UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

ÁREA TÉCNICA

TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Minería de datos para la identificación de similitud en la información de pacientes.

TRABAJO DE TITULACIÓN.

AUTOR: Jimbo Celi, Roger Israel.

DIRECTOR: Reátegui Rojas, Ruth María Mgtr.

LOJA - ECUADOR

2016

Esta versión digital, ha sido acreditada bajo la licencia Creative Commons 4.0, CC BY-NY-SA: Reconocimiento-No comercial-Compartir igual; la cual permite copiar, distribuir y comunicar públicamente la obra, mientras se reconozca la autoría original, no se utilice con fines comerciales y se permiten obras derivadas, siempre que mantenga la misma licencia al ser divulgada. http://creativecommons.org/licenses/by-nc-sa/4.0/deed.es

Septiembre, 2016

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN

Magister.
Ruth María Reátegui Rojas
DOCENTE DE LA TITULACIÓN

De mi consideración:

El presente trabajo de titulación: **Minería de datos para la identificación de similitud en la información de pacientes,** realizado por Jimbo Celi Roger Israel, ha sido orientado y revisado durante su ejecución, por cuanto se aprueba la presentación del mismo.

Loja, noviembre de 2016

f) . . . . . .

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

"Yo Jimbo Celi Roger Israel declaro ser autor (a) del presente trabajo de titulación: **Minería de datos para la identificación de similitud en la información de pacientes**, de la Titulación Sistemas Informáticos y Computación, siendo Ruth María Reátegui Rojas director (a) del presente trabajo; y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales. Además certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad.

Adicionalmente declaro conocer y aceptar la disposición del Art. 88 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que en su parte pertinente textualmente dice: “Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad”

f.....
Autor Jimbo Celi Roger Israel.
Cédula 1104198294

DEDICATORIA

Este trabajo de titulación va dirigido a Dios y la Virgencita del cisne, por permitirme llegar a este punto de mi vida y haberme dado la fortaleza para seguir adelante sobre cualquier obstáculo que se me haya presentado, sobre esta meta importante en mi vida profesional.

A mi súper mamá, que me ha sabido guiar y luchar por mí siempre, por haberme brindado tanto amor y apoyo a lo largo de toda mi vida, por ser una amiga y confidente, mi ejemplo a seguir y sentir por ella un verdadero amor.

Para mis dos grandes amores, mi esposa por a verme apoyado y ayudado a lo largo de toda la carrera universitaria por estar en los momentos buenos y malos y por todo su amor, a mi querida hija por ser la persona por la cual me quiero superar y es el motor de mi vida, por haber descubierto en mi corazón un amor tan bello y puro, por a verme dado una motivo más en mi vida para seguir cosechando triunfos.

A mis queridos abuelitos, más que abuelitos son unos verdaderos padres para mí, siempre me han sabido dar consejos para andar en un camino bueno y ser una mejor persona.

A toda mi familia que siempre estamos unidos para apoyarnos en cualquier circunstancia

Roger Israel Jimbo Celi

AGRADECIMIENTO

Agradezco a Dios y a la Virgencita del Cisne por haberme permitido llegar, a esta etapa de mi vida, y por tantas bendiciones otorgadas.

Estoy eternamente agradecido con mi querida madrecita por todo su esfuerzo, tiempo y dedicación que me sigue brindado cada día y por ser madre y padre al mismo tiempo, por estar siempre conmigo demostrándome su amor y confiar siempre en mí. Te lo debo todo gracias.

Agradezco a mi esposa por darme siempre apoyo y confiar en mis capacidades, por compartir mis alegrías y tristezas. A mi querida hija por ser el regalo más lindo que me pudo dar dios, por ser el motivo para superándome, por ser parte de mi esfuerzo y dedicación para lograr esta meta.

A mis abuelitos agradezco sus cuidados, por haberme criado y dado amor como un hijo más, por apoyarme y dar ánimos para culminar con mis estudios universitarios.

Agradezco especialmente a mi directora del trabajo de fin de titulación Mgtr. Ruth María Reátegui Rojas, por su apoyo, conocimiento, preocupación y ayuda a lo largo del proyecto, siendo fundamental para realizar el mismo.

Roger Israel Jimbo Celi

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

En el presente trabajo de fin de titulación, tiene como objetivo identificar grupos de pacientes similares basados en enfermedades. Guiados por la metodología CRISP-DM, se hicieron algunos experimentos, utilizando métodos de agrupación como BDSCAN y K-means, y métodos semánticos como LSA. Una vez que se obtuvo los diferentes grupos, se pudo identificar, las enfermedades más comunes de cada grupo. Se realizó un análisis individual y grupal sobre resultado de la experimentación. Finalmente se realizó la validación de todos los grupos de pacientes encontrados.

PALABRAS CLAVES: minería de datos, metodología Crisp –Dm, métodos de agrupamiento, método de selección de característica, grupos, RStudio, Plotly, LSA, DBSCAN, K-Means, coseno,

ABSTRACT

In the present work order degree, it aims to identify groups of similar patients based on disease. Guided by the CRISP-DM methodology, some experiments were done using clustering methods as BDSCAN, and K-means clustering and semantic methods as LSA. Once the different groups was obtained, could be identified, the most common diseases in each group. Individual and group analysis result of experimentation was performed. Finally validation of all patient groups was performed found

KEYWORDS: data mining, CRISP-DM methodology, clustering methods, feature selection method, Clusters, RStudio, Plotly, LSA, DBSCAN, K-Means, Cosine

INTRODUCCIÓN

Existen numerosas áreas en donde se trabaja con la minería de datos, una muy importante es la medicina. La información de las enfermedades de los pacientes, es de suma importancia para los especialistas en las distintas áreas médicas. Es necesario investigar y desarrollar trabajos con información de los pacientes considerando el valor ético, ya que este tipo de información es muy sensible. A través de la minería de datos es posible encontrar patrones en la información, estos patrones son útiles para toma de decisiones. Por tales motivos, el presente proyecto pretende, identificar grupos de pacientes similares de acuerdo a sus enfermedades, además por cada grupo identificar cuáles son las enfermedades más comunes.

Durante el proceso de investigación se pudo realizar un análisis de los posibles algoritmos o técnicas que se acoplen a lo que se requiere hacer. Se seleccionó los métodos de agrupación y semánticos. Para la elección de corpus se investigó algunos eventos que ponen a disposición corpus de información médica para su estudio. Finalmente considerando que este trabajo forma parte de un Proyecto de Investigación desarrollado en la UTPL, se trabajó con los datos obtenidos para el proyecto luego de cumplir con el certificado ético necesario para trabajar con este tipo de información.

La metodología que se escogió para este trabajó fue CRISP-DM, pero se decidió modificarla, para que se acople al presente trabajo, por lo tanto no se consideró una fase y algunas otras se adaptaron, pero siempre se conservó su ciclo interactivo.

Como parte de la experimentación se utilizó el algoritmo LSA que forma parte principal de todo el trabajo, luego para realizar los grupos de pacientes se estableció dos algoritmos pero sin duda el más relevante entre los dos es DBSCAN debido a que este algoritmo soporta el tipo de dato que vamos a utilizar y no es sensible al ruido. El segundo algoritmo K-means que se lo empleo para comparación en la en experimentación. Todo este trabajo fue desarrollo en Rstudio y la herramienta Plotly.

El presente trabajo consta de 4 capítulos para su desarrollo: En el primer capítulo se presentará el contexto de la investigación, trata de tres partes principales en donde se

identifica el problema y a su vez se especifica una solución, otra parte trata de establecer el alcance del trabajo y finalmente se define los objetivos a cumplir.

El segundo capítulo está dirigido al estado del arte, conformado por una conceptualización de los temas a ser tratado he implementado, los temas principales a tomar en cuenta son minera de datos y documentos clínicos y una revisión de trabajos relacionados.

El tercer capítulo se enfoca en la metodología de desarrollo. Se estructura de la siguiente forma: definición y establecimiento de la metodología CRISP-DM: especificación de la obtención del corpus: pre procesamiento del corpus escogido; finalmente la experimentación. En este capítulo se obtiene la matriz documento-termino, se establece el algoritmo LSA para realizar una reducción del corpus y trabajar con una matriz de pacientes y enfermedades por separado, luego se implementa DBSCAN y se obtiene 3 grupos, a continuación se compara K-means obteniendo 5 grupos. Por otra parte se indica las herramientas que se usaran para la práctica como es Rstudio, y plotly. Por ultimo para identificar los beneficios brindados por los algoritmos y técnicas que se usaron en la experimentación se realiza una evaluación de cada uno de ellos.

Cuarto capítulo corresponde al análisis de la experimentación, se detallan cada grupo por cada algoritmo de forma individual y grupal, se especifican las enfermedades más comunes que existen en el corpus. Finalmente se realiza una validación sobre los grupos obtenidos por medio de DBSCANy K-means. Con la ayuda del algoritmo Coseno se obteniendo una matriz que sirvió para ver la similitud que existe entre los pacientes.

CAPÍTULO I

CONTEXTO DE LA INVESTIGACIÓN

1.1 Tema

Minería de datos para la identificación de similitud en la información de pacientes.

1.2 Planteamiento del problema

La información de los pacientes como sus enfermedades, medicamentos, tratamientos, son de gran importancia para médicos y demás profesionales del área de la salud. Este tipo de información permite caracterizar a los pacientes. Por tanto el presente proyecto permitirá encontrar relaciones o grupos de pacientes con características comunes. Para esto se ha previsto trabajar con técnicas y métodos de minería de datos que permitan encontrar la similitud entre documentos tomando como base entidades médicas, específicamente enfermedades de los pacientes. A futuro este trabajo ayudará al personal de la salud a mejorar por ejemplo los tratamientos y medicamentos que se puede dar a pacientes.

Dentro de las técnicas de minería de datos a considerar esta la agrupación y métodos semánticos distribuidos. Uno de los métodos semánticos es el análisis semántico latente que permite encontrar las variables significativas dentro de un grupo de datos y por ende la reducción de la información. Estos últimos se han utilizado para inferir el significado de un texto a través de la distribución de sus elementos en elcontexto dado (Jonnalagadda, Cohen, Wu, & Gonzalez, 2012).

Además es importante recalcar que debido a la delicadeza de la información con la que va a tratar, se explorará eventos y organizaciones que hagan disponible dataset para estudios científicos. Es necesario recurrir a estos eventos ya que no es posible conseguir fácilmente la información de parte de centros médicos.

1.3 Alcance

El proyecto se concreta en la identificación de similitud sobre un dataset clínico, este dataset contendrá específicamente enfermedades con sus respectivos pacientes, al mismo que se aplicara técnicas de agrupación y semánticas de minería de datos con el fin de encontrar grupos que muestren las similitud que van a existir entre los pacientes sus

enfermedades. Para esta experimentación se ha decidido trabajar con el lenguaje R y con la metodología CRISP DM.

1.4 Objetivos

1.4.1 Generales.

Identificar similitud en la información de pacientes.

1.4.2 Específicos.

- Crear o identificar un dataset de documentos médicos pertenecientes a diferentes pacientes.
- Aplicar técnicas de minería de datos sobre el corpus médico.
- Identificar grupos de documentos similares.

CAPÍTULO II
ESTADO DEL ARTE

2.1 Minería de datos

La Minería de datos es el proceso de descubrimiento de patrones y conocimiento desde una gran cantidad de datos. La fuente de obtención de los datos puede ser una base de datos, data warehouse, la Web entre otros repositorios de información (Han, Kamber, & Pei, 2011).

Según (Orallo, Ramirez, Quintana, & Jose., 2014) afirman que la minería de datos forma varios técnicas, para el análisis y extracción de los datos, además de su aplicabilidad en diversas áreas, el fin de la minería de datos se dirige a poder extraer patrones, lograr descubrir tendencias, predecir comportamientos todo estos con el objetivo de plasmarlo en un manera más eficiente y al mismo tiempo precisa para poder actuar y tomar decisiones adecuadas.

Algunos de los términos que se utilizan para hacer referencia a la minería de datos están: minería de conocimiento desde datos, extracción de conocimiento, análisis de datos/patrones, arqueología de datos, etc (Han et al., 2011; Orallo et al., 2014).

(Han et al., 2011) indica que la minería de datos es tratada ya sea como todo el proceso de descubrimiento de conocimiento desde datos (KDD por las siglas en inglés) o como un paso del KDD.

2.1.1 Fases del proceso de extracción de conocimiento.

Figura 1. Proceso de KDD
Fuente: Introducción a la Minería de Datos
Elaboración: Orallo et al., (2014)

Como se aprecia en la figura 1, KDD define pasos para la selección, limpieza, transformación, con el fin de analizar datos para extraer patronees y modelos adecuados y finalmente convertirlos en conocimiento (Orallo et al., 2014).

Según (Han et al., 2011; Orallo et al., 2014) coinciden, que el KDD es un proceso iterativo e interactivo que incluye los siguientes pasos:

- Fase de Integración y recopilación de datos: datos de diversas fuentes pueden ser combinados.

- Fase de selección, limpieza y transformación: la selección permite que los datos relevantes para la tarea de análisis son recuperados desde la base datos, la limpieza eliminar el ruido y los datos inconsistentes, la transformación hace que los datos son transformados y consolidados para la tarea de minería, se utilizan las operaciones de resumen y agregación.

- Fase de minería de datos: aplica métodos inteligentes para la extracción de patrones.

- Fase de evaluación e interpretación: aplica medidas para identificar patrones reales e interesantes que representan el conocimiento.

- Fase de difusión: se presenta el conocimiento minado a los usuarios a través de técnicas de visualización y representación.

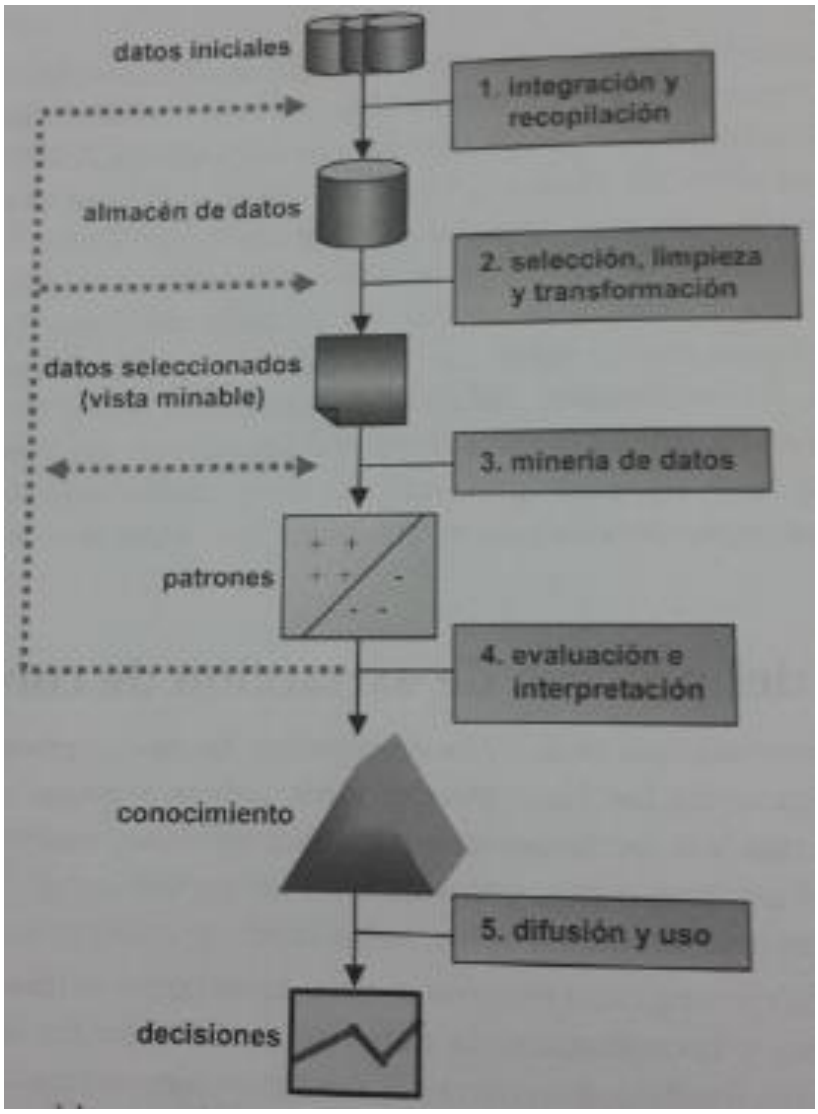

Figura 2. Fase de KDD
Fuente: Introducción a la Minería de Datos
Elaboración: Orallo et al., (2014)

2.1.2 Tareas de la minería de datos.

Al hablar de minería de datos se consideran algunas tareas. Estas tareas son un problema a resolver con algoritmo de minería de datos. Las tareas se clasifican en: 1. tareas predictivas, es estas pueden estar la clasificación y regresión; 2. Tareas descriptivas como el agrupamiento, asociación y correlación (Orallo et al., 2014).

Por lo tanto las tareas de la minería de datos pueden ser categorizadas en:

1. Predictivas: realizan una inducción sobre los datos con el objetivo de hacer predicciones.

2. Descriptivas: las tareas se enfocan en caracterizar propiedades de los datos.

Algunas de estas tareas son detalladas en (Han et al., 2011; Orallo et al., 2014), a continuación un pequeño resumen de las mismas:

- Clasificación: Es una de las tareas más utilizada, trabaja con un conjunto de datos de entrenamiento cuyas clases están etiquetadas. Esta tarea permite encontrar un modelo o función que describa las clases. Luego el modelo puede ser usado para identificar la clase de nuevos ejemplos cuya clase se desconoce.

- Regresión: es un modelo estadístico que permite predecir datos de tipo numérico

- Patrones frecuentes: Estos pueden ser ítems frecuentes o secuencias frecuentes. Los ítems frecuentes se refiere a ítems que aparecen juntos en una transacción como por ejemplo comprar leche y pan. Las secuencias frecuentes son principalmente a patrones de compra que los clientes puedan tener como por ejemplo primero comprar una laptop y luego una cámara fotográfica. La minería de patrones de frecuencia permite descubrir asociaciones y correlaciones en los datos

- Agrupamiento: Se utiliza para generar clases de etiquetas o encontrar grupos de datos. El principio de agrupamiento consiste en formar grupos de objetos que maximicen la similitud entre grupos y que minimice la similitud en los elementos de un mismo grupo. Cada grupo puede ser visto como una nueva clase de objetos

- Reglas de Asociación secuenciales: Son utilizadas con el fin de encontrar patrones secuenciales en los datos.

2.1.3 Tecnologías relacionadas a la minería de datos.

La Minería de datos adopta técnicas de algunos campos, entre ellos están: Estadística,

Sistemas de bases de datos, Data warehouse, Recuperación de la información, Aprendizaje de máquina, Reconocimiento de patrones, Visualización, Algoritmos, Computación de alto rendimiento, entre otros.

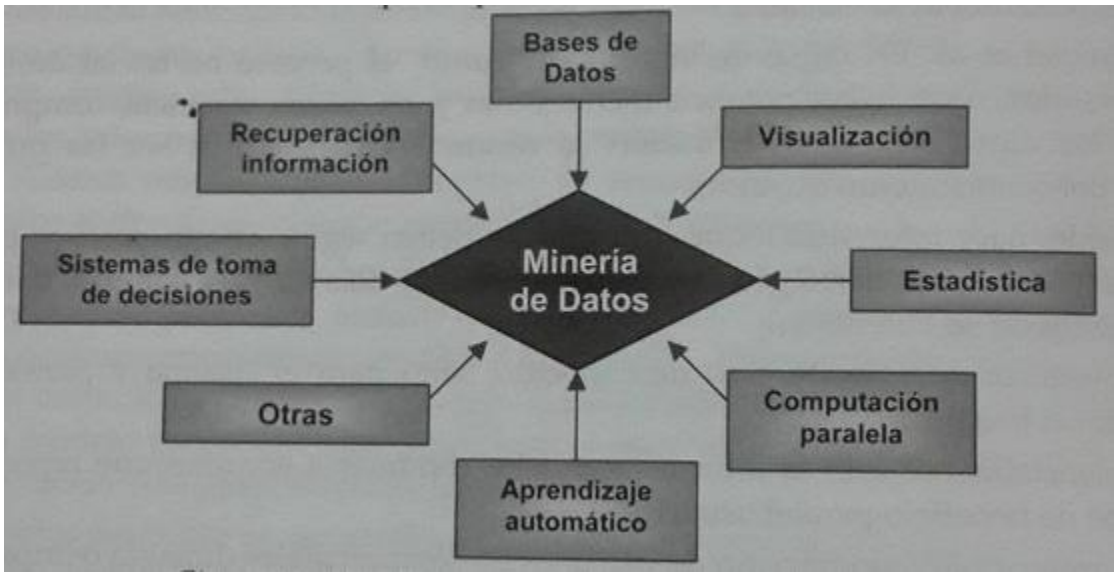

Figura 3. Disciplinas que contribuyen a la minería de datos
Fuente: Introducción a la Minería de Datos
Elaboración: Orallo et al., (2014)

Enfocándonos al aprendizaje de máquina o machine learning, su principal objetivo es que las computadoras aprendan automáticamente a reconocer patrones y tomar decisiones inteligentes basados en datos. Algunas de los problemas de machine learning están relacionados con la minería de datos, (Han et al., 2011) se refieren:

- Aprendizaje supervisado: se puede decir que es un sinónimo de clasificación. La parte supervisada tiene su base en los ejemplos etiquetados que existen en la datos de entrenamiento de un determinado algoritmo

- Aprendizaje no supervisado: es básicamente sinónimo de agrupamiento. El aprendizaje es no supervisado ya que no existe una clase etiquetada en los datos de entrenamiento

- Aprendizaje semi-supervisado: Utiliza ejemplos etiquetados y no etiquetados.

2.2 Métodos de Agrupamiento

La clasificación de los métodos de agrupación puede variar según los autores, en (Han et al., 2011) se indican los siguientes.

- **Métodos Particionales:** Este método particiona los datos en k grupos. La mayoría de estos métodos están basado en distancias donde los objetos en el mismo grupo están cerca, mientras que los objetos en grupos diferentes deben ser más distantes o diferentes. K-means y K-medias son particionales

- **Métodos Jerárquicos:** Estos métodos crean una colección jerárquica de objetos. Estos métodos pueden ser aglomerativos o divisivos

- **Métodos basados en Densidad:** Crean un grupo considerando el número de objetos o densidad en un vecindario. DBSCAN es un algoritmo de esta clase

- **Métodos basados en cuadricula:** Estos métodos cuantifican el espacio de objetos en un número finito de células que forman una estructura de rejilla. Su ventaja es el tiempo de procesamiento.

Para el presente trabajo, basados en la revisión de la literatura, se identificaron tres principales algoritmos utilizados para encontrar similitud y grupos en pacientes o de enfermedades estas son: DBSCAN, K-means y Coseno, a continuación se describen los mismos.

2.2.1 Densidad DBSCAN.

DBSCAN se maneja a través de punto central, este concepto hace referencia a los puntos que conforma la vecindad con una cantidad mayor o igual a un umbral específico. Además DBSCAN trabaja con bordes y ruido (Pascual, Pla, & Sánchez, 2007).

DBSCAN inicia seleccionando un punto central p de forma arbitrario, una vez establecido este punto se procede a elaborar un grupo, dirigiéndose los puntos densamente-alcanzables desde p, los puntos que han quedado sin estar en los grupos formados se los conoce como puntos ruido. Los puntos que no conforma ni punto central ni ruido se los conoce como punto borde. Cabe mencionar que un grupo puede tener más de un punto central (Pascual et al., 2007).

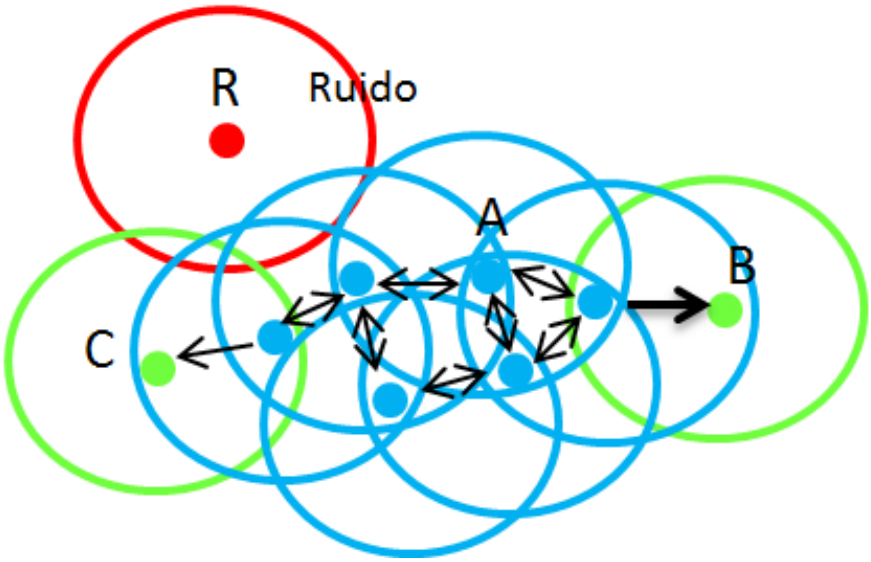

Figura 4. Representación de DBSCAN
Fuente: DBSCAN.
Elaboración: Propia. (Adaptado de wikipedia)
Recuperado de: https://es.wikipedia.org/wiki/DBSCAN

En la figura 5 los puntos señalados como A son puntos centros. Los puntos marcados como B y C son densamente alcanzables desde A y pertenecen al mismo clúster. En cambio el punto R es un punto ruidoso que no es centro ni densamente alcanzable.

DBSCAN según (Antonelli et al., 2013) tiene las siguientes ventajas:

- Es menos sensible a los valores atípicos o ruidos.
- No requiere de un detalle sobre el número de grupos que se van encontrar en los datos.

2.2.2 Distancia K-means.

El algoritmo K-means es una técnica de agrupamiento basada en vecindad. Esta inicia en dos partes principales, la primera establece el número determinado de los centros en el espacio y la segunda define un conjunto de ejemplos a agrupar. K-means normalmente trabaja con la distancia euclidiana (Orallo et al., 2014).

(Orallo et al., 2014) afirma. “Las regiones se definen minimizando la suma de las distancias cuadráticas entre cada vector de entrada y el centro de su correspondiente clase, representado por el prototipo correspondiente” (p.432).

La forma en que trabaja el algoritmo K-means, es ubicando a los centros en el espacio, para que los datos referentes al mismo centro en el espacio, cuenten con características similares (Orallo et al., 2014).

> El algoritmo puede seguir dos enfoques distintos: K medias por lotes (batch) y K medias en línea (on-line). El primero se aplica cuando todos los datos de entrada están disponibles desde un principio, mientras que el segundo se aplica cuando no se dispone de todos los datos desde el primer momento, sino que pueden añadirse ejemplos adicionales más tarde. (Orallo et al., 2014, p.432)

Según (Antonelli et al., 2013) afirmando que existen dos principales limitaciones las cuales son:

- Sensibilidad a los valores atípicos
- Tiene que definir el número esperado de grupos.

2.2.3 Similitud/Distancia Coseno.

Coseno es una unidad de similitud, que se basa en calcular la distancia que existe entre un documento y otro, por tal razón su utilidad puede ser dirigida a la comparación de los documentos que se requiera, esta comparación se la realiza por medio de los dos vectores x, y (Han et al., 2011).

La unidad de similitud Coseno fue selecciono en el presente trabajo, con el objetivo de obtener una matriz en donde se pueda apreciar la similitud de cada paciente, para hacer una validación en la parte final con esta matriz

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Figura 5. Vector término frecuencia
Fuente: Data Mininge: Concepts and Techniques
Elaboración: Han et al., (2011)

Como se puede apreciar en la figura 5 cada documento es un objeto, que va a representar un vector término frecuencia (Han et al., 2011).

2.3 Métodos de selección de características

Dentro de la minería de datos se debe identificar cuáles son las características más importantes que definan a los datos. Existe una variedad de métodos para seleccionar dichas características, una de ellas son los métodos basados en LSA. A continuación se detalla el método de Análisis semántico latente (Aggarwal & Zhai, 2012)

2.3.1 Análisis semántico latente (LSA).

La selección de características permite explícitamente identificar cuales características representarán a todo el conjunto de datos. El análisis semántico latente por lo contrario, realiza una transformación sobre los datos de tal forma que se obtienen nuevas características que reducen la dimensionalidad de los datos originales o una matriz original (Aggarwal & Zhai, 2012; Holzinger, Schantl, & Schroettner, 2014).

El proceso del LSA se debe iniciar representado el texto por medido de una matriz en la cual cada fila es una palabra única y cada columna un documento, una parte del texto o cualquier otro contexto. Por lo tanto cada celda va a contener la frecuencia con la cual la

palabra de su fila se presentara en el paso señalado por su respectiva columna (Landauer, Foltz, & Laham, 1998).

A más de la reducción de dimensionalidad, LSA permite eliminar el ruido para luego aplicar alguna técnica de agrupamiento (Aggarwal et al., 2012).

Por tanto, LSA puede mostrar conceptos similares (Holzinger et al., 2014) en una cantidad de textos o argumentos. Considerando este tipo de datos, LSA primeramente representa los datos a través de una matriz, donde cada fila puede ser por ejemplo una palabra única, mientras que cada columna podría ser el texto u otro argumento. El valor de cada celda representa la frecuencia de la palabra en el texto (Landauer et al., 1998). Luego se utiliza la descomposición de valores singulares para descomponer y reducir la (Holzinger et al., 2014).

2.4 Documentos clínicos

Los documentos clínicos son guías que ayudan un especialista de la salud, a tomar decisiones y auxiliar en el caso de incertidumbre y a disminuir la inestabilidad de la práctica clínica. Hay una variedad de documentos clínicos. También hay que aclarar que los documentos clínicos exponen terminologías específicas, abreviaturas y términos complejos y ambiguos (Raghavan, 2014; Saura-Llamas, Saturno Hernández, Romero Román, Gaona Ramón, & Gascón Cánovas, 2001). Así mismo (Holzinger et al., 2014) aseguran que los documentos clínicos no son gramaticalmente correctos, debido a que utilizan diversas abreviaturas y tienen faltas ortográficas.

(Romero Nieva, 2008) afirma los documentos clínicos son," el soporte de cualquier tipo o clase que contiene un conjunto de datos e informaciones de carácter asistencial" (párr.9).

2.5 Trabajos relacionados

A continuación se detallan algunas investigaciones que han utilizado técnicas de agrupación para encontrar enfermedades o pacientes con características similares.

2.5.1 Primer estudio.

(Antonelli et al., 2013) el objetivo de este proyecto, es descubrir las rutas de examen de pacientes diabéticos. Realizaron un análisis de datos exploratorio, utilizaron un modelo de espacio vectorial para representar la información de pacientes. Con técnicas de agrupamiento, específicamente BDSCAN, clasificaron los pacientes de acuerdo a su historial de examinación, trabajan con una metodología de 5 faces. Los datos sobres los exámenes de los pacientes fueron representados a través del modelo de espacio vectorial (VSM)

Category	Examination	C_{1_1}	C_{2_1}	C_{3_1}	C_{4_1}	C_{5_1}
Routine	Checkup visit	78	96	58	100	100
	Glucose level	78	98	63	–	100
	Urine test	72	97	58	–	–
	Venous blood	96	75	35	–	–
	Capillary blood	72	97	58	–	–
	Haemoglobin	100	–	–	–	–
	Specialistic visit	–	13	100	–	–
Cardiovascular	Electrocardiogram	–	–	100	–	–
Eye	Fundus oculi	–	–	28	–	–
Number of patients		223	1764	43	110	41
Silhouette		0.67	0.55	0.85	0.99	1.0

Figura 6. Frecuencia de examen por medio de DBSCAN
Fuente: http://doi.org/10.1016/j.eswa.2013.02.006
Elaboración: Antonelli et al., (2013)

Category	Examination	C_{6_1}	C_{7_1}	C_{8_1}	C_{9_1}	C_{10_1}	C_{11_1}
Routine	Checkup visit	77	78	66	62	68	97
	Glucose level	74	74	64	62	59	97
	Urine test	74	74	64	57	56	92
	Venous blood	57	60	53	48	44	97
	Capillary blood	74	73	63	55	56	92
	Haemoglobin	–	–	–	14	12	100
	Specialistic visit	–	–	–	–	–	–
	Complete blood count	–	–	–	–	3	3
Cardiovascular	Electrocardiogram	–	100	100	–	–	42
	Cholesterol	–	–	–	–	–	100
	HDL cholesterol	–	–	–	–	–	100
	Triglycerides	–	–	–	–	–	100
Eye	Fundus oculi	100	–	100	–	–	39
Liver	ALT	–	–	–	–	–	100
	AST	–	–	–	–	–	100
Kidney	Creatinine	–	–	–	–	–	3
	Creatinine clearance	–	–	–	–	–	100
	Culture urine	–	–	–	–	–	100
	Microscopic urine analysis	–	–	–	–	–	100
	Uric acid	–	–	–	–	–	100
Carotid	ECO doppler carotid	–	–	–	100	–	–
Limb	ECO doppler limb	–	–	–	–	100	–
Number of patients		294	144	140	42	34	36
Silhouette		0.65	0.74	0.79	0.95	0.97	0.90

Figura 7. Frecuencia de examen por medio de DBSCAN
Fuente: http://doi.org/10.1016/j.eswa.2013.02.006
Elaboración: Antonelli et al., (2013)

Como se puede apreciar en la figura 6 y 7, la primera columna, son las categorías a examinar, la segunda menciona los exámenes por cada categoría, de la tercera c1 hasta la c5 se refiere a los grupos o clusters de los exámenes para vigilar la condición de la diabetes de la c6a la c11, se refiere a los exámenes básicos para diagnosticar complicaciones de diabetes.

2.5.2 Segundo estudio.

Otro trabajo es el de (Lyalina et al., 2013) los cuales en base a una matriz de 45 medicamentos y 78 enfermedades, identificaron fenotipos correspondientes a tres tipos autismo, trastorno bipolar y la esquizofrenia. Los registros fueron extraídos de más de 7000 pacientes en el hospital de Stanford y la Fundación Médica de Palo Alto (PAMF).

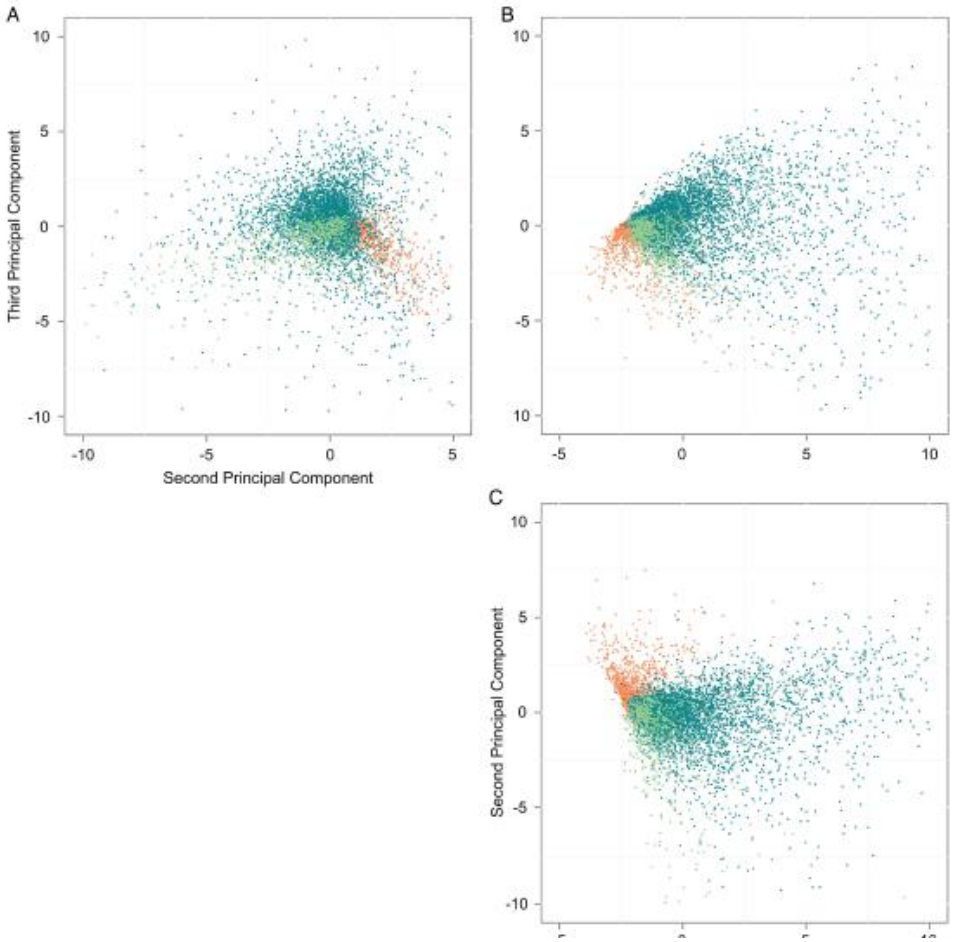

Figura 8. Representación de Los pacientes autistas, esquizofrénicos y bipolares.
Fuente: http://doi.org/10.1136/amiajnl-2013-001933
Elaboración: Lyalina et al., (2013)

En la figura 8 los pacientes autistas se representan por el color naranja, los pacientes esquizofrénicos con color verde claro y los pacientes bipolares en color turquesa oscuro.

2.5.3 Tercer estudio.

Un trabajo reciente es el presentado por (Van der Esch et al., 2015) quienes aplicaron K-means para identificar grupos de pacientes con artrosis de rodilla. Evaluaron cuatro variables o características clínicas de los pacientes: fuerza muscular, el índice de masa corporal (IMC), la gravedad radiológica, y estado de ánimo depresivo. El resultado fue cinco grupos de pacientes con artrosis de rodilla.

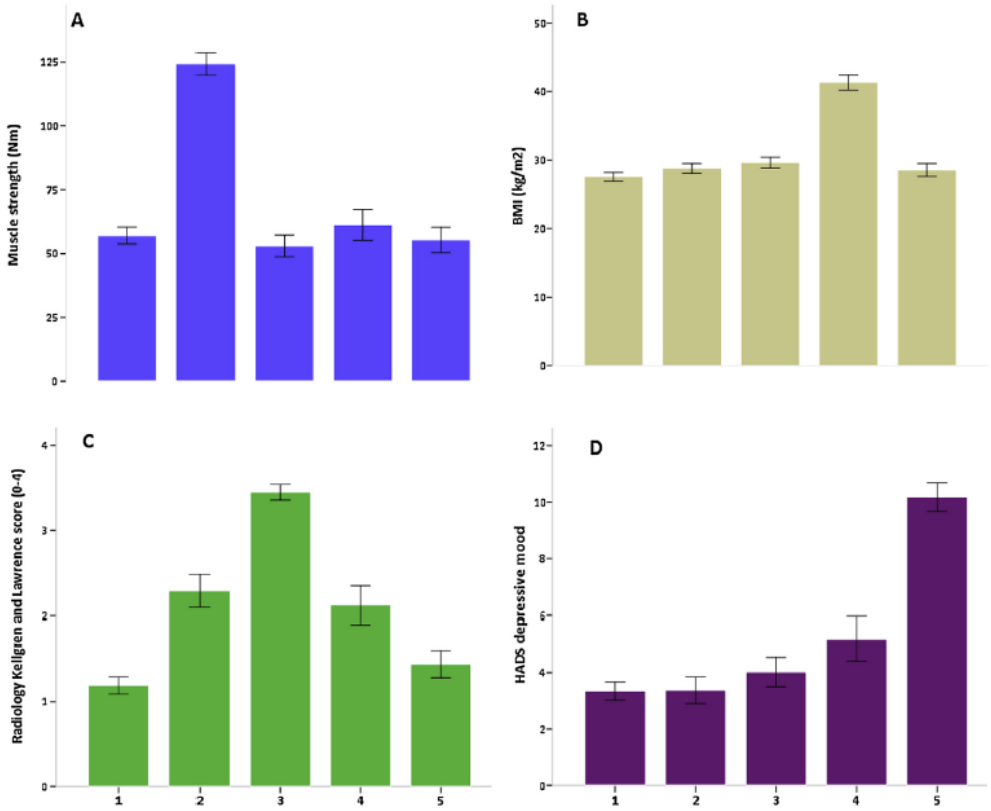

Figura 9. La comparación de los cinco fenotipos sobre las 4 características clínicas por medio de K-means
Fuente: http://doi.org/10.1016/j.joca.2015.01.006
Elaboración: Van der Esch et al., (2015)

2.5.4 Cuarto estudio.

(Cuc & Sosa, 2015) Este trabajo está basado en la aplicación de técnicas de agrupación exclusivamente con los algoritmos K-means y DBSCAN, para realizar un análisis sobre sonidos pulmonares. El objetivo de este proyecto es la visualización de los sonidos que indique la presencia de estertores y la energía contenida en ella.

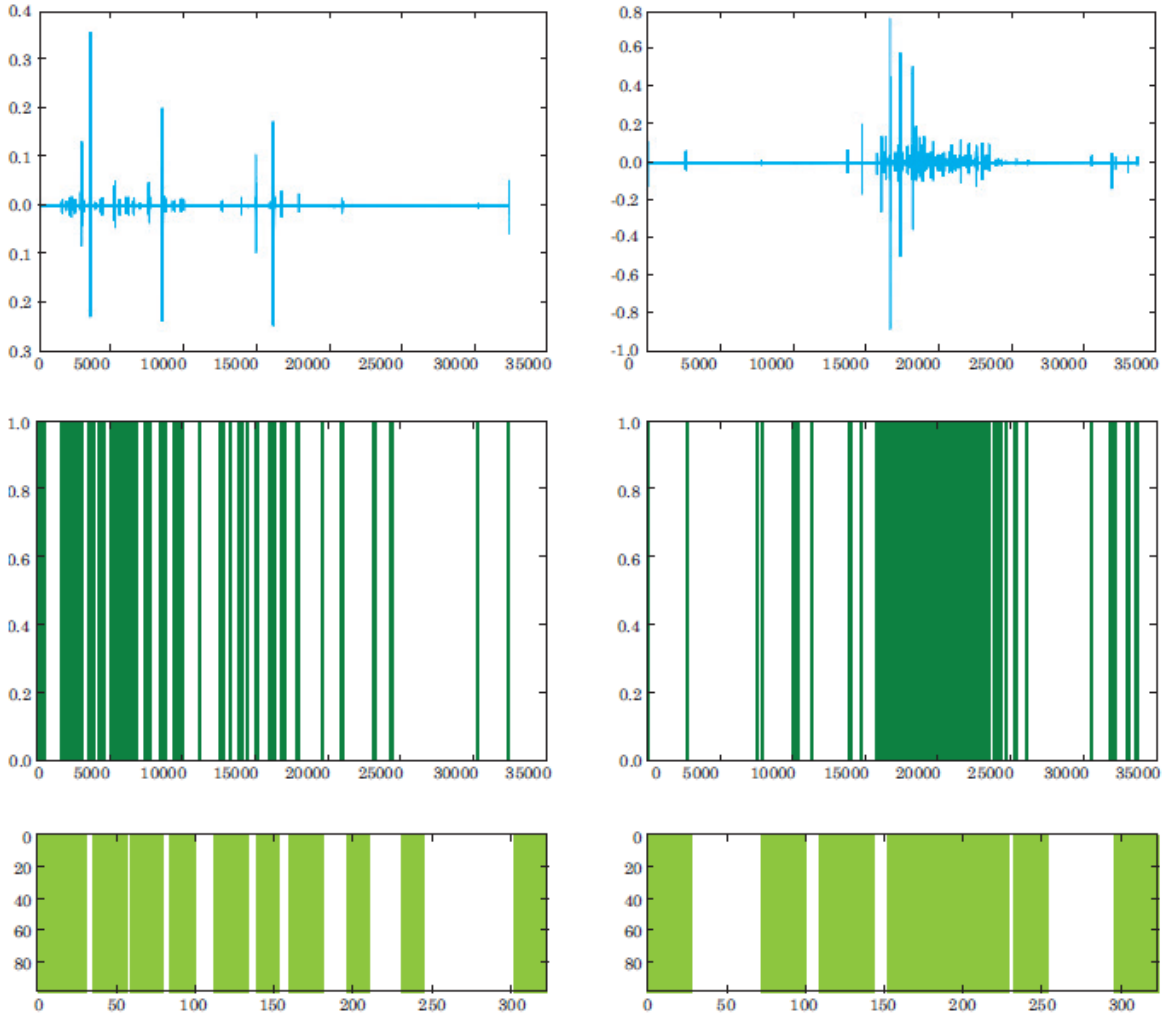

Figura 10. Visualización de los grupos obtenidos por DBSCAN y K-means
Fuente:http://revistascientificas.cuc.edu.co/index.php/ingecuc/article/download/366/2015105
Elaboración: Cuc & Sosa

CAPÍTULO III

METODOLOGÍA DE DESARROLLO

En el presente capítulo se definirá la metodología a trabajar, la elección del corpus, el tipo de algoritmo que se ampliara, la forma y sus parámetros en que se creara la matriz de términos y finalmente la visualización de los datos por agrupación.

3.1 Metodología CRISP-DM

La metodología que se aplico fue CRISP-DM (Proceso de construcción cruzada estándar de minería de datos).

Esta metodología brinda una visión global del ciclo de vida de un proyecto de minería de datos. Consta de faces, tareas y relaciones que se van a realizar dentro de un ciclo (Chapman et al., 2000), por tal razón esta metodología se utiliza en el presente proyecto.

La metodología está dividida por seis faces que van interactuar a lo largo de todo el ciclo de trabajo, estas seis faces son las siguientes.

(Chapman et al., 2000):

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Despliegue.

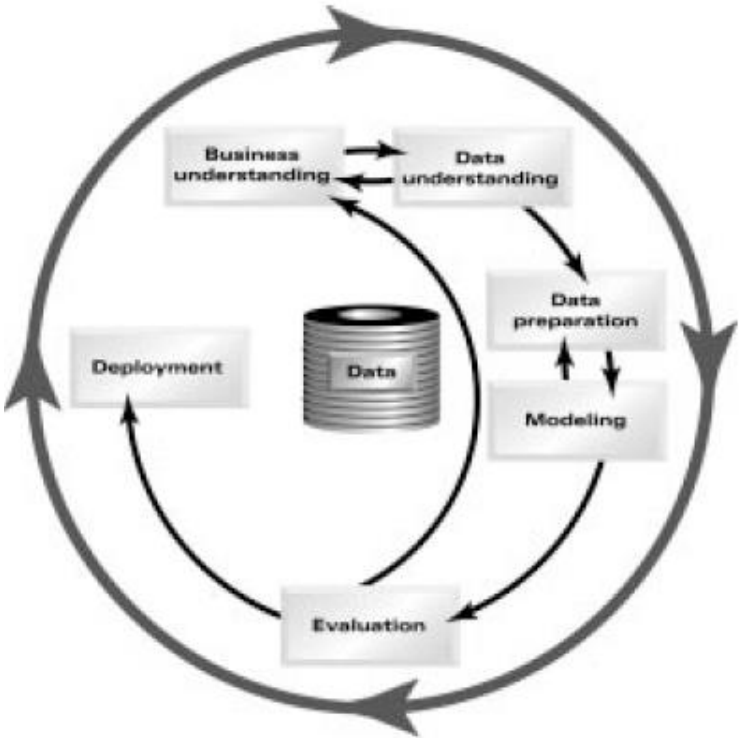

Figura 11. faces de la metodología CRISP-DM
Fuente: Crisp-Dm 1.0
Elaboración: Chapman et al., (2000)

Para plasmar la metodología CRISP-DM hacia este proyecto se tuvo que adecuar algunos puntos y el sentido de los mismos, para que funcione a lo que se requiere hacer.

3.1.1 Comprensión del Proyecto.

Según (Chapman et al., 2000). Esta primera fase está dirigida a comprender cada uno de los objetivos así como los requisitos del proyecto, que forman parte de principal para empezar el proyecto de minería de datos en este caso, para esta primera fase se ha tomado en consideración los siguientes puntos:

- Determinar objetivos del proyecto: Se profundiza en un mayor entendimiento sobre cada uno de los objetivo del proyecto

- Evaluación de la situación: Se describe el acceso que existe sobre la información y los datos para trabajar en el proyecto

- Producir el plan de proyecto: Se efectúa una selección inicial de las herramientas y las técnicas a usar, se indica el camino a seguir durante el proyecto (Chapman et al., 2000).

3.1.2 Comprensión de los datos.

Una vez obtenido los datos que se va a trabajar, esta fase nos indica que se debe realizar un análisis del mismo, con el fin de comprender la información y examinar los datos que sean relevantes al proyecto, en esta fase se ha tomado en consideración los siguientes pasos (Chapman et al., 2000):

- Recopilar datos iniciales: Se lista los corpus o dataset que se ha encontrado, y se explica por qué se los han seleccionado, indicando problemas o beneficios que brindará.

- Descripción de los datos: Se identifica las propiedades de los datos seleccionados.

- Explorar los datos: Se la realiza un análisis más técnico sobre los atributos de los datos, como claves principales, foráneas.

- Verificar la calidad de los datos: Aquí se busca tener muy claro que el dataset seleccionado sea de calidad, cuente con citaciones de trabajos importantes y cuente con un respaldo de organizaciones valederas.

3.1.3 Preparación de los datos.

Esta fase tiene un peso sumamente importante, ya que aquí se escoge el dataset que se va a trabajar durante todo el proyecto y se realiza una limpieza de los datos, Esta fase consta de los siguientes pasos (Chapman et al., 2000):

- Selección de los datos: Para la selección del dataset se tiene que tener en cuenta que cumplan con los objetivos del proyecto, esto conlleva a un análisis tanto de las columnas como filas.

- Limpieza de los datos: Lo impórtate en esta tarea es evitar que los datos a trabajar estén dañados, para ellos se realiza un limpieza tanto de filas como de columnas, y con esto llegar a tener una alta correlación entre estos datos (Chapman et al., 2000).

3.1.4 Implementación.

Esta fase se orienta a encontrar las técnicas o algoritmos y herramientas que se ajusten a la necesidad del proyecto, con el objetivo de resolver el problema que se plantea en el tema de minería de datos. En (Chapman et al., 2000) se definen las siguientes tareas:

- Seleccionar la técnica/ algoritmo y herramientas para la implementación: Se seleccionaran algoritmos que se acoplen a lo que se requiere hacer en el proyecto, sabiendo los beneficios que llevan cada uno de estos de igual forma para las herramientas a usar.

- Aplicación de la implementación: Esta tarea está dirigida al desarrollo de los algoritmos y herramientas seleccionadas y se debe describir lo obtenido por cada una de ellos, así como los problemas que se han encontrado.

3.1.5 Evaluación.

En esta fase se realiza una evaluación, para ver si se ha alcanzado los objetivos que se plantearon, además se evalúa si los algoritmos y herramientas aplicadas dieron el resultado que se deseaba, para esta fase se ha tomo en consideración la siguiente tarea (Chapman et al., 2000):

- Evaluación de los resultados: Se realiza una evaluación de cada uno de los algoritmos y herramientas implementadas y se compara los resultados que ha brindado cada uno de ellos, también se hace una evaluación global de los resultados obtenidos en la implementación para asegurar si te ha llegado a cumplir que el proyecto requería.

3.2 Obtención de corpus

El presente trabajo forma parte del proyecto de investigación “Minería de datos para la identificación de similitud en la información de pacientes”, a cargo de la directora de tesis. Por tal razón, el corpus que fue obtenido para dicho proyecto cuenta con restricciones de acceso por la sensibilidad de la información. A continuación se detalla dicho corpus.

- **MIMIC II**

MIMIC II, es una base de datos, que contiene datos clínicos muy completos. Los datos de los pacientes son miles de datos de unidades de cuidados intensivos (UCI). En algunas áreas, como médica, quirúrgica, cuidados coronarios y neonatal. Los datos fueron recogidos entre el año 2001 y 2008, en un solo hospital. Debido a que la información que contiene MIMIC II es sensible por la información privada de los pacientes, el acceso a esta base, está restringido, solo podrán acceder los usuarios registrados en PhysioNetWorks. Por lo tanto se deberá seguir las respectivas instrucciones que brinda PhysioNetWorks para solicitar el acceso una vez que ya tenga una cuenta (PhysioNet, 2015).

Para hacer las búsquedas en la base de datos clínicos MIMIC II se utiliza comandos SQL. En este caso, se obtuvo de la tabla *COMORBIDITY_SCORES,* 1000 pacientes y 30 enfermedades con las que se va a trabajar.

3.3 Preprocesamiento

Una vez obtenido el corpus en formato csv, se realizó el preprocesamiento, Primeramente se verifico, si todas las comorbilidades son significativas, y cuenten con una alta correlación, para realizar esta experimentación.

Se realizó un filtro al corpus con la herramienta Excel, con el fin de poder identificar las comorbilidades y pacientes que no son relevantes ya sea porque no contiene datos o hay algún error en la información.

Al momento de analizar los datos se pudo identificar 308 pacientes que no tenían ninguna comorbilidad. Estos fueron excluidos, quedando en total 692 pacientes que tienen entre 1 a 9 enfermedades cada uno.

Para el caso de las 30 enfermedades, se pudo apreciar que existía una que no tenía ningún paciente y dos que tenían únicamente 2 paciente. A estas tres se las eliminó, quedando 27 enfermedades.

Se obtuvo como resultado una reducción del dataset, con 692 pacientes y 27 enfermedades por analizar.

3.4 Experimentación y herramientas

En el presente capítulo se explica el proceso de la experimentación, por medio de algoritmos, herramientas y las gráficas como resultado de la aplicación de los mismos.

3.4.1 Experimentación.

Para iniciar se elaboró una matriz documento–término (pacientes-enfermedades) con los datos del corpus obtenido. La matriz tiene 692 filas que representan a los pacientes y 27 columnas que representan las enfermedades. En la figura 12 se puede observar las 10 primeras filas y todas las columnas.

	DOC	SUBJECT_ID	HADM_ID	CATEGORY	CONGESTIVE_HEART_FAILURE
[1,]	4	4700	9475	1	0
[2,]	5	32611	33875	1	0
[3,]	6	8778	22271	1	0
[4,]	7	28530	35701	1	0
[5,]	8	7874	8734	1	1
[6,]	9	7565	28954	1	0
[7,]	10	11467	620	1	0
[8,]	12	22644	28432	1	0
[9,]	16	13314	10199	1	1
[10,]	17	3541	19300	1	0

	CARDIAC_ARRHTMIAS	VALVULAR_DISEASE	PULMONARI_CIRCULA
[1,]	0	0	0
[2,]	0	0	0
[3,]	0	0	0
[4,]	0	0	0
[5,]	0	0	0
[6,]	0	0	0
[7,]	0	0	0
[8,]	0	0	0
[9,]	0	0	0
[10,]	0	0	0

	PERIPHERAL_VASCULAR	HYPERTENCION	PARALYSIS	OTHER_NEUROLOGICAL
[1,]	0	0	0	0
[2,]	0	1	0	0
[3,]	0	1	0	0
[4,]	0	0	1	0
[5,]	0	0	0	0
[6,]	1	0	0	0
[7,]	0	0	0	0
[8,]	0	0	0	0
[9,]	0	0	0	0
[10,]	0	1	0	0

	CRONIC_PULMONARY	DIABETES_UNCOMPLICATED	DIABETES_COMPLICATED
[1,]	0	0	0
[2,]	0	0	0
[3,]	0	0	0
[4,]	0	0	0
[5,]	0	1	0
[6,]	0	0	0
[7,]	0	0	0
[8,]	0	0	0
[9,]	0	0	0
[10,]	0	0	0

	HYPOTHYROIDISM	RENAL_FAILURE	LIVER_DISEASE	LYMPHOMA
[1,]	0	0	0	0
[2,]	1	0	0	0
[3,]	0	0	0	0
[4,]	0	0	1	0
[5,]	1	0	0	0
[6,]	0	0	0	0
[7,]	0	0	0	1
[8,]	0	0	0	0
[9,]	0	0	0	0
[10,]	0	0	0	0

	METASTATIC_CANCER	SOLID_TUMOR	RHEUMATOID_ARTHRITIS	COAGULOPATHY
[1,]	0	1	0	0
[2,]	0	0	0	0
[3,]	0	0	0	0
[4,]	1	1	0	0
[5,]	0	0	0	1
[6,]	0	0	0	0
[7,]	0	0	0	0
[8,]	0	1	0	0
[9,]	0	0	0	1
[10,]	0	0	0	0

	OBESITY	WEIGHT_LOSS	FLUID_ELECTROLYTE	DEFICIENCY_ANEMIA
[1,]	0	0	0	0
[2,]	0	0	1	1
[3,]	0	0	0	0
[4,]	0	0	1	0
[5,]	0	0	1	0
[6,]	0	0	0	0
[7,]	0	0	0	0
[8,]	0	0	0	0
[9,]	0	0	1	0
[10,]	0	0	0	0

	ALCOHOL_ABUSE	DRUNG_ABUSE	PSYCHOSES	DEPRESSION
[1,]	0	0	0	0
[2,]	0	0	0	0
[3,]	0	0	0	0
[4,]	1	0	0	0
[5,]	0	0	0	0
[6,]	0	0	0	0
[7,]	0	0	0	0
[8,]	0	0	0	0
[9,]	0	0	0	0
[10,]	0	0	0	0

Figura 12. Matriz Documento-Temino.
Elaboración: Propia.

.

Esta matriz es numérica, los pacientes están representados por un id en la columna “Doc”, mientras que las enfermedades están representadas por “0” y “1”. Un “0” significa que el paciente no tiene esa enfermedad. Un “1” significa que el paciente presenta esa enfermedad.

3.4.1.1 LSA.

Con la ayuda de LSA se hizo una reducción de la matriz paciente – enfermedad con el fin de obtener dos componentes o características que representen a todo el conjunto de enfermedades. En base a estos componentes principales, se graficó y se obtuvo una aproximación visual de los grupos de pacientes similares

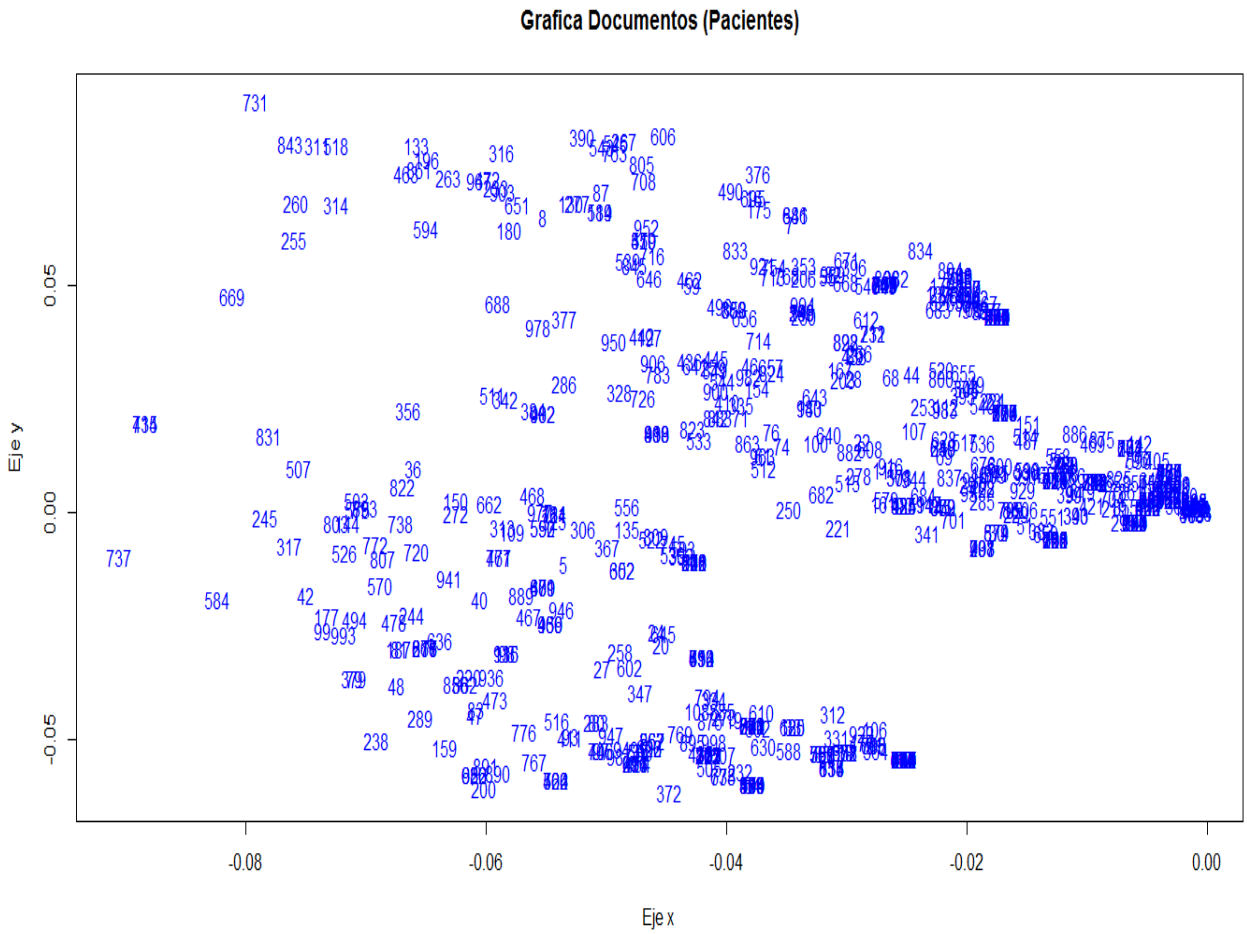

Figura 13. Grafica Matriz de Documentos (Pacientes)
Elaboración: propia.

La grafica 13 muestra los pacientes (números). Los números que se encuentran cercanos muestran caracteristicas similares. Para la grafica se tomaron las dos primeras columas de la matriz LSA, siendo la primera columna el eje “x” y la segunda columna el eje “y”. Los id de los pacientes son los que se indican en color azul en la grafica.

3.4.1.2 Graficas Plotly.

Se utilizó la herramienta Plotly para nuevamente graficar los componentes 1 y 2 de la matriz LSA (matriz reducida pacientes-enfermedades). El proceso para obtener las gráficas se lo puede realizar de dos formas:

1. Exportando un archivo .csv la matriz (paciente – enfermedades) luego en Plothy se importa y se gráfica.

2. Instalando una librería Plotly en Rstudio, para graficar directamente en RStudio.

- Matriz Pacientes:

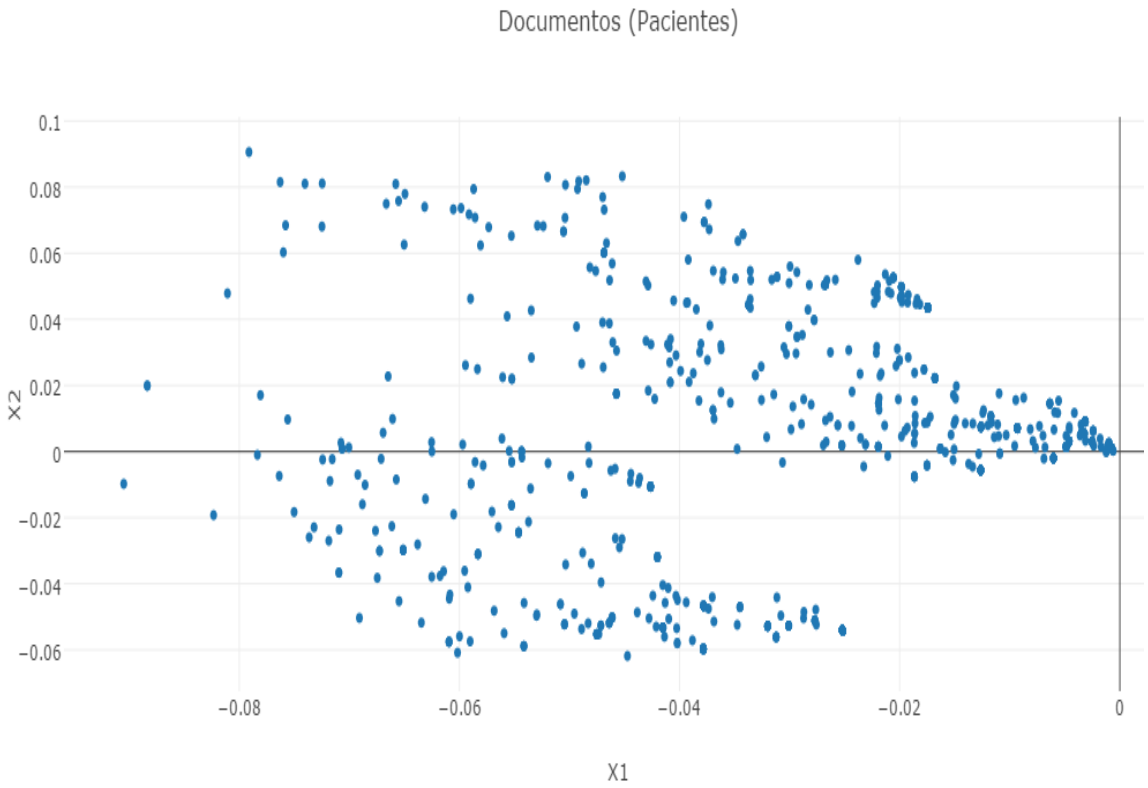

Figura 14. Grafica de matriz de documentos (pacientes) con Plotly
Elaboración: Propia.

Esta grafica igual que la figura 8 se ha tomado de la matriz paciente - enfermedad por medio de LSA, lo único distinto es que están representados por puntos, y lo novedoso aquí es que este tipo de graficas permite mostrar las coordenadas de los ejes “x” y “y” con solo pasar el mouse por encima de cada uno de los puntos.

3.4.1.3 DBSCAN y K-means.

Una vez que se obtuvo las gráficas de la sección 3.4.1.1 y 3.4.1.2, se procedió a identificar grupos de pacientes utilizando dos algoritmos, DBSCAN y K-means sobre la matriz de LSA. Se obtuvo como resultado 5 grupos con DBSCAN siendo uno de ellos el grupo de valores atípicos y 4 grupos con K-means

- Grafica DBSCAN:

Figura 15. DBSCAN con 5 grupos, sobre la matriz de pacientes de LSA, representado por figuras y colores
Elaboración: Propia.

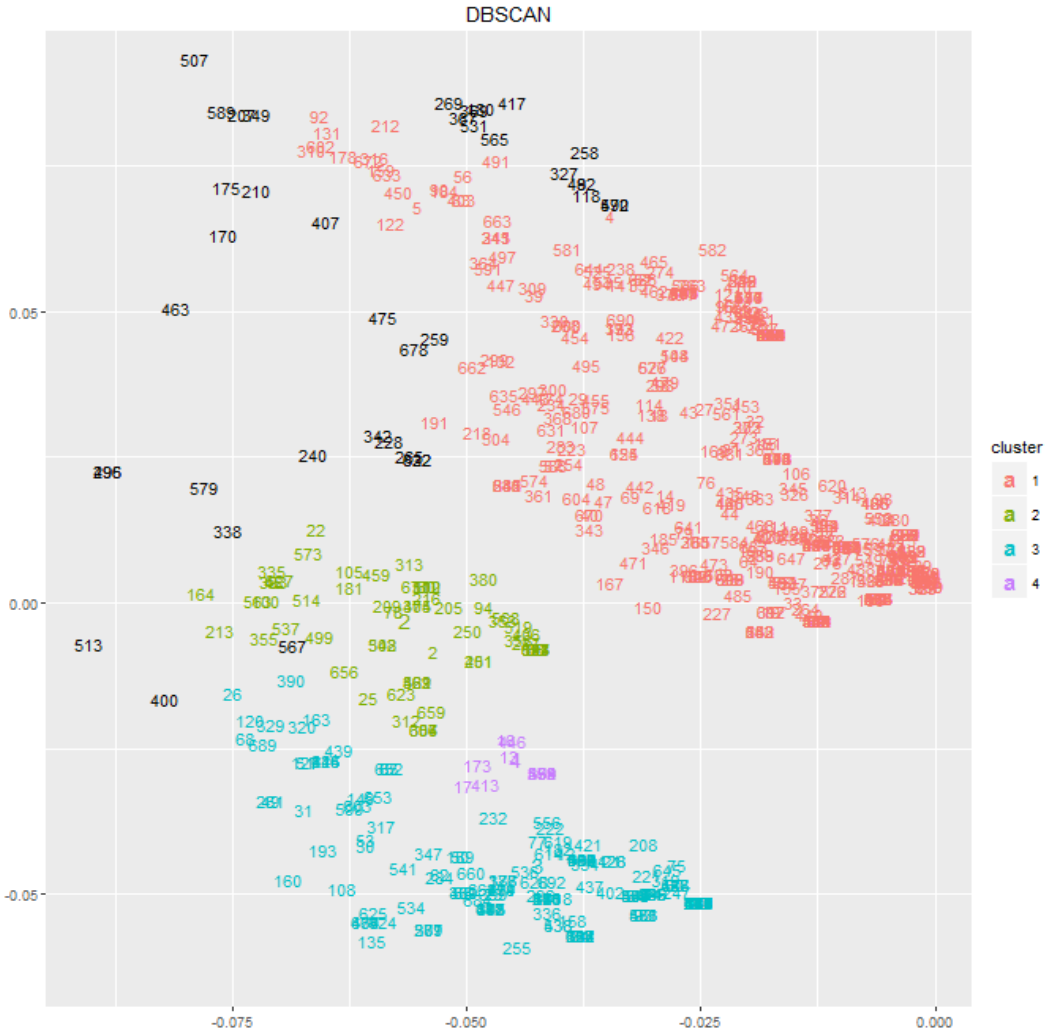

Figura 16. DBSCAN con 5 grupos, sobre la matriz de pacientes de LSA representado por ID de los pacientes y con colores.
Elaboración: Propia.

Una vez que se obtuvo la matriz reducida con LSA, se procedió a utilizar DBSCAN para identificar grupos de pacientes (clusters). Como se pudo observar con las gráficas de los componentes 1 y 2 de LSA, existen valores atípicos que en este caso están representados por el grupo 0 de color negro, que son comunes en dataset que tienen información de enfermedades y pacientes. Por tal razón se decidió aplicar, DBSCAN, ya que este algoritmo trabaja con ruido, y al momento de formar los grupos no causa ningún problema. Además en el ámbito médico los casos atípicos siempre se van a dar porque hay que considerar que los pacientes tienen una enfermedad o síntoma que puede ser muy diferente a otros, por tal motivo no se realizó ningún tratamiento a esos datos atípicos referentes a los pacientes.

El número de grupos obtenido por medio de DBSCAN es de 5 grupos, tomando en cuenta el grupo 0 conformado por datos con ruido, se pueden identificar por el color negro, rojo, verde, azul y purpura.

Tabla 1. Especificación de los datos arrojados por DBSCAN

Dbscan	**Pts=692**	**MinPts=10**	**eps=0.008**		
	0	1	2	3	4
Border	39	20	15	13	1
Seed	0	384	44	166	10
Total	39	404	59	179	11

Elaboración: Propia.

El número de grupos fue obtenido con los dos parámetros principales de DBSCAN, el radio máximo de densidad (esp = 0.008) y el número mínimo de puntos requeridos (MinPts = 10). Estos datos fueron escogidos por medio de la función de DBSCAN llamada kNNdistplot(), la cual permite dibujar la trama k-distancia y encontrar el punto en el que se produce un cambio brusco a lo largo de la curva k-distancia, por medio de esta función se puede definir el valor de épsilon (esp) acompañado del valor (MintPts), como se puede apreciar en la siguiente figura.

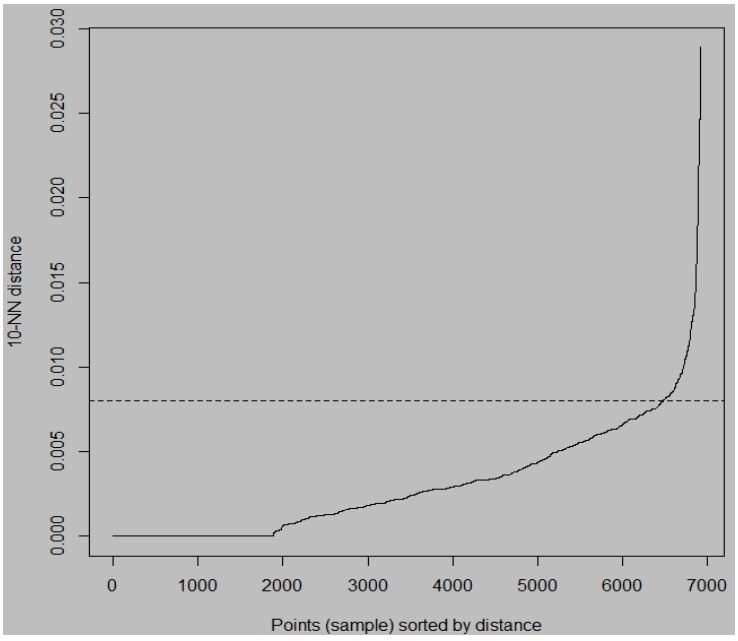

Figura 17. Trama k-distancia para identificar el valor de (esp)
Elaboración: Propia.

Una vez obtenido esta gráfica, se optó por aplicar K-means. Para este algoritmo se obtuvo 4 grupos (clusters) como se puede apreciar en la figura 18. Al igual que DBSCAN, se aplicó K-means con la matriz reducida obtenida con LSA.

- Grafica K-means:

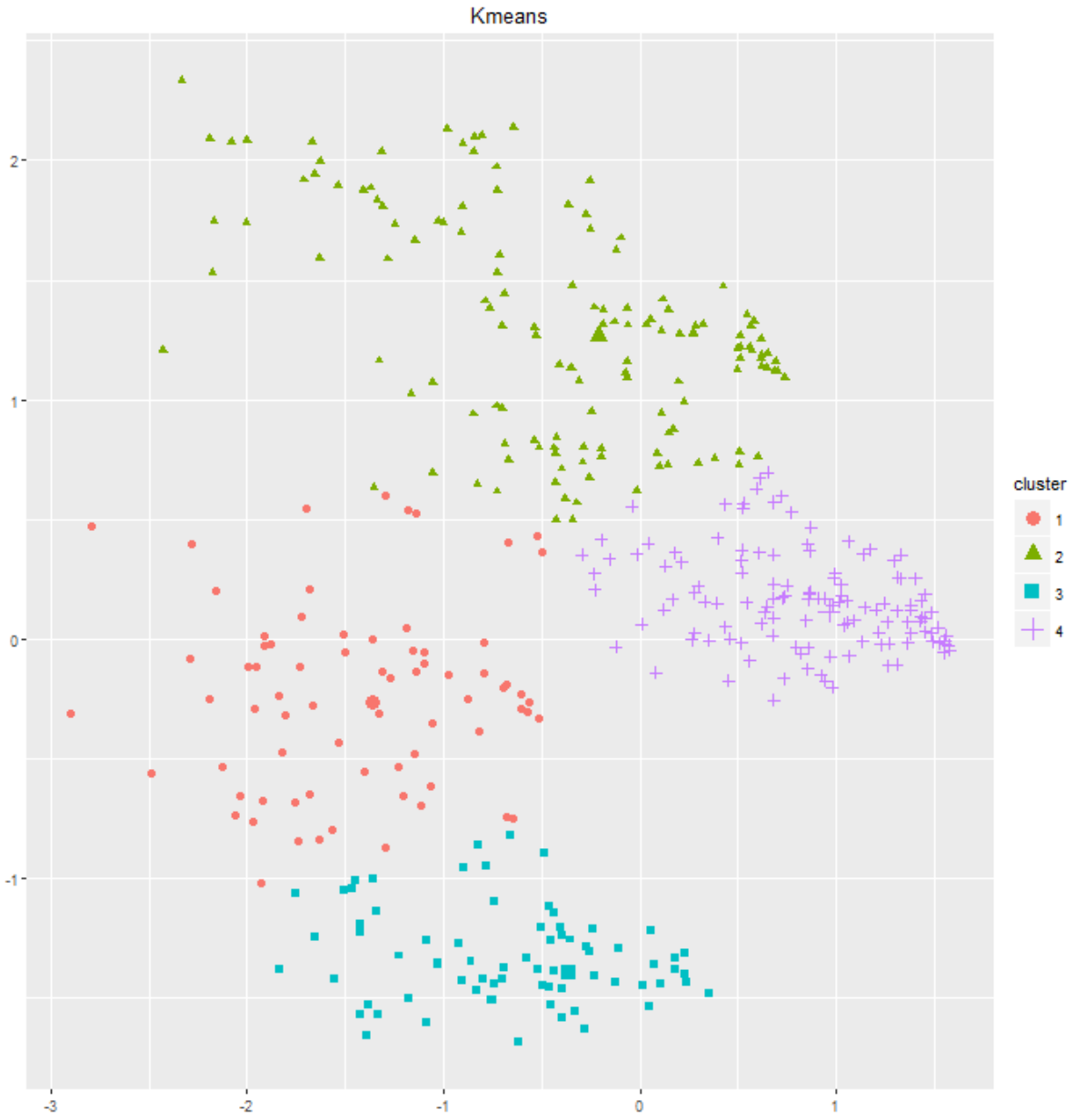

Figura 18. K-means con 4 grupos sobre la matriz de pacientes representados por figuras y con colores
Elaboración: Propia.

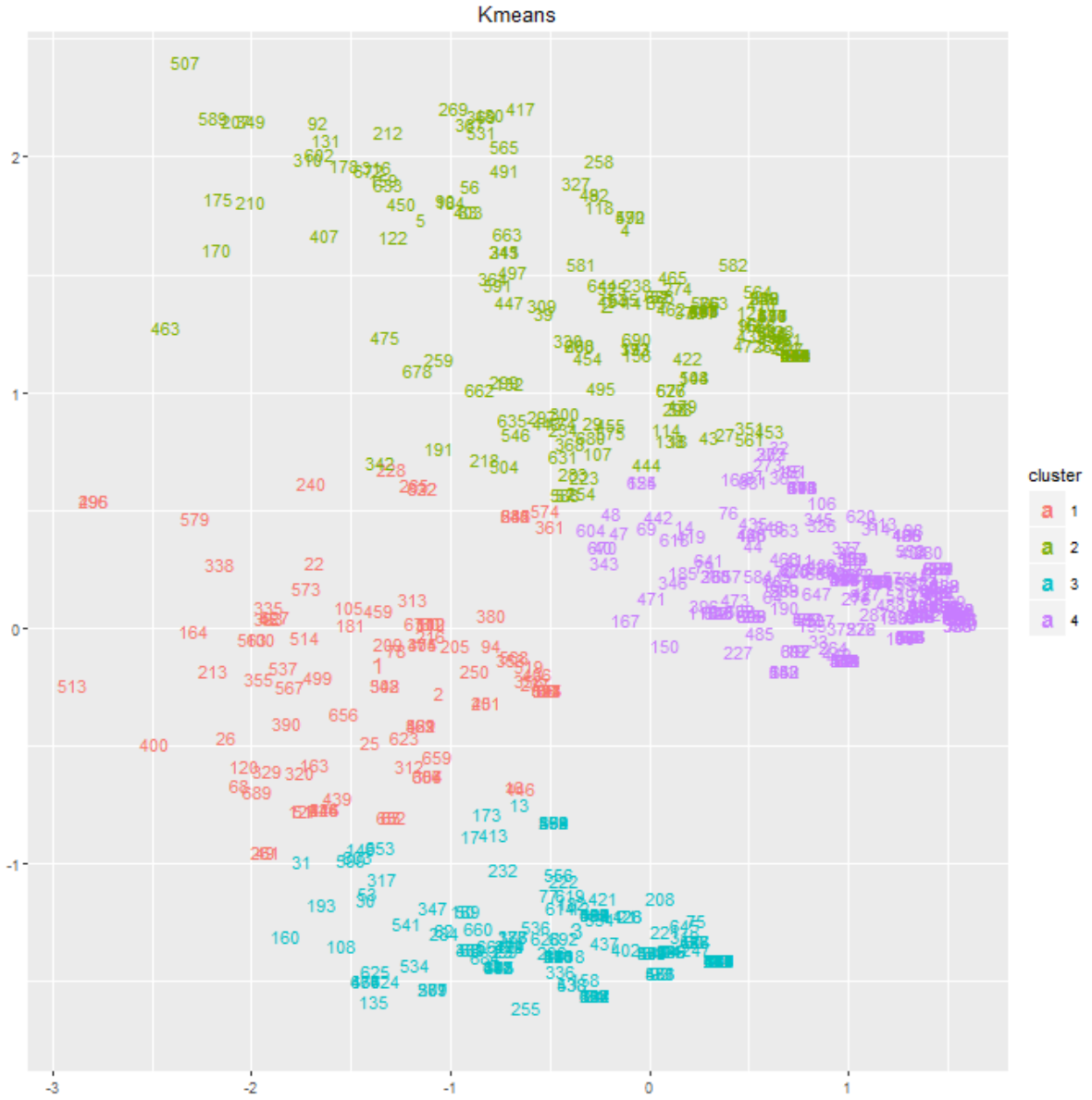

Figura 19. K-means con 4 grupos sobre la matriz de pacientes representados por ID de pacientes y con colores.
Elaboración: Propia.

Cabe señalar que los resultados obtenidos por medio de DBSCAN y K-means estarán influenciados con la selección de los parámetros épsilon y MinPts, en el caso de DBSCAN y K para K-means. Estos resultados por medio de las técnicas de minería de datos aplicadas, contribuirán en una forma automática para encontrar grupos de pacientes similares, convirtiéndose en una ayuda para los especialistas en el área médica, ya que de otra forma tendrán que seguir haciéndolo de forma manual, este tipo de separación de pacientes.

3.4.1.4 Detalle de los grupos.

Con el algoritmo DBSCAN se obtuvo 5 grupos tomando en cuenta al grupo #0 que es el de ruido, con el algoritmo K-means se consiguió 4 grupos.

En esta parte se especifica el número total de pacientes, por cada grupo que se obtuvo con DBSCAN.

Tabla 2. Especificación del número de pacientes de cada grupo con DBSCAN

Grupos	Número de Pacientes
Grupo # 0	39
Grupo # 1	404
Grupo # 2	59
Grupo # 3	179
Grupo # 4	11

Elaboración: Propia.

El número de pacientes que tiene cada grupo por medio de K-means son:

Tabla 3. Especificación del número de pacientes de cada grupo con K-means

Grupos	Número de Pacientes
Grupo # 1	99
Grupo # 2	183
Grupo # 3	242
Grupo # 4	168

Elaboración: Propia.

Además se detalla las enfermedades más comunes, estas fueron consideradas por medio de un porcentaje mayores o iguales al 50 % del total de cada grupo por medio de DBSCAN y K-means.

Porcentaje	Grupos	Enfermedades (27), Documentos (692) DBSCAN
> o = al 50%	Grupo #1	FLUID_ELECTROLYTE
	Grupo #2	CRONIC_PULMONARY, FLUID_ELECTROLYTE, HYPERTENCION
	Grupo #3	CARDIAC_ARRHTMIAS, DIABETES_UNCOMPLICATED, PULMONARI_CIRCULA
	Grupo #4	CONGESTIVE_HEART_FAILURE, HYPERTENCION

Figura 20. Detalle de las enfermedades por cada grupo de DBSCAN
Elaboración: Propia.

En la figura #20 sobre DBSCAN se detallan solo 4 grupos ya que el grupo #0 son datos de ruido por lo que se lo descarta

Porcentaje	Grupos	Enfermedades (27), Documentos (692) KMEANS
> o = al 50%	Grupo #1	CARDIAC_ARRHTMIAS, CRONIC_PULMONARY, CONGESTIVE_HEART_FAILURE, FLUID_ELECTROLYTE, HYPERTENCION
	Grupo #2	CONGESTIVE_HEART_FAILURE, FLUID_ELECTROLYTE
	Grupo #3	
	Grupo #4	DIABETES_UNCOMPLICATED, HYPERTENCION

Figura 21. Detalle de las enfermedades por cada grupo de K-means
Elaboración: Propia.

3.4.1.5 Coseno.

Finalmente se procedió a utilizar el Coseno sobre la matriz original (692 filas y 27 columnas), para poder posteriormente compararla con los grupos que se obtuvo de DBSCAN y K-means. Quedando de la siguiente manera el coseno en una nueva matriz.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	1.0000000	0.0000000	0.0000000	0.4082483	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000	0.5773503	0.0000000	0.5773503	0.0000000
2	0.0000000	1.0000000	0.5000000	0.2041241	0.4472136	0.0000000	0.0000000	0.0000000	0.2886751	0.5000000	0.7071068	0.5773503	0.2500000	0.0000000	0.0000000
3	0.0000000	0.5000000	1.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000	0.7071068	0.5773503	0.5000000	0.0000000	0.0000000
4	0.4082483	0.2041241	0.0000000	1.0000000	0.1825742	0.0000000	0.0000000	0.4082483	0.2357023	0.0000000	0.0000000	0.2357023	0.0000000	0.2357023	0.0000000
5	0.0000000	0.4472136	0.0000000	0.1825742	1.0000000	0.0000000	0.0000000	0.0000000	0.7745967	0.0000000	0.3162278	0.2581989	0.2236068	0.0000000	0.3162278
6	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
7	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
8	1.0000000	0.0000000	0.0000000	0.4082483	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000	0.5773503	0.0000000	0.5773503	0.0000000
9	0.0000000	0.2886751	0.0000000	0.2357023	0.7745967	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000	0.2886751	0.0000000	0.4082483
10	0.0000000	0.5000000	1.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000	0.7071068	0.5773503	0.5000000	0.0000000	0.0000000
11	0.0000000	0.7071068	0.7071068	0.0000000	0.3162278	0.0000000	0.0000000	0.0000000	0.0000000	0.7071068	1.0000000	0.8164966	0.3535534	0.0000000	0.0000000
12	0.5773503	0.5773503	0.5773503	0.2357023	0.2581989	0.0000000	0.0000000	0.5773503	0.0000000	0.5773503	0.8164966	1.0000000	0.2886751	0.3333333	0.0000000
13	0.0000000	0.2500000	0.5000000	0.0000000	0.2236068	0.0000000	0.0000000	0.0000000	0.2886751	0.5000000	0.3535534	0.2886751	1.0000000	0.0000000	0.3535534
14	0.5773503	0.0000000	0.0000000	0.2357023	0.0000000	0.0000000	0.0000000	0.5773503	0.0000000	0.0000000	0.0000000	0.3333333	0.0000000	1.0000000	0.0000000
15	0.0000000	0.0000000	0.0000000	0.0000000	0.3162278	0.0000000	0.0000000	0.0000000	0.4082483	0.0000000	0.0000000	0.0000000	0.3535534	0.0000000	1.0000000

Figura 22. Coseno de la Matriz original
Elaboración: Propia.

En la figura 22 se aprecia una matriz, que contiene la similitud de cada uno de los pacientes con otros, en base a las enfermedades.

3.4.2 Herramientas.

Para el desarrollo del tema se trabajó con el Lenguaje de programación R junto con el entorno de desarrollo integrado (IDE) RStudio, esta elección se la tomó, ya que permite trabajar de una manera sencilla y practica los objetos que se crean, gracias a su interfaz. El presente trabajo de fin de titulación es de interés desarrollar la parte práctica sobre matrices, por lo tanto se decidió llevar acabo con este IDE, ya que permite una visualización y manejo de las mismas de forma eficaz.

Para tener mayor seguridad de las gráficas que se obtuvo con el algoritmo de LSA en el RStudio, se realizó una comprobación de las mismas, con otra herramienta online, gratuita, llamada Plotly, que sirve para graficar plot y otros tipos de gráficas.

3.4.2.1 Rstudio.

RStudio es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, editor de resaltado de sintaxis que soporta la ejecución de código directa, así como herramientas para graficar, el historial, la depuración y la gestión de espacios de trabajo (RStudio, 2016).

3.4.2.1.1 Características:

RStudio es el principal entorno de desarrollo integrado para R. Está disponible en código abierto y las ediciones comerciales en el escritorio (Windows, Mac y Linux) y desde un navegador web a un servidor Linux corriendo RStudio Server o RStudio Server Pro (RStudio, 2016).

- Un IDE que fue construido sólo para R:
  - El resaltado de sintaxis, completado de código, Identación inteligente
  - Ejecutar código R directamente desde el editor de código fuente
  - Saltar rápidamente a las definiciones de funciones

- Trae su flujo de trabajo en conjunto
  - R integra ayuda y documentación
  - Administrar fácilmente múltiples directorios de trabajo utilizando proyectos
  - navegador de espacio de trabajo y visualizador de datos

- Autoría de gran alcance y depuración
  - depurador interactivo para diagnosticar y corregir los errores rápidamente
  - Amplio paquete de herramientas de desarrollo
  - Autoría con Sweave y R Markdown (RStudio, 2016).

3.4.2.2 Plotly.

Plotly es una herramienta diseñada para la visualización de datos y análisis del mismo, algo muy importante que cuenta con una API de gráficos científicos para:

- Python
- R
- Matlab
- Perl
- Julia
- Arduino
- Rest

Además cuenta con una gran variedad de graficas en línea para la colaboración de las mismas (Plotly, 2015).

3.5 Evaluación de los algoritmos

Para hacer una evaluación de las técnicas o algoritmos que se implementó, se desarrolló una comparación con:

- LSA
- Plotly
- Coseno
- DBSCAN
- K-means.

Tabla 4. Beneficios brindados por cada técnicas/algoritmos usados

Beneficios				
LSA	**Plotly**	**Coseno**	**DBSCAN**	**K-means**
Matriz: (Pacientes, enfermedades) de la matriz original.	Graficas de Pacientes	Una Matriz Coseno de la matriz original.	Solución al graficar los datos de pacientes con ruido	Gráfica con 4 grupos de la matriz de Pacientes
Grafica de Pacientes.	Visualización rápida de las condenas (eje x, y) sobre cada paciente	Verificación que se hizo con los grupos de DBSCAN.	Gráfica con 5 grupos de la matriz de Pacientes	
.		Verificación que se hizo con los grupos de K-means.		

Elaboración: Propia.

CAPÍTULO IV

ANÁLISIS

4.1 Análisis de los resultados

Para este análisis se inició exportando los grupos de DBSCAN y K-means en la herramienta Excel, para identificar las enfermedades más comunes de cada grupo de pacientes. Esto se realizó con el objetivo de saber la similitud o diferencia entre ellos. Además se estimaron rangos de pacientes dependiendo del porcentaje de pacientes en cada grupo (100%, 75%, 50% y 25%). Esto se hizo con el objetivo de conocer cuáles son las enfermedades más comunes en los grupos, es decir considerando un número de pacientes igual o mayor al 50 %. Las enfermedades en cada porcentaje y en cada grupo se indican a continuación.

- Grupos de DBSCAN:

	Grupo #1	
	Total de Pacientes = 404	
Rangos de pacientes	Porcentaje	
304 .. 404	100%	
203 ..303	75%	
102 .. 202	50%	FLUID_ELECTROLYTE
0 .. 101	25%	26 restantes

Figura 23. Grupo #1 de las enfermedades consideradas >= al 50% con DBSCAN
Elaboración: Propia.

En el grupo #1, el 50 % de los pacientes poseen la enfermedad FLUID_ELECTROLYTE, mientras que las 26 enfermedades restantes se encuentran en el 25% de los pacientes.

	Grupo #2	
	Total de Pacientes = 59	
Rangos de pacientes	Porcentaje	
45 .. 59	100%	HYPERTENCION, FLUID_ELECTROLYTE
31 .. 44	75%	
16 .. 30	50%	CRONIC_PULMONARY
0 .. 15	25%	24 restantes

Figura 24. Grupo #2 de las enfermedades consideradas >= al 50% con DBSCAN
Elaboración: Propia.

En el grupo #2, el 100% de los pacientes poseen las enfermedades de HYPERTENSION y FLUID_ELECTROLYTE, además el 50% de los pacientes tiene la enfermedad CRONIC_PULMONARY, mientras que las 24 enfermedades restantes se encuentran en el 25% de los pacientes.

.

	Grupo #3	
	Total de Pacientes = 179	
Rangos de pacientes	Porcentaje	
135 .. 179	100%	PULMONARI_CIRCULA
91 .. 134	75%	
46 .. 90	50%	CARDIAC_ARRHTMIAS, DIABETES_UNCOMPLICATED
0 .. 45	25%	24 restantes

Figura 25. Grupo #3 de las enfermedades consideradas >= al 50% con DBSCAN
Elaboración: Propia.

En el grupo #3, el 100% de los pacientes poseen la enfermedad PULMONARI_CIRCULA, además el 50 % de los pacientes tiene CARIAC_ARRHITMIA y DIABETES_UNCOMPLICATED, las 24 enfermedades restantes se encuentran en el 25% de los pacientes.

	Grupo #4	
	Total de Pacientes = 11	
Rangos de pacientes	Porcentaje	
9 .. 11	100%	CONGESTIVE_HEART_FAILURE, HYPERTENCION
7 .. 8	75%	
4 .. 6	50%	
0 .. 3	25%	25 restantes

Figura 26. Grupo #4 de las enfermedades consideradas >= al 50% con DBSCAN
Elaboración: Propia.

En el grupo #4, el 100% de los pacientes poseen las enfermedades CONGESTIVE_HEART_FAILURE y HYPERTENSION, mientras que las 25 enfermedades restantes se encuentran en el 25% de los pacientes.

- Grupos de K-means:

Grupo #1		
Total de Pacientes = 99		
Rangos de pacientes	Porcentaje	
75 .. 99	100%	HYPERTENCION
51 ..74	75%	FLUID_ELECTROLYTE
26 .. 50	50%	CARDIAC_ARRHTMIAS, CRONIC_PULMONARY, CONGESTIVE_HEART_FAILURE
0 .. 25	25%	22 restantes

Figura 27. Grupo #1 de las enfermedades consideradas >= al 50% con K-means
Elaboración: Propia.

En el grupo #1, el 100% de los pacientes poseen la enfermedad de HYPERTENSION, el 75% tiene la enfermedad FLUID_ELECTROLYTE, el 50% poseen las enfermedades CARDIAC_ARRITMIAS, CRONIC_PULMONARY y COONGESTIVE_HEART_FAULURE, mientras que las 22 enfermedades restantes se encuentran en el 25% de los pacientes.

Grupo #2		
Total de Pacientes = 183		
Rangos de pacientes	Porcentaje	
138 .. 183	100%	FLUID_ELECTROLYTE
93 ..137	75%	
47 .. 92	50%	CONGESTIVE_HEART_FAILURE
0 .. 46	25%	25 restantes

Figura 28. Grupo #2 de las enfermedades consideradas >= al 50% con K-means
Elaboración: Propia.

En el grupo #2, el 100% de los pacientes poseen la enfermedad de FLUID_ELECTROLYTE, el 50% tiene CONGESTIVE_HEART_FAILURE, mientras que las 25 enfermedades restantes se encuentran en el 25% de los pacientes.

	Grupo #3	
	Total de Pacientes =242	
Rangos de pacientes	Porcentaje	
183 .. 242	100%	
122 .. 182	75%	
62 .. 121	50%	
0 .. 61	25%	27 enfermedades

Figura 29. Grupo #3 de las enfermedades consideradas >= al 50% con K-means
Elaboración: Propia.

En el grupo #3, el 25% de los pacientes poseen las 27 enfermedades

	Grupo #4	
	Total de Pacientes =168	
Rangos de pacientes	Porcentaje	
127 .. 168	100%	HYPERTENCION
85 .. 126	75%	
43 .. 84	50%	DIABETES_UNCOMPLICATED
0 .. 42	25%	25 enfermedades

Figura 30. Grupo #4 de las enfermedades consideradas >= al 50% con K-means
Elaboración: Propia.

En el grupo #4, el 100% de los pacientes poseen la enfermedad de HYPERTENSION, el 50% tiene la enfermedad de DIABETES_UNCOMPLICATED, mientras que las 25 enfermedades restantes se encuentran en el 25% de los pacientes

4.1.1 Análisis de los grupos.

Para realizar el análisis de los grupos que se obtuvo en la experimentación, se realizó una comparación de los mismos, grupo por grupo, para identificar porque enfermedades son similares y por cuales son diferentes entre sí, para ellos se desarrollara el análisis por separado para DBSCAN y K-means.

- Análisis de los grupos en conjunto de DBSCAN

Tabla 5. Análisis de los grupos, con los porcentajes de cada enfermedad de DBSCAN

Análisis de los grupos de DBSCAN con el porcentaje de cada enfermedad			
Grupo #1	**Grupo #2**	**Grupo #3**	**Grupo #4**
CARDIAC_ARRHTMIAS 25%	CARDIAC_ARRHTMIAS 25%	CARDIAC_ARRHTMIAS 50%	CARDIAC_ARRHTMIAS 25%
CRONIC_PULMONARY 25%	CRONIC_PULMONARY 50%	CRONIC_PULMONARY 25%	CRONIC_PULMONARY 25%
CONGESTIVE_HEART_FAILURE 25%	CONGESTIVE_HEART_FAILURE 25%	CONGESTIVE_HEART_FAILURE 25 %	CONGESTIVE_HEART_FAILURE 100 %
DIABETES_UNCOMPLICATED 25%	DIABETES_UNCOMPLICATED 25%	DIABETES_UNCOMPLICATED 50%	DIABETES_UNCOMPLICATED 25%
FLUID_ELECTROLYTE 50%	FLUID_ELECTROLYTE 100%	FLUID_ELECTROLYTE 25%	FLUID_ELECTROLYTE 25%
HYPERTENSION 25%	HYPERTENSION 100%	HYPERTENSION 25%	HYPERTENSION 100%
PULMONARY_CIRCULA 25%	PULMONARY_CIRCULA 25%	PULMONARY_CIRCULA 100%	PULMONARY_CIRCULA 25%

Elaboración: Propia.

En la tabla 5 se puede apreciar las enfermedades más relevantes (mayores o iguales al 50% de los pacientes) que poseen todos los grupos de DBSCAN, además de una comparación del porcentaje de pacientes, de cada enfermedad por grupo, con el fin de mostrar las diferencias que existen entre todo estos 4 grupos de DBSCAN

- Análisis de los grupos en conjunto de K-means

De igual forma que se hizo el análisis en el caso de los grupos de DBSCAN, se aplica para el caso de K-means. Esto se aprecia en la tabla 6.

Tabla 6. Análisis de los grupos, con los porcentajes de cada enfermedad de K-means

Análisis de los grupos de K-means, con el porcentaje de cada enfermedad			
Grupo #1	**Grupo #2**	**Grupo #3**	**Grupo #4**
CARDIAC_ARRHTMIAS 50%	CARDIAC_ARRHTMIAS 25%	CARDIAC_ARRHTMIAS 25%	CARDIAC_ARRHTMIAS 25%
CRONIC_PULMONARY 50%	CRONIC_PULMONARY 25%	CRONIC_PULMONARY 25%	CRONIC_PULMONARY 25%
CONGESTIVE_HEART_FAILURE 50%	CONGESTIVE_HEART_FAILURE 50%	CONGESTIVE_HEART_FAILURE 25 %	CONGESTIVE_HEART_FAILURE 25%
DIABETES_UNCOMPLICATED 25%	DIABETES_UNCOMPLICATED 25%	DIABETES_UNCOMPLICATED 25 %	DIABETES_UNCOMPLICATED 50 %
FLUID_ELECTROLYTE 75%	FLUID_ELECTROLYTE 100%	FLUID_ELECTROLYTE 25%	FLUID_ELECTROLYTE 25%
HYPERTENSION 100%	HYPERTENSION 25%	HYPERTENSION 25%	HYPERTENSION 100%

Elaboración: Propia.

4.1.2 Verificación de los grupos.

Para tener una mayor certeza del análisis que se hizo sobre los grupos tanto para DBSCAN como para K-means, se decidió verificarlos por medio de la medida de similitud Coseno. Se la obtuvo en la experimentación y fue aplicada a la matriz original. Para la debida validación entre el coseno y los grupos DBSCAN y K-means, se fueron tomando dos id de pacientes por cada grupo que cumplan con el porcentaje de aceptación mayor de 50%, y se fue verificando con la matriz del coseno, de igual forma que cumplan ambos id con el porcentaje de aceptación y se realizó tres verificaciones por cada grupo, de esta forma se pudo culminar la verificación de toda la experimentación.

- Verificación para DBSCAN por medio de Coseno:

Tabla 7. Verificación del grupo #1 DBSCAN con Coseno

Grupo #1 (DBSCAN)			
Verificación	**Id Pacientes 1**	**Id Pacientes 2**	**Porcentaje de similitud**
# 1	758	754	0.5
# 2	915	23	0,70
# 3	12	4	1
# 4	963	61	1

Elaboración: Propia.

Tabla 8. Verificación del grupo #2 DBSCAN con Coseno

Grupo #2 (DBSCAN)			
Verificación	**Id Pacientes 1**	**Id Pacientes 2**	**Porcentaje de similitud**
# 1	126	5	0,7
# 2	848	302	0,82
# 3	571	202	1

#4	461	144	0,87

Elaboración: Propia.

Tabla 9. Verificación del grupo #3 DBSCAN con Coseno

Grupo #3 (DBSCAN)			
Verificación	**Id Pacientes 1**	**Id Pacientes 1**	**Porcentaje de similitud**
# 1	17	6	1
# 2	336	20	1
# 3	561	80	1
# 4	785	91	1

Elaboración: Propia.

Tabla 10. Verificación del grupo #4 DBSCAN con Coseno

Grupo #4 (DBSCAN)			
Verificación	**Id Pacientes 1**	**Id Pacientes 1**	**Porcentaje de similitud**
# 1	430	20	0,71
# 2	645	413	0,67
# 3	653	24	0,71
# 4	793	602	0,82

Elaboración: Propia.

- Verificación para K-means por medio de Coseno:

Tabla 11. Verificación del grupo #1 K-means con Coseno

Grupo #1 (K-means)			
Verificación	**Id Pacientes**		**Porcentaje**
# 1	306	164	0.87

# 2	302	126	0.82
# 3	993	98	0.52
# 4	550	513	0,82

Elaboración: Propia.

Tabla 12. Verificación del grupo #2 K-means con Coseno

Grupo #2 (K-means)			
Verificación	**Id Pacientes**		**Porcentaje**
# 1	841	725	1
# 2	806	350	0,71
# 3	599	162	0.82
# 4	470	359	1

Elaboración: Propia.

Tabla 13. Verificación del grupo #3 K-means con Coseno

Grupo #3 (K-means)			
Verificación	**Id Pacientes**		**Porcentaje**
# 1	554	299	1
# 2	303	148	1
# 3	863	378	0.82
# 4	249	22	0,82

Elaboración: Propia.

Tabla 14. Verificación del grupo #4 K-means con Coseno

Grupo #4 (K-means)			
Verificación	**Id Pacientes**		**Porcentaje**
# 1	257	217	1
# 2	160	6	1

# 3	622	404	0.87
# 4	211	19	0,5

Elaboración: Propia.

- Verificación por medio de las Enfermedades comunes entre DBSCAN y K-means:

Tabla 15. Verificación de enfermedades DBSCAN y K-means

DBSCAN y K-means
CARDIAC_ARRHTMIAS
CRONIC_PULMONARY
CONGESTIVE_HEART_FAILURE
DIABETES_UNCOMPLICATED
FLUID_ELECTROLYTE
HYPERTENSION

Elaboración: Propia.

Como se puede ver en la tabla 15, con BDSCAN and K-means se obtuvieron las mismas enfermedades relevantes o presentes en más del 50% de la población. Además revisando los grupos DBSCAN y K-means, existe una similitud entre los grupos. Por ejemplo el grupo 1 de DBSCAN está contenido en los grupo 2 y 3 de K-means. El grupo 2 de DBSCAN está presente en el grupo 1 de K-means. El grupo 3 de DBSCAN correspondería al grupo 4 de K-means. El grupo 4 de DBSCAN es muy pequeño con solo 11 pacientes este podría estar integrado en el grupo 1 de K-means

CONCLUSIONES

Al finalizar el presente trabajo de fin de titulación se concluye con lo siguiente:

- La identificación de la base de datos MIMIC II, contribuyo satisfactoriamente al desarrollo del trabajo, dando calidad a los datos en este trabajo. Esta base cuenta con prestigio al ser referenciada por trabajos científicos. Los datos se conforman por 30 comorbilidades y 1000 pacientes. Estos datos pasaron por un preprocesamiento antes de ser utilizado en la etapa de experimentación.

- Para el presente trabajo se seleccionó dos tipos de métodos de minería de datos que permitan la identificación de características y agrupación DBSCAN y K-means como métodos de agrupación de los pacientes dieron como resultado grupos similares de pacientes, esto se pudo observar a través de las gráficas y de las enfermedades relevantes en cada grupo.

- La implementación de LSA como método de reducción y selección de características importantes, dio grandes resultados. En base a dos características principales se formaron los grupos en ambos métodos: DBSCAN y K-means.

- Se utilizó DBSCAN, por su desempeño a la hora de trabajar con datos con ruido y los datos médicos poseen este tipo de problema. Por otro lado el algoritmo K-means, se tomó como un algoritmo alternativo ya que ha sido utilizado en otros trabajos relacionados. En ambos casos se obtuvo resultados similares

- Para realizar una verificación sobre los grupos de pacientes obtenidos tanto con DBSCAN y K-means, se obtuvo la matriz coseno de similitud de los pacientes, ayudando a la hora de verificar los pacientes. Esto ayudo a verificar que los pacientes de cada grupo tengan alguna similitud.

- Con ambos métodos se obtuvo las mismas enfermedades relevantes. Además haciendo un análisis de los grupos obtenidos con DBSCAN y K-means, se tiene los mismos grupos. Esto se puede verificar gráficamente y a través de las enfermedades que conforman cada grupo.

RECOMENDACIONES

Al finalizar el presente trabajo de fin de titulación se recomienda lo siguiente:

- Para una continuidad de este proyecto se recomienda buscar otros métodos y técnicas para realizar la agrupación de pacientes, se podría trabajar con métodos jerárquicos por ejemplo.

- Además, para trabajos futuros se recomienda la integración de especialistas médicos que ayuden en la selección de las mejores características o variables y además evalúen los resultados.

- La selección del lenguaje de programación queda abierta ya que se puede realizar tanto en Python o en weka, las matrices son fácilmente exportadas para ser trabajadas en cualquiera de estos lenguajes.

BIBLIOGRAFÍA

Rodríguez, B., A., Simón, C. A., Guevara M. E., y Hojas, M. W. (2015). Modelo de Representación de textos basado en grafo para la minería de texto. *Ciencias de la Información, 46*(1) Enero-Abril, 63-71. Recuperado de http://www.redalyc.org/articulo.oa?id=181439409009

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer-Verlag New York. http://doi.org/10.1007/978-1-4614-3223-4

Antonelli, D., Baralis, E., Bruno, G., Cerquitelli, T., Chiusano, S., & Mahoto, N. (2013). Analysis of diabetic patients through their examination history. *Expert Systems with Applications*, *40*(11), 4672–4678. http://doi.org/10.1016/j.eswa.2013.02.006

Chapman, P. C., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0.

Cuc, I., & Sosa, G. D. (2015). Application of Clustering Techniques for Lung Sounds to Improve Interpretability and Detection of Crackles * Aplicación de Técnicas de Clustering en Sonidos Adventicios para Mejorar la Interpretabilidad y Detección de Estertores, *11*(1), 53–62. Retrieved from http://revistascientificas.cuc.edu.co/index.php/ingecuc/article/download/366/2015105

Han, J., Kamber, M., & Pei, J. (2011). *Data Mininge: Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

Holzinger, A., Schantl, J., & Schroettner, M. (2014). Biomedical text mining: State-of-the-art, open problems and future challenges. *… Discovery and Data …*, 271–300. Retrieved from http://link.springer.com/chapter/10.1007/978-3-662-43968-5_16

Jonnalagadda, S., Cohen, T., Wu, S., & Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, *45*(1), 129–140. http://doi.org/10.1016/j.jbi.2011.10.007

Landauer, T., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2&3), 259–284. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01638539809545028

Lyalina, S., Percha, B., LePendu, P., Iyer, S. V, Altman, R. B., & Shah, N. H. (2013). Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, *20*(e2), e297–305. http://doi.org/10.1136/amiajnl-2013-001933

Orallo, H., Ramirez, F., Quintana, C. R., & Jose., M. (2014). *Introducción a la Minería de Datos.* Madrid: Pearson Prentice Hall.

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. *Unpublished*, 164–174. Retrieved from http://marmota.dlsi.uji.es/WebBIB/papers/2007/1_Pascual-MIA-2007.pdf

PhysioNet. (2015). MIMIC II: Clinical Database Overview. Retrieved January 12, 2016, from https://physionet.org/mimic2/mimic2_clinical_overview.shtml

Plotly. (2015). Plotly. Retrieved April 17, 2016, from https://plot.ly/

Romero Nieva, F. D. O. (2008, February). La historia clínica: Aspectos asistenciales. Retrieved from http://www.revistahospitalarias.org/info_2008/01_191_03.htm

RStudio. (2016). RStudio. Retrieved April 10, 2016, from https://www.rstudio.com/

Van der Esch, M., Knoop, J., van der Leeden, M., Roorda, L. D., Lems, W. F., Knol, D. L., & Dekker, J. (2015). Clinical phenotypes in patients with knee osteoarthritis: A study in the Amsterdam osteoarthritis cohort. *Osteoarthritis and Cartilage*, *23*(4), 544–549. http://doi.org/10.1016/j.joca.2015.01.006

wikipedia. (2016). DBSCAN. Retrieved June 20, 2016, from https://es.wikipedia.org/wiki/DBSCAN

ANEXOS

ANEXO 1

Archivo matriz documento-termino_27_ enfermedades.xlsx

El archivo de la matriz documento-termino está disponible en: https://github.com/rijimbo/Mineria_de_datos_para_la_identificacion_de_similitud_en_la_informacion_de_pacientes/blob/master/Tesis_roger/Data/matriz%20termino-documento_27_enfermedades.xlsx

ANEXO 2

Archivo Matrix tk_pacientes(27 enfer).xlsx

El archivo que contiene la matriz reducida “tk” que se trabajó para graficar los pacientes, se encuentra disponible en: https://github.com/rijimbo/Mineria_de_datos_para_la_identificacion_de_similitud_en_la_informacion_de_pacientes/blob/master/Tesis_roger/Data/Matrix%20tk_pacientes(27%20enfer).xlsx

ANEXO 3

Archivo DBSCAN.xlsx

El archivo que se realizó el análisis de los grupos de BDSCAN se encuentran disponibles en: https://github.com/rijimbo/Mineria_de_datos_para_la_identificacion_de_similitud_en_la_informacion_de_pacientes/blob/master/Tesis_roger/Scripts/DBSCAN.xlsx

ANEXO 4

Archivo Kmeans.xlsx

El archivo que se realizó el análisis de los grupos de k-menas se encuentran disponibles en:
https://github.com/rijimbo/Mineria_de_datos_para_la_identificacion_de_similitud_en_la_informacion_de_pacientes/blob/master/Tesis_roger/Scripts/kmeans.xlsx

ANEXO 5

Archivo Coseno.xlsx

La matriz de coseno que se utilizó para la verificación de los grupos de DBSCAN y K-means se encuentra disponible en:
https://github.com/rijimbo/Mineria_de_datos_para_la_identificacion_de_similitud_en_la_informacion_de_pacientes/blob/master/Tesis_roger/Scripts/Coseno.xlsx

ANEXO 6

Código del proyecto

Código y archivos de todo el proyecto disponible en: https://github.com/rijimbo/Mineria_de_datos_para_la_identificacion_de_similitud_en_la_informacion_de_pacientes/tree/master/Tesis_roger