Metodología CRISP-DM en la gestión de proyecto de Data Mining. Caso enfermedades dermatológicas

Magda Gabriela Sánchez Trujillo
Escuela Superior Tepeji del Rio, Universidad Autónoma del Estado de Hidalgo.
magdags@uaeh.edu.mx

José Ángel Pérez Hernández
Tecnologías de la información y comunicación, Universidad Tecnológica Tula Tepeji
joseangel.perez@uttt.edu.mx

Resumen
Hoy en día todas las organizaciones y/o empresas almacenan una gran cantidad de información, que ha llegado a exceder la habilidad del personal para analizar, resumir e interpretar los datos, dando lugar a la técnica Data mining, con objeto de predecir de forma automatizada tendencias y comportamientos de modelos. El caso se realiza en una clínica dermatológica en el estado de Hidalgo, México, la cual busca determinar a partir de datos históricos el comportamiento de las enfermedades dermatológicas en los próximos años, indicando el lugar donde se presentan el mayor número de casos, edad y sexo. Los datos con los que se integra la gestión del proyecto abarcan los años 2013-2020, lugar de residencia de los pacientes, padecimientos por entidad y municipio. Así, el propósito del presente caso es aplicar la metodología de gestión de proyectos CRISP-DM (Cross Industry Standard Process for Data Mining), la cual proporciona una idea clara de la estructura el ciclo de vida del proyecto de data mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto: análisis del problema, análisis de datos, preparación de datos, modelado, evaluación y seguimiento, a fin de planificar, dirigir y dar seguimiento al proyecto. Los resultados permiten identificar las enfermedades dermatológicas con mayor incidencia en los próximos años (2021-2025) por diagnóstico, edad, sexo.

Palabras Clave: Gestión de proyectos, minería de datos, modelado de datos, regresión lineal, equipo de proyecto.

CRISP-DM methodology in Data Mining project management. Case skin diseases

Abstract
Today all organizations and / or companies store a large amount of information, which has exceeded the ability of staff to analyze, summarize and interpret data, giving rise to the technique of data mining, in order to predict in a way automated trends and model behaviors. The case is performed in a dermatology clinic in the state of Hidalgo, Mexico, which seeks to determine from historical data the behavior of dermatological diseases in the coming years, indicating the place where the largest number present case, age and sex. The data with which the project management is integrated cover the years 2013-2020, place of residence of the patients, conditions by entity and municipality. Thus, the purpose of this case is to apply the project management methodology CRISP-DM (Cross Industry Standard Process for Data Mining), which provides a clear idea of the structure of the life cycle of the data mining project in six phases, that interact with each other iteratively during the development of the project. analysis of the problem, data analysis, data preparation, modeling, evaluation and monitoring, in order to plan, manage and monitor the project. The results allow identifying the dermatological diseases with the highest incidence in the coming years (2021-2025) by diagnosis, age, and sex.

Keywords: Project management, data mining, data modeling, linear regression, project team

1. INTRODUCCIÓN

Actualmente las organizaciones almacenan gran cantidad de información, la cual ha llegado a exceder la habilidad de analizar, resumir e interpretar los datos. El desarrollo tecnológico presenta otras posibilidades para optimizar esos datos con el término de *Data Mining*. Esta herramienta es un proceso complejo de extracción, el cual requiere la aplicación de una metodología estructurada para que su aplicación sea ordenada y eficiente.

Los proyectos con Data Mining son aplicables en cualquier campo o disciplina y tienen como objetivo extraer información útil a partir de grandes cantidades de datos (Molina, 2006). En este capítulo se presenta la aplicación de la metodología de gestión de proyectos CRISP-DM (Cross Industry Standard Process for Data Mining), la cual proporciona una idea clara de la estructura el ciclo de vida del proyecto de data mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto: análisis del problema, análisis de datos, preparación de datos, modelado, evaluación y seguimiento, a fin de planificar, dirigir y dar seguimiento al proyecto (Roman, 2016).

Lo anterior permite establecer un contexto claro de datos para la elaboración del modelo de predicción utilizando el software libre Waikato Environment for Knowledge Analysis (WEKA) mediante técnicas de preprocesado, clasificación, agrupamiento y asociación.

Respecto a las lesiones cutáneas y trastornos dermatológicos, la investigación se orienta a identificación de enfermedades epidérmicas para su tratamiento. La piel tiene abundantes lesiones importantes que suelen reconocerse por medios clínicos. La piel constituye el órgano más extenso y externo, que con sus múltiples funciones contribuye a asegurar el funcionamiento del organismo humano, su vida y su salud, además de proteger del ambiente a todos los órganos y aparatos del cuerpo. Su importancia no radica sólo en su función protectora, sino en su complejo trabajo fisiológico que realiza. Se reconocen diferentes tipos de piel: seca, grasosa, deshidratada, hidratada y mixta (Arenas, 2015). Estos tipos están

dados por el grado de hidratación, la edad, sexo y por factores individuales, nutricionales y externos que la afectan, como la contaminación y el sol.

De acuerdo a datos del estudio de (Canul, 2020), los padecimientos dermatológicos más frecuentes en la población mexicana es el cáncer de piel: carcinoma basocelular (65-74%), seguido del carcinoma epidemoide o de células escamosas (14-23%) seguido por el melanoma (3-6.5%), el resto de las neoplasias malignas cutáneas escila entre 5.5 a 9%. La incidencia es mayor en mujeres, caso contrario a lo encontrado en la literatura internacional donde se presentan más casos en la población masculina. Estudios recientes, han identificado factores psicológicos relacionados con enfermedades de la piel, reconociendo que el estrés, ansiedad y depresión pueden originar trastornos como el acné, la rosácea y las verrugas (Antuña y García, 2020), psoriasis (González 2008, Pezzarossa, Ciani y Bassi, 2015) y casos de alopecia (González, Méndez, Sánchez, 2019). Otro aspecto que afecta la salud cutánea es la exposición repentina del cuerpo a cambios o excesos en el régimen alimenticio (Murieta, 2014).

El caso se realiza en una clínica dermatológica en el estado de Hidalgo, México, con el objetivo de aplicar la metodología CRISP-DM a partir de datos históricos y pronosticar el comportamiento de las enfermedades dermatológicas en los próximos años por entidad federativa, indicando el lugar donde se presentan el mayor número de casos, edad y sexo.

Los datos con los que se integra la gestión del proyecto abarcan los años 2013-2020, contiene lugar de residencia de los pacientes, padecimientos por entidad y municipio. De esta forma, se busca obtener conclusiones y hacer predicciones lo más fiables posibles que ayuden a la toma de decisiones en cuanto a los servicios y atención que ofrece la clínica a sus pacientes.

Para cumplir con el objetivo del proyecto, el presente documento está integrado de la siguiente manera. En la primera parte se muestran los conceptos básicos sobre minería de datos y sus aplicaciones en el proceso de manejo y extracción de información. Posteriormente se explica la metodología seleccionada, donde se relatan las actividades llevadas a cabo, finalmente se analizan e interpretan los resultados.

2. Revisión de Modelos Data mining

En este apartado se presentan las principales metodologías utilizadas para la realización de proyectos Data minning. SEMMA (Sample, Explore, Modify, Model, Assess) y CRISP-DM, ambas metodologías de acuerdo a la encuesta KDnuggets (2018) son las más utilizadas (71%) por analistas de proyectos que tienen como objetivo extraer información útil a partir de grandes cantidades de datos.

SEMMA

Utiliza para su desarrollo cinco fases:

1. Extracción de la muestra representativa, que permita validar el modelo y los resultados. En este punto, se considera además el nivel de confianza.

2. Exploración, permite examinar la información disponible a fin de optimizar la eficiencia del modelo, aquí se hace uso de técnicas estadísticas para visualizar la relación entre las variables de estudio.

3. Manipulación, con base en la exploración se definen las variables de entrada del modelo

4. Modelado, en este punto se establece la relación entre variables que posibiliten inferir el valor de las mismas con el nivel de confianza establecido, aquí se pueden utilizar técnicas de regresión, análisis discriminante, arboles de decisión, etc).

5. Valoración, consiste en realizar o corroborar los resultados contrastando con medidas de bondad de ajuste del modelo estadístico.

CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) (Chapman, 2000), es un modelo de referencia que describe de forma general las fases, tareas generales y salidas de un proyecto de Data Mining en general. Por otro lado, la guía del usuario proporciona información a detalle sobre la aplicación del modelo a proyectos de Data Minning específicos, se proporcionan guías y listas de cotejo sobre las actividades y tareas correspondientes a cada fase. De esta manera se estructura el ciclo de vida de un proyecto en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del mismo.

Comprensión del negocio
Es la fase de inicio, se identifica el objetivo del negocio, cual es problema y que se busca resolver las tareas, propósitos y requisitos del proyecto desde una perspectiva de negocio, que se traducen en objetivos técnicos para plasmarlos en un plan del proyecto.

Comprensión de los datos
Se establece el primer acercamiento para familiarizarse con el problema, en esta fase de recolectan los datos iniciales, se describe su significado, se procede a explorar los datos, es decir aplicar pruebas estadísticas básicas e identificar la calidad de los datos y plantear las hipótesis.

Preparación de datos
Después de la recolección inicial, se seleccionan, limpian, formatean e integran los datos

Modelado
En esta fase se elije la técnica en función de: disposición de datos, conocimiento de la técnica, pertinencia con el problema y cumplir con los requisitos. Se realiza un plan de prueba para generar el modelo y evaluarlo.

Evaluación
Se evalúan los resultados, se revisa el proceso para identificar deficiencias. Si el modelo generado es válido en función de los criterios establecidos en la fase anterior, se procede a la implementación del modelo

Implementación
En esta fase se propone contar con estrategias para monitorear y mantener el modelo.

Herramientas

En este apartado se describe la herramienta de apoyo para la minería de datos que en este caso fue WEKA (Waikato Environment for Knowledge Analysis) (Waikato, 2011) es un software libre, escrito en lenguaje Java. WEKA integra un conjunto de herramientas de visualización y algoritmos para el análisis de datos y modelado predictivo, a su vez cuanta con una interfaz gráfica para poder acceder fácilmente a sus funcionalidades. Una de las ventajas de WEKA es que se puede ejecutar en cualquier plataforma y resulta altamente portable, además, WEKA contiene un amplio repertorio de técnicas para preprocesamiento de datos, como, son: selección de atributos, tratamiento de valores desconocidos y transformación de atributos numéricos.

A partir de información que se encuentra organizada en una base de datos, se llevan a cabo varios procesos iterativos que se denomina fase de extracción de conocimiento Knowledge Discovery from Databases (KDD) (Hernández, Ramírez y Ferri, 2004), que consta de cinco etapas centrándose para efectos del presente trabajo en la minería de datos, incluyendo, cómo se seleccionan, limpian, transforman y almacenan estos datos, a fin de evaluar e interpretar las conclusiones del estudio. A continuación, se describen las etapas de extracción de datos:

1. Integración y recopilación de datos. Consiste en obtener datos de fuentes
2. Selección, limpieza y transformación. Se realiza selección de datos relevantes, corrección y se eliminan datos incompletos
3. Minería de datos. Se decide cuál es la tarea (agrupar, clasificar, etc.) que se va a realizar y se elige el método que se va a utilizar para ello.
4.- Evaluación e interpretación. Se identifican e interpretan patrones
5. Difusión y uso. Se utilizan los resultados

Una vez que se obtienen los patrones, existen diversas tareas y requisitos de minería de datos, las que se dividen en, predictivas o descriptivas. En las predictivas se estiman valores futuros o desconocidos de la(s) variable(s) de interés a partir de otras variables independientes (predictivas). El objetivo de las tareas descriptivas es identificar patrones en los datos que los explican o bien resumen.

DATA MINNING BERRY Y LINOFF

Es un proceso de negocio desarrollado por Berry y Linoff en 1997 que consta de cuatro etapas: identificar el problema, transformar datos en información, tomar acciones, medir resultados.

KDD (Knowledge Discovery and Data Mining)

Es una metodología asiática propuesta en 1996 por Motoda y consta de cinco fases: selección, pre procesamiento de datos, transformación, minería de datos, evaluación e implementación.

Comparación de las metodologías

Las metodologías utilizan Data mining y sus fases se interrelacionan, de tal forma que son procesos iterativos e interactivos. De esta manera la principal diferencia radica en que SEMMA y BERRY y LINOFF y KDD, se ajustan en las características técnicas del proceso, en tanto que CRISP-DM presenta una perspectiva mas completa al incorporar el conocimiento y los objetivos del negocio. De esta manera se comienza analizando el problema de la organización, empresa o proyecto, lo que hace mas fácil integrarla como una metodología de gestión de proyectos que contemple tareas técnicas y administrativas. Aunado a lo anterior CRISP-DM a diferencia de SEMMA, BERRY y LINOFF y KDD puede

utilizar herramientas de software libre como WEKA utilizado en el presente estudio. Mientras que las otras metodologías se limitan a los productos de SAS (Analítica, inteligencia artificial y gestión de datos) por sus siglas en inglés.

3. METODOLOGIA Y RESULTADOS

En esta etapa se aplica cada una de las fases de la metodología CRISP-DM al problema descrito.

1.- Objetivos del negocio

El objetivo del negocio es determinar, a partir de datos históricos, el comportamiento de las enfermedades dermatológicas en los próximos años, indicando el lugar en donde se presentan más casos, así como la edad y el sexo con mayor afectación.

1.1Evaluación de la situación

Para esta fase un aspecto importante es contar con información suficiente para tener una visión clara y poder ofrecer una solución al problema, así la clínica proporcionó La base de datos con información de 6 años.

1.2.- Objetivos de la minería de datos

- Predecir el número de casos por enfermedad dermatológica por año
- Predecir las enfermedades de mayor incidencia por entidad federativa
- Identificar los rangos de edad de los casos más comunes
- Identificar la afectación de hombres y mujeres

1.3.- Plan del Proyecto

Etapa 1. Análisis de la estructura y base de datos
Etapa 2. Ejecución con muestras
Etapa 3. Preparación de datos (limpieza, selección, etc.)
Etapa 4. Elección de técnica de modelado
Etapa 5. Análisis de resultados
Etapa 6. Elaboración de informes
Etapa 7. Presentación final

1.4.-Paso seguido se recolectaron los datos iniciales para separar la problemática en variables, es decir el diagnóstico, la entidad federativa sexo y edad (tabla 1).

Tabla 1. Variables

Diagnostico	Entidad federativa	Sexo	Edad
Fecha de diagnostico Enfermedad diagnosticada, número de pacientes que lo padecen	Lugar donde fue diagnosticado Número de pacientes diagnosticados	Numero de pacientes femeninos, masculinos, enfermedades según sexo	Numero de pacientes por rango de edad Relación edad-diagnóstico

Fuente. Elaboración propia

2. Descripción de los datos

Pacpaciente. Esta tabla se compone de los campos necesarios para identificar el paciente a quien se le dio cierto diagnóstico. Los datos a ocupar son: id y sexo

Exp-diagnostico. Cada diagnóstico registrado cuenta con un id, valor diagnóstico, fecha diagnóstico y edad.
Para la edad se manejan los siguientes valores:

0-1
1-5
6-10
11-20
21-30, etc.

Comdirección. Esta tabla almacena las direcciones del paciente, lo cual servirá para conocer en qué entidad federativa se dan más los diagnósticos. Y de esta tabla se tomará id y entidad federativa.

2.1. -Exploración de datos

Una vez que se mapeó la base de datos transaccional a ocupar, se fue analizando el contexto de la problemática a resolver. Se hizo una revisión a cada uno de los registros para poder determinar que no existieran valores nulos, es decir, que el campo tuviera información válida.

3. Preparación de los datos

Aquí se preparan datos de la información obtenida el número de pacientes, diagnósticos encontrados, edad de paciente, fecha del diagnóstico, entidad federativa, etc., para que en la siguiente etapa pueda llevarse a cabo el proceso, evitando cualquier situación a causa de datos inválidos y así poder aplicar las técnicas de minería de datos.

3.1 Seleccionar los datos

De acuerdo al caso planteado se seleccionaron los datos útiles para cada punto planteado (diagnósticos, entidad federativa, edad y sexo, además de número de casos por año).

3.1.2 Limpiar datos- Consiste en eliminar campos que contengan información espuria, es decir para efectos de los análisis eliminados fueron: idComDirección, idPacPaciente, dteFechaCreaciónExp y idExpExpediente. Dado que para el análisis no es relevante integrar datos personales como RFC del paciente, dirección postal y fecha de apertura de expediente médico.

Integrar datos- Para realizar la integración de los datos, es importante mencionar que se utilizó Data Warehouse, en donde se crearon las tablas necesarias para trabajar el minado de los datos a partir del proceso que se llevó a cabo en las etapas anteriores.
Integrar datos

3.1.3 Formateo de datos. En esta fase de la metodología, no se requirió llevarla a cabo, puesto que la base de datos nos brindaba el formato correcto y con ello se podía trabajar con el algoritmo de minería de datos.

4. Descripción del Modelo

A continuación, se procede a ejecutar el resultado de ejecutar cada uno de los modelos para cada objetivo elegido sobre los datos seleccionados.

Para probar la calidad y eficiencia del modelo se utilizará la técnica predictiva de regresión lineal, la cual relaciona la distribución aleatoria de la variable dependiente.

Para probar la calidad y eficiencia del modelo se utilizarán las medidas de Coeficiente de correlación (correlation coefficient) error cuadrático medio (root mean squared error) y el error absoluto medio (mean absolute error). Por error se entiende, la diferencia entre el valor estimado y el valor real. El error medio, es otra forma de evaluar la calidad en los modelos de regresión.

Estos datos, los calcula automáticamente Weka al ejecutar el modelo de regresión lineal para cada predicción.

Modelo de predicción Objetivo 1. Este modelo ha devuelto los siguientes resultados para el algoritmo Regresión Lineal

Coeficiente de correlación (Correlation coefficient) para el algoritmo tiene un valor de 0.7754.

Modelo de predicción Objetivo 2

Este modelo ha devuelto los siguientes resultados para el algoritmo Regresión Lineal:

Coeficiente de correlación (Correlation coefficient) para el algoritmo tiene un valor de 0.7826.

Modelo de predicción Objetivo 3

Este modelo ha devuelto los siguientes resultados para el algoritmo Regresión Lineal:

Coeficiente de correlación (Correlation coefficient) para el algoritmo tiene un valor de 0.3184.

Modelo de predicción Objetivo 4

Este modelo ha devuelto los siguientes resultados para el algoritmo Regresión Lineal:

Coeficiente de correlación (Correlation coefficient) para el algoritmo tiene un valor de 1.

5. Evaluación del Modelo

Acorde a la minería de datos, una manera de evaluar la efectividad de los modelos es utilizar los dos indicadores que se establecieron en el plan de pruebas de este documento, dichos indicadores son el error cuadrático medio (root mean squared error) que compara un valor predictivo con un valor observado o conocido, el error absoluto medio (mean absolute error) que calcula el promedio de la diferencia absoluta entre el valor observado y los valores predichos, lo que significa que todas las diferencias individuales se ponderan por igual. En la tabla 2 se resumen los distintos indicadores.

Tabla 2. Indicadores de modelo

	Predicción	Error Absoluto Medio	Error cuadrático
Modelo 1	0.7754	0,212	0,271
Modelo 2	0.7826	0, 161	0,206
Modelo 3	0.3184	0, 380	0, 327
Modelo 4	1	0	0

Fuente. Elaboración propia

El objetivo 1 tiene un valor de .77, tanto el valor del error absoluto medio (0,21) como el del error cuadrático medio (0,27) es menor para el algoritmo SVM, que presenta los valores 0,89 y 1,08 respectivamente, por lo que se emplearía este algoritmo para resolver el objetivo 1.

El objetivo 2 tiene un valor de .78, tanto el valor del error absoluto medio (0,16) como el del error cuadrático medio (0,20) es menor para el algoritmo SVM, por lo que se emplearía este algoritmo para resolver el objetivo 2.

El objetivo 3 tiene un valor de .31, tanto el valor del error absoluto medio (0,38) como el del error cuadrático medio (0,33) es mayor para el algoritmo SVM, por lo que se empleará bajo reserva para resolver el objetivo 3.

El objetivo 4 tiene un valor de 1, tanto el valor del error absoluto medio (0,0) como el del error cuadrático medio (0,0) es menor para el algoritmo SVM, por lo que se emplearía este algoritmo para resolver el objetivo 4.

6. Implementación

6.1 Predicción de enfermedades dermatológicas por entidad federativa para 2021-2025

6.1.1. Predicción 1. Por entidad federativa

La entidad federativa con mayor cantidad de casos es Hidalgo, esto puede ser ocasionado por la cantidad de fábricas e industrias que radican en este estado, entre estas se destacan la Refinería Miguel Hidalgo, Planta de Tratamiento de Aguas Residuales Atotonilco, Empresas Textiles, Parques industriales, entre otros (ver tabla 3).

Tabla 3. Entidades federativas con mayor incidencia para los años 2021

Año	Entidad federativa	Casos
	Hidalgo	7211.2
	México	1096.8

2021	Ciudad de México	106.4
	Querétaro	32.2
	Veracruz	17.3
	Puebla	16.5
	Guerrero	8.6
	Guanajuato	4.2

Fuente. Elaboración propia

Predicción 2 por sexo

A partir de los datos obtenidos, se determina que la enfermedad dermatológica con más casos en los próximos años es enfermedad de las glándulas, por ello es importante mencionar que dicha enfermedad se presenta más en el sexo femenino (tabla 4).

Tabla 4. Casos dermatológicos por año en mujeres

Enfermedad de glándulas por año	Numero de casos sexo femenino
2021	1,512.5
2022	1,656
2023	1,799.5
2024	1,943
2025	2,086.5

Fuente. Elaboración propia.

Predicción 3. Enfermedades por edad

Con el análisis realizado para las edades de los pacientes en relación a las enfermedades dermatológicas, se pudo identificar una gran variación en el número de casos. Por ejemplo, las enfermedades del pelo, se presentan más en pacientes de 21 a 30 años, seguido de pacientes entre las edades 31 a 40 años. Esta información se toma con reserva dado el resultado del modelo de predicción.

Predicción de enfermedades dermatológicas por diagnóstico

Se identificaron 5 enfermedades que tienen mayor número de casos por diagnóstico de los 10 existentes. La enfermedad con mayor número de casos para los años de 2021 a 2025 es en las glándulas, le sigue enfermedad por el sol, las enfermedades reaccionales, tumores y por último enfermedades por alta pigmentación. En la tabla 5 se observan las cinco enfermedades que tienen los casos más altos.

Tabla 5. Enfermedades con mayor incidencia para los años 2021-2025

Año	Diagnóstico

2021-2025	Enfermedad de las Glándulas
	Ocasionadas por rayos solares
	Enfermedades alérgicas
	tumores
	ALT Pigmentación (hiperpigmentación)

Elaboración propia Fuente.

Revisar el Proyecto

En esta etapa de la metodología, se evalúan las actividades que se realizaron de forma correcta y las que tienen áreas de mejora en futuras ejecuciones. Lo positivo fue contar con los datos para poder hacer los análisis, no obstante, existen múltiples factores que pudieran incluirse para incrementar la fiabilidad de los modelos, tales como factores ambientales, nivel socioeconómico, antecedentes heredofamiliares, entre otros.

4. CONCLUSIONES

Los resultados obtenidos han permitido alcanzar el objetivo planteado del proyecto, aplicar la metodología CRISP-DM a partir de datos históricos de casos de incidencia de enfermedades dermatológicas que permitió hacer predicciones fiables sobre su comportamiento para los próximos 5 años (2021 a 2025), las entidades con mayor número de casos, así como el sexo con mayor afectación y la edad a partir de los modelos obtenidos. De esta manera, se busca facilitar la toma de decisiones en cuanto a atención oportuna, tratamientos, seguimiento y servicios complementarios que ofrece la clínica a sus pacientes.

La presentación de las diversas fases de la metodología CRISP-DM para explorar, analizar y visualizar datos, proporciona una idea más completa respecto a la estructuración y desarrollo d proyectos aplicando data Mining, ya que es posible resaltar las principales ventajas que aporta la minería de datos en el modelado, en este caso de enfermedades de la piel, así como la versatilidad para adaptarse incorporar diversas funciones (entidad federativa, diagnóstico, edad y sexo), además de garantizar efectividad para procesos no del todo lineales.

Por su parte el desarrollo de bases de datos y su proceso de análisis requiere de una metodología estructurada y organizada para realizar proyectos que manejan grandes cantidades de información, con el objetivo d encontrar patrones repetitivos, tendencias, comportamiento de datos, y uso de estadística, para devolver resultados pertinentes, oportunos y específicos fáciles de interpretar, por lo que el proceso de data mining facilita la planificación, dirección, control y seguimiento del proyecto, principalmente para toma de decisiones, incremento de la eficiencia, para identificar algunos del bien común, como es el caso del diagnostico de enfermedades.

Finalmente, las herramientas de software de data mining que existen actualmente en el mercado son variadas y de excelente aplicación. Su correcta elección depende de la necesidad de la empresa y de los objetivos a corto y largo plazo que pretenda alcanzar.

5. REFERENCIAS

Antuña B. y García V. E. (2020). Perfil psicológico y calidad de vida pacientes con enfermedades dermatológicas. *Psicothema,* Vol. 12, Suplem.2, pp. 30-34
Recuperado de: http://digibuo.uniovi.es/dspace/bitstream/10651/27477/1/Psicothema.2000.12%28S.2%29.30-34.pdf

Arenas, Ro. (2015). *Dermatología, Atlas, diagnóstico y tratamiento*. (6ª. Edición). México, D.F.:Mc. Graw Hill. ISBN 978-607-15-1269-7.

Canul, A. y Rocher, C., (2020). Cáncer de piel en Yucatán: Un Estudio Epidemiológico de 10 Años. *Dermatología Cosmética, Médica y Quirúrgica* pp.8-10**.**
Recuperado de: https://pdfs.semanticscholar.org/cb37/7a59cc40519a9ae5c34d0a8b6c0baf40ef01.pdf.

Chapman, P. (2000). *CRISP-DM, 1.0*, Step-by-step Data Mining Guide, 2000.

Datawarehouse. (2020). Recuperado de:
https://www.sinnexus.com/business_intelligence/datawarehouse.aspx

Hernández, O., Ramírez, Q., y Ferri, C. (2004). *Introducción a la Minería de Datos*.
México, D.F: Ed. Pearson Educación.

Fundación Piel Sana. (2020). ¿Qué es la dermatología? Recuperado
https://dermatologia.almirallmed.es/webs-apps-pacientes/fundacion-piel-sana

Garassino, A. (2015). ¿Para qué sirve la minería de datos? *Ciencia, pp. 20-22*
Recuperado de:
https://www.syloper.com/blog/recursos/para-que-sirve-la-mineria-de-datos/

González-Hernández, Méndez J. A, Sánchez-Álvarez I. (2019). Tratamientos emergentes de la alopecia areata. *Dermatol Rev Mex*. 2019 septiembre octubre; 63(5):469-480.

González, M. C, Rojo, G. (2008). Infecciones bacterianas de la piel. *Pediatría Integral*;16(1) pp.7-31.

Gorbea, P. (2020). Diseño de un data Waterhouse para medir el desarrollo disciplinar en instituciones académicas. Recuperado de: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2017000200161

Minería de Datos en Base de Datos de Servicios de Salud. (2020). Ebook.
Recuperado de: http://www.conaiisi.unsl.edu.ar/portugues/2013/132-505-1-DR.pdf

Molina, L. (2006). Técnicas de Análisis de Datos. Universidad Carlos III de Madrid, Recuperado
http://matema.ujaen.es/jnavas/web_recursos/archivos/weka%20master%20recursos%20naurales/apuntesAD.pdf

Román, J. V. (2016). CRISP-DM: La metodología para poner orden en los proyectos. Recuperado de:
https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Suárez, S. J., y Colín, R. L. (2020). Una aproximación al diagnóstico de enfermedades de la piel por medio de aprendizaje profundo. Recuperado de: http://fcqi.tij.uabc.mx/usuarios/revistaaristas/numeros/N12/articulos/13-16.pdf

WEKA The University of Waikato. Recuperado de: WEKA The University of Waikato: http://www.cs.waikato.ac.nz/ml/weka/