos métodos estadísticos ayudan a representar, visualizar y analizar, ya sea de forma gráfica (histogramas, diagramas de dispersión) o conceptualmente (media, mediana, desviación) los datos sin la necesidad de utilizar procedimientos complejos o que requieran de grandes recursos, hablando de hardware y software, para lograrlo. Pueden ser utilizados durante todo el transcurso de la investigación, desde la visualización de la historia previa y actual, hasta las proyecciones futuras, además su uso en la medicina es ampliamente reconocido y aceptado como herramienta para la descripción de poblaciones y la toma de decisiones, por lo que es indispensable como parte de los primeros pasos de la investigación (Esper y Machado 2008). Por otra parte, a medida que la complejidad del análisis va aumentando y la estadística básica no es suficiente para encontrar o visualizar las relaciones y tendencias implícitas entre los datos almacenados en las bases de datos de los HIS, será necesario la utilización de, en este caso, minería de datos y otras herramientas para el descubrimiento de conocimiento en los registros en bruto.

➢ Tipos de variables

En el entendido de que una variable representa alguna de las características de la población de estudio, estas van a poder tomar distintos tipos de valores, ya sean cuantitativos o cualitativos, por lo que es importante saber qué tipos de variables existen según la escala de medición que se utilice, a continuación, se observan los distintos tipos:

- **Nominal**: Los valores no pueden ser ordenados por su relevancia, sólo sirven para clasificar en clases. Por ejemplo: Sexo, Ocupación, Trabajo.
- **Ordinal**: Los valores pueden ser clasificados y también ordenados según una jerarquía dependiendo de la característica que se evalúa. Por ejemplo: Grado de Estudios, Nivel de Aceptación (Alto, Medio, Bajo).
- **Intervalo**: Son valores de orden natural, sirven para cuantificar, por lo general tiene una unidad de medida (km, ml), puede ser un intervalo discreto cuando se utiliza para contar, por ejemplo: Número de Personas, Autos en una Familia,

entre otros; o puede ser continuo cuando se utiliza para medir, por ejemplo: Estatura, Peso, IMC.

- **Razón**: Posee las mismas características que la escala de medición por intervalos, con la diferencia que también puede contener al cero absoluto entre sus valores. Por ejemplo: Saldo de una Cuenta Bancaria.

Con estas escalas de medición tanto las variables cuantitativas como las cualitativas podrán ser clasificadas correctamente para su posterior trabajo y estudio mediante las distintas técnicas de procesamiento de datos.

➢ La distribución Normal en los datos

Otro de los aspectos que se deben de verificar cuando se va a realizar algún análisis que implica el uso de elementos de la estadística, es conocer el tipo de distribución que describen los datos con que se trabaja.

La distribución Normal o Gauss, permite caracterizar y medir la dispersión de los datos que conforman una muestra, la forma de la distribución representa la probabilidad de que un elemento tenga un valor cercano al de la media, y cómo este va disminuyendo a medida que se alejan de la misma. Si los datos muestrales se distribuyen normalmente, se pueden inferir teorías y realizar predicciones que de otra manera no serían válidas. Por ejemplo, los elementos de la muestra que se encuentran entre la media y ± 2 desviaciones estándar, constituyen exactamente al 95% de los casos, lo que deja sólo un 5% de datos que no se comporta como el resto (Esper y Machado 2008), entonces se podrán hacer inferencias y generalizaciones que abarquen a una gran parte de la población de donde se obtuvo la muestra con un bajo nivel de error.

7.8. KDD

Como suele ocurrir con este tipo de ramas de las ciencias, no existe una definición única para ellas. El término KDD se acuño en 1989 para enfatizar el hecho de que el producto final del análisis de datos es el conocimiento mismo.

7.8.1. Los 9 pasos de KDD

1. **Entendimiento de la importancia del conocimiento:** Aprender y entender la importancia que el conocimiento tiene, así como identificar el o los objetivos que se tengan al utilizar KDD.
2. **Definición del grupo de datos objetivo:** Selección del conjunto de datos con el que se trabajará, provendrá de las fuentes o bases de datos.
3. **Limpieza de datos y preprocesado:** Realización de las tareas de filtrado, remoción de ruido, campos vacíos, duplicados, incompletos, y demás características no deseadas en los datos.
4. **Reducción y proyección de datos:** Se buscará disminuir el número de variables aleatorias del conjunto de datos mediante métodos de transformación o reducción.
5. **Selección de la tarea de Minería de Datos apropiada:** Asegurarse de que el o los objetivos establecidos se adapten a alguno de los métodos de minería de datos como lo son los de clasificación, regresión, árboles de decisión, entre otros.
6. **Análisis exploratorio, modelo y selección de enfoque:** Se selecciona el o los algoritmos y métodos de minería de datos que se adaptaron a los objetivos definidos.
7. **Minería de Datos:** En este paso los patrones en los datos son encontrados, usualmente en forma de reglas de clasificación, modelos de regresión, y árboles de decisión.
8. **Interpretación de patrones:** Aquí se realizará la visualización, a través de distintos medios, de los patrones y modelos encontrados.
9. **Consolidar el conocimiento descubierto:** El paso final es integrar el conocimiento descubierto dentro de algún otro sistema que permita su almacenamiento y acceso para futuras acciones, o reportar dicho conocimiento a los individuos interesados y que pueden hacer uso de él.

7.9. Minería de Datos

Aun y cuando se han aplica los distintos tipos de estadística a un grupo de datos lo bastantemente significativo, los resultados obtenidos estarán limitados por la capacidad de procesamiento, el dinamismo, complejidad y volumen que los mismos

presenten. Cuando se sobrepasa alguno de estos límites es necesario la utilización de técnicas y métodos que estén diseñados para hacer frente a los retos que representa el trabajar con gran cantidad de registros, pero aun nivel de mayor profundidad al integrar múltiples disciplinas para ello, permitiendo acceder a conocimiento que de otra forma sería muy tardado o simplemente imposible de obtener.

Al tratarse de un conjunto de varias disciplinas, fórmulas matemáticas, software, y otras técnicas, aun no se tiene una definición única para describirla, pese a que este término surgió entre la comunidad académica en 1995.

7.9.1. Técnicas de Minería de datos

Regresión: Función que busca mapear a los elementos de una muestra para obtener sus características he inferir en base a ellas. Su objetivo es el de predecir el valor que tomará una variable cuantitativa en base a la información que se ha ganado de muestras previas similares a las de donde proviene la variable por predecir.

Asociación: Busca identificar patrones o hechos frecuentes que ocurren en un grupo de datos determinado (aprendizaje máquina). Su objetivo es encontrar relaciones no explícitas entre variables. Por ejemplo, si una persona compra un boleto de avión y reserva una habitación, por asociación, es probable que también rente un automóvil.

Agrupación (Clustering): Se emplea para la obtención de grupos naturales de los datos a partir de criterios, por lo general distancia o similitud entre ellos; no es necesario que exista una definición previa de los criterios para poder realizar el agrupamiento (aprendizaje máquina). Forma parte de las tareas descriptivas, y su objetivo es el de encontrar un número finito de categorías para, precisamente, describir a los datos dentro de cada grupo creado.

Correlación: Se utiliza para obtener el grado de dependencia entre una determinada variable y otra dentro de un grupo de datos, esta puede ser positiva, negativa o simplemente no existir dependencia entre las variables.

Resumen: Involucra métodos para encontrar maneras compactas de describir los subconjuntos de características en un grupo de datos. Por ejemplo, se pueden tabular las medias y desviaciones estándar para todos los campos del grupo, así tener una visión rápida de estas características del grupo de datos.

Análisis de atípicos: Estas técnicas se utilizan para la identificación de comportamientos fuera de lo normal y que no cumplen con los criterios generales dentro del grupos de datos. Un ejemplo es su utilización en los bancos para la detección de compras fuera de lo normal con tarjetas de crédito, que posiblemente fueron robadas o clonadas.

7.9.2. Algoritmos más utilizados para el diagnóstico y predicción de enfermedades

En la tabla 7.6 se muestran algunos de los algoritmos de minería de datos clasificados según su tipo.

Supervisados	No Supervisados
Árboles de Decisión	Detección de Desviaciones
Inducción Neuronal	Segmentación
Regresión	Agrupamiento (Clustering)
Series Temporales	Reglas de Asociación
	Patrones Secuenciales

Tabla 7.6. *Clasificación de algoritmos según su tipo (Moreno García et al., 2001).*

De este ejemplo de clasificación y de la revisión de trabajos de varios autores, se describirán algunos de los algoritmos más utilizados en la predicción y diagnóstico de enfermedades, dada la naturaleza de la presente investigación. Esfandiari et al. (2014) por ejemplo, dicen que las tareas de minería de datos que mejores aplicaciones en la salud tienen son las de Asociación, Regresión, Agrupación, entre otras, por lo que los algoritmos que pertenezcan a ellas tendrán preferencia en este caso.

- **Arboles de Decisión**

Existe un gran número de algoritmos de este tipo, uno de los más ampliamente utilizados por su versatilidad y capacidad para manejar valores faltantes o creación de

reglas es el llamado C4.5 (conocido como J48 en Weka), el cual es una extensión del que se conoce como ID3. Es una máquina de aprendizaje para la predicción con una variable dependiente que busca llegar a un objetivo deseado basándose en los atributos disponibles. Los nodos internos están conformados por los distintos atributos enseñados, los nodos finales u hojas son las posibles clasificaciones a la que puede pertenecer el dato de entrada. Se basa en una serie de iteraciones, cada vez que se realiza una corrida el algoritmo aprende y mejora su precisión, este se detiene una vez que todos los datos han sido clasificados. El árbol de decisión J48 posee una alta precisión en comparación con otros clasificadores como el de Naive Bayes (Abdar *et al.*, 2015; Suca *et al.*, 2016).

- **K Vecinos Próximos (K-NN)**

Es uno de los algoritmos más utilizados para reconocimiento de patrones, minería de datos, bases de datos y aprendizaje máquina, esto en gran medida debido a su simplicidad, alta precisión y a que reemplaza valores faltantes con los del vecino más cercano, se encuentra en el top 10 de los algoritmos de minería de datos. Se trata de un clasificador que se basa en la similitud de los datos de entrada, está en la categoría de “perezoso”, ya que no realiza una etapa de entrenamiento, sino que aprende hasta que se realiza una corrida (Thirumal y Nagarajan 2015). Funciona midiendo las distancias de nuevas entradas con los datos ya existentes basándose en su similitud, los que se encuentran más cerca son llamados “Vecinos” y se agregan a la clase donde existan más vecinos con la menor distancia entre ellos. El algoritmo también es capaz de calcular valores continuos al utilizar el valor promedio de los vecinos más cercanos y realizando una predicción sobre la nueva entrada (Abdar et al. 2015).

- **Regresión**

Este tipo de algoritmos tiene como objetivo el definir una función (fórmula matemática) que permita predecir el valor de una variable continua (real) basándose en otros atributos de un grupo de datos. Existen formas de calcular valores discretos con la ayuda de algoritmos genéticos o algoritmos de enumeración refinados. Algunas de las

aplicaciones son el predecir una cantidad de habitantes de una localidad basándose en las cantidades de otras localidades conocidas y comparables, estimar la probabilidad de que un paciente no muera dados los resultados de pruebas de diagnóstico, predecir la demanda de un consumidor por un nuevo producto, entre otros (Beltrán-Martínez 2003). Algunos tipos son de Regresión Lineal Global (clásica), Regresión Lineal Ponderada Localmente, Regresión No Lineal.

- **Redes Neuronales**

Basado en la forma en que se interconectan las neuronas del cerebro humano, se trata de un algoritmo de procesamiento de datos que incluye un gran número de “pequeños” procesos que da soporte a este procesado de datos. Actúa como una red interconectada de procesos paralelos uno con el otro para la solución de un problema (Abdar et al. 2015). Se realiza un aprendizaje mediante el entrenamiento para realizar una predicción, este aprendizaje es autónomo y no programado como en otros algoritmos. Se pueden tener múltiples capas, y para que la información pase de una neurona a otra se debe de llevar a cabo una cierta “estimulación” que desencadene una reacción basada en lo aprendido. Se puede ver a este algoritmo como una serie de nodos interconectados en capas, al menos una de entrada, una capa oculta y una de salida que se basan en programación de retroceso (el resultado de salida es comparado con la salida esperada para calcular una señal de error el cual es regresado hacia atrás a la capa oculta hasta que todas las neuronas conocen este error y aprenden, propiciando un reorganización y ayudando a que cada neurona reconozca distintas características de los nuevos datos de entrada), en cada una de las neuronas intermedias se realizan procesos, los cuales determinan la ruta que deberá de seguir la información hacia otra de las neuronas conectadas.

- **Clasificadores Bayesianos**

El algoritmo busca clasificar los elementos del grupo de datos aplicando el teorema de Naive Bayes, en el cual intervienen una serie de probabilidades condicionadas y distribuciones de probabilidad. Los elementos se clasificarán en una determinada clase

según su probabilidad de pertenecer a esta (Thirumal y Nagarajan 2015). Se requiere una pequeña cantidad de datos para realizar el entrenamiento y determinar los parámetros que definen a cada una de las clases, es ideal para grandes cantidades de datos por su velocidad de ejecución y alto nivel de precisión.

El algoritmo de Naive Bayes mide las relaciones entre los datos de un grupo determinado. Este algoritmo ayudará a crear un modelo para la clasificación y predicción, suele utilizarse para resolver problemas, identificar y clasificar datos entre distintas clases basándose en probabilidad (Sa'di et al. 2015).

7.10. Software para Minería de Datos

WEKA: Es una herramienta que agrupa una colección de algoritmos de aprendizaje máquina para tareas de minería de datos. Es sumamente flexible, ya que los algoritmos se pueden utilizar directamente en un grupo de datos o incluirlos en un código propio JAVA (lenguaje de programación), también permite el desarrollo de nuevos esquemas de aprendizaje máquina. Dentro de las técnicas soportadas se encuentran aquellas para:

- Preprocesamiento de datos.
- Clasificación.
- Regresión.
- Agrupamiento (Clustering).
- Asociación.
- Visualización de resultados.

Otra de las ventajas es que es de acceso libre y puede ser descargado desde https://www.cs.waikato.ac.nz/ml/weka/downloading.html. Esto ha causado que se convierta en uno de los softwares más utilizados en investigaciones que involucran minería de datos, así que existe amplia documentación y gran cantidad de tutoriales y cursos disponibles.

Según un análisis realizado por la empresa española Stratebi (2010), especialista en soluciones tecnológicas Open Source (software libre), se evaluaron 6 distintos softwares tomando en cuenta su capacidad de soportar alguno de los 198 algoritmos que se revisaron, como resultado WEKA soporta 168 de ellos, convirtiéndola en una de las herramientas más utilizadas en investigaciones de este tipo.

MATLAB: Se trata de un software de paga (requiere la compra de una licencia de uso) que integra la programación, matemáticas y visualización para el análisis de datos o el desarrollo de algoritmos personalizados. Cuenta con un lenguaje de programación propio (llamado "M") que puede trabajar con vectores y matrices, funciones, programación orientada a objetos, entre otros. Permite su utilización en un entorno de escritorio visual o mediante un script (archivo que contiene una serie de órdenes para su procesamiento). Se puede descargar una versión de prueba de en https://www.mathworks.com/programs/trials/trial_request.html?procode=ML.

Cuenta con "cajas de herramientas" que agrupan distintas características y programas especiales para el desarrollo de actividades en particular, algunas de estas cajas son:

- Matemáticas, Estadística y Optimización.
- Sistemas de Control.
- Generación de Código.
- Simulación en Tiempo Real y Gráfica.

En minería de datos cuenta con múltiples cajas de herramientas que han sido desarrollada y probadas exhaustivamente por profesionales (ya que es software de paga) para un análisis de calidad, estas van desde herramientas para usuarios de Excel, hasta aquellos que buscan encontrar conocimiento en campos como el aeroespacial, farmacéutica, y demás. Entre algunas de las técnicas de minería de datos que maneja se encuentran:

- Análisis exploratorio.
- Clasificación.

- Árboles de Regresión.
- Redes Neuronales Artificiales.

Otra de las características de este software es que permite ver cómo funcionan distintos algoritmos sobre un determinado grupo de datos, realizar tantas iteraciones como sea necesario hasta alcanzar los resultados esperados y generar automáticamente un programa de MATLAB para reproducir o automatizar el trabajo.

ORANGE: Es un software para minería de datos que se centra en la visualización de datos para encontrar patrones ocultos, proveer de un análisis intuitivo, permitir la comunicación de resultados, y más. Algunas de las herramientas de visualización son:

- Diagramas de dispersión.
- Diagrama de caja e histograma.
- Dendrograma (árboles de decisión).
- Diagrama de siluetas.

Otras herramientas que pueden ser agregadas e incluir aquellas para representar redes, mapas geográficos, entre otras.

Entre las herramientas que soporta se encuentra aquellas para el modelado de datos y preprocesamiento, algunas de las técnicas para minería de datos con las que trabaja son:

- Clasificación.
- Regresión.
- Agrupación (Clustering).
- Proyección y Evaluación.

Es ideal para principiantes, pero con las capacidades que cualquier experto puede requerir. Es software libre y se puede descargar de su sitio web para distintos sistemas operativos https://orange.biolab.si/download/, junto con toda su documentación,

cursos y ejemplos prácticos. Otra de sus ventajas es la elaboración automática de reportes (en .pdf, .html) para su impresión o utilización de las gráficas elaboradas.

7.11. GIS

Los GIS también puedes ser utilizados para el perfilamiento de una comunidad o zona en una ciudad, es decir, conocer las características y carencias del lugar, y así tenerla catalogada y enfocar los esfuerzos específicos que esta necesite. Por ejemplo, el perfilamiento permitirá determinar las mejores ubicaciones para llevar a cabo campañas de salud, toma de decisiones sobre los servicios de salud que se necesitan en una comunidad, mejor aprovechamiento de infraestructura para el mejoramiento de la calidad de vida (parques, centros deportivos, gimnasios), entre otros aspectos.

7.12. Uso de redes Sociales en la salud

Otro de los usos de las RS por parte de las instituciones de salud es la realización de investigaciones en diversas líneas, en su mayoría existen dos principales tipos, el primero involucran la utilización de las RS para difundir contenido informativo y recomendaciones para lograr un impacto positivo en la salud a través del cambio de hábitos poco beneficiosos y la prevención de enfermedades como la diabetes tipo 2 y problemas de sobrepeso u obesidad; el segundo es el monitoreo de las RS para llevar a cabo vigilancia del contenido que circula por las distintas plataformas, y mediante el uso de técnicas de análisis como minería de datos, visualizar tópicos sobre temas de salud que estén afectando a una determinada población, e inclusive detectar el brote de epidemias basándose en la información que circula por RS (Stetson et al. 2017; Paul et al. 2016; Napolitano et al. 2013). El número de estudios donde se utilizan RS para influenciar cambios en los hábitos de las personas que favorezcan su salud es muy grande, por lo que esta tendencia a utilizar medios de comunicación de interacción social electrónicos, como Apps, tiene aún bastante por aportar.

Las aplicaciones móviles (Apps), son paquetes precargados de contenido interactivo instalado en teléfonos inteligentes u otros dispositivos que los usuarios llevan consigo, y que a diferencia de los sitios web tradicionales que necesitan de una conexión constante a internet, las Apps permiten acceder a ellas aun y cuando no se cuenta con

conexión. Al igual que en las RS, los temas de salud también están siendo utilizado para el desarrollo de Apps, esto debido a que se enfocan en la personalización de contenido para el usuario (Brusse et al. 2014). En un estudio realizado por Kumar et al. (2015) donde se buscaba alentar el cuidado y visitas al médico para el diagnóstico en personas con tendencia a padecer diabetes tipo 2 mediante la recepción de recordatorios vía dispositivos móviles, se tuvo una mejora significativa en el seguimiento del diagnóstico de aquellos que recibieron los recordatorios, en comparación con los sujetos de control que no los recibían.

7.12.1. Ventajas, desventajas y métricas del uso de redes sociales

Primeramente, se observan los pros y contras más comunes que se pueden presentar cuando se decide utilizar alguna red social para la difusión de contenido.

Ventajas.

- Mejoran la interacción paciente-médico, y en general con las instituciones de salud al permitir el intercambio de puntos de vista y comentarios.
- Estimulan la motivación del paciente.
- Permite la comunicación de alertas oportunas.
- Proveen información verídica sobre el cuidado de la salud.
- Crean grandes comunidades que enriquecen y favorecen la generación de conocimiento.
- Promueven la comunicación entre profesionales de la salud en todo el mundo.
- Influyen en el comportamiento de los pacientes a través de la influencia social, imitando hábitos saludables.

Desventajas:

- Publicación de material y contenido ofensivo, inapropiado para el público o que sea poco sensible al tratar un tema en particular.
- Problemas con el control de la privacidad, invasión de esta.
- Relaciones paciente-médico más allá de lo profesional, lo que puede llegar a afectar el criterio del profesional de la salud.

- Pueden existir conflictos de intereses entre quienes publican y las bases de la institución a las que representan.
- Pérdida de tiempo al invertirlo en acciones poco productivas o de ocio.

Estas son sólo una muestra de las cualidades e inconvenientes que implica la utilización de RS en campañas publicitarias de cualquier tipo.

➢ Métricas para la evaluación de la difusión de contenido en RS

Una vez que las ventajas y desventajas de la utilización de aplicaciones móviles y redes sociales se conocen y son comprendidas, debe de existir una manera de determinar el impacto, ya sea positivo o negativo, que se tuvo en el público objetivo con la finalidad de evaluar si el mensaje fue comprendido y tiene la posibilidad de ser aplicado en el cambio de hábitos y mejora de la salud. Chung (2015) después de una revisión considerable de artículos, donde el tema central era la forma de evaluar el impacto de las RS, logró sintetizar tres de las métricas clave para medir y evaluar las campañas basadas en RS, las cuales fueron adaptadas a la terminología estándar que se maneja en distintas plataformas de RS (figura 7.2).

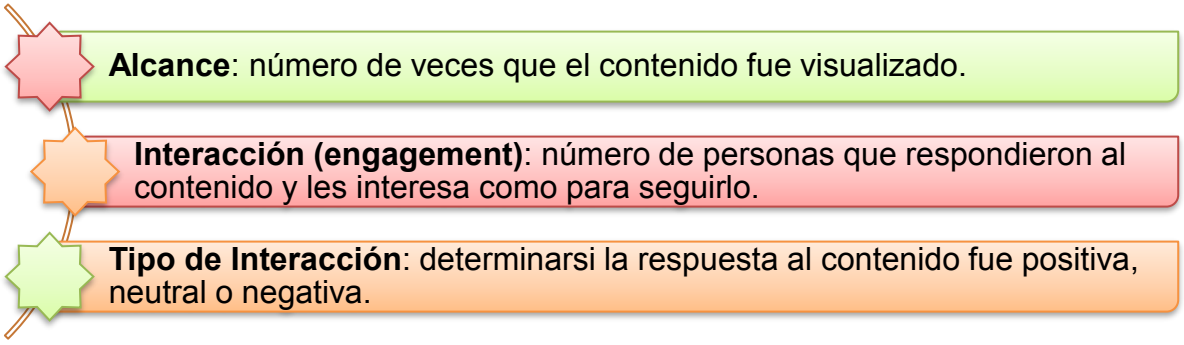


Figura 7.2. *Principales Métricas para la evaluación de campañas basadas en RS.*

El Alcance está directamente relacionado con el número de usuarios que utilizan la plataforma elegida y han visualizándolo el contenido difundido, independientemente del dispositivo utilizado o si el contenido causó alguna reacción; la Interacción representa a todo tipo de respuesta o reacción que el usuario tuvo con el contenido

publicado (número de “Me Gusta”, “Retweets”, “Compartir”) y las convierte en valores numéricos procesables (Veale et al. 2015). La utilización de una o algunas de ellas dependerá del tipo de plataforma(s) utilizada(s) para la difusión de las campañas, así como del tipo de información que se requiera recopilar para la evaluación de resultados.

Las tres principales métricas presentadas son aplicables a cualquier tipo de material que haya sido difundido vía RS, por lo que pueden ser utilizadas para evaluar el impacto que una campaña de difusión ha tenido, los resultados obtenidos ayudarán a mejorar el material y contenido informativo, el trabajo que se esté realizando en el diseño, tipo de acercamiento con el público objetivo y el nivel de respuesta o aceptación obtenido. Existen muchas otras métricas que también pueden ser utilizadas, como por ejemplo aquellas que hacen referencia a las “horas pico” o de mayor tránsito en las plataformas, los días de la semana de mayor visualización del contenido, el rango de edades de usuarios que fueron expuestos o reaccionaron al contenido, entre otras.

7.12.2. Consideraciones para la difusión y uso de redes sociales

Los criterios que se deben de considerar cuando se van a utilizar RS, en especial cuando se trata de su uso en instituciones públicas que manejan información sensible, van desde aquellos que surgen del sentido común de quien genera y publica el contenido, hasta normas gubernamentales que rigen la difusión de contenido electrónico y de salud.

Cada una de las instituciones cuentan con criterios propios sobre lo que se publica y cómo se publica, haciendo necesaria una revisión de sus reglamentos o acudiendo al área de Comunicación Social, Prensa, o quien se encargue de aprobar el material informativo en busca de los lineamientos que deben de ser cumplidos. También se necesita asegurar que lo creado sea lo suficientemente bueno para tener un impacto en el público, por lo que el diseño juega un papel importante.

Una vez que se cumplen con los criterios de contenido de la institución, se debe de considerar la forma en que este será presentado ante los distintos grupos de enfoque que son objetivo de las campañas de salud, lograr aceptación y un impacto positivo. Entre estas consideraciones se encuentran las que Veale et al. (2015) enumeran en su estudio y están relacionadas con la publicación constante de contenido, la interacción directa con los usuarios al responderles de forma individual, alentar a la interacción al generar discusión e intercambio de puntos de vista, utilizar multimedia y acceso a fuentes adicionales para ampliar la información publicada.

7.13. Clasificaciones utilizadas para los afiliados

La tabla 7.7 muestra las distintas clasificaciones en las que un afiliado puede ser encasillados.

#	CLASIFICACIÓN DE AFILIADOS
1	HIJO(A)
2	ESPOSA(O)
3	TRABAJADOR(A)
4	PENSION DERIVADA
5	ESTUDIANTE
6	JUBILADO
7	PENSIONADO
8	MADRE SIN ARANCEL
9	CONCUBINA
10	INCAPACITADO
11	VIGENTE C./ GOCE DE SERV. MEDICO
12	ESPOSO INC.
13	PADRE SIN ARANCEL
14	PENSION EN TRAMITE
15	PENSION DER. POR OTROS ORGS.
16	PENSION POR VIUDEZ EN TRAMITE
17	SERV. MED. APROB. POR JUNTA DIRECTIVA
18	PENSION POR OTROS ORGS.
19	JUBILACION EN TRAMITE
20	PENSION POR ORFANDAD EN TRAMITE
21	MADRE (SISMP)
22	MADRE ARANCELADA
23	PADRE ARANCELADO
24	PADRE (SISMP)
25	PENSION POR ASCENDENCIA EN TRAMITE
26	PADRE CON PENSION/ASCENDENCIA
27	PENSION POR INVALIDEZ TEMPORAL

Tabla 7.7. *Clasificación institucional de afiliados.*

7.14. Georreferenciación de colonias con más diagnósticos

La figura 7.3 muestra el mapa con las 10 colonias de mayor incidencia de obesidad en Hermosillo.

7.14.1. Obesidad

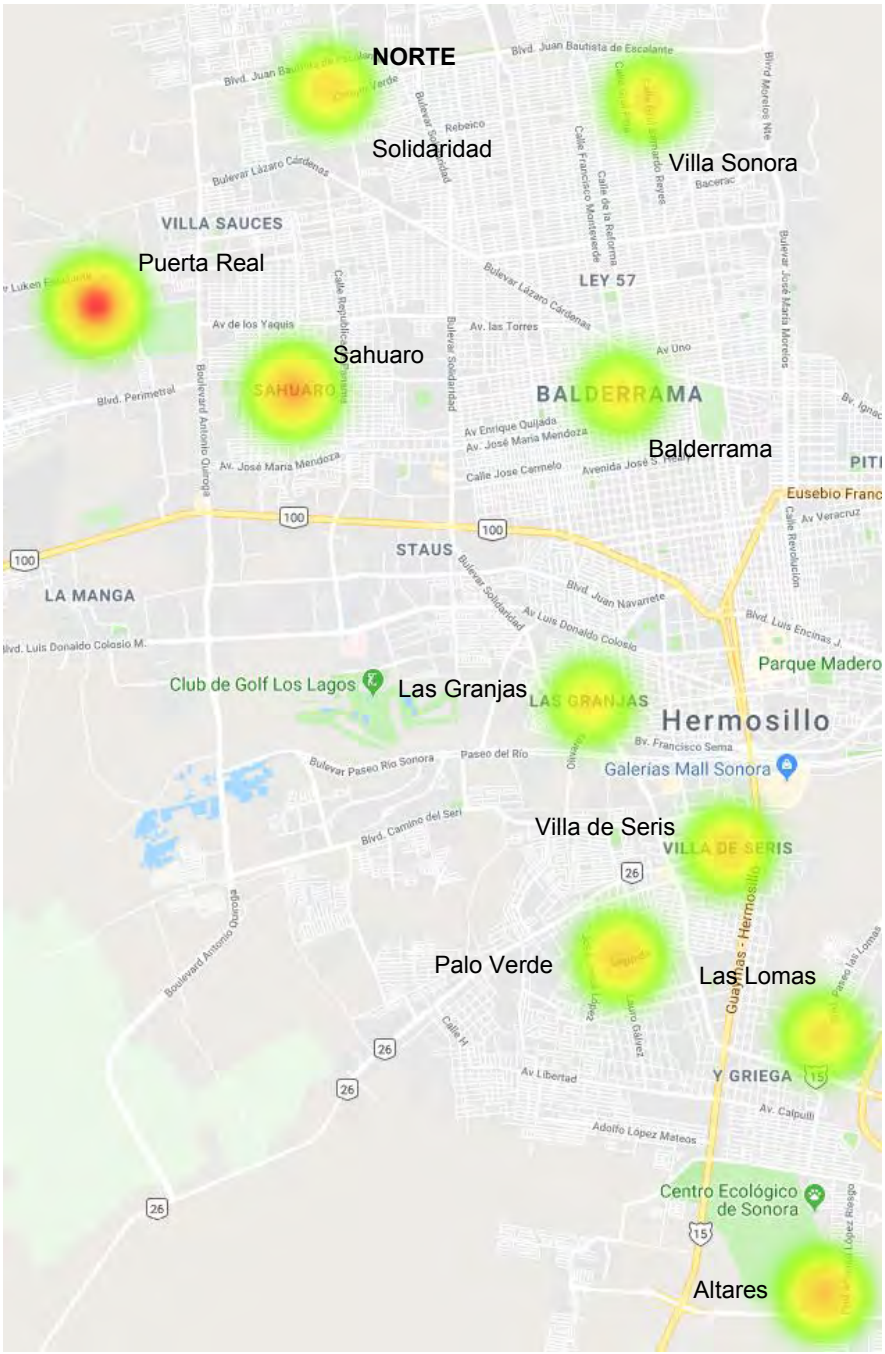


Figura 7.3. *10 colonias con mayor incidencia de obesidad en Hermosillo.*

7.14.2. Diabetes

La figura 7.4 muestra el mapa con las 10 colonias de mayor incidencia de diabetes en Hermosillo.

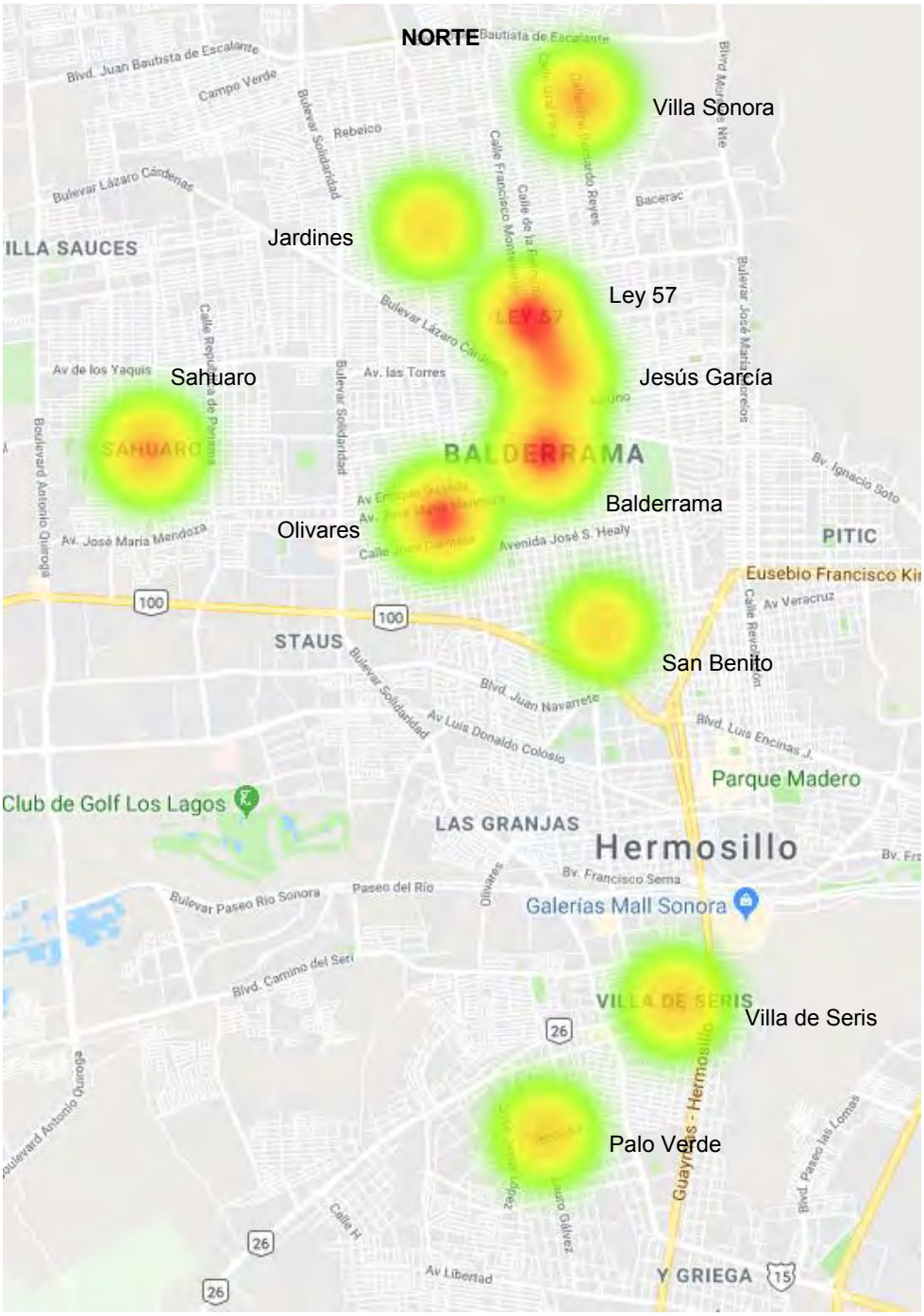


Figura 7.4. *10 colonias con mayor incidencia de diabetes en Hermosillo.*