MINERÍA DE DATOS PARA EL DESCUBRIMIENTO DE PATRONES EN ENFERMEDADES RESPIRATORIAS EN BOGOTÁ, COLOMBIA

ERIKA ANDREA ROJAS GUTIÉRREZ
JUAN SEBASTIÁN AGUILAR

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
MODALIDAD TRABAJO DE INVESTIGACIÓN
BOGOTÁ D.C., COLOMBIA
2017

MINERÍA DE DATOS PARA EL DESCUBRIMIENTO DE PATRONES EN ENFERMEDADES RESPIRATORIAS EN BOGOTÁ, COLOMBIA

ERIKA ANDREA ROJAS GUTIÉRREZ
JUAN SEBASTIÁN AGUILAR

TRABAJO DE GRADO PARA OPTAR AL TÍTULO DE

INGENIERO DE SISTEMAS

Director (a):
JOHN ALEXANDER VELANDIA
Título (Prof. M.Sc. Eng.)

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
TRABAJO DE INVESTIGACIÓN
BOGOTÁ D.C., COLOMBIA
2017

Nota de Aceptación

_____

_____

_____

_____

_____

_____

Jurado

_____

John Alexander Velandia Vega

Director

_____

Revisor Metodológico

Bogotá, 14, Noviembre, 2017

AGRADECIMIENTOS

A Dios por permitirnos ser guía en cada objetivo que nos proponemos.

A nuestros padres por su apoyo constante e incondicional durante toda la carrera y en nuestras vidas.

A nuestro director de trabajo de grado el profesor John Alexander Velandia por ser de guía durante la elaboración de este proyecto generador de conocimiento.

A nuestros compañeros y profesores de la Universidad Católica de Colombia los cuales dejan enseñanzas en nuestras vidas.

Al profesor Raúl Menéndez por la colaboración en el acceso a las fuentes de datos.

CONTENIDO

LISTA DE TABLAS

LISTA DE ILUSTRACIONES

PÁG.

LISTA DE ANEXOS

GLOSARIO

DANE: Departamento Administrativo Nacional de Estadística((DANE), 2017)

Flunet: Es una herramienta global basada en la web para la vigilancia

virológica de la influenza.(Organization, 2017)

OLAP: Base de datos para la toma de decisiones o procesamiento analítico.(A. O. Rodríguez, 2013)

OLTP: Bases de datos operacionales o procesamiento de transacciones.(A. O. Rodríguez, 2013)

PM10: Cantidad de partículas menores a 10 micrómetros (Galvis & Rojas, n.d.)

PM2.5: Cantidad de partículas menores a 2.5 micrómetros (Galvis & Rojas, n.d.)

RIPS: Registros Individuales de Prestación de Servicios de Salud(Individuales et al., 2000)

SISPRO: Sistema Integral de Información de la Prestación Social(M. de Salud, 2017)

SIVIGILA: Sistema Nacional de Vigilancia en Salud Pública(I. N. de Salud, n.d.)

TOGAF: Es un marco de arquitectura. TOGAF proporciona los métodos y herramientas para ayudar a aceptación, producción, uso y mantenimiento de una arquitectura empresarial.(Standard & Group, 2011)

RESUMEN

El presente proyecto se basa en la aplicación de minería de datos mediante el algoritmo de clustering K- means que permita la generación de un modelo descriptivo con el análisis de los datos y con el objetivo de identificar posibles comportamientos en enfermedades respiratorias en la ciudad de Bogotá. El conjunto de clústeres generados por la herramienta RapidMiner es la recopilación de datos de un periodo de cinco años de 2012 a 2016, en donde se contemplan el número de casos asociados a 184 diagnósticos de enfermedades respiratorias y la edad de los pacientes corresponde de 0 a 5 años la fuente de información seleccionada de acuerdo a la completitud y asociación de las variables es la del cubo SISPRO – Sistema Integral de Información de la Protección Social.

Asimismo, se ve la necesidad de realizar análisis de la información mediante clustering, debido al crecimiento exponencial de datos que generan los sistemas y plataformas del sector salud con el paso del tiempo, permitiendo que estos datos organizados en los clústeres finalmente se conviertan en información y sea un insumo para la toma de decisiones enfocado a los diagnósticos de las enfermedades respiratorias.

A continuación, se relaciona la metodología que consiste una serie de pasos que permiten lograr la aplicación de minería de datos. Primero, consiste en la obtención de bases de datos y sus variables, segundo la selección de la herramienta de análisis de datos, tercero la aplicación de las bases de datos en la herramienta de análisis, luego la aplicación algoritmos de clustering que permitan identificar los patrones de comportamiento, seguido de la obtención de resultados mediante los patrones ya identificados y por ultimo realizar la documentación sobre los resultados obtenidos.

Finalmente, cabe aclarar que las variables que se contemplan para el desarrollo, son datos que no son sensibles y que tampoco son suministradas por las diferentes fuentes de información con el fin de mantener la seguridad e integridad de los pacientes.

Palabras Claves: Arquitectura de datos, Clustering, Calidad de datos, Diagnósticos, Fuentes de extracción, K-MEANS, técnica de minería de datos.

INTRODUCCIÓN

Con el constante crecimiento exponencial de información sobre sistemas que son utilizados por áreas de la salud, es importante realizar un análisis acerca de los volúmenes de información que estos sistemas generan y que es cada vez más deficiente el análisis sin herramientas de minería de datos que permitan identificar comportamientos o patrones para que las entidades de salud sean más eficientes en la toma de decisiones.

Según la revista COMPUTERWORD(“COMPUTERWORD,” 2017) con el artículo titulado Tecnologías para el crecimiento de la salud publicada en abril de 2017, indica que la industria de la salud se dirige hacia un modelo de cuidados personalizados apoyado de las nuevas tecnologías, en donde llegan a la conclusión que identifican 18 tecnologías que impactaran esta industria para el año 2025 entre ellas la necesidad de analizar datos en gran volumen que excedan la capacidad humana y el análisis de la salud de la población.

En el transcurso del documento se evidenciarán conceptos asociados que comprenden la minería de datos como el funcionamiento, las extensiones de la minería de datos que permitirán dar una idea al desarrollo de la investigación.

Básicamente el enfoque del proyecto es el análisis de datos de la población de la ciudad de Bogotá para identificar patrones sobre las enfermedades respiratorias, lo anterior es aplicado con herramientas de minería de datos, con el objetivo de generar nuevo conocimiento a partir del análisis de la información y posteriormente realizar planes de acción que permitan mejorar los servicios prestados por los hospitales.

Por lo anterior el análisis de esta información conlleva a la generación de nuevo conocimiento para la industria de la salud por medio del análisis de datos.
Según el artículo(Frsf & Conicet, n.d.-a) Minería de Datos en Base de Datos de Servicios de Salud, indica que se adelantan investigaciones y se realizan trabajos para el mejoramiento de algoritmos de minería de datos y el desarrollo de los mismos para el descubrimiento de patrones de comportamiento en áreas como la medicina.

En referencia a las estrategias de comunicación para la divulgación del trabajo, es necesario indicar que para la obtención de datos e información relevante sobre el problema planteado es obtenida de artículos científicos encontrados en bases de datos especializadas académicamente que indican información actualizada con respecto al tema de nuestro proyecto.

1. GENERALIDADES

1.1 ANTECEDENTES

A continuación, se darán a conocer los diferentes trabajos aplicados en otras partes de acuerdo con el tema tratado en esta investigación:

1. **Reconocimiento de patrones aplicado a la predicción de series temporales:** Representado por la Universidad Pablo de Olavide y la Universidad de Sevilla, buscaron encontrar patrones de acuerdo a algoritmos de predicción y de *clustering* para encontrar el comportamiento en variables como: Temperatura, precio y Ozono.(Alvarez, Troncoso, & Riquelme, n.d.)

2. **Minería de Datos en Base de Datos de Servicios de Salud:** Representado por la Universidad Tecnológica Nacional Facultad Regional Santa Fe. Buscaron en bases de datos relaciones donde lograron demostrar la distribución correspondiente de cada variable dentro del conjunto de datos aplicando e concepto de minería de datos.(Frsf & Conicet, n.d.-b)

3. **Sistema prototipo para la estimación del comportamiento del índice de calidad del aire usando técnicas de aprendizaje computacional:** Representado por la Universidad Nacional de Colombia. Este trabajo consistió en buscar comportamientos de acuerdo a variables como temperatura, presión atmosférica, velocidad del viento, entre otras, aplicando algoritmo de *clustering* para realizar el agrupamiento de estas variables y obtener los comportamientos deseados a encontrar.(Anaya, 2015)

1.2 PLANTEAMIENTO DEL PROBLEMA

1.2.1 Descripción del Problema. Identificando que las enfermedades respiratorias son una de las principales causas de mortalidad en Bogotá, especialmente en menores de cinco años y adultos mayores, y de acuerdo con la secretaria general de la salud- SDS, el 70% de las muertes por enfermedad respiratoria aguda- ERA durante el periodo entre 2010 y 2015, fueron muertes evitables en niños y niñas menores de 5 años(Secretaría Distrital de Salud, 2015). según la SDS estas causas se concentran principalmente en escasa cobertura de estrategias de promoción de prácticas de autocuidado y prevención de la ERA.(Secretaría Distrital de Salud, 2015)

1.2.2 Formulación del Problema. Se tiene información almacenada en bases de datos que cuentan con tendencias no conocidas acerca de enfermedades respiratorias. Estas bases de datos fueron obtenidas de entidades de salud del Distrito, específicamente de Hospitales y del Ministerio de Salud. Los datos son pertenecientes a la ciudad de Bogotá.

Debido al gran volumen de datos por analizar, se atrasa la toma de decisiones en donde se relacionan con los patrones encontrados para la prevención de las enfermedades respiratorias de la ciudad de Bogotá, entre ellas las más comunes Infecciones Respiratorias Agudas- IRA,ERA, es por ello que con este trabajo de investigación se entreguen resultados que ayuden al área de la Salud en Bogotá a crear estrategias de promoción y prevención para disminuir muertes por enfermedades respiratorias a niños en edades entre cero y cinco años.

Para ello se realizan las siguientes preguntas que se tendrían en cuenta para un análisis de las enfermedades respiratorias que se generan en la ciudad de Bogotá:

¿Cómo la técnica de minería de datos puede identificar patrones que permitan mejorar los programas de prevención para las enfermedades respiratorias en Bogotá?

¿Qué comportamiento han tenido las enfermedades respiratorias en los últimos cinco años de la ciudad de Bogotá?

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Descubrir patrones de enfermedades respiratorias en la ciudad de Bogotá utilizando técnicas de minería de datos.

2.2 OBJETIVOS ESPECÍFICOS

- Definir las fuentes de extracción de datos y sus variables clasificándolos a través de criterios de selección.
- Diseñar y construir una arquitectura de datos para asegurar la implementación adecuada de los algoritmos de *clustering*.
- Seleccionar e implementar algoritmos de *clustering* para el análisis de los datos.
- Reconocer los patrones de comportamiento de las enfermedades respiratorias a partir de la información recopilada.

3. JUSTIFICACIÓN

El problema actual acerca de las enfermedades respiratorias, es dar a conocerse como una de las principales causas de mortalidad en Bogotá, especialmente en menores de cinco años y adultos mayores, y de acuerdo con la secretaria general de la salud- SDS, el 70% de las muertes por enfermedad respiratoria aguda- ERA durante el periodo entre 2010 y 2015, fueron muertes evitables en niños y niñas menores de 5 años(Secretaría Distrital de Salud, 2015). según la SDS estas causas se concentran principalmente en escasa cobertura de estrategias de promoción de prácticas de autocuidado y prevención de la ERA.(Secretaría Distrital de Salud, 2015)

En esta investigación se busca encontrar patrones de comportamiento que tienen las enfermedades respiratorias en edades de 0 a 5 años. Para esto se cuenta con técnicas de minería de datos como *clustering*, donde a través de algoritmos de agrupación se realizará la búsqueda de dichos comportamientos.

Con este trabajo de investigación se pretende generar resultados que den a conocer los comportamientos asociados a las enfermedades respiratorias y así la Secretaria de Salud en Bogotá pueda crear estrategias de promoción y prevención para disminuir muertes por enfermedades respiratorias en la población infantil de cero a cinco años, específicamente en la ciudad de Bogotá.

También se propone este método de investigación para profundizar los análisis que se realizan, tomando como ejemplo, los boletines epidemiológicos donde da a conocer que el 57,2 % de las muertes por IRA en menores de cinco años se notificó en el sexo masculino, el 67,4 % en residentes de la cabecera municipal, el 63,6 % pertenecían al régimen subsidiado. Por pertenencia étnica, el 28,9 % corresponde a población indígena, donde en este análisis no se llega a conocer muy claro cómo actúan las enfermedades. Teniendo como novedad, resultados de comportamientos en enfermedades respiratorias de manera entendible y eficaz.

4. DELIMITACIÓN

4.1 ALCANCE

Este trabajo de investigación tendrá como alcance mostrar resultados de análisis de datos para específicamente la ciudad de Bogotá y enfocadas únicamente enfermedades respiratorias.

Teniendo como limitaciones la obtención de los datos a través de base de datos del ministerio de salud, secretaria de salud y un hospital de la ciudad de Bogotá para realizar el respectivo análisis de patrones de comportamiento de las enfermedades respiratorias en el caso que se menciona anteriormente.

4.2 ESPACIO

El desarrollo del proyecto será elaborado en las instalaciones de la Universidad y en la residencia de cada uno de los estudiantes.

4.3 TIEMPO

El tiempo estimado para el desarrollo de trabajo de grado es de seis meses aproximadamente, se llevará a cabo en el segundo semestre del año 2017.

4.4 CONTENIDO

El resultado del análisis de los datos, permitirá identificar comportamientos de las enfermedades respiratorias, para niños de 0 a 5 años de edad, evaluando los datos en el periodo comprendido de 2012 a 2016.

El análisis de los datos se realiza mediante clústeres, que son los generados a partir de la herramienta RapidMiner.

5. MARCO REFERENCIAL

5.1 MARCO TEÓRICO

5.1.1 Historia de Data Mining. La data mining surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos. De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.(L. C. Molina, 2002)

5.1.2 Aplicación de Data Mining. La aplicación de técnicas de minería de datos es de suma importancia para las organizaciones en donde su minería de datos crece exponencialmente. Por tanto la minería de datos es un tema de investigación que puede proporcionar una gran utilidad en la búsqueda de patrones.(Díaz Arévalo & Pérez García, 2002)

La minería de datos es el proceso de Seleccionar, Explorar, Modificar, Modelizar y valorar grandes cantidades de datos con el objetivo de descubrir conocimiento.(Daniel, 2006)

En el área de las ciencias de la salud, se utilizan para la detección precoz y prevención de enfermedades, para el análisis de marcadores genéticos, para prever la probabilidad de una respuesta satisfactoria a un tratamiento médico, como por ejemplo las reglas de asociación utilizadas en [Marchán et al., 2011] para determinar factores de riesgo epidemiológico de transmisión de enfermedades, para asistir al médico en el diagnóstico, por ejemplo detectar patrones anormales en los análisis bioquímicos o en las pruebas de imagen y diagnóstico digital.(Gutiérrez, 2016) (Marchán E, Salcedo J, Aza T, Figuera L, Martínez de Pisón F, 2011)

Los mineros de datos son programas que de forma automática encuentran similitudes y desviaciones en una base de datos.

El concepto de minería de datos apareció hace más de 10 años. El interés en este campo y su explotación en diferentes especialidades como por ejemplo (negocios, finanzas, ingeniería, banca, salud, sistemas de energía, entre otros).

Actualmente el análisis de datos ha permitido realizar investigaciones sobre identificación de comportamientos a través de patrones donde se muestra el cómo actúa un conjunto de objetos o seres que pueden ser diferentes pero pueden llegar a tener cosas en común de manera oculta.(Berkhin, 2002)

Para ello se implementaron algoritmos que ayudaron en la búsqueda de los datos en común e identificar así los comportamientos que se llegan a tener

frente a una situación, o escenario específico(Berkhin, 2002). Para llegar a estas conclusiones es necesario tener diversas técnicas y algoritmos usados comúnmente para poder implementarse y así obtener buenos resultados(Berkhin, 2002) (Wu et al., 2008).

Como se menciona en el documento "*Top 10 algorithms in datamining XindongWu*"(Wu et al., 2008) y "*Survey of Clustering Data Mining Techniques*"(Berkhin, 2002) la implementación de las buenas practicas utilizando técnicas y algoritmos base que se utilizan frecuentemente pueden ayudar en la búsqueda de los resultados esperados en una investigación. Sin embargo también depende de lo que se quiere buscar, en este caso y de acuerdo con el documento "Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software"(D. Rodríguez, Cuadrado, & Sicilia, 2007a) da a conocer que depende de lo que se quiera buscar existe un algoritmo y este actúa diferente o se comporta diferente frente a propiedades de los datos como el tamaño, el peso, la longitud, entro otros haciendo que actúen de manera óptima o no optima en los resultados de la investigación(Berkhin, 2002) (Wu et al., 2008) (D. Rodríguez, Cuadrado, & Sicilia, 2007b).

Para el caso del documento donde se comparan los algoritmos este concluye que no mejor el algoritmo conocido como COBWEB donde su definición de acuerdo al documento dice "Se trata de un algoritmo de clustering jerárquico. COBWEB, se caracteriza porque utiliza aprendizaje incremental, esto es, realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada"(D. Rodríguez et al., 2007b) aplicado para el caso de costos en un desarrollo de software, pero a su vez concluye que el algoritmo de K-media es más adecuado dando como definición lo siguiente "Se trata de un algoritmo clasificado como Método de Particionado y Recolocación. El método de las k-medias, es hasta ahora el más utilizado en aplicaciones científicas e industriales"(D. Rodríguez et al., 2007b).

Uno de estos algoritmos mencionados anteriormente K-media llega a estar en el top de los algoritmos que se mencionan en el documento "*Top 10 algorithms in datamining XindongWu*"(Wu et al., 2008) el cual da a conocer una mejor perspectiva e importancia para el uso de este algoritmo sobre los demás(Wu et al., 2008).

En ambos documentos se menciona también el algoritmo EM el cual se define como "EM pertenece a una familia de modelos que se conocen como *Finite Mixture Models*, los cuales se pueden utilizar para segmentar conjuntos de datos"(D. Rodríguez et al., 2007b) dando también una importancia del uso de este algoritmo pero no mucho en el uso para la investigación de costos sobre proyectos de desarrollo de software(D. Rodríguez et al., 2007b).

5.1.3 Investigaciones en Data Mining. Con base en la investigación asociada a la identificación de patrones de comportamientos de enfermedades respiratorias, en otros países como Hong Kong se ha realizado una

investigación similar donde se encontraron varios causas comunes sobre las enfermedades respiratorias agudas(T. Wang et al., 2003).

En esa investigación llamada "*A Cluster of Cases of Severe Acute Respiratory Syndrome in Hong Kong*"(T. Wang et al., 2003) se identificaron síntomas comunes de acuerdo a varias muestras que en este caso son los mismos pacientes los cuales tenían una enfermedad en común donde se trata de averiguar que comportamientos comunes se encuentran entre la muestra para poder identificar una causa especifica dentro de las enfermedades que se trataron en la investigación(T. Wang et al., 2003).

Para ello realizaron sobre los pacientes varias evaluaciones tales como *"microbiologic assessment", "radiologic assessment" y "microbiologic evaluation"* donde tomaban los datos de los resultados de estas y los agrupaban para ver específicamente resultados comunes dentro de estas(T. Wang et al., 2003).

También observaron el contacto histórico entre los pacientes para saber o tener más claro de cómo se propagaban estas enfermedades respiratorias agudas, debido a que se deseaba prevenir una epidemia ya que las enfermedades respiratorias que estaban afectando a Hong Kong se transmitían por contacto el cual se estudió y se dieron resultados sobre la frecuencia de transmisión de estas enfermedades en diferentes sitios de la ciudad(T. Wang et al., 2003).

Después de haber hallado o identificado los comportamientos o la causa común de las enfermedades respiratorias agudas en Hong Kong según el documento realizan diferentes tratamientos para la prevención y manejo de estas enfermedades para poder prevenir como anteriormente se menciona una epidemia dentro de esta ciudad y lo cual es el objetivo de estas investigaciones en cuestión de salud a nivel de ciudades, países o mundial(T. Wang et al., 2003).

5.1.4 Arquitectura de datos. Se tuvieron en cuenta las consideraciones clave para la arquitectura de datos según metodología TOGAF(Standard & Group, 2011) que consta de lo siguiente:

Gestión de datos: Cuando una empresa ha optado por llevar a cabo una transformación arquitectónica a gran escala, importante para entender y abordar los problemas de gestión de datos. Una estrategia y enfoque integral de la gestión de datos permite el uso eficaz de los datos para capitalizar sus ventajas competitivas.(Standard & Group, 2011)

Migración de datos: Cuando se reemplaza una aplicación existente, habrá una necesidad crítica de migrar datos (maestro, transaccional y de referencia) a la nueva aplicación. (Standard & Group, 2011)

Gobernabilidad de datos: Las consideraciones de gobernabilidad de datos aseguran que la empresa tenga las dimensiones para permitir la transformación. (Standard & Group, 2011)

Estructura: Esta dimensión se refiere a si la empresa tiene la estructura organizativa y los organismos de normalización para gestionar los aspectos transformación.(Standard & Group, 2011)

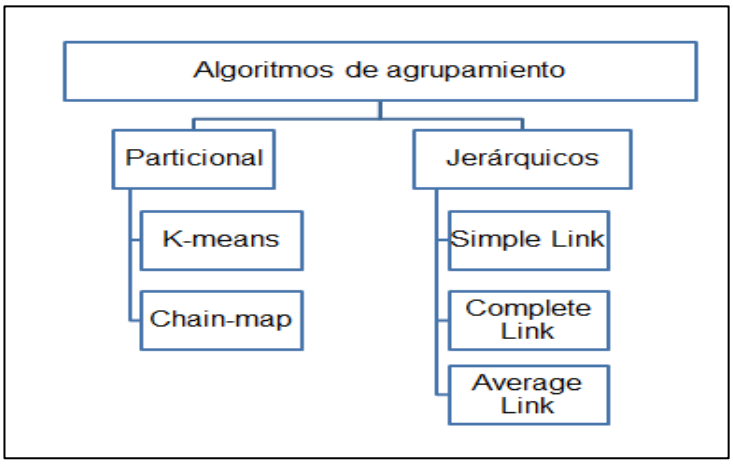
Sistema de Gestión: Aquí las empresas deben tener el sistema de gestión necesario y programas relacionados con los datos para gestionar los aspectos de gobernanza de las entidades de su ciclo de vida.(Standard & Group, 2011)

Personas: En esta dimensión se abordan las habilidades y roles relacionados con los datos de la empresa requiere para la transformación. Si el empresario carece de tales recursos y debe considerar la adquisición de esas habilidades críticas o la formación interna existente recursos para satisfacer los requisitos a través de un programa de aprendizaje bien definido.(Standard & Group, 2011)

5.1.5 Clasificación de Algoritmos de Clustering. Clustering es el proceso de encontrar grupos significativos en los datos. Independientemente de los tipos de aplicaciones de clúster, la tarea de minería de datos de la agrupación busca encontrar los agrupamientos en datos, de tal manera que los datos un clúster son más "similares" entre sí que a los puntos de datos en otros clústeres (Witten y Frank, 2005)(Elkan, 2010)

A continuación, se encuentra la clasificación general con los algoritmos de clustering a trabajar:

Ilustración 1 Clasificación de Algoritmos de Clustering



Fuente: Los autores

En la siguiente tabla se indica un cuadro comparativo del algoritmo particional y jerárquico:

Tabla 1 Comparación algoritmos de clustering

Comparación de algoritmos de clustering	
Separa un clúster ya existente para dar origen y formar un nuevo clúster.	Realiza una distribución de los elementos entre el número establecido de clústeres.
Encuentra clústeres sucesivos utilizando los ya preestablecidos.	No hay estructura jerárquica
La salida es un diagrama de árbol que muestra diferentes clústeres.	Determinan todos los clústeres a la vez, seleccionando el número de grupos establecidos
	Son eficientes con un conjunto de datos amplios.

Fuente: Los autores

5.1.5.1 Algoritmo de Clustering Particional. Un algoritmo de agrupamiento particional es aquel que obtiene como resultado una única partición de los datos iniciales, en lugar de una estructura de agrupamiento con varios niveles de particiones. Un algoritmo particional asigna a un conjunto de objetos K grupos sin estructura jerárquica, siendo K un número real menor que el número total de objetos. Este tipo de algoritmos son muy eficientes en aquellas aplicaciones con conjuntos de datos de amplia dimensión, pero presentan el problema de que es necesario escoger el número de grupos deseados.(Campos, 2009)

5.1.5.1.1 Algoritmo K-Means. Esta técnica está basada en el clustering particional que intenta encontrar un número de clústeres (K) especificados por el usuario, los cuales son representados por sus centroides. El algoritmo básico se describe a continuación: Primero se eligen K centroides iniciales, donde K es un parámetro especificado por el usuario y corresponde al número de clústeres deseados. Cada punto es asignado a su centroide más cercano y cada colección de puntos asignado a un centroide representa un clúster. El centroide de cada clúster se actualiza basado en la asignación de puntos al clúster. Se repiten los pasos de asignación y actualización hasta que los puntos dentro del clúster no cambien, o equivalentemente, hasta que los centroides dejen de cambiar. (Eduardo & Medina, 2014)

El algoritmo de clustering k-means se basa en los trabajos de Stuart Lloyd y E.W. Forgy (Lloyd, 1982) ya veces se denomina el algoritmo de Lloyd-Forgy o Algoritmo de Lloyd. Visualmente, el algoritmo k-means divide el espacio de datos en k particiones o límites, donde el centroide en cada partición es el prototipo de los clústeres.(Elkan, 2010)

El algoritmo k-means particional, donde cada centro de clúster está representado por el valor medio de los objetos en el grupo.

1. **Entrada:**
   - k: el número de clúster,
   - D: un conjunto de datos que contiene n objetos.

2. **Salida: Un conjunto de k clústeres.**
3. **Método:**
   - Elegir arbitrariamente k objetos de D como los centros de agrupación inicial;
   - Repetir
   - (Re) asignar cada objeto al clúster al que el objeto es el más similar,
   - Basado en el valor medio de los objetos del clúster;
   - Actualizar los medios de agrupamiento, es decir, calcular el valor medio de los objetos para cada grupo.
   - Hasta que no cambie.

5.1.5.1.2 Algoritmo de las Distancias Encadenadas O Chain-Map. El algoritmo chain-map, es un algoritmo de agrupamiento en donde dado el conjunto de objetos, en forma de vectores, cada elemento es el valor de una característica algoritmo comienza seleccionando un objeto cualquiera de los disponibles.(Benítez & Díez, 2005)

El algoritmo chain-map es recomendable aplicarlo como un paso previo para la iniciación de otros algoritmos de clustering, como por ejemplo k-means en donde se utiliza para la estimación inicial del número de grupos a buscar.(Benítez & Díez, 2005)

5.1.5.2 Algoritmos Jerárquicos. Los algoritmos jerárquicos encuentran clústeres sucesivos utilizando los previamente establecidos, mientras que los algoritmos de partición determinan todos los clústeres a la vez. Los algoritmos jerárquicos pueden ser aglomerados o divisivos; los algoritmos aglomerativos comienzan con cada elemento como un grupo separado y combinan los clústeres obtenidos en grupos sucesivamente mayores.(Magdaleno, Miranda, Fuentes, & García, 2015)

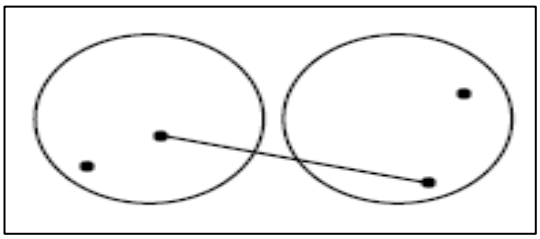
La agrupación jerárquica es un proceso en el que una la jerarquía de clúster se crea en función de la distancia entre los puntos de datos. La salida de una agrupación jerárquica es un diagrama de árbol que muestra diferentes clústeres en cualquier punto de precisión que es especificado por el usuario.(Elkan, 2010)

Los algoritmos divisivos, comienzan con todos los puntos en un solo clúster y los va dividiendo en dos grupos donde no tengan ninguna semejanza.(Cumming et al., 2011)

5.1.5.2.1 Simple Link. Se basa en la división o agrupación de grupos de manera que cada grupo tenga menor diferencia o mayor semejanza, es decir, se tiene dos grupos los cuales para el caso de simple link en tipo de aglomeración y divisivo este agrupara de acuerdo a la mayor semejanza entre los pares de grupos.(Cumming et al., 2011)(Manning, Raghavan, & Schütze, 2009)

Ejemplo en caso de aglomeración:

Ilustración 2 Relación grupos simple link



Fuente Clustering 2: Hierarchical Clustering (Cumming et al., 2011)(Manning et al., 2009)

Se tienen varios grupos donde tienen diferentes datos, el algoritmo buscara la mayor similitud entre los datos para encontrar el par de grupos que tengan relación y unirlo, realizando el mismo proceso para los demás grupos, encontrando la mayor similitud posible y la pareja de cada grupo.

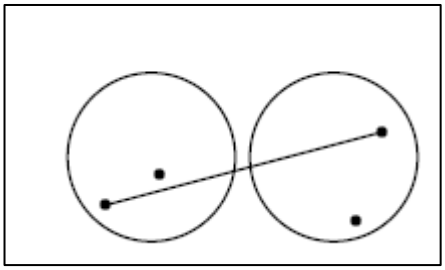
En el caso del tipo divisivo realiza el proceso de dividir en pares de grupos acorde a la mayor semejanza de los datos en el grupo único entrante, es decir, entra un grupo único, busca dentro del grupo datos semejantes y los divide en dos grupos.

5.1.5.2.2 Complete Link. Se basa en la división o agrupación de grupos de manera que cada grupo tenga menor diferencia o mayor semejanza, es decir,

se tiene dos grupos los cuales para el caso de simple link en tipo de aglomeración y divisivo este agrupara de acuerdo a la menor semejanza entre los pares de grupos.(Cumming et al., 2011)(Manning et al., 2009)

Ejemplo en caso de aglomeración:

Ilustración 3 Relación grupos complete link



Fuente Clustering 2: Hierarchical Clustering (Cumming et al., 2011)(Manning et al., 2009)

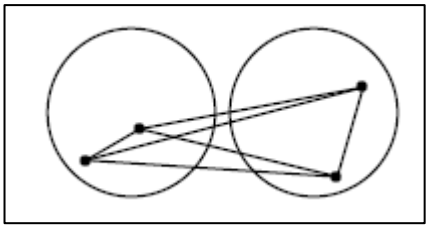
Se tienen varios grupos donde tienen diferentes datos, el algoritmo buscara la mayor diferencia entre los datos para encontrar el par de grupos que tenga más relación y unirlo, realizando el mismo proceso para los demás grupos encontrando la mayor similitud posible y la pareja de cada grupo.

En el caso del tipo divisivo realiza el proceso de dividir en pares de grupos acorde a la mayor diferencia de los datos en el grupo único entrante, es decir, entra un grupo único, busca dentro del grupo datos semejantes y los divide en dos grupos.

5.1.5.2.3 Average Link. Se basa en la división o agrupación de grupos de manera que cada grupo tenga menor diferencia o mayor semejanza, es decir, se tiene dos grupos los cuales para el caso de simple link en tipo de aglomeración y divisivo este agrupara de acuerdo al promedio de semejanza entre los pares de grupos.(Cumming et al., 2011)(Manning et al., 2009)

Por ejemplo, como se muestra en el siguiente gráfico:

Ilustración 4 Relación grupos Average link



Fuente Clustering 2: Hierarchical Clustering (Cumming et al., 2011)(Manning et al., 2009)

El algoritmo buscara el promedio de semejanzas entre todos los datos donde, para el caso de aglomeración, agrupara en un grupo el par de grupos que tenga la mayor relación en todos los puntos. En el caso del tipo divisivo dividirá en dos grupos de acuerdo a la relación de todos los puntos, contando que cada punto es un dato dentro del conjunto.

5.2 MARCO CONCEPTUAL

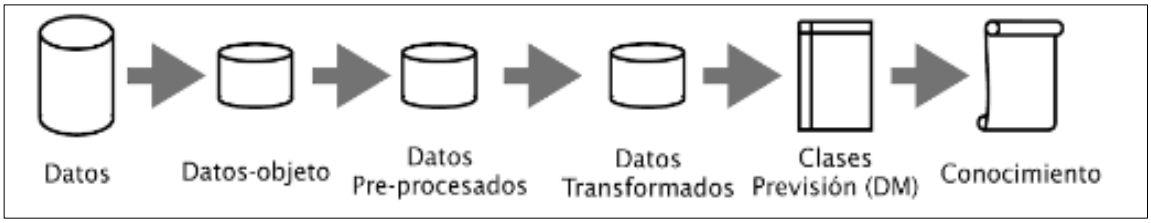
5.2.1 Minería de Datos. La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar cantidades de datos.

La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining (López, 2017)

5.2.2 Proceso KD. La sigla KDD – Knowledge Discovery in Databases fue creada en 1995 para designar el conjunto de procesos, técnicas y abordajes que propician el contexto en el cual la minería de datos tendrá lugar.(Viera, n.d.)

La minería de datos está incluida en un proceso mayor denominado Descubrimiento de Conocimientos en Base de Datos, *Knowledge Discovery in Database* (KDD), que es el proceso de extracción de conocimiento de bases de datos.(Viera, n.d.)

Ilustración 5 Esquema para la generación de conocimiento en bases de datos KDD



Fuente: Introducción a la Minería de Datos.

5.2.2 Proceso de minería de datos. El proceso de minería de datos se divide en las siguientes etapas:

1. **Selección de datos:** Es la etapa más dispendiosa, ya que consiste en la recolección y preparación de los datos. En este proceso se comprende la problemática asociada a la base de datos y se establecen objetivos. A la vez, se identifican las variables que serán consideradas para la construcción del modelo de minería de datos (MD).(Uiaf, n.d.)

2. **Pre procesamiento de datos:**

- **Integración de datos:** Se analiza si la base de datos requiere incluir o integrar información o variables que reposan en otras bases de datos, y que será relevante para el modelo de minería de datos.(Uiaf, n.d.)
- **Reconocimiento y limpieza:** Se depura el conjunto de datos respecto a valores atípicos, faltantes y erróneos (eliminación de ruido e inconsistencias).(Uiaf, n.d.)

3. **Selección de características:**
    - **Exploración y limpieza de datos:** Aplicando técnicas de análisis exploratorio de datos (estadístico, gráfico, entre otros), se busca identificar la distribución de los datos, simetría, pruebas de normalidad y correlaciones existentes entre los datos.(Uiaf, n.d.)
    - **Transformación:** Se estandariza o normaliza la información (colocarla en los mismos términos de formato y forma). La selección de la técnica a aplicar dependerá del algoritmo que se utilizará para la generación de conocimiento.(Uiaf, n.d.)
    - **Reducción de datos:** Se disminuye el tamaño de los datos mediante la eliminación de características redundantes.

4. **Minería de Datos:** La minería de datos según Esteban (2008) et. al. (1991 / 1995), se puede definir como un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos, que a su vez, facilita la toma de decisiones y emplea técnicas de aprendizaje supervisado y no-supervisado.(Uiaf, n.d.)

    A continuación, se presentan las técnicas empleadas, las cuales pueden ser de tipo supervisado o no supervisado.
    - **Identificación:** Evidenciar la existencia de objetos, eventos y actividades en el conjunto de datos (análisis factorial, discriminante, regresivo, de correlaciones).(Uiaf, n.d.)
    - **Clasificación:** Particionar los datos de acuerdo a las clases o etiquetas asignadas al conjunto de datos.(Uiaf, n.d.)
    - **Agrupación:** Permitir la maximización de similitudes y minimización de diferencias entre objetos, mediante la aplicación de algún criterio de agrupación.(Uiaf, n.d.)
    - **Asociación:** Tener presente que las reglas de asociación buscan descubrir conexiones existentes entre objetos identificados.(Uiaf, n.d.)
    - **Predicción:** Descubrir el comportamiento de ciertos atributos en el futuro.(Uiaf, n.d.)
5. **Interpretación y Resultados:** Se analizan los resultados de los patrones obtenidos en la fase de MD, mediante técnicas de visualización y de representación, con el fin de generar conocimiento que aporte mayor valor a los datos.(Uiaf, n.d.)

5.2.3 Técnicas de Minería de datos. Inicialmente las técnicas de minería de datos pueden clasificarse en técnicas de modelado originado por la teoría (en las que las variables pueden clasificarse en dependientes e independientes), técnicas de modelado originado por los datos (en las que todas las variables tienen inicialmente el mismo) y técnicas auxiliares.
Las técnicas de modelado originado por la teoría especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido.(Pérez, 2014)

5.2.3.1 Técnicas Predictivas. Especifican el modelo para los datos en base a un conocimiento técnico previo. El modelo supuesto para los datos debe contratarse después del proceso de la minería de datos antes de aceptarlo como válido. Por ejemplo las predicciones se usan para prever el comportamiento futuro de alguna entidad mientras que una descripción puede ayudar a su comprensión.(J. Molina & García, 2008)
5.2.3.2 Técnicas Descriptivas. En esta tecnica no se asigna ningun papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existenia de un modelo previo para los datos. Los modelos se crean automaticamente partiendo del reconocimiento de patrones.(López, 2017)

5.2.3.3 Técnicas auxiliares. Son herramientas más superficiales y limitadas. Son nuevos métodos basados en técnicas estadísticas descriptivas e informes.(Pérez, 2014)

5.2.4 Herramientas de minería de datos. Permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y para predicir comportamientos futuros. (López, 2017)

Por ejemplo,entre las herramientas de minería de datos más habituales tenemos actualmente el software de IBM y el software de SAS. IBM dispone de las herramientas IBM SPSS Statistics e IBM SPSS Modeler. La primera de ellas contiene varios procedimientos de minería de datos y la segunda es una herramienta específica de minería de datos sucesora de SPSS Clementine. Por su parte SAS dispone del software estadístico general, de SAS Enterprise Guide para el trabajo con procedimientos estadísticos y de minería por menús y del software SAS Enterprise Miner, específico de minería de datos.(Pérez, 2014)

5.2.5 Almacén de datos. Es el sistema de informacion central en todo el proceso de mineria de datos. Un almacen de datos es una colección de datos orientada a un dominio, integrada, no volatil y variante en el tiempo, para ayudar a la toma de desiciones. Un almacen de datos es un conjunto de datos historicos, internos o externos y descriptivos de un contexto o area de estudio,

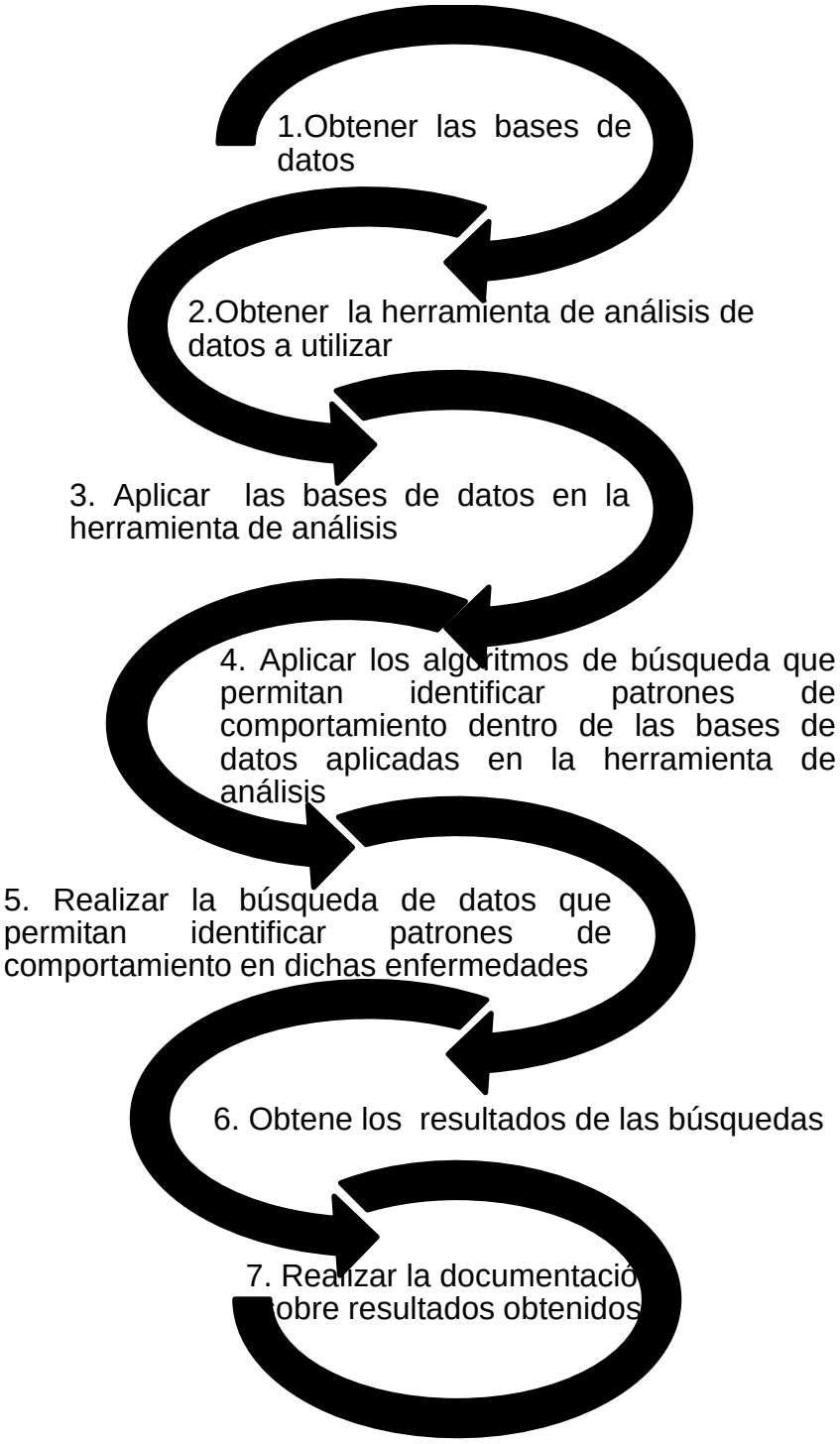
que estan integrados y organizados de tal forma que permiten aplicar efiientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la tarea de desiciones estrategicas.

6. METODOLOGÍA

La metodología que se va a implementar va a permitir tener un orden y un procedimiento adecuado para realizar el respectivo análisis y poder obtener los resultados adecuados que se quieren para este trabajo de investigación.

La metodología consta de los siguientes procedimientos:

Ilustración 6 Metodología con sus procedimientos



Fuente: Los autores

La metodología consta de 7 pasos los cuales se muestran en el diagrama de metodología:

1. **Obtener las bases de datos:** Se buscan las fuentes de extracción, posteriormente se califican las fuentes de acuerdo a algunos criterios de selección que están asociados a calidad de datos, para la finalmente realizar la definición de las variables.
2. **Obtener la herramienta de análisis de datos a utilizar:** Se busca la herramienta de análisis de datos con la cual se va a trabajar, para el caso se escogió la herramienta RapidMiner.
3. **Aplicar las bases de datos en la herramienta de análisis:** Se aplican los datos obtenidos de las fuentes de extracción dentro de la herramienta de análisis de datos.
4. **Aplicar los algoritmos de búsqueda:** Con el objetivo que permita identificar patrones de comportamiento dentro de las bases de datos aplicadas en la herramienta de análisis, por tal motivo, se selecciona un algoritmo de Clustering para ejecutarla en RapidMiner.
5. **Realizar la búsqueda de datos que permitan identificar patrones de comportamiento en dichas enfermedades:** Se ejecuta el algoritmo sobre los datos a analizar dentro de la herramienta de análisis de datos.
6. **Obtener los resultados de las búsquedas:** Se extraen los resultados obtenidos de la búsqueda en la herramienta de análisis de datos para los patrones identificados.
7. **Realizar la documentación sobre resultados obtenidos:** Se realiza la documentación de acuerdo con los resultados obtenidos de la herramienta de análisis de datos.

7. FUENTES DE EXTRACCIÓN Y SUS VARIABLES

Se cuentan con diferentes fuentes de extracción las cuales comprenden en total 18 variables a utilizar, que van a permitir ser materia prima para la búsqueda de datos. Por tal motivo, se procede analizar las fuentes de extracción y sus variables.

7.1 FUENTES DE EXTRACCIÓN

El sector salud es uno de los sectores que genera grandes volúmenes de información y captura de datos de sus pacientes, es por ello que para la pertinente recolección de datos se consideran inicialmente las siguientes fuentes de extracción:

1. Organización Mundial de la Salud – OMS.
2. Artículos de investigación.
3. Datos Abiertos en la categoría Salud y Protección Social.
4. Instituto Nacional de la Salud
5. Encuesta Nacional de Demografía y Salud (ENDS)
6. Organización Panamericana de la Salud
7. SISPRO
8. DANE
9. SIVIGILA
10. FluNet
11. Secretaria Distrital de Planeación
12. Ministerio de Salud
13. Famisanar EPS
14. Hospital Kennedy
15. UNICEF
16. Cruz Roja
17. Sura EPS
18. EPS Sanitas
19. Medimas EPS
20. Secretaria distrital de salud
21. Nueva EPS
22. Páginas Web sobre temas de salud
23. Documentos web sobre temas de salud
24. Salud Vida EPS
25. Dinámica IPS
26. Hospital San José
27. Hospital San Cristóbal

7.1.1 Criterios de Selección. Por consiguiente, de las anteriores fuentes de información y con el objetivo de definir cuales se van a trabajar, se tienen en cuenta los siguientes criterios de selección porque se encuentran enfocados a la calidad de datos:

- **Accesibilidad:** Ser capaz de llegar a los datos, en donde el usuario tiene los medios y el privilegio para obtener los datos.(R. Y. Wang, Reddy, & Kon, 1995) Por ejemplo: Se solicitó información a la Secretaria Distrital de Salud sobre el reporte e morbilidad en Bogotá por enfermedades respiratorias y fue enviado sin ninguna restricción por un funcionario de la entidad.

- **Utilidad:** Los datos pueden utilizarse como una entrada para la decisión del usuario.(R. Y. Wang et al., 1995). Por ejemplo: Los datos asociados al número de casos para la categoría de enfermedades respiratorias, permiten identificar la frecuencia con que se registran y el tipo de enfermedad. Ver sección 7.2 Definición de Variables

- **Credibilidad:** Medida en que él usuario puede utilizar los datos como una entrada de decisión que tenga validez.(R. Y. Wang et al., 1995) . Por ejemplo: En la ejecución de los algoritmos de *clustering*, se identifiquen patrones con los datos que se están asignando como entrada.

- **Cantidad:** Capacidad de tener un número considerable de datos dentro de una fuente de información. Por ejemplo: El rango de los datos consolidados a evaluar es del periodo 2012 a 2016.

- **Precisión:** Implica que los datos tienen detalles suficientes y apropiados.(Oms & Ops, 2014) Por ejemplo: La obtención de casos de enfermedades tienen la fecha detallada indicando el día y el año.

De igual modo, a cada fuente de extracción se le dio una calificación de acuerdo al contenido y calidad de los datos, con la finalidad de obtener la fuente de datos definitiva. La calificación que se le dio a cada criterio de selección, está comprendida con los puntajes 0 y 1, siendo 1 si cumple con el criterio de selección y 0 si no cumple.

La siguiente tabla que indica las fuentes de extracción frente a los criterios de selección:

Tabla 2 Calificación Fuentes de extracción

Ítem	Fuente/Criterio	Accesibilidad	Utilidad	Credibilidad	Cantidad	Precisión	Total
1	Encuesta Nacional de Demografía y Salud	1	1	1	1	1	5

	(ENDS)						
2	SISPRO	1	1	1	1	1	5
3	DANE	1	1	1	1	1	5
4	SIVIGILA	1	1	1	1	1	5
5	FluNet	1	1	1	1	1	5
6	Secretaria Distrital de Planeación	1	1	1	1	1	5
7	Ministerio de Salud	1	1	1	1	1	5
8	Datos Abiertos en la categoría Salud y Protección Social	1	1	1	0	1	4
9	Organización Mundial de la Salud - OMS	1	1	1	0	0	3
10	Artículos de investigación	1	1	1	0	0	3
11	Organización Panamericana de la Salud	1	1	1	0	0	3
12	UNICEF	1	1	1	0	0	3
13	Secretaria distrital de salud	1	1	1	0	0	3
14	Instituto Nacional de la Salud	0	1	1	0	0	2
15	Hospital Kennedy	0	1	1	0	0	2
16	Cruz Roja	0	1	1	0	0	2
17	Hospital San José	0	1	1	0	0	2
18	Hospital San Cristóbal	0	1	1	0	0	2
19	Famisanar EPS	0	0	1	0	0	1
20	Sura EPS	0	0	1	0	0	1
21	EPS Sanitas	0	0	1	0	0	1
22	Medimas EPS	0	0	1	0	0	1
23	Nueva EPS	0	0	1	0	0	1
24	Salud Vida	0	0	1	0	0	1

	EPS						
25	Dinámica IPS	0	0	1	0	0	1

Fuente: Los autores

7.1.2 Selección fuentes de Extracción. De acuerdo a la calificación anterior se eligieron las siete primeras fuentes de extracción de datos, que recibieron la calificación total de 5, para lo cual, las fuentes cumplen con todos los criterios de selección.

Por tanto, la información corresponde a RIP's públicos del sistema SISPRO referente a prestaciones, también se tiene información de SIVIGILA y de FluNet, proporcionadas para identificar patrones con el fin de prevenir enfermedades respiratorias para que las instituciones puedan realizar un mejor plan de contingencia contra ERA e IRA.

Estos RIP's vienen agrupados en las variables que se van a analizar en archivos de hojas de cálculo ya que la información original proveniente de SISPRO se otorgan como archivos de texto plano, en donde es necesario tener una agrupación de datos para posteriormente realizar el proceso de minería de datos frente a los datos correspondientes.

Por otra parte, se extrajo información del DANE relacionada con datos estadísticos como la población e índice de mortalidad. También se tuvo en cuenta las fuentes secretaria distrital de planeación y artículos que contienen información de la ciudad de Bogotá.

En conclusión: se tomaron las fuentes de extracción de datos que representaba mayor puntaje de acuerdo a la calidad de los datos.

7.2 DEFINICIÓN DE VARIABLES

Las variables que se contemplaron para el desarrollo, son datos que no son sensibles y que tampoco son suministradas por las diferentes fuentes de información, ya que comprometen la identidad e integridad de los pacientes, por mencionar un ejemplo el número de identificación y nombres de los pacientes.

Los datos recolectados fueron obtenidos mediante las fuentes de extracción mencionadas anteriormente, las cuales permiten identificar comportamientos posteriormente realizado el análisis de los datos.

También se cuentan con variables que son útiles para validar una medición de la calidad del aire e identificar cómo y en qué momento afecta en la originalidad

de las enfermedades proporcionando una mejor entrega en los resultados, a continuación, se presentan las variables a utilizar:

1. Altitud: Contiene la altura de Bogotá, se utilizará porque se pretende identificar si la altitud es un factor determinante para la generación de algunas enfermedades respiratorias. El tipo de dato es flotante y el rango se encuentra entre 2600 y 2640 metros sobre el nivel del mar.

2. Ciudad: Contiene la ubicación geográfica donde se genera el número de casos, esta variable es de tipo *String* y el rango que aplica es Bogotá.

3. CO: Contiene la cantidad de monóxido de carbono encontrado en una fecha. El tipo de dato es flotante y sus valores están registrados para los años 2012 a 2016. Se utiliza, porque es una partícula base en la medición de la calidad del aire la cual afecta dependiendo a su cantidad a la generación de enfermedades respiratorias. (Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

4. EPS: Contiene el segmento de la población que está afiliado a las diferentes EPS. Esta variable es de tipo cadena y el rango es la lista de EPS descritas por el Ministerio de Salud.

5. Estrato: Contiene información que servirá para comparar el número de personas que hay en diferentes estratos con respecto a las enfermedades respiratorias. Esta variable se utilizará porque es útil el saber en qué estratos ataca más una enfermedad. El tipo de dato para esta variable es entero y comprende de 1 a 6, siendo 1 el bajo y 6 el alto.

6. Fecha: Comprende el tiempo en el cual se obtuvieron los registros asociados a los casos de enfermedades respiratorias y se utilizara porque permite determinar predicciones frente a un volumen de datos significativo. Esta variable es de tipo fecha y el rango corresponde a los años de 2012 a 2016.

7. Fecha medición calidad del aire: Esta fecha contendrá el día específico en el cual se realizaron las respectivas mediciones sobre la calidad del aire, se utilizará por que puede determinar en qué fecha con exactitud se pueden producir mayor número de casos sobre enfermedades. El tipo de dato corresponde a fecha y el rango corresponde a los años 2012 a 2016.

8. Habitantes: Contiene la población de la ciudad de Bogotá, se tiene en cuenta porque es necesario ver cuántas personas hay actualmente en la ciudad y ver la relación e impacto que han tenido con las enfermedades respiratorias. Es una variable de tipo entero y el rango se encuentra entre 8.000.000 y 8.100.000 millones de habitantes.

9. Mortalidad: Contiene el índice de mortalidad de la población y se utilizará porqué permite realizar el análisis sobre causa de mortalidad y la relación de enfermedades respiratorias. El tipo de dato es flotante y el rango de fecha para esta variable aplica para los años 2012 a 2016.

10. NO2: Cantidad de dióxido de nitrógeno medido en una fecha. Se define, porque es una partícula base en la medición de la calidad del aire la cual afecta dependiendo a su cantidad a la generación de enfermedades respiratorias. El tipo de dato es flotante y sus valores están registrados para los años 2012 a 2016 (Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

11. Número de casos: Contiene el número de casos presentados para enfermedades de tipo respiratorio, se utilizará porqué determina con qué frecuencia es registrado un caso para la categoría de las enfermedades respiratorias. El tipo de dato de esta variable es numérico y el rango corresponde de 0 a 200 casos semanales.

12. PM10: Cantidad de partículas menores a 10 micrómetros medidas y se utiliza porqué hay estudios que indican esta asociación de partículas con índices de mortalidad y morbilidad.(Galvis & Rojas, n.d.) . Esta variable es de tipo flotante y el rango de fechas sobre esta variable corresponde a los años 2012 a 2016.

13. PM2.5: Cantidad de partículas menores a 2.5 micrómetros medidas, se utilizará porque está asociada con índices de mortalidad y morbilidad. Esta variable es de tipo flotante, sus valores se usarán en el rango de años 2012 a 2016. (Galvis & Rojas, n.d.)

14. Precipitación: Precipitación obtenida en una fecha. Se utilizará, porque afecta a la calidad del aire y de acuerdo a la precipitación en la que se encuentre una población, puede llegar a generar enfermedades respiratorias. El tipo de dato es flotante y sus valores están registrados para los años 2012 a 2016.(Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

15. SO2: Cantidad de dióxido de azufre medido en una fecha. Se utiliza, porque es una partícula base en la medición de la calidad del aire la cual afecta dependiendo a su cantidad a la generación de enfermedades respiratorias. El tipo de dato es flotante y sus valores están registrados para los años 2012 a 2016. (Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

16. Temperatura máxima: Temperatura máxima obtenida en una fecha. Se utilizará, porque afecta a la calidad del aire y de acuerdo a la temperatura en la que se encuentre una población, puede llegar a generar enfermedades respiratorias. También se elige para saber a qué temperatura máxima en la medición puede llegar a afectar los factores anteriormente mencionados.

El tipo de dato es flotante y sus valores están registrados para los años 2012 a 2016.(Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

17. Temperatura mínima: Temperatura mínima obtenida en una fecha. Se utilizará, porque afecta a la calidad del aire y de acuerdo a la temperatura en la que se encuentre una población, puede llegar a generar enfermedades respiratorias. También se elige para saber a qué temperatura mínima en la medición puede llegar a afectar los factores anteriormente mencionados. El tipo

de dato es flotante y sus valores están registrados para los años 2012 a 2016. (Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

18. Temperatura promedio: Promedio de las temperaturas obtenidas en el día de una fecha. Se utilizará, porque afecta a la calidad del aire y de acuerdo a la temperatura en la que se encuentre una población, puede llegar a generar enfermedades respiratorias. También se elige para saber a qué temperatura promedio en la medición puede llegar a afectar los factores anteriormente mencionados. El tipo de dato es flotante y sus valores están registrados para los años 2012 a 2016.(Instituto de Hidrologia, 2010)(Hernández-Flórez et al., 2013)

Según la información obtenida, las variables sobre calidad del aire se repiten para la zona de Kennedy y Carvajal las cuales se agrupan para analizar en función de un solo conjunto de zonas y se analizan individualmente para ver el impacto obtenido en cada zona.

8. DISEÑO

En esta sección se dará a conocer el diseño de cómo se manejan los datos a través del punto de vista de información con sus correspondientes diagramas, de acuerdo al modelo de datos planteado por TOGAF como: Diagrama de ciclo de vida, diagrama de diseminación de datos y diagrama de origen de datos.

8.1 DISEÑO DE ARQUITECTURA DE DATOS

Es necesario definir una arquitectura de datos, para identificar aspectos a tener en cuenta en el empleo adecuado de los datos: su recolección, agrupación y proceso. Para cumplir con el objetivo se propone utilizar el marco de referencia TOGAF(Standard & Group, 2011) que proporciona métodos y herramientas dentro de una arquitectura empresarial la cual, en conjunto con la minería de datos, facilita la toma de decisiones:

A continuación se presentan consideraciones clave, propias de la metodología TOGAF(Standard & Group, 2011) específicamente en la Fase C, que comprende Arquitecturas de Sistemas de Información para el desarrollo de arquitectura de datos:

1. **Gestión de Datos:** Esta administración permite el uso eficaz de los datos, el cual basados en datos de la salud se puedan determinar patrones mediante *RapidMiner*.
2. **Migración de Datos:** Cuando se decida reemplazar en algún momento la herramienta, es necesario identificar que datos se van a migrar y si es necesario realizar una limpieza de datos o incluso realizar una actualización sobre los datos para que cumplan con nuevos objetivos propuestos si fuese necesario.
3. **Gobernanza de Datos:** El enfoque dado para esta consideración es el recurso humano el cual aborda las habilidades del ingeniero para la transformación de los datos en información.

Uno de los pasos a utilizar de la fase C de la arquitectura de datos es el siguiente:

1. **Seleccionar modelos de referencia, puntos de vista y herramientas:** Los puntos de vista dentro de la Arquitectura de datos, comprenden las partes interesadas, es decir, la Facultad de Ingeniería de la Universidad Católica de Colombia, los estudiantes correspondientes a la Investigación y el Ministerio de Salud.

Por otra parte, los siguientes catálogos se deberán tener en cuenta para el diseño de la arquitectura de datos, diagrama del ciclo de vida de los datos y diagrama de diseminación de datos.

1. Diccionario de datos:

Tabla 3 Diccionario de datos

Nombre Variable	Tipo de Dato
1.Ciudad	Cadena
2.Fecha	Fecha
3.Numero casos	Numérico
4.Altitud	Flotante
5.EPS	Cadena
6.Mortalidad	Flotante
7.Habitantes	Entero
8.Estrato	Entero
9.Fecha medición calidad aire	Fecha
10.PM10	Flotante
11.PM2.5	Flotante
12.Temperatura máxima	Flotante
13.Temperatura mínima	Flotante
14.Temperatura promedio	Flotante
15.Precipitación	Flotante
16.NO2	Flotante
17.CO2	Flotante
18.CO	Flotante

Fuente: Los autores

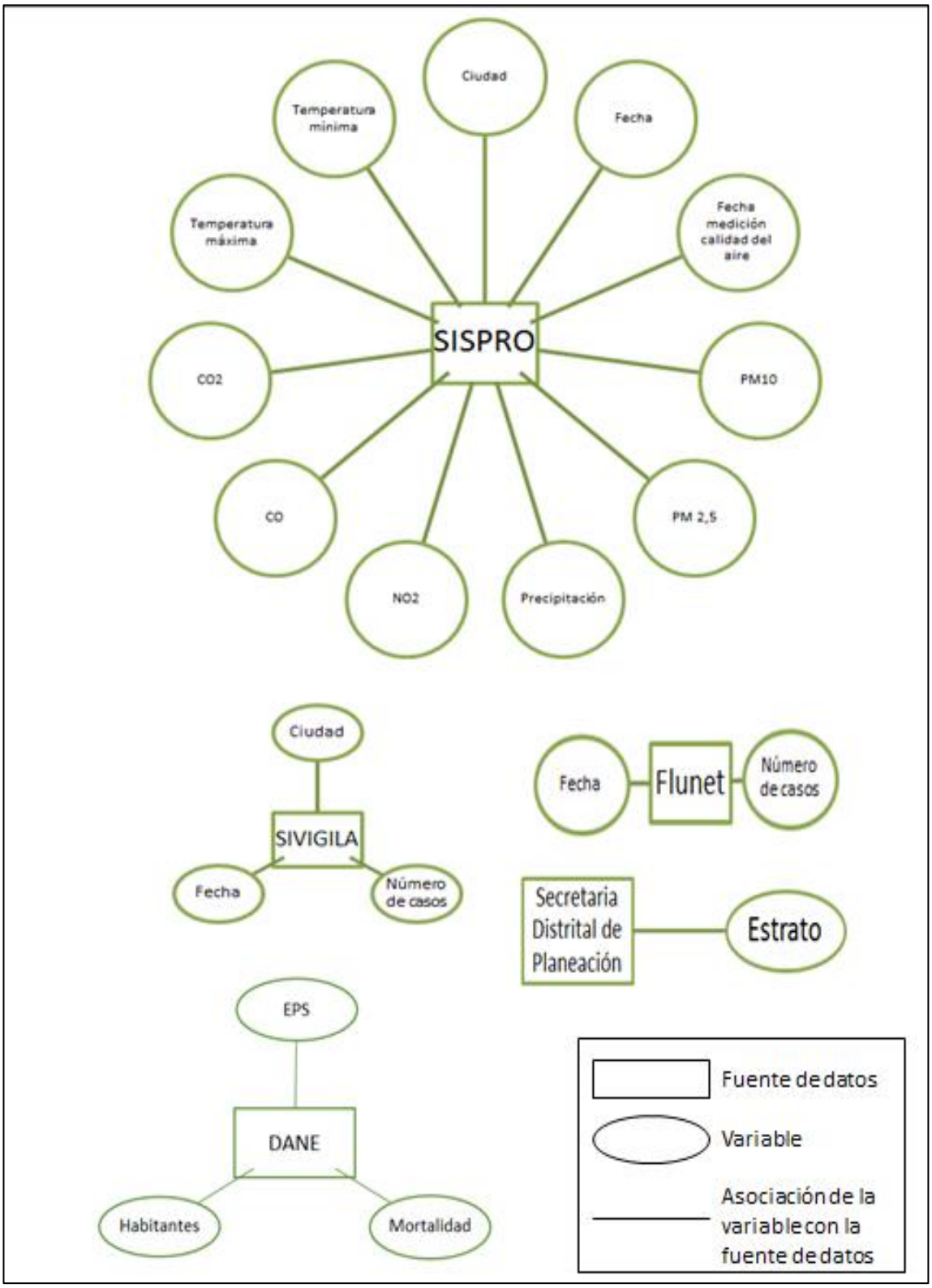
En la gráfica anterior se cuentan con 18 variables en total. La especificación de cada una de las variables esta descrita en la sección 6.2 Definición de variables de este documento.

2. Punto de vista de información

En este punto de vista se dará a conocer como es la composición y el manejo que se le dará a la información obtenida a través de las fuentes de extracción ya mencionadas.

El origen de los datos para cada una de las variables se puede ver a continuación:

Ilustración 7 Diagramas de origen de datos



Fuente: Los autores

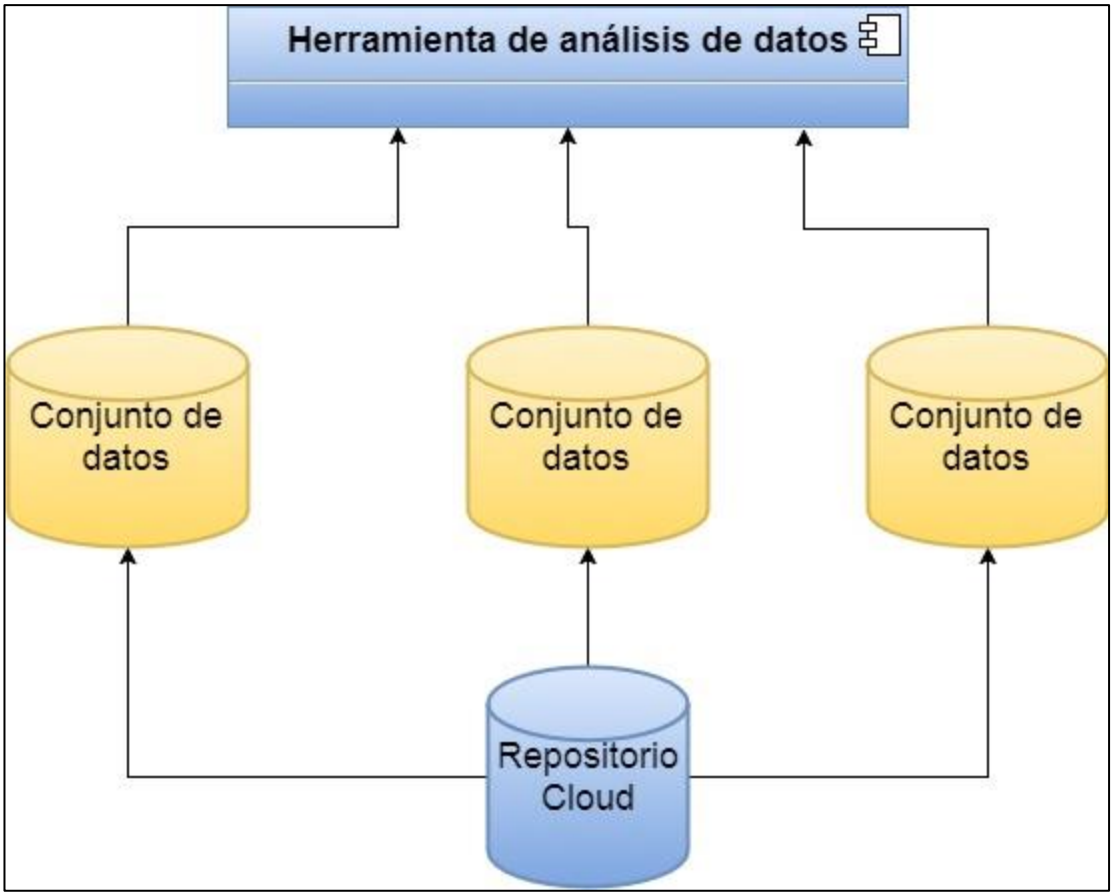
En los diagramas anteriores se pueden identificar los elementos encerrados en rectángulo como fuentes de extracción y los círculos como las variables. Cada variable apunta a una fuente de extracción lo cual significa que la variable está asociada a esa fuente de extracción.

Como resumen las fuentes de Artículos y secretaria distrital de planeación tienen una variable. La fuente de Flunet tiene dos variables mientras que las fuentes del DANE y SIVIGILA tienen 3 variables. Por último, la fuente SISPRO tiene 13 variables concluyendo que la fuente de SISPRO es la principal fuente por la razón de que cubre la mayor parte de las variables definidas.

3. Diagrama de Diseminación de datos:

En este diagrama se presenta que los datos están distribuidos en conjuntos de datos almacenados en un repositorio *cloud* que son utilizados por la herramienta de análisis de datos para realizar el respectivo análisis de datos para uno o varios conjuntos de datos.

Ilustración 8 Diagrama de diseminación de datos



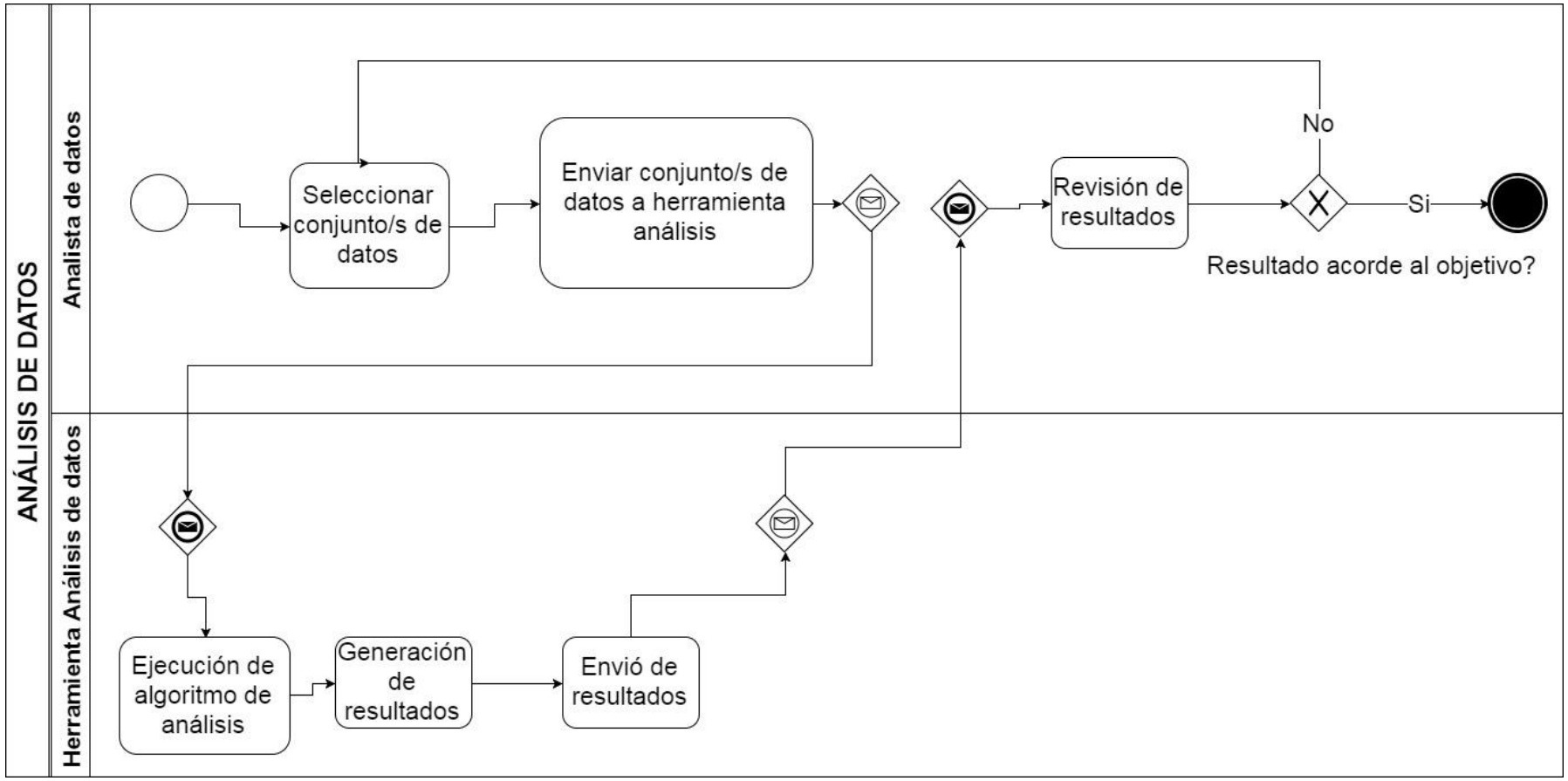
Fuente: Los autores

4. Diagrama de ciclo de vida:

En este diagrama se muestra el proceso del análisis de datos consta de los actores: Analista de datos y la herramienta de análisis de datos, los cuales interactúan para obtener el objetivo de la investigación. Para ello el analista de datos envía los datos a la herramienta de análisis de datos, la herramienta aplica el algoritmo de *Clustering* y el analista revisa los resultados obtenidos de *RapidMiner*.

El analista, después de revisar los resultados verifica si cumplen con el objetivo propuesto, si no cumplen con el objetivo propuesto, se repite el proceso hasta encontrar un resultado adecuado con el objetivo; si el analista obtiene un resultado acorde al objetivo, el proceso finaliza.

Ilustración 9 Diagrama de ciclo de vida



Fuente: Los autores

StakeHolders:

Los *StakeHolders* que interactuaran sobre este punto de vista son:

- Analista de datos: Se encargará de ejecutar el algoritmo de *clustering* en la herramienta de análisis de datos, los conjuntos de datos para la muestra de resultados.
- Arquitecto de datos: Se encargará de diseñar y mostrar a través de un modelo de arquitectura de datos como se va a conformar la información a utilizar.
- Investigador: Se encarga de revisar los resultados obtenidos en la herramienta de análisis de datos y verificar si satisfacen el objetivo principal.

Estos roles los tienen asignados los estudiantes que realizan la investigación.

9. SELECCIÓN DE ALGORITMOS DE CLUSTERING

Es necesario aplicar técnicas de *clustering* porque hay que realizar una clasificación de los datos obtenidos en las fuentes de extracción de una manera homogénea, para la identificación de patrones en enfermedades respiratorias.

9.1 ALGORITMOS DE CLUSTERING

Existen una gran cantidad de algoritmos de clustering, y posiblemente se generen más a medida que pasa el tiempo de acuerdo a las necesidades generadas, es por ello que para la pertinente selección de algoritmos de clustering se consideran inicialmente los siguientes algoritmos de clustering:

1. CLARA
2. AGNES
3. ROCK
4. K-MEANS
5. CURE
6. CHAIN-MAP
7. SIMPLE LINK
8. COMPLETE LINK
9. AVERAGE LINK
10. CHAIN-MAP
11. DBSCAN
12. COBWEB
13. EM
14. AUTOCLASS

9.2 CRITERIOS DE SELECCIÓN

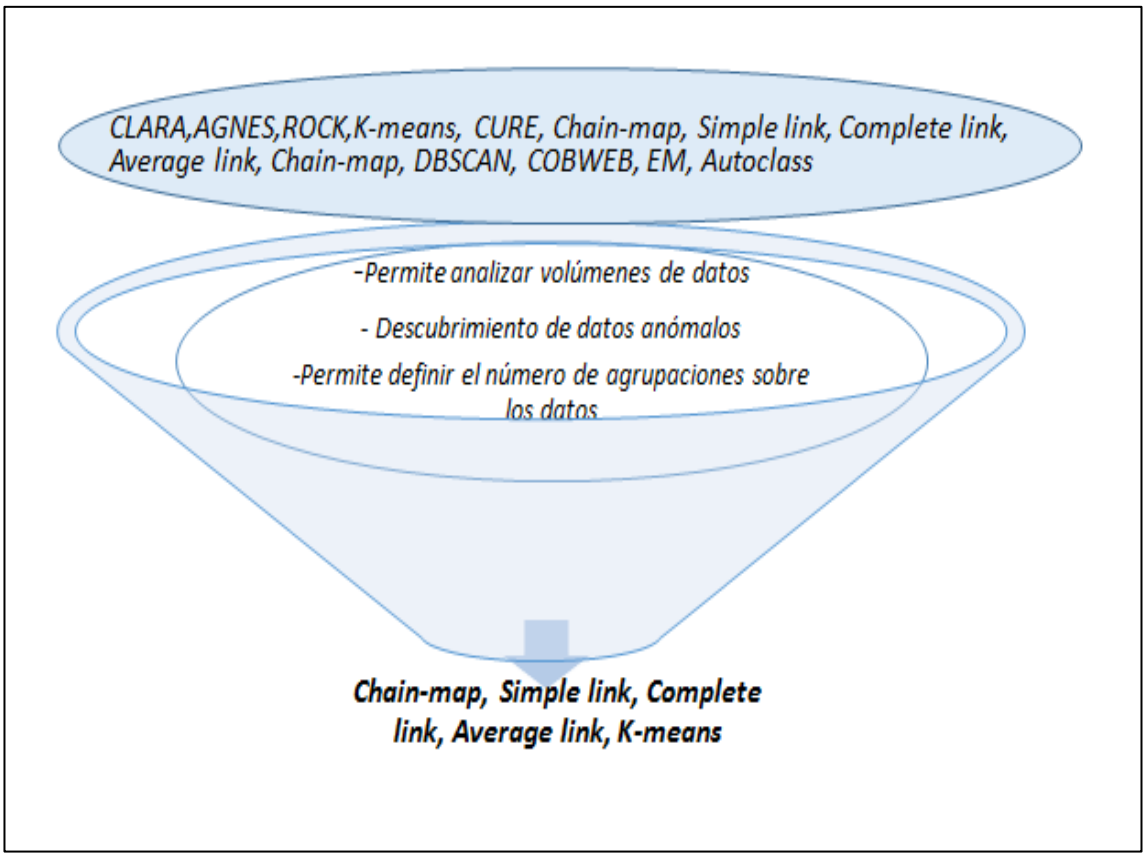
A continuación, se definen los criterios de selección sobre los algoritmos de clustering:

- Permite analizar volúmenes de datos. Por ejemplo: El número de años a validar en la herramienta es información de cinco años.
- Descubrimiento de datos anómalos de las fuentes de información. Por ejemplo: No se generan agrupamiento de los datos.
- Permite definir el número de agrupaciones sobre los datos. Por ejemplo: Se realiza la ejecución inicial de un algoritmo para identificar el número de clúster.

9.3 SELECCIÓN DE ALGORITMOS DE CLUSTERING

Previamente revisados los criterios de selección de los algoritmos de *clustering*, se procede a evaluar cuales cumplen con los criterios para una óptima selección.

Ilustración 10 Selección de algoritmos de Clustering



Fuente: Los autores

De acuerdo a la gráfica anterior se eligieron los algoritmos de clustering de acuerdo a dos tipos de agrupamiento los basados en particiones y jerárquicos, los cuales cumplen con todos los criterios de selección.

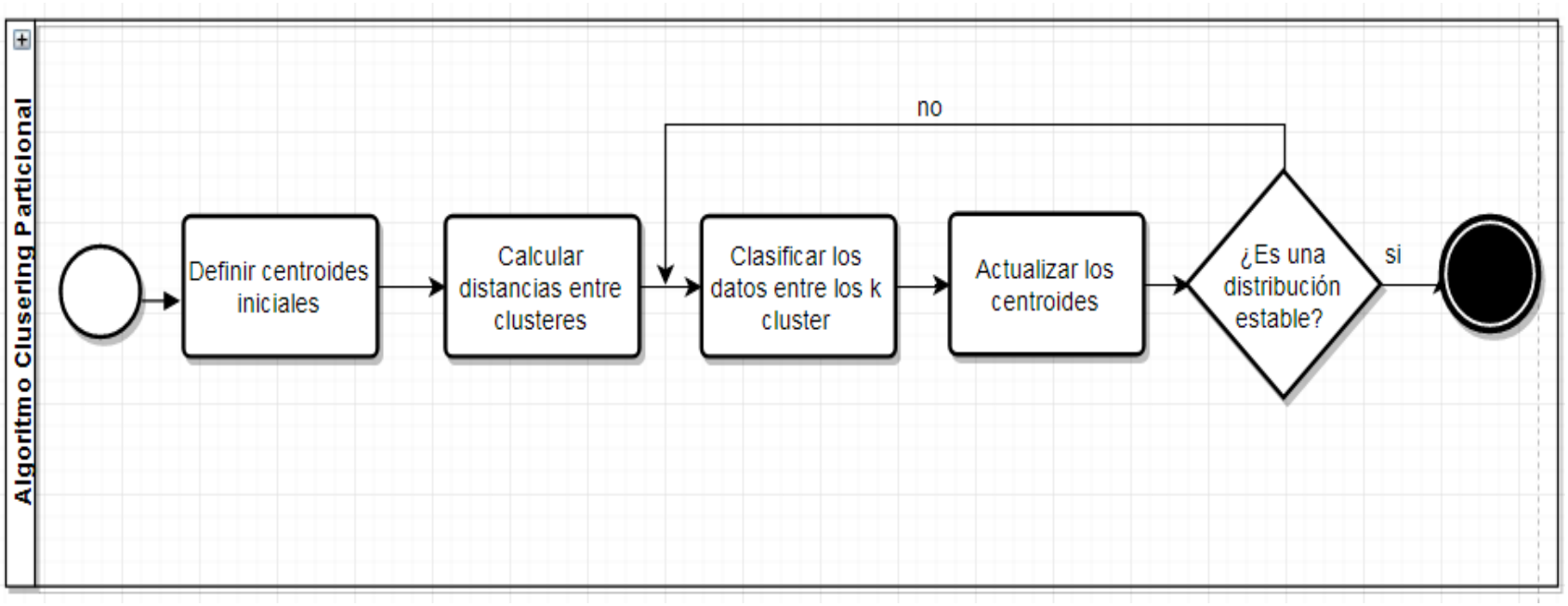
En primer lugar se encuentra el algoritmo de tipo particional *chain-map Ver sección* **5.1.5.1 Algoritmo de Clustering Particional**, en donde se aplica inicialmente ya que determina un número de clúster a buscar.

Posteriormente y luego de tener los clústeres a ejecutar obtenidos del algoritmo *chain-map* se procede a indicar el número de clúster en el algoritmo k-means el cual será el algoritmo a utilizar, porque cumple con los criterios planteados. Este algoritmo se encargará de realizar las particiones respectivas al mismo nivel.

Seguidamente se realiza la ejecución del algoritmo k-means que comienza a realizar el agrupamiento de los datos dependiendo el tipo de relación entre ellos.

A continuación, se muestra el diagrama de flujo para el algoritmo particional *k- means*:

Ilustración 11 Diagrama de flujo particional k-means



Fuente: Los autores

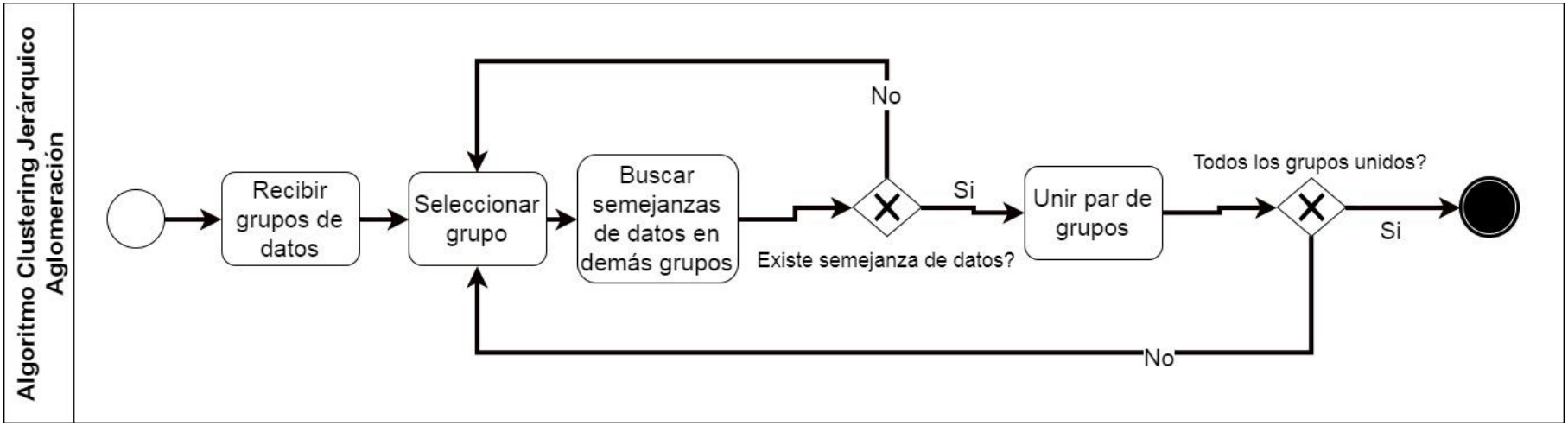
El diagrama de *clustering* particional *k-means* indica un flujo en donde comienza a definir los centroides iniciales o el número de clúster con el cual se va a trabajar, luego el algoritmo calcula las distancias entre clústeres para identificar la similitud en los datos, posteriormente realiza una clasificación de los datos entre los k clúster dados inicialmente. Seguidamente realiza la actualización de los centroides dependiendo de los grupos formados, si es una distribución estable, es decir, hay unos grupos previamente formados finaliza el algoritmo, si no, vuelve a calcular las distancias entre los clústeres.

También se utilizan los algoritmos de *clustering* jerárquico aglomerado como: simple link y complete link, para agrupar en jerarquías de acuerdo a las semejanzas en los grupos donde se está realizando la ejecución.

La utilización del *clustering* de aglomeración permite ver las semejanzas o diferencias que existan en los grupos de *clustering* que se van generando, y realiza las divisiones en forma jerárquica, por tanto, para la identificación de patrones se utiliza en la búsqueda de semejanzas que se obtiene del algoritmo.

Se da a conocer gráficamente de cómo es el proceso del algoritmo de aglomeración jerárquica en el siguiente diagrama:

Ilustración 12 Diagrama de flujo algoritmo jerárquico aglomeración

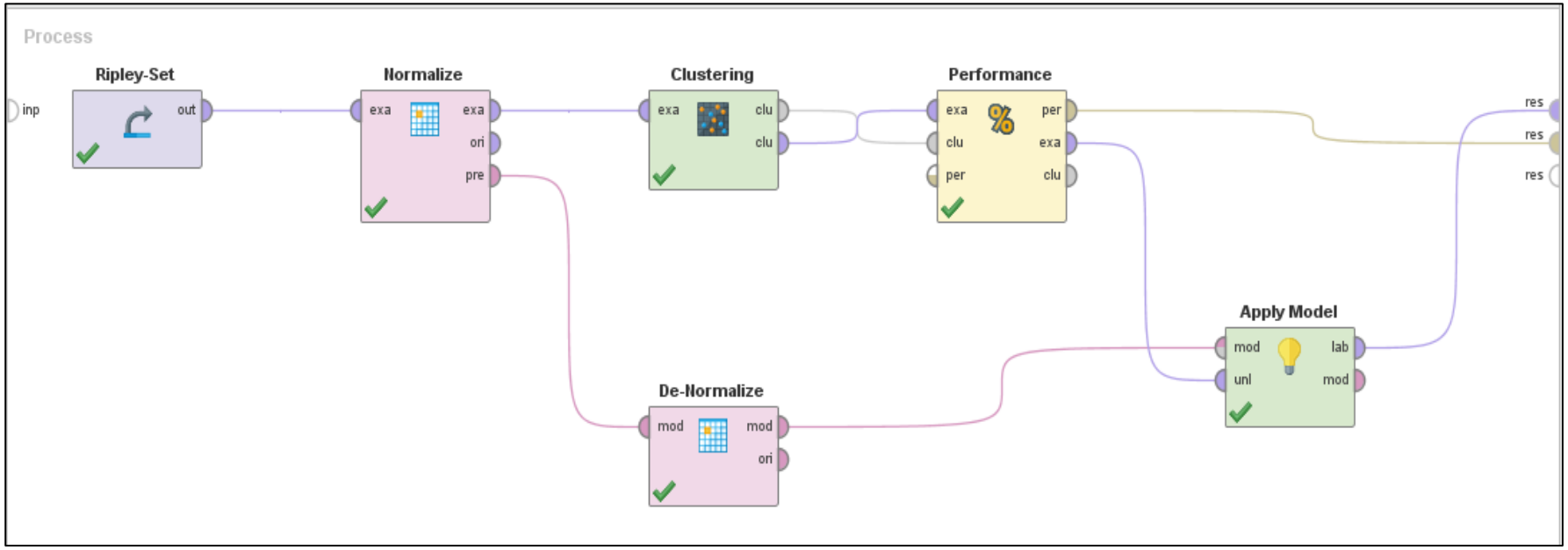


Fuente: Los autores

El algoritmo para el caso de tipo aglomeración recibe varios conjuntos de datos los cuales selecciona uno y compara contra los demás grupos buscando la mayor semejanza en datos, donde si encuentra el grupo indicado para el grupo seleccionado, estos grupos se unen, vuelve y realiza el proceso con el siguiente grupo, así sucesivamente hasta que todos los grupos estén unidos terminando el proceso del algoritmo. En caso contrario de que el grupo seleccionado no tenga semejanzas con los demás, iniciara con el siguiente continuando con el proceso normal del algoritmo.

A continuación, se detalla el proceso para el algoritmo de clustering k-means aplicado en RapidMiner:

Ilustración 13 Proceso para el algoritmo de clustering k-means



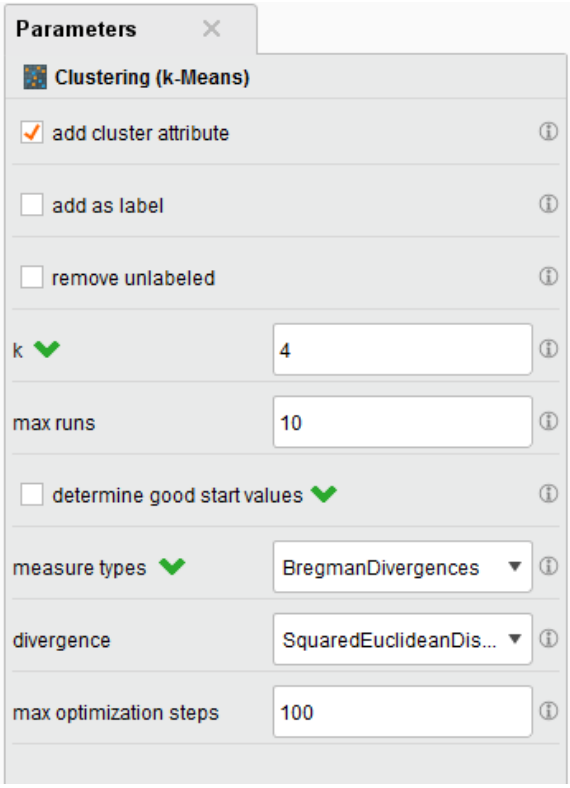
Fuente: Los autores

En este proceso se utilizaron componentes de normalización, clustering, rendimiento, des normalización y aplicación de un nuevo modelo. Comenzando el proceso se toma un *Ripley- Set* que contiene los datos a ser analizados enviándose al proceso de normalización el cual se encargara de ajustar los valores que tienen los datos para que no hallan valores más grandes que otros lo cual causaría que el algoritmo de clustering no ejecutara adecuadamente porque al tomar valores que

son más altos que otros, por ejemplo: Año (2016) contra número de casos (14) el algoritmo realizara el agrupamiento acorde a estos y sus resultados quedan dispersos.

El siguiente proceso a realizar es él envió de la tabla normalizada al proceso de *clustering* el cual ejecutara el algoritmo de *k-means* de acuerdo a los siguientes parámetros:

Ilustración 14 Atributos del proceso para el algoritmo de clustering k-means



Fuente: Los autores

Donde se realiza check la agregación de un atributo para mostrar que datos quedaron asignados a cada clúster, k como número de clústeres el cual este número fue elegido de acuerdo al proceso de rendimiento que se mencionara más adelante.

Las siguientes variables la herramienta de análisis recomienda las siguientes opciones con valores por defecto de acuerdo a que la aplicación dentro del algoritmo utilizando esos subprocesos es la más óptima para recibir parámetros y para la muestra de resultados:

- ***Max runs****:* Es el número de iteraciones que realizara el algoritmo, por defecto la herramienta recomienda 10 iteraciones.
- ***Measure types*****:** Es el tipo de medida que usara en la ejecución del algoritmo, el cual, *Bregman divergences* es el recomendado frente a opciones como medidas mixtas, medidas numéricas, medidas nominales, ya que con esta opción se obtienen resultados más exactos ya que maneja datos tipo flotante.
- ***Divergence*****:** Se utiliza como recomendado la distancia euclidiana que se encargara de medir la distancia más óptima entre las variables comparándose por medio de vectores.
- ***Max optimitation steps*****:** Este parámetro especifica el número máximo de iteraciones realizadas para una ejecución de *k-Means*.

Al mismo tiempo que se ejecuta el proceso de clustering se ejecuta el proceso de des normalización el cual se encarga de regresar la tabla a sus valores normales para ser enviada a la aplicación de un modelo.

Para que los datos sean correctamente tomados dentro del algoritmo de clustering se tuvieron que convertir el tipo de dato de las variables enfermedad y genero a tipo numérico.

Después de la ejecución del proceso del algoritmo de clustering se envía al proceso de rendimiento el cual se encarga de medir distancias con respecto a sus centroides. Este proceso es necesario para obtener el número de clústeres ya que a medida que se ajusta el número de clústeres la distancia va acercándose al centroide correspondiente de cada clúster, lo cual evidencia una mejor agrupación de cada clúster con las respectivas variables.

Finalmente, la tabla normalizada y el modelo de clúster pasan por el proceso de agregar modelo el cual se encarga de recibir un modelo y ajustarlo frente a un proceso que corresponde en este caso al algoritmo de clustering para que los resultados se realicen sobre los valores de los datos originales y no sobre los valores que traducen los datos originales a una forma normalizada.

10. RECONOCER PATRONES A PARTIR DE LA INFORMACIÓN RECOPILADA

Para el análisis se tuvieron en cuenta cuatro clústeres que permitieron realizar la distribución de los datos de acuerdo a sus parámetros en común. Para la ejecución del clúster se tuvieron en cuenta 10.000 registros y se contó con cinco variables que son: tipo de enfermedad, año, edad, número de casos y género serán analizadas en cada uno de los clústeres.

A continuación, se encuentra el análisis correspondiente a cada clúster de acuerdo a la distribución realizada por el algoritmo k-means asociado con el tipo de enfermedad:

10.1 ANÁLISIS DE RESULTADOS

10.1.1 Análisis Clúster 0

Este clúster presenta una agrupación total de 2.426 registros, los cuales tienen asociados 115 tipos de enfermedades y 59.345 casos. En general se puede evidenciar que presenta las edades de 3, 4 y 5 años, siendo el porcentaje más representativo la edad de 5 años. Por otra parte, generará una proporción similar en la distribución de género con un porcentaje del 51% para el género masculino y un 49% para el género femenino. Ver Ilustración 18 Clúster vs Número de casos, Ilustración 16 Clúster vs Edad, Ilustración 17 Clúster vs Genero.

Este clúster se caracteriza por tener la menor incidencia en el tipo de enfermedad “RINOFARINGITIS CRONICA” representando tan solo un 0,09% con 88 casos para las edades de 5 años entre género femenino y masculino.

Dentro de la evaluación de los datos también se puede evidenciar que para el año 2014 se genera una gran incidencia de casos para el diagnóstico J459 - ASMA, NO ESPECIFICADA (representando un porcentaje de 19% para todos los años analizados) para niños entre las edades 4 y 5 años; generando 1.763 casos para el género femenino y 2.155 casos para el género masculino, originando un decrecimiento posterior de esta enfermedad para los años 2015 a 2016 llegando en este último año a 694 casos para el género femenino con una disminución del 39% y 874 casos para el masculino con una disminución del 41%. Se podría inferir que para este tipo de enfermedad se crearon planes de acción por que el número de casos cayó significativamente del 2014 a los años 2016. Ver Ilustración 14 Clúster Vs Enfermedad, Ilustración 16 Clúster vs Edad, Ilustración 17 Clúster vs Genero, Ilustración 18 Clúster vs Número de casos.

Referente al diagnóstico J980 - ENFERMEDADES DE LA TRAQUEA Y DE LOS BRONQUIOS, NO CLASIFICADAS EN OTRA PARTE, representa el 25,30% de los casos con 15.012. Para esta enfermedad se encuentran todos los años de 2012 a 2016 y los registros reportados correspondes a edades 3, 4

y 5 años de ambos géneros. Esta enfermedad tiene un comportamiento con crecimiento exponencial para el género femenino y masculino, pasando de 1211 a 1738 generando un aumento del 69,68% de los casos. Ver Ilustración 14 Clúster Vs Enfermedad.

Finalmente, este clúster tiene un total de 59.345 casos, distribuidos en cada una de los tipos de enfermedades. Ver Ilustración 19 Distribución de números de casos en cada clúster.

Conclusión: 115 enfermedades respiratorias actúan en edades de 3,4 y 5 años desde los últimos 5 años donde afecta en mayor parte a la población de género masculino.

Hipótesis: Si se aplican estrategias de promoción y prevención para el año 2017 sobre el diagnostico J980 - Enfermedades de la tráquea y de los bronquios, no clasificadas en otra parte, entonces habrían casos evitables, porqué de acuerdo a el clúster analizado representa un crecimiento exponencial desde el año 2014 a 2016, comenzando con 2.572 casos en el 2014 y generando para este último año 3819 casos registrados para Bogotá, lo cual indica que no se hayan realizado planes de acción este periodo.

10.1.2 Análisis Clúster 1

Este clúster presenta una agrupación total de 2.520 registros y 55.404 casos reportados, los cuales tiene asociadas 122 tipos de enfermedades de 184 analizadas en total. Adicionalmente, este clúster se caracteriza por tener asignadas todas las edades estimadas de 0 a 5 años con una agrupación para los últimos años correspondientes a 2014, 2015 y 2016. Ver Ilustración 18 Clúster vs Número de casos, Ilustración 15 Clúster vs Año, Ilustración 16 Clúster vs Edad, Ilustración 14 Clúster Vs Enfermedad.

De acuerdo al clúster generado, el tipo de enfermedad menos representativa para esta agrupación es la del diagnóstico “J684 - AFECCIONES RESPIRATORIAS CRONICAS DEBIDAS A INHALACION DE GASES, HUMOS, VAPORES Y SUSTANCIAS QUIMICAS” con 0 casos reportados para el año 2016. Ver Ilustración 14 Clúster Vs Enfermedad.

Por otra parte, el tipo de enfermedad más generada en este clúster es el diagnostico J303 - OTRAS RINITIS ALERGICAS, en donde tiene asociado 8112 casos con una representación del género femenino del 46,72% en donde se encuentran todas las edades, en cambio, el género masculino representa para este diagnóstico el 53,98% de los casos para todas las edades. Para el año 2015 tiene su punto máximo para los dos géneros, con 1771 casos para las niñas y 2332 para los niños. Ver Ilustración 14 Clúster Vs Enfermedad, Ilustración 17 Clúster vs Genero, Ilustración 18 Clúster vs Número de casos.

Presenta una distribución similar con el clúster 0 con respecto a la variable género, en donde el género masculino representa el 49.5% de los casos mientras que el género femenino el 50.5%. Ver Ilustración 17 Clúster vs Genero.

Finalmente, este clúster tiene un total de 132.172 casos, distribuidos en cada una de los tipos de enfermedades. Ver Ilustración 19 Distribución de números de casos en cada clúster.

Conclusión: 122 enfermedades respiratorias actúan en edades de 0 a 5 años desde los últimos 3 años donde afecta más a la población de género femenino.

Hipótesis: Si hubiesen aplicado planes de prevención entre el año 2014 y 2015 para el diagnóstico J303 - OTRAS RINITIS ALERGICAS para niños entre 0 y 5 años, entonces muchos casos habrían podido ser evitables, porque reportaron un crecimiento del 334% de los casos durante este periodo.

Estas enfermedades comenzaron a actuar desde los últimos 3 años afectando la población femenina, haciendo advertencia a que antes de los últimos 3 años estas enfermedades tenían menos impacto y que en la actualidad su frecuencia aumento.

10.1.3 Análisis Clúster 2

Este clúster presenta una agrupación total de 2.509 registros, los cuales tiene asociadas 116 tipos de enfermedades de 184 analizadas en total, donde los tipos de enfermedad que agrupa corresponden a los códigos 47 a 176 y 184 de acuerdo a la nomenclatura de códigos mencionada anteriormente para cada tipo de enfermedad. Dentro de la agrupación de tipos de enfermedad que tiene este clúster, el tipo de enfermedad menos representativa para esta agrupación son el diagnostico “J301 - RINITIS ALERGICA DEBIDA AL POLEN” y “J30- OTRA RINITIS ALERGICA ESTACIONAL”. Ver Ilustración 14 Clúster Vs Enfermedad, Ilustración 18 Clúster vs Número de casos.

Adicionalmente este clúster se caracteriza por contener edades estimadas de 0 a 2 años con una agrupación para los últimos años correspondientes a 2012, 2013, 2014, 2015 y 2016. Ver Ilustración 15 Clúster vs Año, Ilustración 16 Clúster vs Edad.

Presenta una distribución similar con el clúster 0 con respecto a la variable género, en donde el género masculino representa el 50.29% de los datos mientras que el género femenino el 49.71%. Ver Ilustración 17 Clúster vs Genero.

Finalmente, este clúster tiene un total de 60.136 casos, distribuidos en cada una de los tipos de enfermedades. Ver Ilustración 19 Distribución de números de casos en cada clúster.

Conclusión: Las enfermedades correspondientes a los códigos 47 a 176 y 184 tienden a afectar a niños con edades de 0 a 2 años desde los últimos 5 años afectando más a la población de género masculino.

Hipótesis: Si se toman medidas para prevención en edades de 0 a 2 años para las enfermedades con códigos 47 a 176 y 184, entonces empezaría a disminuir el número de casos y afectados por estas enfermedades porque estas enfermedades afectan principalmente a este segmento de población.

10.1.4 Análisis Clúster 3

Este clúster presenta una agrupación total de 2.541 registros, los cuales tiene asociadas 115 tipos de enfermedades de 184 analizadas en total, en donde los tipos de enfermedad que agrupa corresponden a los códigos 0 a 119 de acuerdo a la nomenclatura de códigos mencionada anteriormente para cada tipo de enfermedad. Dentro de la agrupación de tipos de enfermedad que tiene este clúster, los tipos de enfermedad menos representativas para esta agrupación son los siguientes diagnósticos:

1. J47X - BRONQUIECTASIA
2. J628 - NEUMOCONIOSIS DEBIDA A OTROS POLVOS QUE CONTIENEN SILICE
3. J64X -NEUMOCONIOSIS, NO ESPECIFICADA
4. J660 - BISINOSIS
5. J668 - ENFERMEDAD DE LAS VIAS AEREAS DEBIDAS A OTROS POLVOS ORGANICOS ESPECIFICOS
6. J670 - PULMON DEL GRANJERO

Adicionalmente este clúster se caracteriza por contener edades estimadas de 0 a 5 años con una agrupación para los últimos años correspondientes a 2012, 2013, 2014. Donde se contiene menos cantidad de datos para los años 2015 con un 0.47% y 2016 con un 0.07%. Ver Ilustración 14 Clúster Vs Enfermedad, Ilustración 15 Clúster vs Año, Ilustración 16 Clúster vs Edad, Ilustración 18 Clúster vs Número de casos.

Presenta una distribución similar con el clúster 0 con respecto a la variable género, en donde el género masculino representa el 50.13% de los datos mientras que el género femenino el 49.86%. Ver Ilustración 17 Clúster vs Genero.

Finalmente, este clúster tiene un total de 375.300 casos, distribuidos en cada una de los tipos de enfermedades. Ver Ilustración 19 Distribución de números de casos en cada clúster.

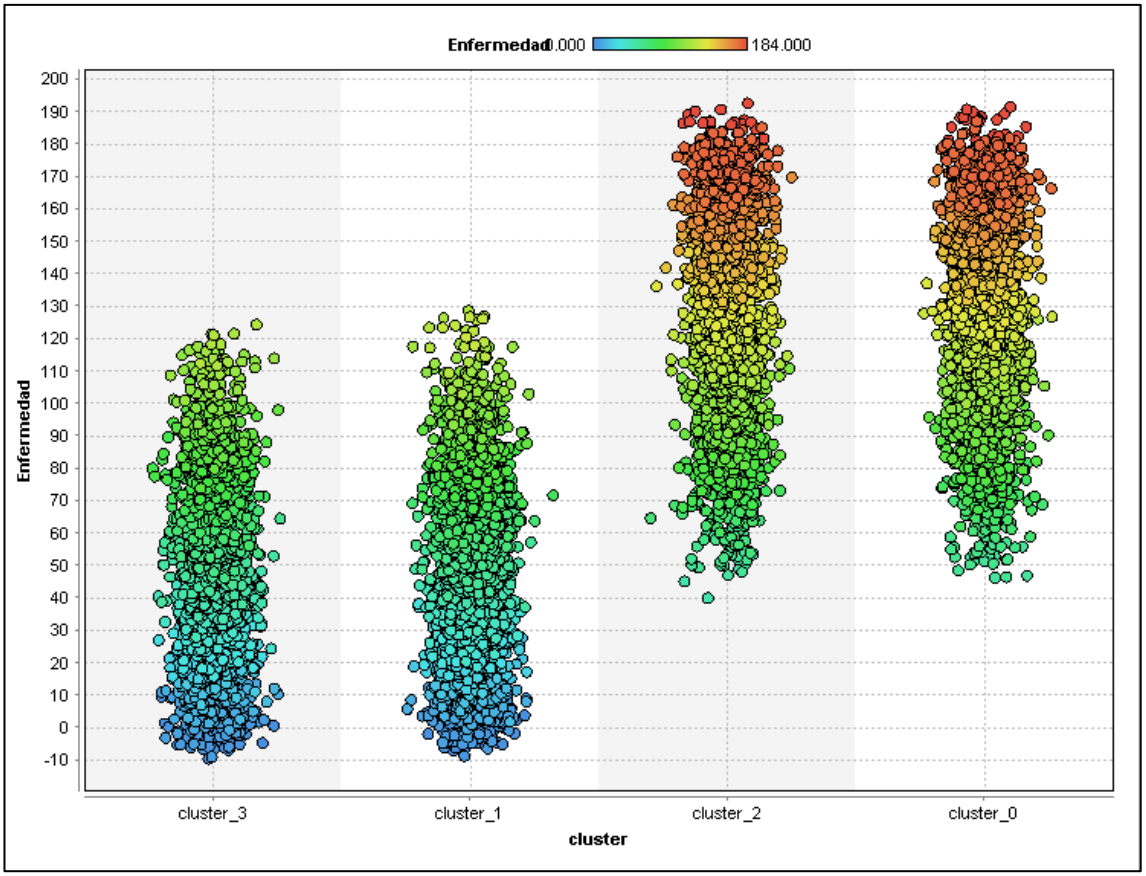
Conclusión: Las enfermedades correspondientes a los códigos 0 a 119 tienden a afectar a niños con edades de 0 a 5 años especialmente a la población de género masculino, el cual este comportamiento se ha ido disminuyendo en los últimos años.

Las enfermedades Bronquiectasia, Neumoconiosis debida a otros polvos que contienen Sílice, neumoconiosis no especificada, Bisinosis, enfermedad de las vías aéreas debidas a otros polvos orgánicos específicos, pulmón del granjero fueron las que menos afectaron a los niños de 0 a 5 años donde en los últimos años pocos casos existen.

Hipótesis: Si enfermedades como bronquiectasia, neumoconiosis debida a otros polvos que contienen sílice, neumoconiosis no especificada, Bisinosis, enfermedad de las vías aéreas debidas a otros polvos orgánicos específicos, pulmón del granjero y las demás enfermedades de códigos 0 a 119 tuvieron poca presencia en los últimos años para esta cantidad de casos, entonces significa que se están realizando planes de prevención pero con falta de más ejecución, porque se siguen frecuentando estas enfermedades a pesar de su disminución.

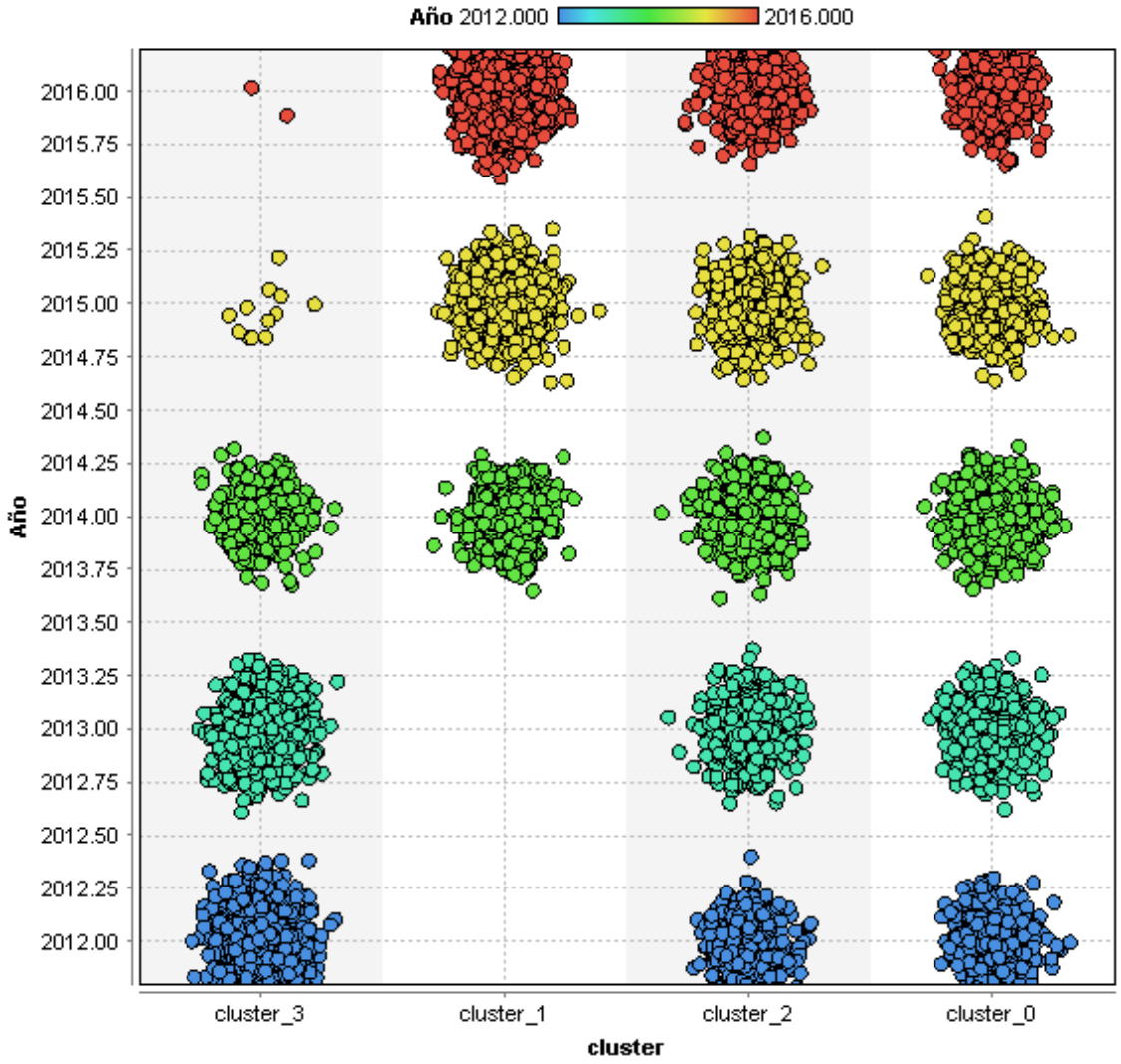
A continuación, se presentan las gráficas que muestran el agrupamiento de cada una de las variables en los clústeres generados por la herramienta de análisis:

Ilustración 14 Clúster Vs Enfermedad



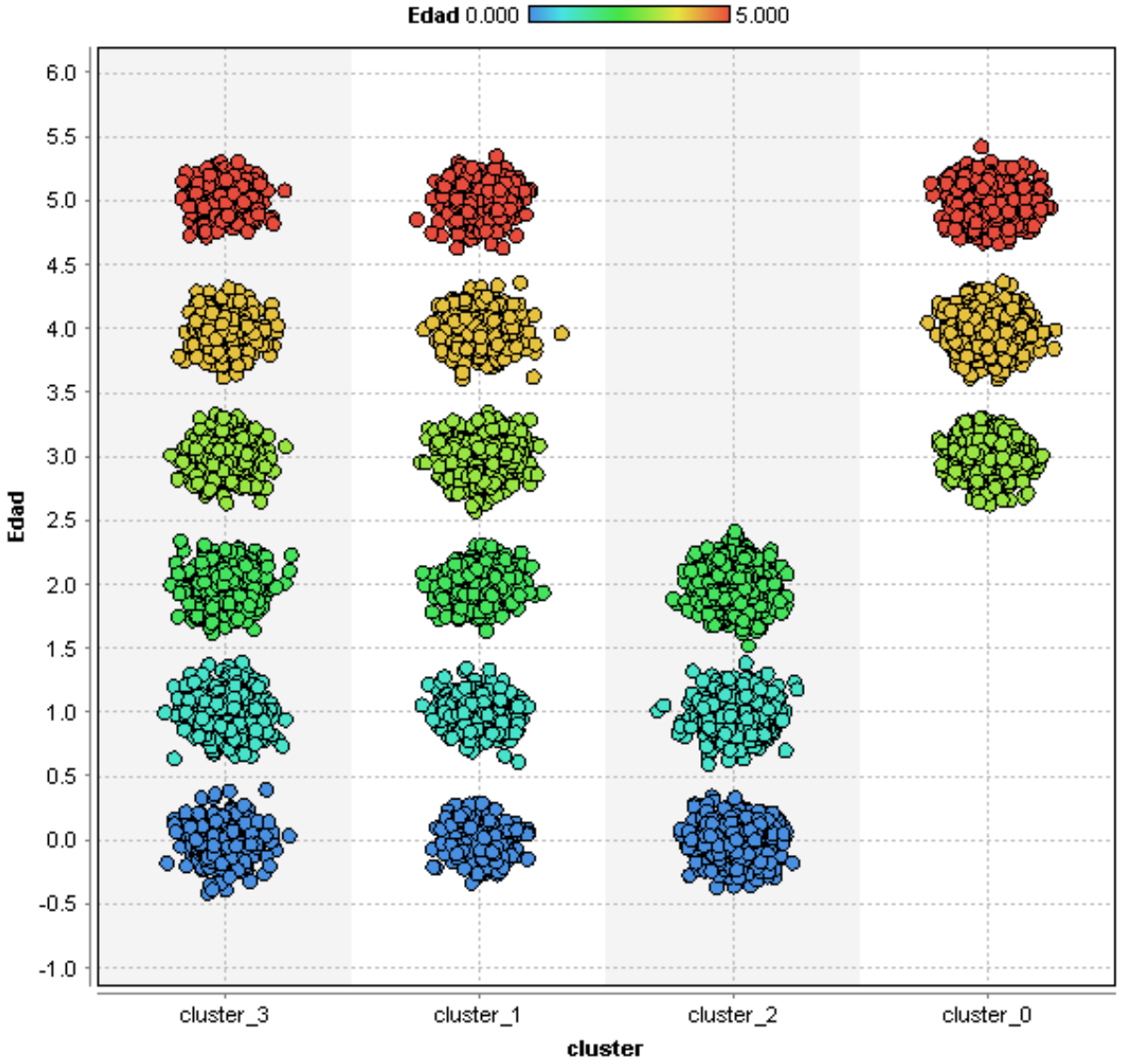
Fuente: Los autores

Ilustración 15 Clúster vs Año



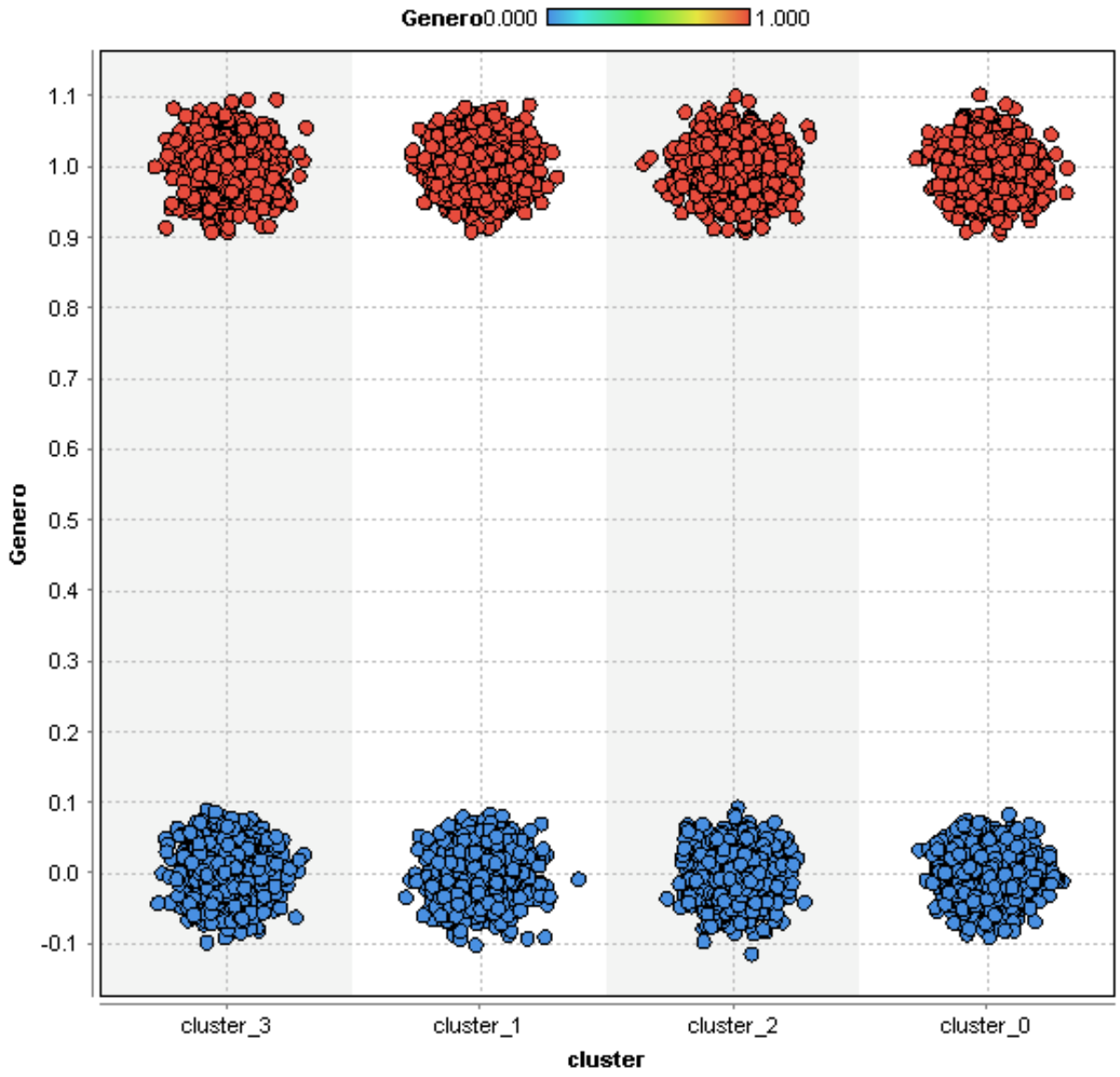
Fuente: Los autores

Ilustración 16 Clúster vs Edad



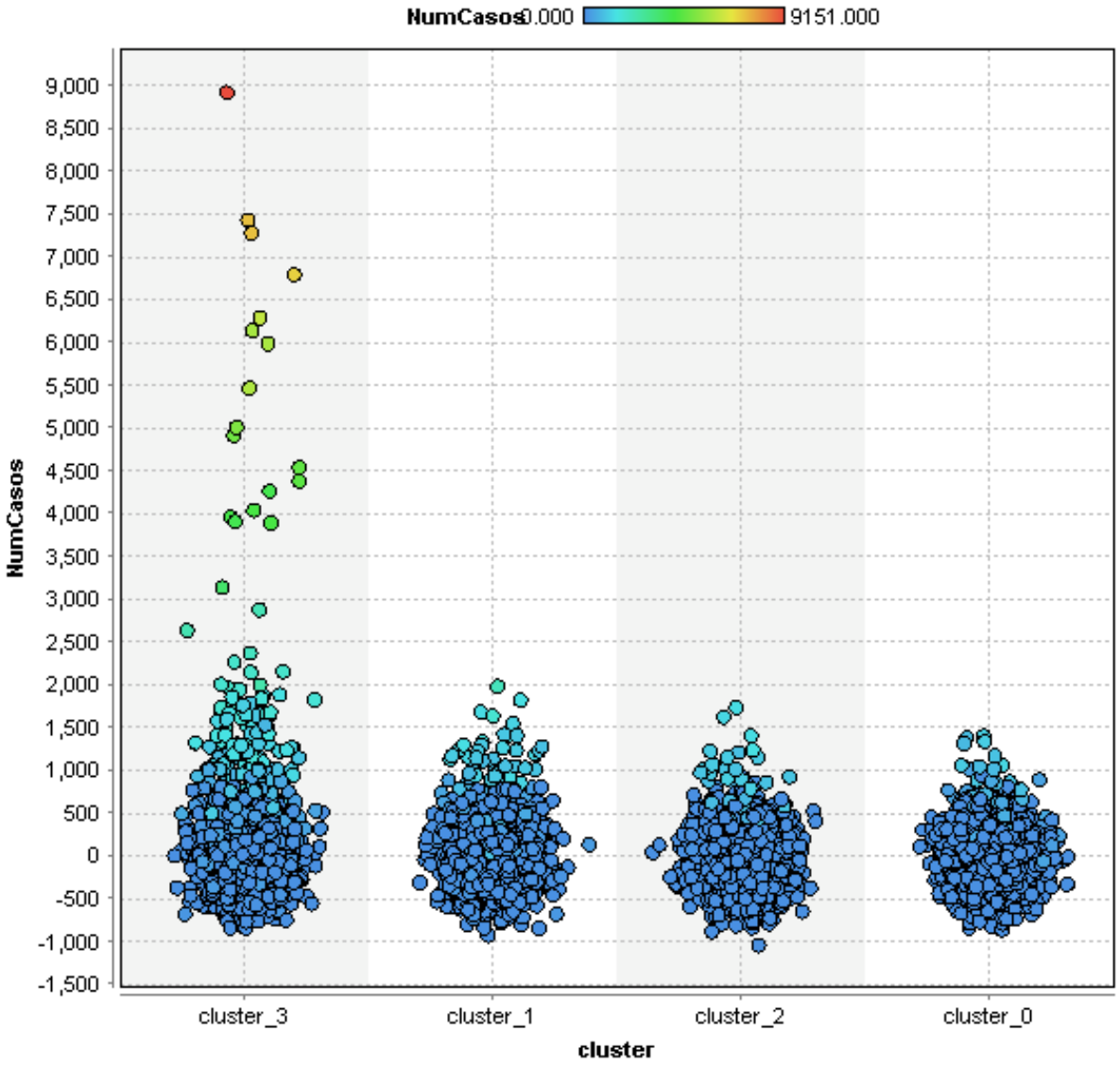
Fuente: Los autores

Ilustración 17 Clúster vs Genero



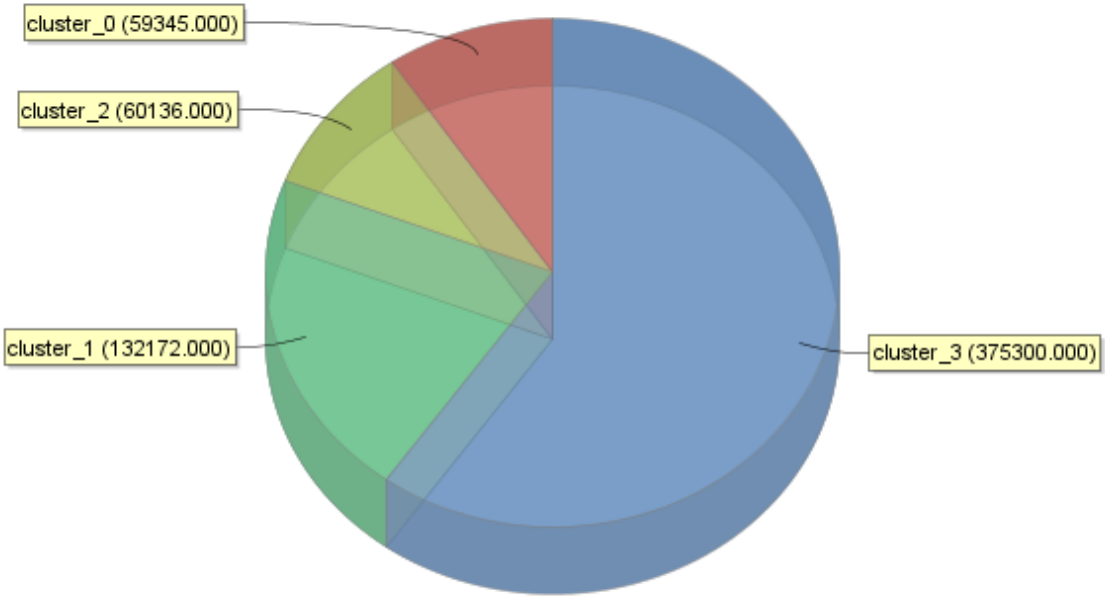
Fuente: Los autores

Ilustración 18 Clúster vs Número de casos



Fuente: Los autores

Ilustración 19 Distribución de números de casos en cada clúster



Fuente: Los autores

11. CONCLUSIONES

1. Como resultado de las fuentes de extracción se obtuvieron los datos correspondientes a cada una de las variables definidas; sin embargo, los datos de las fuentes que se tuvieron en cuenta inicialmente, no contaban con suficientes registros para aplicar minería de datos, ya que algunas variables de las fuentes de extracción presentaban datos resumidos.
   No obstante, se tuvo acceso finalmente a la fuente del cubo SISPRO, que permitió realizar la extracción de manera integral de las variables que se acercaban al objetivo general. Las variables extraídas de esta fuente de extracción son: tipo de enfermedad, año, edad, número de caso y género para la ciudad de Bogotá.

2. La gestión de los datos permitió darle a los mismos una estructura y posteriormente convertirlos en información mediante la aplicación de minería de datos RapidMiner, y finalmente realizar un análisis sobre un volumen de 10.000 registros para la aplicación de clustering. Por otro lado, las arquitecturas de datos permitieron tener una idea general de las fuentes de información, conocer donde se puede extraer datos para la realización de futuras investigaciones y las variables correspondientes a cada fuente de datos.
   A las variables tipo de enfermedad y género se realizó normalización, debido a que el modelo solo recibe datos tipo numéricos.

3. La selección e implementación de algoritmos permitió conocer algoritmos de acuerdo a sus funcionalidades con respecto a la clasificación de algoritmos de *clustering,* en el que se tomaron diferentes métricas para evaluar dichos algoritmos, ayudando en la búsqueda y elección de estos. Esta selección tiene como aprendizajes el conocimiento de clasificación en los algoritmos de *clustering*, conocer sus funcionalidades y ver cual puede llegar a ser el adecuado para la aplicación de un caso de análisis.
   También permitió conocer las herramientas que aplican análisis de datos donde permiten la ejecución de varios algoritmos de búsqueda de datos, a través de un proceso que se diseña en la herramienta para que esta pueda ejecutar de acuerdo al orden establecido en el diseño del proceso y dar resultados

4. El reconocimiento de patrones se puede llegar a realizar a través del análisis de datos por medio de los resultados obtenidos los cuales dan a conocer diferentes comportamientos a través de la agrupación obtenida en los clústeres generados, ya que cada clúster contiene un caso con relación de las variables trabajadas. Reconociendo como un comportamiento con respecto a los resultados obtenidos que la mayor parte de los casos en todas las edades presentaron enfermedades con códigos entre el rango de 0 a 60 aproximadamente.

Los clústeres obtenidos realizaron la correspondiente agrupación de acuerdo a los parámetros iniciales que pide el algoritmo, donde se pudo evidenciar una agrupación que busco la relación posible entre las variables y donde su cantidad también se distribuyó en los clústeres de manera equitativa. También se pudo observar en cada clúster el manejo de las variables que realizo cada uno de estos, donde para las variables de género, año se presentan agrupaciones separadas por valor, eso da a demostrar que el algoritmo si agrupa de acuerdo a los valores de las variables.

Respondiendo a la pregunta ¿Cómo la técnica de minería de datos puede identificar patrones que permitan mejorar los programas de prevención para las enfermedades respiratorias en Bogotá?, se puede concluir que a través de la minería de datos y sus aplicaciones se pueden identificar patrones dentro del análisis de los resultados obtenidos por los clústeres, en el cual se da a conocer por medio de representaciones graficas el comportamiento que tiene cada uno de los datos de las variables con respecto a los demás, en donde posteriormente se podrán tomar decisiones con base en los resultados analizados generando posiblemente programas de prevención y promoción.

Respondiendo a la pregunta ¿Qué comportamiento han tenido las enfermedades respiratorias en los últimos cinco años de la ciudad de Bogotá? Se identificaron cuatro tipos de comportamientos, asociados a la generación de clústeres, los cuales están más especificados en la sección

BENEFICIOS

1. Realizar planes de prevención para niños entre 0-5 años aplicando técnicas de minería de datos, minimizando el tiempo en la toma de decisiones de acuerdo a los resultados obtenidos en los análisis.
2. Encontrar información oculta sobre enfermedades respiratorias para conocer el comportamiento de estas.

12. TRABAJOS FUTUROS

Se recomienda extraer las variables a trabajar con la mínima cantidad de fuentes, teniendo como referencia una o dos fuentes en donde sus variables se puedan relacionar.

También se recomienda aplicar otros algoritmos de *clustering,* con el objetivo de encontrar otros patrones de comportamiento y otro tipo de análisis descriptivo o predictivo que con k-meas no se hayan encontrado.

No se recomienda tener dentro del análisis, variables de tipo constante ya que no permiten ningún tipo de comparación al ser un solo dato fijo, por lo tanto, no va a generar agrupamiento de datos.

Los datos que sean de tipo descriptivo se deben normalizar para poderlos ejecutar los algoritmos de *clustering,* ya que el algoritmo trabaja sobre variables tipo cuantitativas.

13. REFERENCIAS BIBLIOGRÁFICAS

(DANE), D. A. N. D. E. (2017). DANE. Retrieved September 10, 2017, from http://www.dane.gov.co/

Alvarez, F. M., Troncoso, A., & Riquelme, J. C. (n.d.). Reconocimiento de patrones aplicado a la predicción de series temporales.

Anaya, J. J. (2015). Jhon Jairo Anaya Díaz.

Benítez, I., & Díez, J. L. (2005). Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos, (September 2017), 1–48.

Berkhin, P. (2002). Survey Of Clustering Data Mining Techniques. *Accrue Software, San Jose, CA*, 1–56.

Campos, G. (2009). Aplicación de técnicas de clustering para la mejora del aprendizaje.

COMPUTERWORD. (2017). Retrieved May 20, 2017, from www.computerworld.com

Cumming, G., Fidler, F., Vaux, D. L., Ioannidis, J. P. A., R Development Core Team, R., Hanahan, D., … Gersbach, C. a. (2011). Graph Kernels. *Bioinformatics (Oxford, England)*. https://doi.org/10.1038/sdata.2014.31

Daniel, P. L. C. y S. G. (2006). *Data Mining Soluciones con Enterprise Miner*. (A. R.- Ma, Ed.).

Díaz Arévalo, J. L., & Pérez García, R. (2002). Estado Del Arte En La Utilización De Tecnicas Avanzadas Para La Busqueda De Información No Trivial a Partir De Datos En Los Sistemas De Abastecimiento De Agua Potable. *Departamento de Ingeniería Hidráulica Y Medio Ambiente*.

Eduardo, J., & Medina, T. (2014). Facultad De Ciencias Físicas Y Matemáticas. Retrieved from http://repositorio.uchile.cl/bitstream/handle/2250/103717/Separacion-de-renio-por-electrodialisis-a-partir-de-soluciones-acidas-con-presencia.pdf?sequence=3

Elkan, C. (2010). *Predictive analytics and data mining*. https://doi.org/10.4018/978-1-4666-9562-7.ch019

Frsf, C. U. T. N., & Conicet, I. U.-. (n.d.-a). Minería de Datos en Base de Datos de Servicios de Salud. Retrieved from http://conaiisi.unsl.edu.ar/2013/132-505-1-DR.pdf

Frsf, C. U. T. N., & Conicet, I. U.-. (n.d.-b). Minería de Datos en Base de Datos de Servicios de Salud.

Galvis, B., & Rojas, N. Y. (n.d.). ciudad de Bogotá, *3*, 336–353.

Gutiérrez, J. (2016). Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas. *Pdfs.Semanticscholar.Org*, (1), 1–17.

Hernández-Flórez, L. J., Aristizabal-Duque, G., Quiroz, L., Medina, K., Rodríguez-Moreno, N., Sarmiento, R., & Osorio-García, S. D. (2013). Contaminación del aire y enfermedad respiratoria en menores de cinco años de Bogotá, 2007. *Revista de Salud Pública*, *15*(4), 503–516. Retrieved from http://www.revistas.unal.edu.co/index.php/revsaludpublica/article/view/38719/44829

Individuales, L. R., General, S., Social, S., Frecuentes, P., Individual, R., & Versi, R. (2000). Preguntas frecuentes, 1–17.

Instituto de Hidrologia, M. y estudios ambientales. (2010). *Calidad del Aire*.

López, C. P. (2017). *Minería de datos: técnicas y herramientas*. España.

Magdaleno, D., Miranda, Y., Fuentes, I. E., & García, M. M. (2015). Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents. https://doi.org/10.4114/ia.v18i55.1098

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Hierarchical clustering. *Introduction to Information Retrieval*, (c), 377–401. https://doi.org/10.1017/CBO9780511809071.017

Marchán E, Salcedo J, Aza T, Figuera L, Martínez de Pisón F, G. P. (2011). Reglas de asociación para determinar factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas. *Revista de Ciencia E Ingeniería*, (January), 55–60.

Molina, J., & García, J. (2008). Técnicas de Minería de Datos basadas en Aprendizaje Automático. *Técnicas de Análisis de Datos*, 96–266.

Molina, L. C. (2002). Data mining: torturando a los datos hasta que confiesen. *Fuoc*, 1–11. Retrieved from http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

Oms, & Ops. (2014). Modulo 4: Análisis de la calidad del dato. *Herramientas Para El Monitoreo de Coberturas de Intervenciones Integradas En Salud Pública.*

Organization, W. H. (2017). FluNet. Retrieved September 9, 2017, from http://www.who.int/influenza/gisrs_laboratory/flunet/en/f

Pérez, M. (2014). *MINERÍA DE DATOS A TRAVÉS DE EJEMPLOS*.

Rodríguez, A. O. (2013). Guía práctica para Arquitecturas de Datos Empresariales, 1–9.

Rodríguez, D., Cuadrado, J., & Sicilia, M. (2007a). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Ingeniería Del Software*, *3*(1), 6–22. Retrieved from http://en.scientificcommons.org/44226406

Rodríguez, D., Cuadrado, J., & Sicilia, M. (2007b). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Ingeniería Del Software*, *3*(1), 6–22.

Salud, I. N. de. (n.d.). SIVIGILA. Retrieved from http://www.ins.gov.co/lineas-de-accion/Subdireccion-Vigilancia/sivigila/Paginas/sivigila.aspx

Salud, M. de. (2017). SISPRO. Retrieved October 15, 2017, from http://www.sispro.gov.co/

Secretaría Distrital de Salud. (2015). Diagnóstico sectorial de salud, 62. Retrieved from http://www.saludcapital.gov.co/Empalme del Sector Salud 20122016/DIRECTIVA 09 DE 2015/1 DIAGNOSTICO SECTORIAL DE SALUD.pdf

Standard, O. G., & Group, T. O. (2011). *Open Group Standard The Open Group*.

Uiaf, D. (n.d.). DETECCIÓN Y PREVENCIÓN Y LA FINANCIACIÓN.

Viera, L. P. (n.d.). *Introducción a la Minería de Datos*.

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Sunx Toward quality data : An attribute-based approach, *13*, 349–372.

Wang, T., Chan-yeung, M., Lam, W. K., Wong, P. C., Lam, B., Ip, M. S., … Sc, D. (2003). A Cluster of Cases of Severe Acute Respiratory Syndrome in Hong Kong, 1977–1985.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). *Top*

10 algorithms in data mining. *Knowledge and Information Systems* (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2

14. Anexos

Anexo A Asignación de códigos para variable Enfermedad

Enfermedades Respiratorias	Código
J100 - INFLUENZA CON NEUMONIA, DEBIDA A VIRUS DE LA INFLUENZA IDENTIFICADO	**0**
J101 - INFLUENZA CON OTRAS MANIFESTACIONES RESPIRATORIAS, DEBIDA A VIRUS DE LA INFLUENZA IDENTIFICADO	**1**
J108 - INFLUENZA, CON OTRAS MANIFESTACIONES, DEBIDA A VIRUS DE LA INFLUENZA IDENTIFICADO	**2**
J110 - INFLUENZA CON NEUMONIA, VIRUS NO IDENTIFICADO	**3**
J111 - INFLUENZA CON OTRAS MANIFESTACIONES RESPIRATORIAS, VIRUS NO IDENTIFICADO	**4**
J118 - INFLUENZA CON OTRAS MANIFESTACIONES, VIRUS NO IDENTIFICADO	**5**
J120 - NEUMONIA DEBIDA A ADENOVIRUS	**6**
J121 - NEUMONIA DEBIDA A VIRUS SINCITIAL RESPIRATORIO	**7**
J122 - NEUMONIA DEBIDA A VIRUS PARAINFLUENZA	**8**
J128 - NEUMONIA DEBIDA A OTROS VIRUS	**9**
J129 - NEUMONIA VIRAL, NO ESPECIFICADA	**10**
J13X - NEUMONIA DEBIDA A STREPTOCOCCUS PNEUMONIAE	**11**
J14X - NEUMONIA DEBIDA A HAEMOPHILUS INFLUENZAE	**12**
J150 - NEUMONIA DEBIDA A KLEBSIELLA PNEUMONIAE	**13**
J151 - NEUMONIA DEBIDA A PSEUDOMONAS	**14**
J152 - NEUMONIA DEBIDA A ESTAFILOCOCOS	**15**
J153 - NEUMONIA DEBIDA A ESTREPTOCOCOS DEL GRUPO B	**16**
J154 - NEUMONIA DEBIDA A OTROS ESTREPTOCOCOS	**17**
J156 - NEUMONIA DEBIDA A OTRAS BACTERIAS AEROBICAS GRAMNEGATIVAS	**18**

J157 - NEUMONIA DEBIDA A MYCOPLASMA PNEUMONIAE	19
J158 - OTRAS NEUMONIAS BACTERIANAS	20
J159 - NEUMONIA BACTERIANA, NO ESPECIFICADA	21
J160 - NEUMONIA DEBIDA A CLAMIDIAS	22
J168 - NEUMONIA DEBIDA A OTROS MICROORGANISMOS INFECCIOSOS ESPECIFICADOS	23
J170 - NEUMONIA EN ENFERMEDADES BACTERIANAS CLASIFICADAS EN OTRA PARTE	24
J171 - NEUMONIA EN ENFERMEDADES VIRALES CLASIFICADAS EN OTRA PARTE	25
J178 - NEUMONIA EN OTRAS ENFERMEDADES CLASIFICADAS EN OTRA PARTE	26
J180 - BRONCONEUMONIA, NO ESPECIFICADA	27
J181 - NEUMONÍA LOBAR, NO ESPECIFICADA	28
J182 - NEUMONIA HIPOSTATICA, NO ESPECIFICADA	29
J188 - OTRAS NEUMONIAS, DE MICROORGANISMO NO ESPECIFICADO	30
J189 - NEUMONIA, NO ESPECIFICADA	31
J200 - BRONQUITIS AGUDA DEBIDA A MYCOPLASMA PNEUMONIAE	32
J201 - BRONQUITIS AGUDA DEBIDA A HAEMOPHILUS INFLUENZAE	33
J202 - BRONQUITIS AGUDA DEBIDA A ESTREPTOCOCOS	34
J203 - BRONQUITIS AGUDA DEBIDA A VIRUS COXSACKIE	35
J204 - BRONQUITIS AGUDA DEBIDA A VIRUS PARAINFLUENZA	36
J205 - BRONQUITIS AGUDA DEBIDA A VIRUS SINCITIAL RESPIRATORIO	37
J206 - BRONQUITIS AGUDA DEBIDA A RINOVIRUS	38
J207 - BRONQUITIS AGUDA DEBIDA A VIRUS ECHO	39
J208 - BRONQUITIS AGUDA DEBIDA A OTROS MICROORGANISMOS ESPECIFICADOS	40
J209 - BRONQUITIS AGUDA, NO ESPECIFICADA	41
J210 - BRONQUIOLITIS AGUDA DEBIDA A VIRUS SINCITIAL RESPIRATORIO	42
J218 - BRONQUIOLITIS AGUDA DEBIDA A OTROS MICROORGANISMOS ESPECIFICADOS	43
J219 - BRONQUIOLITIS AGUDA, NO ESPECIFICADA	44
J22X - INFECCION AGUDA NO ESPECIFICADA DE LAS VIAS RESPIRATORIAS INFERIORES	45
J300 - RINITIS VASOMOTORA	46

J301 - RINITIS ALERGICA DEBIDA AL POLEN	47
J302 - OTRA RINITIS ALERGICA ESTACIONAL	48
J303 - OTRAS RINITIS ALERGICAS	49
J304 - RINITIS ALERGICA, NO ESPECIFICADA	50
J310 - RINITIS CRONICA	51
J311 - RINOFARINGITIS CRONICA	52
J312 - FARINGITIS CRONICA	53
J320 - SINUSITIS MAXILAR CRONICA	54
J321 - SINUSITIS FRONTAL CRONICA	55
J322 - SINUSITIS ETMOIDAL CRONICA	56
J323 - SINUSITIS ESFENOIDAL CRONICA	57
J324 - PANSINUSITIS CRONICA	58
J328 - OTRAS SINUSITIS CRONICAS	59
J329 - SINUSITIS CRONICA, NO ESPECIFICADA	60
J330 - POLIPO DE LA CAVIDAD NASAL	61
J331 - DEGENERACION POLIPOIDE DE SENO PARANASAL	62
J338 - OTROS POLIPOS DE LOS SENOS PARANASALES	63
J339 - POLIPO NASAL, NO ESPECIFICADO	64
J340 - ABSCESO, FURUNCULO Y ANTRAX DE LA NARIZ	65
J341 - QUISTE Y MUCOCELE DE LA NARIZ Y DEL SENO PARANASAL	66
J342 - DESVIACION DEL TABIQUE NASAL	67
J343 - HIPERTROFIA DE LOS CORNETES NASALES	68
J348 - OTROS TRASTORNOS ESPECIFICADOS DE LA NARIZ Y DE LOS SENOS PARANASALES	69
J350 - AMIGDALITIS CRONICA	70
J351 - HIPERTROFIA DE LAS AMIGDALAS	71
J352 - HIPERTROFIA DE LAS ADENOIDES	72
J353 - HIPERTROFIA DE LAS AMIGDALAS CON HIPERTROFIA DE LAS ADENOIDES	73
J358 - OTRAS ENFERMEDADES CRONICAS DE LAS AMIGDALAS Y DE LAS ADENOIDES	74

J359 - ENFERMEDAD CRONICAS DE LAS AMIGDALAS Y DE LAS ADENOIDES, NO ESPECIFICADA	75
J36X - ABSCESO PERIAMIGDALINO	76
J370 - LARINGITIS CRONICA	77
J371 - LARINGOTRAQUEITIS CRONICA	78
J380 - PARALISIS DE LAS CUERDAS VOCALES Y DE LA LARINGE	79
J381 - POLIPO DE LAS CUERDAS VOCALES Y DE LA LARINGE	80
J382 - NODULOS DE LAS CUERDAS VOCALES	81
J383 - OTRAS ENFERMEDADES DE LAS CUERDAS VOCALES	82
J384 - EDEMA DE LARINGE	83
J385 - ESPASMO LARINGEO	84
J386 - ESTENOSIS LARINGEA	85
J387 - OTRAS ENFERMEDADES DE LA LARINGE	86
J390 - ABSCESO RETROFARINGEO Y PARAFARINGEO	87
J391 - OTROS ABSCESOS DE LA FARINGE	88
J392 - OTRAS ENFERMEDADES DE LA FARINGE	89
J393 - REACCION DE HIPERSENSIBILIDAD DE LAS VIAS RESPIRATORIAS SUPERIORES, SITIO NO ESPECIFICADO	90
J398 - OTRAS ENFERMEDADES ESPECIFICADAS DE LAS VIAS RESPIRATORIAS SUPERIORES	91
J399 - ENFERMEDAD DE LAS VIAS RESPIRATORIAS SUPERIORES, NO ESPECIFICADA	92
J40X - BRONQUITIS, NO ESPECIFICADA COMO AGUDA O CRONICA	93
J410 - BRONQUITIS CRONICA SIMPLE	94
J411 - BRONQUITIS CRONICA MUCOPURULENTA	95
J418 - BRONQUITIS CRONICA MIXTA SIMPLE Y MUCOPURULENTA	96
J42X - BRONQUITIS CRONICA NO ESPECIFICADA	97
J430 - SINDROME DE MACLEOD	98
J431 - ENFISEMA PANLOBULAR	99
J432 - ENFISEMA CENTROLOBULAR	100
J438 - OTROS TIPOS DE ENFISEMA	101

J439 - ENFISEMA, NO ESPECIFICADO	102
J440 - ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON INFECCION AGUDA DE LAS VIAS RESPIRATORIAS INFERIORES	103
J441 - ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON EXACERBACION AGUDA, NO ESPECIFICADA	104
J448 - OTRAS ENFERMEDADES PULMONARES OBSTRUCTIVAS CRONICAS ESPECIFICADAS	105
J449 - ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA, NO ESPECIFICADA	106
J450 - ASMA PREDOMINANTEMENTE ALERGICA	107
J451 - ASMA NO ALERGICA	108
J458 - ASMA MIXTA	109
J459 - ASMA, NO ESPECIFICADA	110
J46X - ESTADO ASMATICO	111
J47X - BRONQUIECTASIA	112
J60X - NEUMOCONIOSIS DE LOS MINEROS DEL CARBON	113
J628 - NEUMOCONIOSIS DEBIDA A OTROS POLVOS QUE CONTIENEN SILICE	114
J634 - SIDEROSIS	115
J64X - NEUMOCONIOSIS, NO ESPECIFICADA	116
J660 - BISINOSIS	117
J668 - ENFERMEDAD DE LAS VIAS AEREAS DEBIDAS A OTROS POLVOS ORGANICOS ESPECIFICOS	118
J670 - PULMON DEL GRANJERO	119
J677 - NEUMONITIS DE LA VENTILACION DEBIDA AL ACONDICIONADOR Y HUMIDIFICADOR DEL AIRE	120
J678 - NEUMONITIS DEBIDA A HIPERSENSIBILIDAD A OTROS POLVOS ORGANICOS	121
J679 - NEUMONITIS DEBIDA A HIPERSENSIBILIDAD A POLVO ORGANICO NO ESPECIFICADO	122
J680 - BRONQUITIS Y NEUMONITIS DEBIDAS A INHALACION DE GASES, HUMOS, VAPORES Y SUSTANCIAS QUIMICAS	123
J682 - INFLAMACION RESPIRATORIA SUPERIOR DEBIDA A INHALACION DE GASES, HUMOS, VAPORES Y SUSTANCIAS QUIMICAS, NO CLASIFICADAS EN OTRA PARTE	124
J683 - OTRAS AFECCIONES RESPIRATORIAS AGUDAS Y SUBAGUDAS DEBIDAS A INHALACION DE GASES, HUMOS, VAPORES Y SUSTANCIAS QUÍMICAS	125

	J988 - OTROS TRASTORNOS RESPIRATORIOS ESPECIFICADOS	**179**
	J989 - TRASTORNO RESPIRATORIO, NO ESPECIFICADO	**180**
	J990 - ENFERMEDAD PULMONAR REUMATOIDE (M05.1†)	**181**
	J991 - TRASTORNOS RESPIRATORIOS EN OTROS TRASTORNOS DIFUSOS DEL TEJIDO CONJUNTIVO	**182**
	J998 - TRASTORNOS RESPIRATORIOS EN OTRAS ENFERMEDADES CLASIFICADAS EN OTRA PARTE	**183**
	J920 - PAQUIPLEURITIS CON ASBESTOSIS	**184**

Fuente: Los autores

Anexo B Asignación de códigos para variable Género

Genero	Código
Masculino	0
Femenino	1

Fuente: Los autores