Generación De Patrones En Tratamientos Farmacológicos Utilizando Modelos Descriptivos Mediante Técnicas De Clustering

Yerfeison Fernando Bolaños Hoyos [1 *], Arley Perafán Buitrón [2]
Director: Mg. Julián Eduardo Hoyos – Ing. Daniela Iboth Gutiérrez
Facultad de Ingenierías, Ingeniería de Sistemas, Fundación Universitaria de Popayán.
** fernando.bolanios@estudiante.fup.edu.co- arley.perafan@estudiante.fup.edu.co*
Grupo de investigación: Logiciel

Resumen— Este artículo analiza la segmentación en productos farmacológicos a partir de las prescripciones médicas, en un instituto prestador de salud, en la región del Cauca; se usaron datos diarios provenientes del registro de las consultas médicas de los pacientes cubriendo un periodo de seis meses, en el cual se clasificaron 152188 observaciones o registros con 17 variables. Se implemento un algoritmo K-Means que permitió establecer tres clusters, con el fin de descubrir similitudes y algunas características más representativas de las variables. En los resultados obtenidos muestran que el primer grupo se caracteriza por la administración de dosis altas y un menor número de días de tratamiento, estas características indican que se utilizan agentes farmacológicos efectivos, como antiparasitarios o antibacterianos, que son adecuados para el tratamiento de poblaciones jóvenes, incluyendo a niños y jóvenes adultos, que tienen un menor riesgo de complicaciones. El segundo grupo tiene dosis bajas y días de tratamiento más altos, se sugiere el uso de medicamentos antihipertensivos, terapias con insulina tradicional o fármacos orales para tratar enfermedades agudas o graves que requieren un tratamiento intensivo, especialmente en las personas mayores y con un mayor riesgo de complicación. El tercer grupo se sitúa en una posición intermedia en cuanto a las anteriores variables, lo que sugiere que podrían requerir el uso de fármacos pertenecientes tanto al primer como al segundo grupo previamente mencionados. Este grupo se compone principalmente de personas jóvenes y adultos mayores, que presentan un riesgo moderado de complicaciones. Finalmente se contribuye al desarrollo de prevención y manejo de patologías crónicas, para mejorar la prestación del servicio de manera que se pueda implementar, medidas preventivas y de control por parte del instituto para el mejoramiento de la calidad de vida de los usuarios (pacientes), satisfaciendo las necesidades de forma más acertada y eficiente.

Palabras Claves: Algoritmo, Clustering, Modelos, Segmentación

***Abstract*-- This article analyzes the segmentation of pharmacological products based on medical prescriptions, in a health provider institute, in the Cauca region; Daily data from the records of the patients' medical consultations were used, covering a period of six months, in which 152188 observations or records with 17 variables were classified. A K-Means algorithm was implemented that allowed establishing three clusters, in order to discover similarities and some more representative characteristics of the variables. The results obtained show that the first group is characterized by the administration of high doses and a lower number of days of treatment, these characteristics indicate that effective pharmacological agents are used, such as antiparasitics or antibacterials, which are suitable for the treatment of young populations, including children and young adults, who have a lower risk of complications. The second group has low doses and longer treatment days, the use of antihypertensive drugs, traditional insulin therapies or oral drugs is suggested to treat acute or severe diseases that require intensive treatment, especially in the elderly and at higher risk. of complication. The third group is located in an intermediate position in terms of the above variables, which suggests that they might require the use of drugs belonging to both the first and second groups previously mentioned. This group is made up mainly of young people and older adults, who present a moderate risk of complications. Finally, it contributes to the development of prevention and management of chronic pathologies, to improve the provision of the service so that preventive and control measures can be implemented by the institute to improve the quality of life of users (patients), satisfying the needs in a more accurate and efficient way.**

I. INTRODUCCIÓN

Los tratamientos Farmacológicos, hacen parte del conjunto de estrategias que usan los profesionales de la salud, en la toma de decisiones e incluye las recomendaciones farmacológicas y no farmacológicas, tales como: estilos de vida saludables, hábitos de higiene, recomendaciones nutricionales y medicamentosas, estas últimas con su correcta dosificación [1, 2].

Las estrategias están orientadas al manejo integral de las patologías[1], con el objetivo de impactar la historia natural de las enfermedades[2], modificarlas, disminuir sus síntomas y en algunos casos con el propósito curativo, de manera que influyan positivamente en la calidad de vida de los pacientes[2,3].

La importancia de analizar y clasificar los tratamientos farmacológicos que se les administra a individuos y poblaciones, es de suma importancia en el desarrollo del análisis de las interacciones y en la intervención de los factores de riesgo, de manera que las diferentes patologias se puedan intervenir [2]. Es por esta razón que el presente estudio permite diseñar estrategias mediante la metodología de clustering, para detectar riesgos y las probables complicaciones, y por lo tanto intervenirlos. De manera que pueda impactar en la calidad de atención y en consecuencia en la seguridad del paciente [2].

En la búsqueda de la literatura no se encuentran estudios locales que analicen estos factores, sin embargo, hay estudios similares

[1]*Patología:* síntoma o grupo de síntomas de una enfermedad [2].

[2]*Historia natural de la enfermedad:* Evolución de una enfermedad [3].

a nivel nacional e internacional, en los cuales las entidades prestadoras de salud, buscan herramientas que contribuyan a diseñar estrategias innovadoras y precisas para intervenir a sus usuarios a partir de la caracterización de la población y de sus necesidades particulares y generales, de manera que se pueda brindar una atención personalizada y de mejor calidad [4,5,6]. Entre las diferentes alternativas que ofrece el mercado, la mayoría han optado por la segmentación y la aplicación de modelos como, árboles de decisión, los cuales facilitan la manipulación de los datos, buscando segmentos óptimos que proporcionen las herramientas de análisis a fin de obtener la clasificación de los consumidores [7,8,9].

La caracterización de clientes a través de las metodologías anteriormente mencionadas, han permitido hacer estudios sobre poblaciones e individuos específicos que padecen enfermedades, la mayoría de tipo crónico, tales como la Hipertensión Arterial [3] [6]. Cuyo propósito es identificar factores que determinan e influyen en que una persona desarrolle esta enfermedad a lo largo de su vida [6].

Los modelos descriptivos permiten conocer los posibles comportamientos de las diferentes patologías en los individuos o en poblaciones específicas [10]. Por lo tanto, son una herramienta importante para poder intervenir los factores que influyan en la historia natural de la enfermedad [3]. Por consiguiente, esta propuesta metodológica, es una de las mejores alternativas que se le puede ofrecer a las empresas prestadoras en salud, ya que se relaciona con la segmentación y permite la generación de patrones aplicables a los modelos descriptivos.

II. MATERIALES Y MÉTODOS

Antes de trabajar con alguna técnica, el paso más importante es la obtención de los datos. Estos pueden ser de tipo nominal y/o categóricos, los cuales pueden ser transformados para adaptarse a la técnica elegida si así se requiere.

1. Obtención De Datos

Se parte de este proceso, el cual implica la identificación y extracción de datos de una o más fuentes. Posteriormente, se selecciona un subconjunto de ellos, en caso de ser necesario, siempre se debe enfocar en aquellos que resulten más relevantes y suficientes para lograr el objetivo deseado [7]. Con estos procesos, se puede garantizar una mejor toma de decisiones y un desarrollo más eficiente [11].

Para el desarrollo del presente trabajo se cuenta con una colección de 152188 registros y 17 variables, la cuales se aprecian en Tabla. I.

TABLA I
CLASIFICACIÓN DE VARIABLES

Número de variable	**Tipo De Variables**	
	Cualitativas	**Descripción**
1	Tipo_id_paciente	T. identificación
2	Sexo_id	Genero (f, m)
3	descripcion	Ubicación
4	descripcion.1	Nombre medicamento
5	tipo_diagnostico_id	Código de diagnostico
6	diagnostico_nombre	Nombre diagnostico
Cuantitativas		
7	codigo_producto	Identificador de medicamento
8	fecha_nacimiento	Fecha nacimiento
9	peso	Peso
10	talla	Talla
11	taalta	Tensión arterial alta
12	tabaja	Tensión arterial baja
13	per_abdominal	Perímetro abdominal
14	Fecha	Fecha de atención medica
15	dosis	Dosis del medicamento
16	días_tratamiento	Días de tratamiento del medicamento
17	cantidad	Cantidad en medicamento

2. Metodología

La metodología SEMMA permite seleccionar, explorar, modificar, modelar y evaluar altos volúmenes de información para la generación de patrones de modelos descriptivos, de manera que se pueda identificar los posibles comportamientos de las patologías[12].

Para la aplicación de la metodología SEMMA, en el desarrollo de modelos descriptivos mediante la técnica de clustering, se parte del historial médico de los pacientes, teniendo en cuenta, variables demográficas como la edad, antropométricas[4] como talla y peso, de intervención como tiempos de tratamiento [1]. Este tipo de modelo facilita la descripción de patrones de comportamiento de factores que influyen positivamente o negativamente en la salud de los pacientes [10].

El clustering o segmentación de acuerdo al objetivo que se busque, permite separar en pequeños grupos el alto volumen de datos que se ha recolectado según el propósito deseado [8,9,13]. Para el desarrollo de la propuesta se emplea una segmentación no dirigida, en la cual no se tiene variables implícitas en otra [7,14].

El éxito para alcanzar el objetivo de la segmentación y el diseño de un modelo de descripción, no solo depende de la selección de técnicas a utilizar, sino también requiere tener presente la

[3]Hipertensión Arterial: medida de fuerza que ejerce la sangre en cada artería del cuerpo, dicha fuerza puede subir o bajar [6].

[4] Antropométrica: referente a medidas del cuerpo humano [1].

obtención de datos, el modelado y el uso de resultados. Las cuales se explican paso a paso en Fig. 1. Estos pueden determinar el logro o el fracaso de un buen modelo [7].

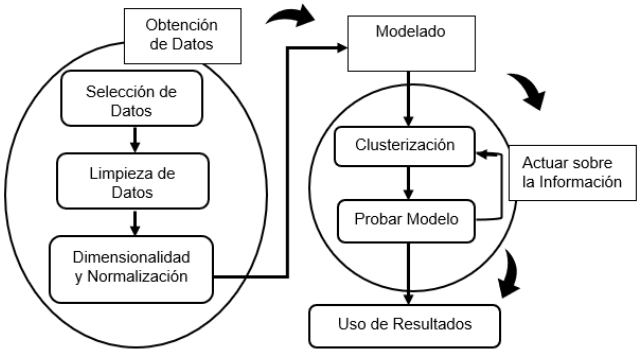

Fig. 1. Metodología de SEMMA

3. Fases

Para llevar a cabo cada fase del análisis, se utilizó la herramienta RStudio, que ofrece la posibilidad de realizar limpieza de datos y reducción de variables utilizando diversas funciones y métodos. Además, esta herramienta proporciona gráficos y visualizaciones para facilitar la comprensión de los datos y el análisis [15].

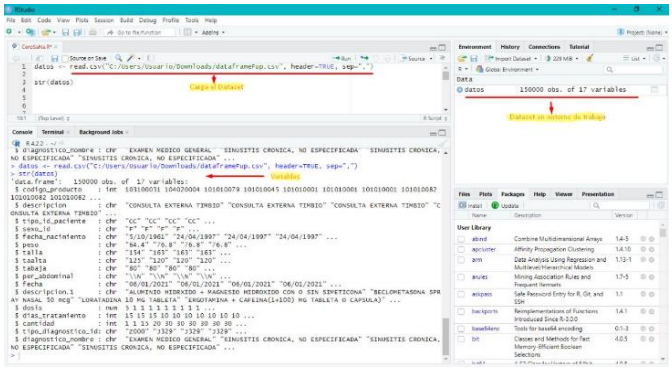

Fig. 2. Entorno inicial de trabajo -Rstudio

Para dar inicio a las fases se procede a cargar la fuente de datos desde su formato original (.csv). Como se observa en la anterior Fig. 2.

i. Obtención de datos

Como se comentó anteriormente se obtuvo un dataset con 150000 registros y 17 variables. A estos datos se procede a realizar un preprocesamiento, como se indica en los siguientes ítems:

a) Limpieza De Datos

Se identifica los datos anómalos, este proceso valida los datos verificando que todos cuenten con el mismo formato, que los datos sean acordes a la característica de la variable o campo, tanto para las cualitativas, como para cuantitativas, verificando datos faltantes [7]. Es decir, que no haya campos vacíos o registros fuera de rangos. Tales datos anómalos se remplazan por el valor de la moda en registros descriptivos y la media para datos numéricos[11]. Este proceso busca mejorar la calidad y fiabilidad de los datos utilizados en el análisis estadístico.

b) Dimensionalidad

La dimensionalidad, es el tamaño o cantidad de variables y registros que contiene el dataset [13]. Con el fin de evitar redundancia en esos datos, es importante verificar el tipo de variables que se ha obtenido y que cada una de ellas contenga información no equivalente, en caso contrario se procede a la reducción de la dimensionalidad [11]. Esta técnica permite eliminar variables que contienen información de igual significado, sin que esto afecte los futuros análisis [13].

c) Normalización

Esta consiste en encontrar los valores medios, evitando que haya datos demasiado distantes, estos afectan la posición de los centroides, por lo tanto, la calidad de los clusters resultantes. Con la normalización se busca un mayor equilibrio en la varianza, ya que esta puede influir erráticamente en la creación del modelo descriptivo [13].

Para proceder a la normalización del dataset, se debe convertir las variables categóricas a cuantitativas, como se indica en la Fig. 3. Esto facilita realizar gráficas y generar rangos de dichos componentes.

```
datos$descripcion <- as.numeric(factor(datos$descripcion))
datos$tipo_id_paciente <- as.numeric(factor(datos$tipo_id_paciente))
datos$sexo_id <- as.numeric(factor(datos$sexo_id))
datos$tipo_diagnostico_id <- as.numeric(factor(datos$tipo_diagnostico_id))

#visualiza las primeras 6 filas
 head(datos)
```

```
dosis dias_tratamiento cantidad peso talla taalta tabaja per_abdominal
    5              15        1 64.4   154    125     80           96
    1              15        1 76.8   163    120     80           96
    1              15       15 76.8   163    120     80           96
    1              10       20 76.8   163    120     80           96
    1              10       30 58.0   145    124     74           91
    1              10       30 58.0   145    124     74           91

  edad  codigo_producto descripcion tipo_id_paciente sexo_id tipo_diagnostico_id
59.25479       103100031          2                1       1              1657
23.70411       104020004          2                1       1               718
23.70411       101010079          2                1       1               718
23.70411       101010045          2                1       1               718
74.57534       101010001          1                1       1               991
74.57534       101010001          1                1       1              1356
```

Fig. 3 Conversión de variables categóricas a cuantitativas.

d) Implementación del PCA

El PCA (Análisis de Componentes Principales) es una técnica de análisis estadístico utilizada para identificar patrones en datos numéricos. Permite hacer un análisis de las variables en base a la matriz de covarianza, mostrando su mejor tendencia de variabilidad. Es decir que reduce las observaciones o registros de cada variable, con lo cual se puede tener una mejor apreciación de la mayor o menor representación de ellas [16].

El PCA se utiliza comúnmente en el análisis de datos para reducir la dimensión del conjunto de datos. También puede ayudar a identificar las variables más importantes y la relación entre estas, identificando la variabilidad en los datos y del aporte que puede tener cada una de estas variables para los posibles análisis [13].

Con las variables cuantitativas y haciendo uso de la función prcomp(), se procede a realizar un PCA con el dataset (datosPCA), como se muestra en Fig. 4. Esto genera un resumen estadístico de los componentes principales, lo cual nos proporciona la contribución de cada componente.

```
pca1<- prcomp(datosPCA[,1:9])
pca1
>  pca1
Standard deviations (1, ..., p=9):
[1] 49.666070 21.004659 18.655090 16.096286 14.588756 10.674227
    6.772638  6.103743  5.278926

Rotation (n x k) = (9 x 9):
                         PC1         PC2         PC3         PC4         PC5
dosis              0.002589263 -0.02270114  0.009068986 -0.005620852 -0.01971599
dias_tratamiento -0.539702131  0.11367180 -0.682991364  0.373440307 -0.29597028
cantidad         -0.830358796 -0.20862034  0.464779441 -0.194053929  0.11430149
peso             -0.028995133  0.49302306  0.386762162  0.522418710 -0.02761269
talla            -0.029274639  0.31803219  0.194320789  0.326749751  0.19357664
taalta           -0.032640466  0.43064497  0.081518132 -0.465163569 -0.60650714
tabaja            0.004757739  0.21075130  0.127565120 -0.191213446 -0.33086090
per_abdominal    -0.003536113  0.17478280  0.156974621  0.170040179 -0.09241565
edad             -0.128144115  0.58172545 -0.287255959 -0.401638790  0.61225254

                         PC6         PC7         PC8         PC9
dosis            -0.031913926 -0.018561282  0.053050774 -0.997395603
dias_tratamiento  0.017812306  0.036087392  0.024631878 -0.006383830
cantidad         -0.008560991 -0.009568929 -0.006489044  0.005759740
peso             -0.220152353  0.151708489 -0.510138468 -0.033091148
talla             0.710557520 -0.192425707  0.417149904 -0.008182565
taalta            0.134455940 -0.427070693 -0.148785221  0.001197055
tabaja            0.075825376  0.822250570  0.333759271  0.004017685
per_abdominal    -0.632366128 -0.274510331  0.654737216  0.058476133
edad             -0.146687003  0.066160400 -0.008268240 -0.023001568
>
```

Fig. 4 PCA del Dataset.

ii. Modelado

El modelado es la aplicación de técnicas, para el proceso de obtener un conjunto de información, adecuada y estandarizada. De manera que estos segmentos de datos sean relevantes [14,7]. Para el modelado se requiere el desarrollo de los siguientes pasos:

a) Clusterización

Esta es una técnica de aprendizaje no supervisado que consiste en agrupar un conjunto de objetos en subconjuntos (grupos o clusters) homogéneos en función de sus similitudes o afinidades. En otras palabras, es la división de un conjunto de datos en grupos o clusters, de manera que los objetos dentro de cada grupo sean más similares entre sí. Es muy utilizado en campos como: la minería de datos, la segmentación de mercados entre otros [13,17].

Para la ejecución de la técnica de Clusterización se requiere de hacer algunas métricas (medidas de distancias) o procedimientos para validar el tamaño más apropiado de clusters a crear.

- Creación de un nuevo PCA

A diferencia del caso anterior donde se hizo un PCA únicamente con variables cuantitativas y se excluyó las de tipo categórico, para el modelo general se incluyen todas las variables del dataset y sus registros. Como ya se había mencionado anteriormente, el objetivo de realizar un PCA es reducir dimensiones y observaciones [13]. Permitiendo seleccionar las mejores componentes y obtener representaciones de menor tamaño, pero manteniendo un alto porcentaje de información necesaria para la generación de modelos o para la realización de análisis [16].

Ejecutado los pasos para la realización del PCA, como se indicó en la figura 3. Se han seleccionado 13 PC (Componentes Principales) o variables con 20000 observaciones. Con estos datos se continúa el proceso del modelo.

- Cálculo de la Matriz de Distancias

La matriz de distancias, es una matriz cuadrada en la que la entrada (i, j) es la distancia entre los elementos i y j. La diagonal principal de la matriz suele estar compuesta de ceros, ya que la distancia de un elemento a sí mismo es cero. En el análisis de datos, una matriz de distancias se utiliza a menudo para medir la similitud entre pares de objetos [17,18].

En conclusión, la matriz de distancias es una herramienta útil en el análisis de datos para medir la similitud entre los objetos y se puede utilizar como entrada para diversas técnicas de análisis de datos.

El cálculo de la matriz de distancias se realiza empleando métodos de vecindad. El criterio de comparación principal utilizado, es la distancia [19]. Por lo tanto, es importante mencionar las diferentes formas con las que se puede calcular, como, la distancia euclidiana clásica, los métodos de Manhattan, de Chebychev y del Coseno [17,19].

Los métodos de vecindad, dependen de la distancia, permiten establecer que tan cerca o lejos están los elementos y la similitud entre estos [17,18]. Son útiles en el análisis, para la exploración de patrones, estructuras en los datos y se aplican en diferentes contextos, como el clustering, el análisis de series temporales, el procesamiento de imágenes y la reducción de dimensionalidad [13,19].

Para cálculo de distancias, existe la distancia delta que se utiliza para calcular la longitud entre atributos nominales, los cuales son muy comunes en la minería de datos. Cabe destacar que la distancia euclidiana es una de las medidas más utilizadas en el espacio de vectores y cumple con las propiedades de una métrica [19].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Ec. 1 Distancia Euclidiana [19].

Tomando el método euclidiano se procede a calcular la matriz de distancia, con esta técnica se busca elementos en función de su proximidad con el propósito de agruparlos. Se procede a emplear una variable previamente declarada, la cual recoge y almacena el resultado del mencionado cálculo de la matriz. Dicha variable, denominada "mdist". Como indica la Fig.5.

```
mdist <- dist(x = datos20, method = "euclidean")
round(mdist, digits = 3)
```

data metodo

Fig. 5 Calculo de Matriz de Distancia

En Fig. 5, se observa la ejecución de la función “dist” de Rstudio, donde permite calcular la matriz de distancia aplicando el método euclidiano.

- Agrupación de Clusters

La metodología utilizada para llevar a cabo la agrupación de los datos en clústeres, consiste en la búsqueda de los centros de los mismos y en la posterior agrupación de las observaciones de acuerdo a su cercanía a dichos centros [8]. En este sentido, es fundamental resaltar que hay una gran variedad de algoritmos disponibles para la implementación de este método, siendo el más comúnmente utilizado el conocido como "k-means"[13, 20]. Dicho algoritmo ha demostrado ser sumamente efectivo en la agrupación de datos y resulta de gran utilidad para el caso de estudio en cuestión [13]. Siendo éste el método elegido para llevar a cabo el análisis correspondiente.

Entre los métodos de agrupación Clustering, se utilizó, el método de, vecino más cercano y vecino más lejano. Son dos técnicas comunes utilizadas para este proceso [17].
El método de vecino más cercano, que también se le conoce como "single-linkage", se inicia considerando que cada observación es un cluster por sí misma. Luego, se unen los dos clusters más cercanos entre sí, repitiendo este proceso hasta que todas las observaciones se agrupan en un solo cluster [17]. La distancia hallada entre dos clusters, se define como, distancia más corta entre dos observaciones, una menor distancia significa que tienen mayor probabilidad de pertenecer a la misma clase. [17,18].

El método de vecino más lejano, también se le conoce como "complete-linkage", se inicia considerando que cada observación es un cluster por sí misma. Luego, se unen los dos clusters más lejanos entre sí, repitiendo este proceso hasta que todas las observaciones se agrupan en un solo cluster [17]. La distancia hallada entre dos clusters se define como, distancia más larga entre dos observaciones, Esto implica que, a mayor distancia, mayor probabilidad de pertenecer a la misma clase [17,18].

- Selección del Número óptimo de Clusters

En la búsqueda de la literatura se encontró, que no se tiene el tamaño definido como referencia para indicar cuantos clusters debe contener el nuevo clustering resultante [21]. Sin embargo, existe una serie de métodos que facilitan el cálculo del número más apropiado, entre los métodos utilizados para este proceso se tiene, wss y silhouette [15,22].

El método de "wss" (Within-Cluster Sum of Squares) es una técnica comúnmente utilizada en análisis de clustering para estimar un número óptimo de clusters de un conjunto de datos [15]. El método implica calcular la suma de cuadrados de las distancias entre cada elemento o punto y su centroide en un número determinado de clusters, y luego graficar la curva de "wss" óptimo de clusters, se busca un "codo" en la curva, que indica un punto en donde agregar clusters adicionales ya no proporciona una mejora significativa en la varianza explicada [15,22]. Como se indica en la Fig. 6.

```
fviz_nbclust(datoss,kmeans,method="wss")
```

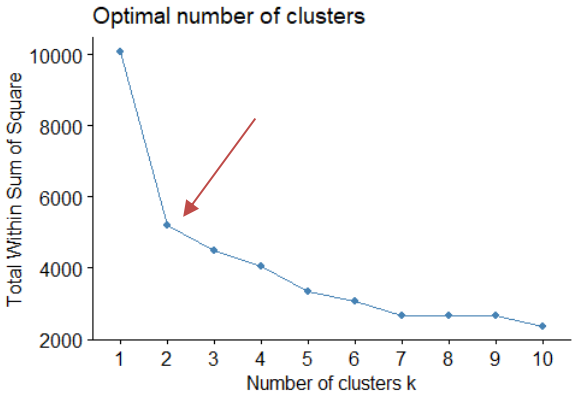

Fig. 6 Cluster óptimos

El método silhouette, es otra técnica de evaluación de clusters, utilizada en análisis de datos para evaluar la calidad de agrupación realizada por algoritmos de clustering como, k-means [15,22]. Este método mide la similitud entre un punto de datos respecto otros puntos en un mismo cluster, haciendo comparación con puntos de datos de clusters vecinos, el valor de coeficiente Silhouette, varía de -1 a 1. Si este valor es cercano a 1 indica que un punto de datos está correctamente ajustado en su propio clúster. Mientras que los valores cercanos a -1 indican que un punto de datos podría haberse asignado a un clúster diferente [22]. Luego genera la gráfica que indica el brazo de quiebre. Este punto representa el número de cluster más indicado [15]. Como se indica en Fig. 7.

```
fviz_nbclust(datoss,kmeans,method="silhouette")
```

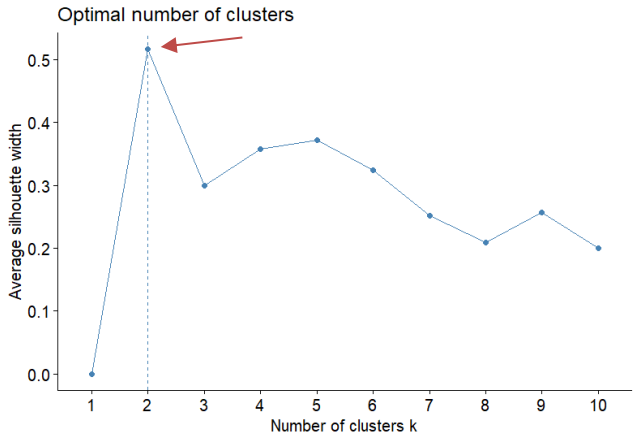

Fig. 7 Cluster óptimos en método silhouette

b) Construcción del Modelo

Para la construcción del modelo, se hace uso del método del centroide, el cual es un algoritmo de clustering o agrupamiento, utilizado para encontrar grupos o clusters de objetos o instancias similares en un conjunto de datos [13,18]. Funciona en tres pasos: Paso uno, selecciona aleatoriamente K centroides iniciales en el espacio de características.
Paso dos: asigna cada instancia del conjunto de datos al centroide más cercano.
Paso tres: calcula los nuevos centroides para cada grupo a partir de la media aritmética de las instancias del grupo.
Repite los pasos dos y tres hasta que no haya más cambios en las asignaciones de instancias [13,18]. El resultado final es un conjunto de K clusters, cada uno representado por su centroide [15].

c) Ejecución Del Modelo

Aplicando el K-means, que como ya se había comentado, es un un conjunto de datos en k grupos distintos [13,20]. La función kmeans(), toma dos argumentos principales: los datos que deseas agrupar y el número de grupos (k) que desear crear. Como se observa en Fig. 8.

```
# Clasifica los elementos usando el algoritmo K-means con k=3
km <- kmeans(datosN, centers = 3, nstart = 10, iter.max=1000, algorithm="Forgy")
```
Fig. 8 Modelo K-means-RStudio

III. RESULTADOS

En esta sección, se presentan los resultados obtenidos al aplicar diversas técnicas de análisis de datos, tales como: Limpieza, Normalización de datos, así como la Reducción de la dimensionalidad y la implementación de Modelos. El objetivo de este análisis es procesar, clasificar y agrupar conjuntos de datos relacionados con productos farmacológicos y sus variables, con el fin de obtener una comprensión más profunda y significativa de los mismos.

A) Resultado Del Preprocesamiento

En los siguientes ítems, se presentan dichos resultados.

a) Resultados Limpieza De Datos

Se obtuvo un dataset de 15000 observaciones o registros con 17 variables, el cual se realizó limpieza de datos en las variables cualitativas y cuantitativas que contenían datos anómalos (campos sin registros o vacíos y datos por fuera de rango), ver Fig. 9 y Fig.10.

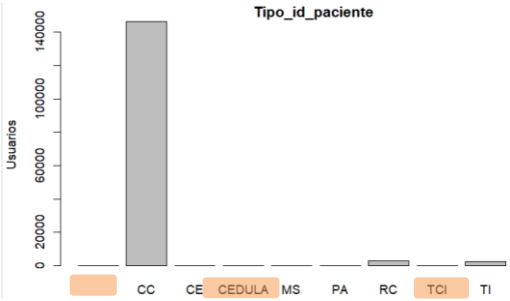

Fig. 9 Datos anómalos variable cualitativas

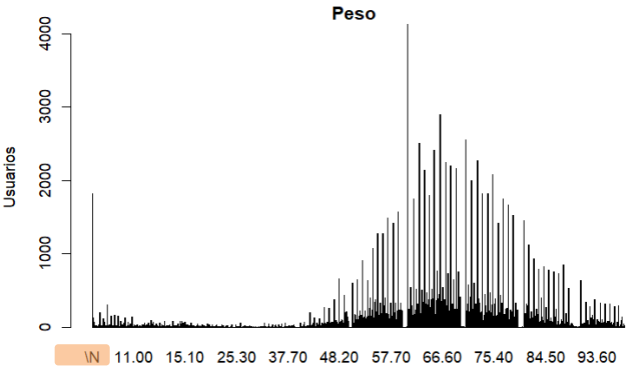

Fig.10 Datos anómalos variable cuantitativa

En Fig. 9, se observan, registros vacíos, “CEDULA” y “TCI”. En Fig. 10, se observan datos anómalos como, \N. Es importante en cuenta que también es un dato anómalo, cuando no tiene el mismo formato o el mismo tipo de los demás registros. Para realizar la corrección en este tipo de variables se procede en dos pasos. Paso uno: reemplazar los valores anómalos por NA, Ver Fig. 11.

```
# reemplazando valores anómalos por NA
datos$peso[datos$peso == "\\N"] <- NA
# muestra cuantos valores NA  tenemos en la columna
sum(is.na(datos$peso))
> sum(is.na(datos$peso))
[1] 1826
```
Fig. 11 Valores reemplazados por NA

Como se puede ver en la anterior Fig. 11, en esta variable se ha encontrado 1826 valores anómalos, los cuales han sido reemplazados y ahora toman el valor de NA. Seguidamente con los datos válidos de esta columna se procede a calcular la media, para reemplazar esos valores. Como indica la Fig. 12.

```
# pasamos los datos de caracteres a numericos para
# realizar procesos matemáticos
datopeso<-as.numeric(datos$peso)
# se omite valores NA
colpeso<-na.omit(datopeso)
# calcular la media
media = round(mean(colpeso))
# Reemplazo NA por la media
datos$peso[is.na(datos$peso)] = media
```

Fig. 12 Reemplazo de NA por la media

Para variables cualitativas, al igual que para las cuantitativas primero se remplazan los valores anómalos por NA. Luego se calcula la moda, para cambiar nuevamente el valor NA por dicha moda. Como se indica en Fig. 13.

```
sum(is.na(datos$tipo_id_paciente))
 > sum(is.na(datos$tipo_id_paciente))
 [1] 11

#Funcion para calcular la moda
moda <- getmode <- function(v){
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
# Omitir los datos faltantes NA
datoslimpios<- na.omit(datos)
# Cáculo de la moda para reemplazar el valor NA.
modacolm = moda(datoslimpios$tipo_id_paciente)
```

modacolm	"CC"

```
# reemplazar datos NA.
datos$tipo_id_paciente[is.na(datos$tipo_id_paciente)]<- modacolm
```

Fig. 13 Reemplazo de NA por la moda

Como se puede ver en la anterior Fig. 13, en esta variable se remplazaron 11 valores anómalos. Seguidamente se continuó con el segundo paso, para hallar la moda de los valores válidos de esa columna y reemplazar los NA.

En resumen, la limpieza de datos es esencial para asegurar la calidad, precisión y consistencia de los datos utilizados en análisis posteriores, mejorar la eficiencia del proceso. Cumplir con los requisitos normativos y facilitar la visualización y presentación de resultados. Por lo tanto, es una etapa crucial en cualquier proyecto de análisis de dato

b) Resultados De Dimensionalidad

Como se observa en Fig. 14, las variables tipo chr son variables categóricas y las de tipo int y num son variables cuantitativas con las que se pueden realizar operaciones matemáticas. De estas variables se toma la fecha_nacimiento y fecha, estas se reducen en la variable edad, esto con el propósito de seleccionar la información no redundante.

```
fecha_nac <- format(as.Date(datos$fecha_nacimiento),"%Y/%m/%d")
fecha_nu <- format(as.Date(datos$fecha),"%Y/%m/%d")
# convertir las variables en tipo Date
fn1= as.Date(fecha_nac)
f1 = as.Date(fecha_nu)
# Calcular edad
edad <- age_calc(fn1,f1,units = "years")
# Agregar edad al dataframe
datos$edad <- edad
# Eliminamos las Variables Fecha_nacimiento y Fecha
datos <- select(datos,-fecha_nacimiento, -fecha)

# Verificando que se modificó el dataframe
str(datos)
```

```
> str(datos)
'data.frame':   152188 obs. of  16 variables:
 $ codigo_producto    : int  103100031 104020004 101010079 101010045 1010
 $ descripcion        : chr  "CONSULTA EXTERNA " "CONSULTA URGENCIA " "CON
 $ tipo_id_paciente   : chr  "CC" "CC" "CC" "CC" "TI" ...
 $ sexo_id            : chr  "F" "F" "F" "F" "M" "M" ...
 $ peso               : num  64.4 76.8 76.8 76.8 58 58 58 58 58 58 ...
 $ talla              : int  154 163 163 163 145 145 145 145 145 145 ...
 $ taalta             : int  125 120 120 120 124 124 124 124 124 124 ...
 $ tabaja             : int  80 80 80 80 74 74 74 74 74 74 ...
 $ per_abdominal      : int  96 96 96 96 91 91 91 91 91 91 ...
 $ descripcion.1      : chr  "ALUMINIO HIDRXIDO + MAGNESIO HIDROXIDO CON
 $ dosis              : num  5 1 1 1 1 1 1 1 1 1 ...
 $ dias_tratamiento   : int  15 15 15 10 10 10 10 10 10 10 ...
 $ cantidad           : int  1 1 15 20 30 30 30 30 30 30 ...
 $ tipo_diagnostico_id: chr  "Z000" "J329" "J329" "J329" ...
 $ diagnostico_nombre : chr  "EXAMEN MEDICO GENERAL" "SINUSITIS CRONICA,
 $ edad               : num  59.3 23.7 23.7 23.7 74.6 ...
```

Fig. 14 Reducción de Variables.

c) Resultados De Normalización

Se procede a la normalización de los datos para evitar que algunos valores extremos incidan en el análisis, además de poder realizar escalas comunes que faciliten la comparación de las variables de manera que se tenga un buen modelo con el fin de facilitar su interpretación [13]. Para normalizar el dataset se hace uso de otras de las funciones de Rstudio, en este caso la min-max(). Como se indica en Fig. 15.

```
# funcion para normalizar
  min_max_norm <- function(x) {
    (x - min (x)) / (max(x) - min(x))
  }

# Normalizar el dataset
  datosNorm<- as.data.frame(lapply(datos2, min_max_norm))
```

dosis	dias_tratamiento	cantidad	peso	talla	taalta	ta
0.016502750	0.078212291	0.000000000	0.3796407	0.8138298	0.5414847	0
0.003167195	0.078212291	0.000000000	0.4538922	0.8617021	0.5196507	0
0.003167195	0.078212291	0.028056112	0.4538922	0.8617021	0.5196507	0
0.003167195	0.050279330	0.058116232	0.3413174	0.7659574	0.5371179	0
0.003167195	0.162011173	0.058116232	0.4305389	0.8457447	0.6855895	0

Fig. 15 Variables normalizadas

La siguiente Fig. 16, muestra que existe una proporción de la población que se encuentra por encima o por debajo del promedio en cuanto a su peso y talla se refiere. En otras palabras, hay individuos cuyas medidas corporales se alejan significativamente del promedio de la población. Esta información puede ser valiosa para identificar grupos de individuos que puedan necesitar atención especial en términos de salud y bienestar, ya que pueden tener algunos problemas de desnutrición y otros problemas de obesidad y que este factor se presenta en todos los diferentes rangos de edad.

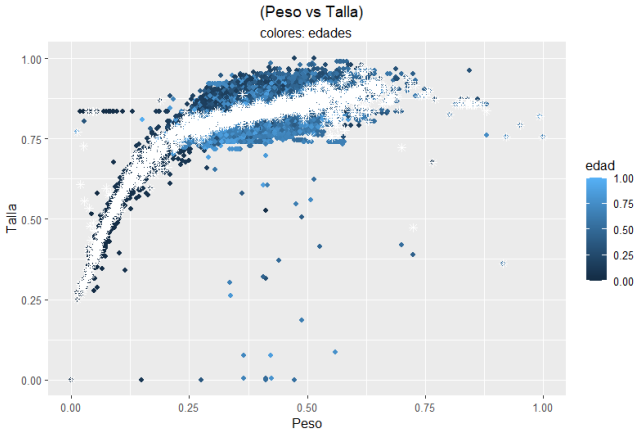

Fig. 16 Peso Respecto A Talla

d) Resultados Del PCA

Para realizar el análisis del PCA, se puede utilizar la función summary() de Rstudio. Con esto se visualizan valores de desviación estándar, de varianza y de varianza acumulada, estos datos se reflejan en Fig. 17. De acuerdo a esos resultados se decide cuantos y cuáles son los componentes más representativos que se pueden tomar para continuar con el análisis.

```
summary(pca1)
  >  summary(pca1)
  Importance of components:
         Standard deviation   Proportion of Variance   Cumulative Proportion
  PC1        49.666                0.624                  0.624
  PC2        21.0047               0.1116                 0.7357
  PC3        18.65509              0.08804                0.82371
  PC4        16.09629              0.06555                0.88925
  PC5        14.58876              0.05384                0.94310
  PC6        10.67423              0.02883                0.97192
  PC7         6.7726               0.0116                 0.9835
  PC8         6.10374              0.00943                0.99295
  PC9         5.27893              0.00705                1.00000
```

Fig. 17 Representación Bidimensional de cada Componente.

En la anterior figura 17, se aprecia, que con los dos primeros componentes se puede obtener el 73.57% de la información suficiente para desarrollar cualquier análisis, tan solo se pierde un 26.4% de toda la información. Para una mejor visualización se refleja el diagrama de la Fig. 18.

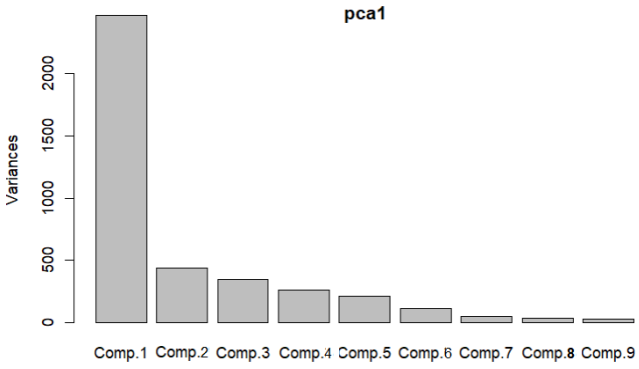

Fig. 18 Varianza de Componentes.

Una vez reducida la dimensionalidad se obtiene un nuevo archivo con los componentes principales. Como se indica en Fig. 19, con el cual se realiza algunos análisis.

	PC1	PC2
101010021	139.253349	-15.8184594
103060006	60.106456	-53.9214198
101010088	-135.60221	-32.173522
101010324	-138.118670	-18.6126876
101010076	-104.671795	-2.4678225
103100028	74.7321364	-150.731673
105010007	-285.059033	-69.793287
101010092	-4.960526	7.8559618
105010003	-210.329331	-50.994755
101010083	-139.586419	-16.633329
101010044	137.174558	25.470604
103010016	67.903703	-60.0830800

Fig. 19 Componentes Principales

B) Conclusión Del Análisis De Las Componentes Principales:

CP1: Se tiene cuatro grupos de medicamentos con más dosis prescritas, los antiepilépticos, antidiabéticos, hipotiroideos y antihipertensivos.

CP2: Se encuentra con tres grupos de medicamentos con mayor trascendencia, los cuales han sido prescritos para problemas digestivo gastrointestinal, antidiabéticos y analgésicos.

De la Fig. 19, se puede observar que existe una clasificación de seis grupos de medicamentos. Otra consideración que se puede obtener es respecto a los dos componentes en donde el indicativo de crecimiento no está en dependencia, es decir que el factor de crecimiento de uno no influye en el crecimiento del otro. En Fig. 20, se puede apreciar mucho mejor dicha apreciación.

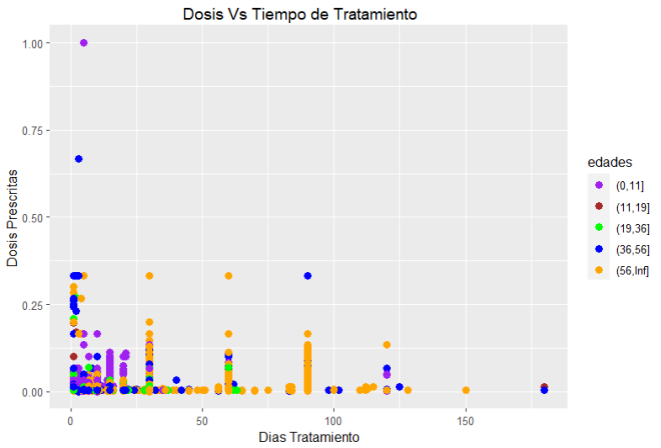

Fig. 20 Dosis Respecto a Tiempo de Tratamiento

En la Fig. 20, se observa claramente que el tiempo del tratamiento no incide en la cantidad de medicamentos que se les suministra a los pacientes. Además, otro aspecto que se puede resaltar es la edad, en donde se aprecia que, a menor edad, más alta es la posibilidad de adquirir cualquier dolencia, aunque sus tratamientos no requieren periodos largos, mientras que en personas mayores a 56 años los tratamientos se extienden, en donde la causa puede estar relacionada con alguna posible enfermedad crónica.

A partir de los resultados obtenidos en la gráfica 19, se ha obtenido un listado de medicamentos que han sido mayormente prescritos. Como se aprecia en Fig. 21.

	descripcion.1
101010021	CARBAMAZEPINA 200MG
103060006	SALES DE REHIDRATACION ORALFORMULA OMS POLVO PA...
101010088	METFORMINA 850 MG TABLETA
101010324	METFORMINA + LINAGLIPTINA 2.5MG/1U/1000MG/1U
101010076	LEVOTIROXINA SODICA 100 McG TABLETA
103100028	ALBENDAZOL 400MG/10ML SUSP ORAL FCOx10ML
105010007	AGUJAS PARA PEN DE INSULINA
105010003	TIRILLAS PARA GLUCOMETRIA
101010083	LOSARTAN 50 MG TABLETA O TABLETA RECBIERTA
101010044	ENALAPRIL MALEATO 5 MG TABLETA
103010016	ACETAMINOFEN JARABE POR 120ML

Fig. 21 Medicamentos con Mayor Prescripción

Para análisis de otras patologías se presenta la siguiente tabla II, en ella, los rangos de códigos referentes a los distintos diagnósticos, de manera que facilite la interpretación de las gráficas de aquellas variables que se tomen para análisis.

TABLA II
RANGOS DE DIAGNÓSTICOS[5]

Rango de Códigos DX	Enfermedad
A00 - B99	Parasitarias e infecciosas
C00 – D48	Neoplasias
D50 – D89	Afecciones a la sangre y afecciones de inmunidad
E00 – E90	Indoctrinas y metabólicas
F00 – F99	Trastornos mentales y de comportamiento
G00 – G99	Afecciones al sistema nervioso
H00 – H59	De los ojos
H60 – H95	De los oídos
I00 – I99	Afecciones al sistema circulatorio
J00 – J99	Afecciones al sistema respiratorio
K00 – K93	Afecciones al sistema digestivo
L00 – L99	Afecciones de la piel
M00 – M99	Afecciones musculares
N00 – N99	Afecciones urinarias
O00 – O99	Embarazo y gestación
P00 – P96	Afecciones perinatales
Q00 _ Q99	Congénitas
R00 _ R99	clínicos de laboratorio
S00 – T98	Traumas, envenenamiento y otras externas
V01 – Y98	Morbilidad y mortalidad
Z00 – Z99	servicios de salud
U00 – U99	Situaciones especiales

Para reducir nombre de código de diagnóstico, se crea otra columna como idDX, se cambia por pocas letras para facilitar su manipulación y asignarles un color para mejorar la apreciación de las gráficas, en su posterior análisis. Como se indica en Fig. 22.

```
datosN <- datos
datosN <- select(datosN,-edades)
datosN$idDX <- gsub("[^A-Za-z]","",datosN$tipo_diagnostico_id)
unique(datosN$idDX)
```

tipo_diagnostico_id	idDX	diagnostico_nombre
Z000	Z	EXAMEN MEDICO GENERAL
J329	J	SINUSITIS CRONICA, NO ESPECIFICADA
J329	J	SINUSITIS CRONICA, NO ESPECIFICADA
J329	J	SINUSITIS CRONICA, NO ESPECIFICADA
M199	M	ARTROSIS, NO ESPECIFICADA

```
# Agregar color a cada valor, en este caso hay 43 valores
my_colors <- terrain.colors(43)
my_colors[1:4]<- "#8F4C1F"
```

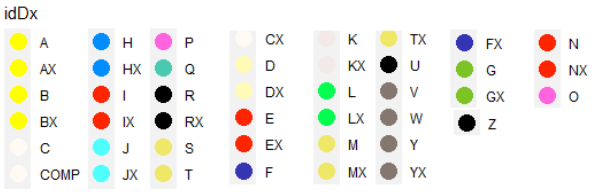

Fig. 22 Reducción de Nombre de Variable

Se procede a realizar un análisis, entre las variables, peso y tensión arterial alta, con el objetivo de conocer la incidencia que estos pueden tener en la salud de cada individuo.

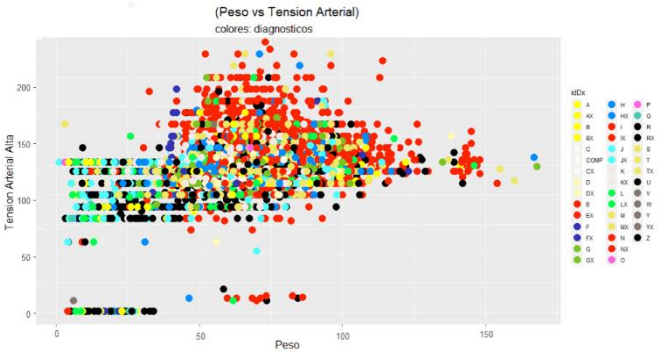

Fig. 23 Peso vs Tensión Arterial Alta

En la Fig. 23, permite apreciar la existencia de una relación positiva, entre la presencia de una presión arterial elevada y un mayor índice de masa corporal, esta relación a su vez, aumenta la probabilidad de que un individuo sea diagnosticado con

[5] Fuente: https://estrucplan.com.ar/que-es-el-cie10/

diversas enfermedades que afectan su sistema cardiovascular, endocrino y genitourinario, poniendo en grave riesgo su estado de salud en general. Por tanto, se resalta la importancia en mantener niveles apropiados de tensión arterial y peso corporal. Con esta medida se puede prevenir la aparición de las mencionadas patologías y preservar la salud en óptimas condiciones.

La siguiente gráfica; Fig. 24, permite observar un análisis de la incidencia entre la edad y la tensión arterial. Con ello se puede contribuir al mejoramiento de la salud, y actuar de forma preventiva.

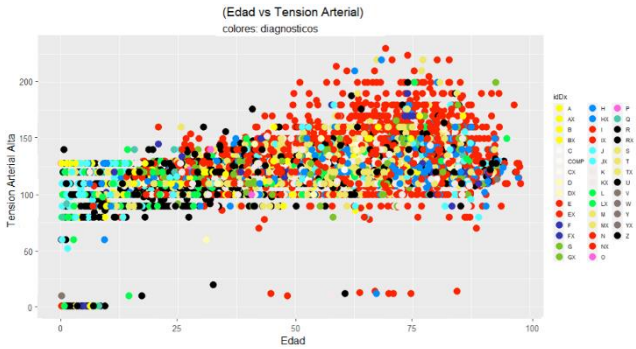

Fig. 24 Peso vs Tensión Arterial Alta

En la Fig. 24, permite apreciar claramente que los usuarios mayores de 50 años son más propensos a riesgo de ser diagnosticados con enfermedades cardiovasculares, endocrinas o genitourinarias debido a sus altos niveles de tensión arterial u otros factores asociados. Por otro lado, en el grupo de 0 a 25 años, se pueden observar usuarios que presentan niveles de tensión arterial dentro de los parámetros normales, por lo tanto, tienen diagnósticos que no representan una amenaza para su salud. Es importante destacar la relevancia de estos hallazgos en cuanto a la prevención y el manejo de estas patologías en diferentes grupos de edad, así como la necesidad de mantener un control regular de los niveles de tensión arterial para asegurar un estado de salud óptimo.

Otro aspecto relevante, son aquellos usuarios que están en el período de primera infancia, en esta etapa se es más propenso a presentar enfermedades relacionadas con el sistema respiratorio, mientras que, en contraste, la población adulta suele experimentar patologías vinculadas con el oído y la visión. Esta tendencia indica que existe una clara diferencia en las condiciones de salud de los distintos grupos etarios, lo cual es un factor importante que se debe tener en cuenta en la prevención y el tratamiento de enfermedades.

e) Resultado Modelado-Clustering

1) Cálculo de la matriz de distancias

Los datos que se almacenaron en la variable, “mdist”, son de gran utilidad para realizar posteriormente diversas operaciones y también para análisis en los datos obtenidos, a partir de la matriz de distancia. Es importante destacar que la correcta asignación y uso de variables en la programación es fundamental para garantizar la eficiencia y precisión de los procesos realizados. Como se indicó en, Fig. 5.

2) Agrupación de clustering

De la Fig. 6, que correspondió a la gráfica del método wss, se obtuvo que del rango de valores de K (número de clusters), la gráfica sugiere 2 y máximo 3 clusters.

De la Fig. 7, que correspondió a la gráfica del método silhouette, se obtuvo que del rango de valores de K (número de clusters), el número de clusters recomendado es 2. Por tanto, de las dos gráficas se obtiene que k= 3, y ese es el número de clusters más indicado para agrupar los datos.

En resumen, al llevar a cabo el análisis correspondiente, se ha obtenido como resultado la identificación de dos clústeres que presentan similitudes notables en cuanto a sus características. Esta información resulta de gran importancia y utilidad a la hora de tomar decisiones con relación a la selección del método K-Medias y el número de clusters indicado, lo cual contribuye a lograr los objetivos específicos planteados en el análisis en cuestión. Es fundamental tener en cuenta que el conocimiento detallado de las similitudes y diferencias entre los clústeres identificados permite la identificación de patrones y tendencias que pueden ser de suma importancia en la toma de decisiones acertadamente [12,21].

3) Construcción del modelo

- Modelamiento con el método centroide.

Se hace uso de la matriz de distancias y con dos clústeres.

```
# construcción del modelo CENTROID
modelo1 <- hclust(mdist, method = "centroid")
#dendrograma
plot(modelo1,hang = -1,main = "Metodo Centroid" )
# Dividir el dendograma en dos cluster
rect.hclust(modelo1, k=2, border= c("red","green"))
```

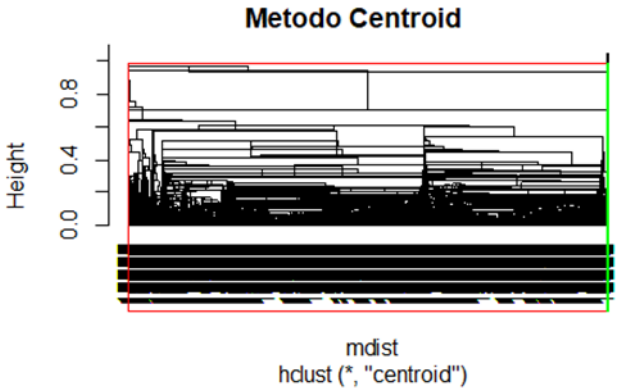

Fig. 25 Modelo con Centroide

En Fig. 25, se observan las línea roja y verde, indican el agrupamiento del clúster en 2 grupos, como lo indicado, en la elección de selección de número de óptimo de clúster del método silhouette.

- Modelamiento con el método vecino más cercano.

Se hace uso de la matriz de distancias y con tres clústeres.

```
# construcción Método del vecino más cercano
modelo2 <- hclust(mdistt, method = "single")
# Generar dendrograma
plot(modelo2,hang = -1, main = "Método vecino más cercano")
# Dividir el dendograma en tres clusters
rect.hclust(modelo2, k=3, border= c("red","green","blue"))
```

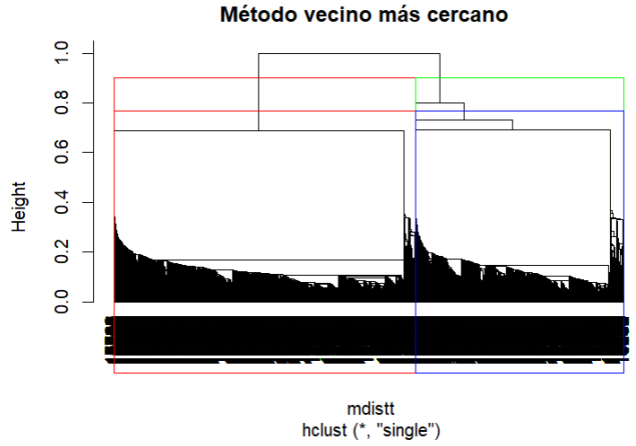

Fig. 26 Modelo con Vecino Más Cercano

En Fig. 26, se observan las línea roja, verde y azul que indican el agrupamiento de los 3 clúster, como lo indicado en la elección de selección de número óptimo de clúster.

El método de vecino más cercano es útil para encontrar clusters con formas irregulares y puede funcionar bien cuando los clusters tienen tamaños desiguales. Sin embargo, también puede ser sensible a valores atípicos o ruido en los datos, lo que puede resultar en la unión de clusters que no son realmente similares, es adecuado para un conjunto de datos moderado, o como un punto de partida antes de aplicar técnicas más avanzadas [21].

- Modelamiento con el método vecino más lejano.

Se hace uso de la matriz de distancias y con tres clusters.

```
# construcción Método del vecino más Lejano
modelo3 <- hclust(mdistt, method = "complete")
# Generar dendrograma
plot(modelo3,hang = -1, main = "Método vecino más Lejano")
# Dividir el dendograma en tres clusters
rect.hclust(modelo3, k=3, border= c("red","green","blue"))
```

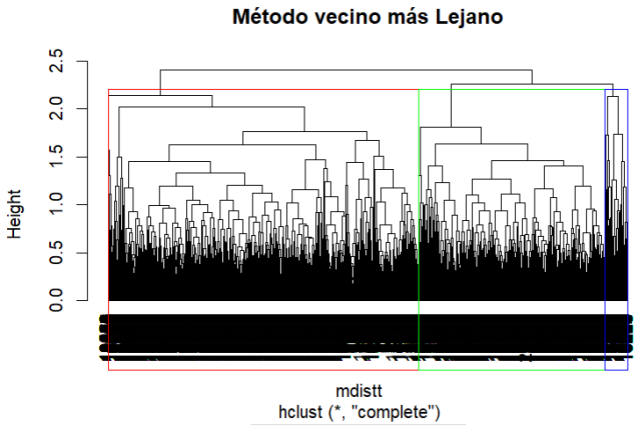

Fig. 27 Modelo con Vecino Más Lejano

En Fig. 27, se observan las línea roja, verde y azul que indican el agrupamiento de los 3 clúster.

El método vecino más lejano es muy resistente a los valores atípicos. Funciona bien con grandes conjuntos de datos, no requiere una matriz de distancia completa y puede utilizar solo una fracción de los datos para identificar los vecinos más cercanos [21].

4) Ejecución del modelo

Para la ejecución del modelo se ha optado por utilizar el método k-means con tres clusters, como lo indicado en selección de número de óptimo de clúster. Como se indica en Fig. 28.

```
K-means clustering with 3 clusters of sizes 1004, 7635, 11360
                                            cantidad de registros por clúster
Cluster means:
  codigo_producto     dosis dias_tratamiento  cantidad
1      0.28540163 0.04391760      0.04525474 0.01973998
2      0.07880862 0.01646323      0.23220101 0.13421842
3      0.08465139 0.01551503      0.22780313 0.12844830

      peso     talla    taalta    tabaja per_abdominal
1 0.1970340 0.5588329 0.1729394 0.2841064   0.3926310
2 0.4604054 0.8367447 0.2846694 0.4352755   0.4093318
3 0.4171078 0.7617084 0.2834803 0.4292455   0.4009396

        edad tipo_id_paciente sexo_id diagnostico
1 0.09007719     0.8326693227 0.5488048  0.5734584
2 0.67221676     0.0006548788 1.0000000  0.2987967
3 0.63785028     0.0000000000 0.0000000  0.3127544

Numero de clúster
```

Fig. 28 k-means de clustering

Luego se procede a graficar el clustering, mediante un mapa de calor. Como se indica en Fig.29.

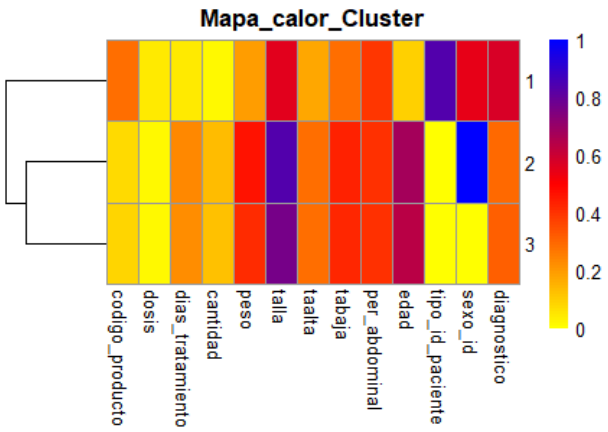

Fig. 29 Mapa de calor del clustering

En Fig. 28 y Fig. 29, los resultados mostraron que el algoritmo identificó tres clúster o grupos de medicamentos. Cada clúster se caracteriza por un conjunto de valores promedio para diferentes variables como la dosis, los días de tratamiento, la cantidad, la edad, el peso, la talla, entre otros.

El primer clúster está compuesto por 1004 registros y se caracteriza por la administración de dosis altas y un menor número de días de tratamiento en promedio. Estas características sugieren la utilización de fármacos como antiparasitarios y antibacterianos, que están indicados para el tratamiento de enfermedades diarreicas y respiratorias que requieren un tratamiento corto y de fácil manejo en la conducta de los pacientes. Además, este grupo se compone principalmente de pacientes más jóvenes, con un menor número de medicamentos en uso, un peso adecuado y una presión arterial normal, lo que podría indicar una menor prevalencia de enfermedades crónicas en comparación con otros grupos. Es esencial subrayar que, a pesar de que este grupo está formado por pacientes con un riesgo reducido de enfermedades crónicas, es crucial realizar un seguimiento riguroso. Esto se debe a que las manifestaciones sintomáticas pueden estar relacionadas con factores climáticos y condiciones sociales. En este sentido, es importante recordar que es necesario llevar a cabo investigaciones para determinar los factores específicos que influyen en la aparición de estas enfermedades en esta población. En resumen, el primer clúster está compuesto por pacientes jóvenes y con menor riesgo de enfermedades crónicas, que reciben tratamientos con dosis altas y de corta duración con antiparasitarios y antibacterianos para el manejo de enfermedades diarreicas y respiratorias agudas.

El segundo clúster está compuesto por 7635 registros con dosis bajas y días de tratamiento más prolongados, relacionados con el aumento de la presión arterial. Esta característica sugiere que los medicamentos utilizados pueden estar dirigidos al tratamiento de enfermedades agudas o graves, como los antihipertensivos orales, terapias de insulina tradicional o la combinación de ambos fármacos, que requieren un tratamiento intensivo a largo plazo. Además, se observa que este grupo tiende a ser manejado con fármacos antiinflamatorios y analgésicos para tratar dolencias musculares. Cabe destacar que la edad y el peso de los pacientes de este clúster pueden indicar una mayor prevalencia de enfermedades agudas en este grupo. Por lo tanto, es fundamental realizar una evaluación detallada y cuidadosa para determinar el tratamiento adecuado para cada paciente, evitar el aumento de dosis con otros fármacos y posibles complicaciones. Es importante tener en cuenta que la prescripción de medicamentos para la hipertensión y la diabetes debe ser controlada y precisa, ya que estas enfermedades pueden ser crónicas y requerir un tratamiento a largo plazo.

En resumen, el análisis de los registros del segundo clúster permite identificar patrones de prescripción de medicamentos y características de los pacientes que pueden ser útiles para mejorar la atención médica y prevenir complicaciones.

El tercer clúster es el mayor en los registros por que cuenta 11360. lo cual es importante destacar que este grupo está conformado por una amplia variedad de edades y medicamentos, requiere un enfoque personalizado y vigilancia constante para asegurar que los pacientes reciban el tratamiento adecuado. Además, el hecho de que este grupo esté en una posición intermedia en las variables analizadas puede ser indicativo de que los pacientes se encuentran en una fase temprana de una enfermedad o de que están en riesgo de desarrollar una enfermedad crónica. Por lo tanto, es crucial que los profesionales de la salud presten atención y monitoreen de manera rigurosa el tratamiento de estos pacientes, para evitar complicaciones de salud y prevenir la progresión de enfermedades de alto riesgo. Asimismo, es necesario brindar una atención médica adecuada y personalizada a cada uno de ellos, teniendo en cuenta sus características particulares y necesidades específicas. En conclusión, el tercer grupo de registros es el más grande y diverso, lo que lo convierte en un grupo de alto riesgo para desarrollar enfermedades crónicas. Por ello, es fundamental prestar atención y brindar un seguimiento constante a los pacientes para garantizar su bienestar y prevenir complicaciones graves en su salud.

IV. CONCLUSIONES

En la investigación, se buscó establecer la relación entre variables en cada registro asignado a un grupo de pacientes, con el objetivo de describir el tratamiento adecuado según las características individuales. Para lograr esto, se planteó el uso de la técnica de extracción de información k-means, así como el método de Ward para encontrar patrones en los datos. Sin embargo, se determinó que el método de Ward no era eficiente para el análisis debido a su falta de generación de información relevante, por lo que se implementó un algoritmo descriptivo para describir la población y determinar la relación entre las frecuencias del uso de la formulación médica. Con base en lo anterior, se elaboró un procedimiento de extracción de información que emplearía dichos valores junto con algunas características para determinar el tipo de complicación que se manifestaba en los pacientes.

Respecto a los hallazgos obtenidos, se encontraron características interesantes en relación con la posible complicación de enfermedades crónicas en ciertos grupos de pacientes, como la dosis, los días de tratamiento y la edad. Se sugiere llevar a cabo un estudio más detallado de esta población con el propósito de identificar los elementos que puedan tener un impacto en la aparición de enfermedades crónicas y frecuentes.

En el transcurso de la investigación, se encontraron obstáculos al manipular los datos debido a su gran volumen y alta calidad, lo que incidió en el tiempo de implementación y la calidad de los modelos generados. Es importante contar con una infraestructura adecuada para manejar este tipo de datos y evitar retrasos en el desarrollo de la investigación. Se recomienda el uso de herramientas como, Google Colab, para este propósito.

Es relevante destacar igualmente la fase de preparación de los datos, dado que es una de las etapas cruciales y que consume más tiempo. Resulta esencial efectuar una depuración rigurosa de los datos a fin de evitar variaciones significativas en los resultados al corregir algunas inconsistencias. Por último, cabe destacar la relevancia de la validación de los resultados y la intervención de un experto en minería de datos en el proceso de investigación.

Las investigaciones venideras podrían centrarse en la utilización de modelos predictivos para prever el impacto de los fármacos en patologías de bajo y alto riesgo, adicionalmente se podrían evaluar decisiones sobre otras variables relacionadas con la eficacia de los medicamentos en estas patologías. Estos análisis permitirán optimizar el tratamiento de los pacientes, especialmente aquellos en situaciones de mayor vulnerabilidad.

AGRADECIMIENTOS

Agradecemos en primera instancia a Dios, por darnos la oportunidad de cumplir con una de las metas más importantes en nuestra vida profesional, por darnos la sabiduría y fortaleza necesarias en cada uno de los momentos que pasamos para hoy culminar felizmente esta etapa. Agradecer también, a la Universidad Fundación Universitaria de Popayán, por permitirnos adelantar en ella nuestra carrera universitaria. Igualmente, a nuestras familias, quienes estuvieron pendientes de nuestro proceso, gracias a su amor incondicional, supieron entender y brindar su apoyo en todo momento para poder alcanzar este gran triunfo.

.

REFERENCIAS

[1] S. RCano, J. Soriano del Castillo, and J. Merino-Torres, “Causas Y Tratamiento De La Obesidad,” *Nutr. Clin. y Diet. Hosp.*, vol. 37, no. 4, pp. 87–92, 2017, doi: 10.12873/374rodrigo.

[2] B. Durán, B. Rivera, and E. Franco, “Apego Al Tratamiento Farmacológico En Pacientes Con Diagnóstico De Diabetes Mellitus Tipo 2,” *Salud Publica Mex.*, vol. 43, no. 3, pp. 233–236, 2001, doi: 10.1590/s0036-36342001000300009.

[3] J. Alvarez, N. Buriticá, J. Herrera, D. Ortiz, and K. Salazar, “Uso De La Historia Natural De La Enfermedad Como Herramienta En La Gestión De La Patología Laboral En Colombia,” 2020.

[4] J. Solórzano, “Ingreso Y Comportamiento Del Consumidor Como Base Para Segmentación Y Elaboración De Perfiles De Mercado.,” pp. 37–45, 2013.

[5] E. Ortiz, C. Galarza, F. Cornejo, and J. Ponce, “Acceso A Medicamentos Y Situación Del Mercado Farmacéutico En Ecuador,” vol. 36, no. 1, pp. 57–62, 2014.

[6] G. Amézquita, N. Patiño, and C. Salamanca, “Minería De Datos Para La Identificación De Factores De Riesgo En Pacientes Con Hipertensión Arterial,” 2019. [Online]. Available: https://repositoriocrai.ucompensar.edu.co/handle/compensar/2293

[7] G. Páez, “Metodología Para El Desarrollo De Modelos De Segmentación Y Su Aplicación Al Mercadeo,” 2014.

[8] L. Bautista, “Uso De Minería De Datos En La Detección Temprana Y Prevención De Complicaciones De Enfermedades En El Sistema De Salud Colombiano,” 2010.

[9] N. Mejía and C. Arias, “Los Clusters Como Mecanismo Para La Generación De Economías A Escala En El Sector De Medicamentos De Alto Costo En Colombia,” 2014. [Online]. Available: https://eje.bioscientifica.com/view/journals/eje/171/6/727.xml

[10] R. Escribano, F. Martínez-de-Pisón, M. Castejón, A. Sanz, and R. Fernández, “Modelos Descriptivos Y Predictivos Para La Estimación De Costes En Proyectos Informáticos,” *XIV Internartional Congr. Proj. Eng.*, pp. 2590–2600, 2010, [Online]. Available: http://dspace.aeipro.com/xmlui/bitstream/handle/123456789/2171/CIIP10_2590_2600.PDF?sequence=1&isAllowed=y

[11] B. Beltrán, “Minería De Datos,” *Planet. Space Sci.*, vol. 30, no. 1, 2001, doi: 10.1016/0032-0633(82)90071-X.

[12] F. Pollo *et al.*, “Ingeniería De Proyectos De Explotacion De Informacion,” *XII Work. Investig. en Ciencias la Comput.*, pp. 172–176, 2010.

[13] I. Martínez, “Modelamiento De Confiabilidad Y Análisis Para Flotas: Un Enfoque Basado En Clustering Para Manejo De Datos No Homogéneos,” 2017.

[14] E. Cabanillas and P. Martinez, “Diseño De Un Modelo Computacional Basado En Técnicas De Minería De Datos Para El Pronostico De La Demanda De Productos Farmacéuticos.,” 2014. doi: 10.4045/tidsskr.16.0688.

[15] L. Vargas and E. Mesa, “Introducción Al Análisis De Datos Con RStudio,” *Cenipalma*, pp. 1–65, 2021, [Online]. Available: www.cenipalma.org

[16] D. Garcia-Alvarez, “Estudio Comparativo De Técnicas De Detección De Fallos Basadas En El Análisis De Componentes Principales (PCA),” *RIAI - Rev. Iberoam. Autom. e Inform. Ind.*, vol. 8, no. 3, pp. 182–195, 2011, doi: 10.1016/j.riai.2011.06.006.

[17] P. Ventura, “Métodos de Análisis Cluster Difusos,”

2022.
[18] J. Vicente, “Introduccion Al Analisis De Cluster,” *Introd. Al Análisis Clust.*, p. 22, 2007, [Online]. Available: http://benjamindespensa.tripod.com/spss/AC.pdf
[19] J. Rodríguez, E. Rojas, and R. Franco, “Clasificación De Datos Usando El Método K-Nn,” *Vínculos*, vol. 4, no. 1, pp. 4–18, 2013.
[20] M. Martínez, “Segmentación De Usuarios En La Oficina De Farmacia Mediante Algoritmos Bioinspirados,” 2015. [Online]. Available: http://hdl.handle.net/10433/2361
[21] J. Cerón, “Análisis De Algoritmos De Clustering Para Datos Categóricos,” 2018.
[22] E. León, “Métricas Para La Validación De Clustering,” 2019, [Online]. Available: http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_validacion_Clustering.pdf