Realizado por:

Juan Camilo Restrepo Velez

Wilder Valencia Ocampo

|  |  |
| --- | --- |
| **Base de datos** | Medicamentos |
| **Descripción del negocio y de los datos** | La rinitis alérgica afecta al 35% de la población en Colombia. Los tratamientos y fármacos son variados por lo que se requiere un estudio para identificar la medicina adecuada para cada persona |

**Tabla de contenido**

[I. PREPARACIÓN DE LOS DATOS 2](#_Toc52457309)

[II. MODELAMIENTO 3](#_Toc52457310)

[Clustering 4](#_Toc52457311)

[a.(0.5) Establecer y sustentar la cantidad de clusters ideal para el conjunto de datos mediante el método del codo. 4](#_Toc52457312)

[b.(0.5) Aplicar un método de clustering según la cantidad determinada en el punto anterior. Incluir en el informe la configuración del método y la cohesión de la agrupación. 4](#_Toc52457313)

[c.(0.5) Describir cada uno de los perfiles encontrados y asignar un nombre a cada grupo. 5](#_Toc52457314)

[Asociación 7](#_Toc52457315)

[d.(0.5) Crear reglas de asociación que permitan determinar patrones en los datos. 7](#_Toc52457316)

[e.(0.5) Explicar las 3 reglas más relevantes. 7](#_Toc52457317)

[Selección de Factores 9](#_Toc52457318)

[f.(2.5) Cuáles son las 3 variables que más influyen sobre la variable objetivo de los datos? Aplicar 5 técnicas diferentes. 9](#_Toc52457319)

# I. PREPARACIÓN DE LOS DATOS

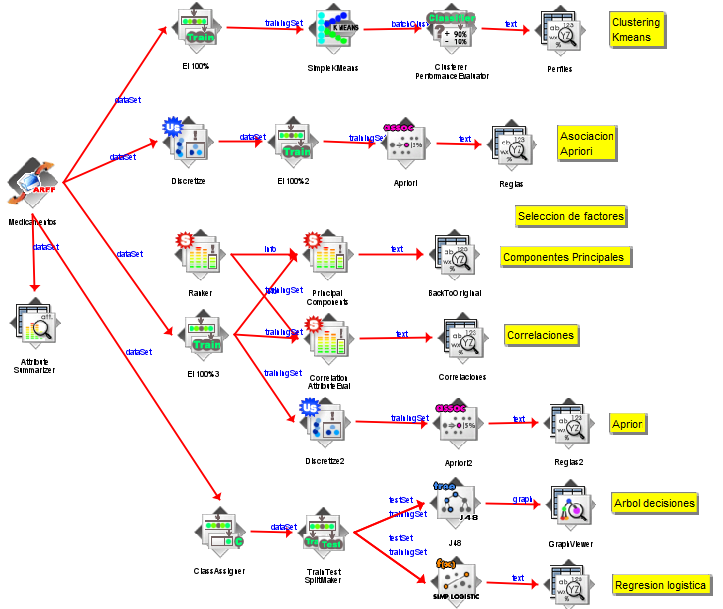
Ya realizada en la práctica anterior “Practica Análisis descriptivo”, algunos de los ítems mas importantes encontrados en dicha preparación:

* La variable ‘ID’ no es relevante, por lo tanto, se elimina
* Se encuentran algunas irregularidades en:
  + ‘EDAD’ (2 edades fuera de rango)
  + ‘SEXO’ (categoría extra, producto de mala digitación)
  + ‘PRESIÓN SANGUÍNEA’ (presencia de nulos)
  + ‘COLESTEROL’ (categoría extra, producto de mala digitación)
  + ‘SODIO’ (dato fuera de rango)
  + ‘POTASIO’ (dato fuera de rango)
* No se presentan registros duplicados
* Se presentan bajos niveles de nulos, como consecuencia de “datos atípicos”:
  + ‘EDAD’ (1%)
  + ‘PRESIÓN SANGUÍNEA’ (1%)
  + ‘SODIO’ (1%)
  + ‘POTASIO’ (2%)
* No se presentan altas correlaciones que haga pensar que hay presencia de variables redundantes.
* El tipo de datos de las variables son:
* EDAD à variable Numérica
* SEXO à variable Categórica
* PRESIÓN SANGUÍNEA à variable Categórica
* COLESTEROL à variable Categórica
* SODIO à variable Numérica
* POTASIO à variable Numérica
* MEDICAMENTO (variable objetivo) à variable Categórica

**NOTA:** Para las reglas de asociación se discretizaron las variables *EDAD*, *SODIO* y *POTASIO* en tres rangos

# II. MODELAMIENTO

Se crea el siguiente modelo de con las tres técnicas de análisis descriptivo



Donde se utiliza el 100% de los datos como conjunto de entrenamiento para las técnicas:

* Clustering, con el método Kmeans.
* Asociación, para el método Apriori, que además se realiza la discretización de las variables numéricas.
* Componentes Principales, al cual se le asigna el ranker.
* Correlaciones, al cual se le asigna el ranker.

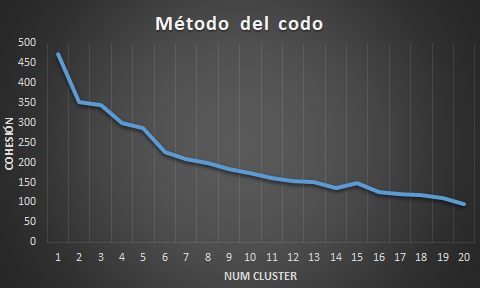
Para los métodos que crean un modelo predictivo (aunque estos se utilizarán para analizar los atributos que consideran como relevantes) se realiza la división 70-30, estos son:

* Árbol de decisiones, el componente J48.
* Regresión logística, por medio del SimpleLogistic

## Clustering

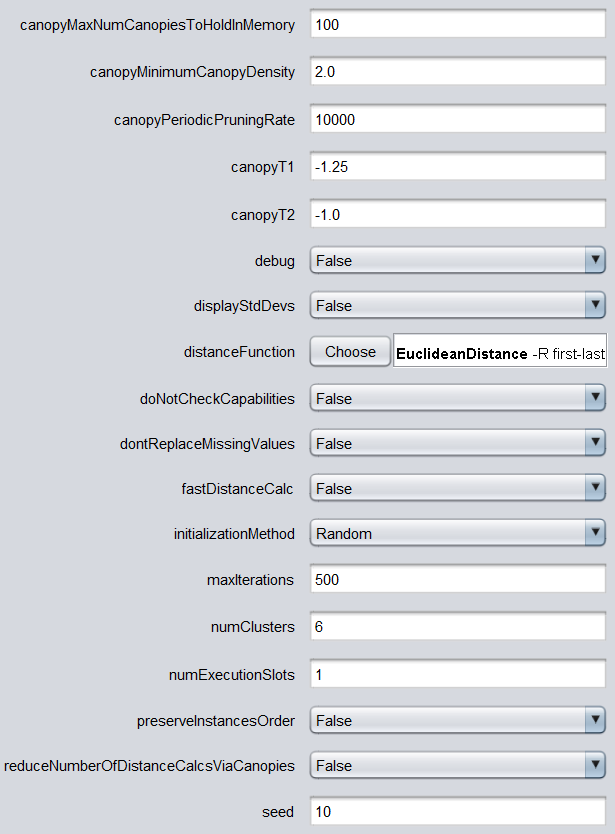
### a.(0.5) Establecer y sustentar la cantidad de clusters ideal para el conjunto de datos mediante el método del codo.

Se realiza el método del codo variando la cantidad de clusters entre 1 y 20, estableciendo que la cantidad ideal son 6 cluster, que es cuando se empieza aplanar la curva



### b.(0.5) Aplicar un método de clustering según la cantidad determinada en el punto anterior. Incluir en el informe la configuración del método y la cohesión de la agrupación.

Se establecen 6 clusters, según lo indicado en el punto anterior, también se deja la medida de distancia euclidiana y demás configuraciones como se observa a continuación.

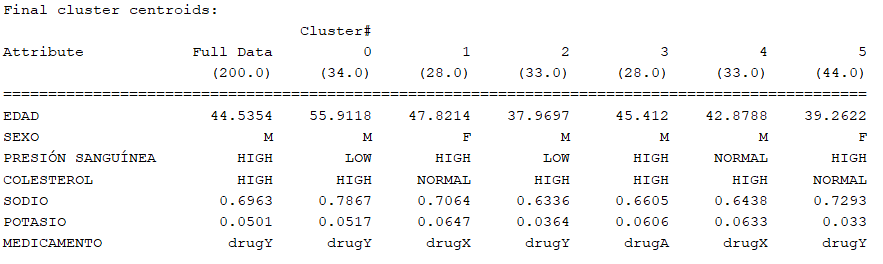


En cuanto al resultado se tiene una medida de cohesión de 227 y que el método solo tuvo que realizar 6 iteraciones para ajustar los centroides de los cluster.



### c.(0.5) Describir cada uno de los perfiles encontrados y asignar un nombre a cada grupo.

Los centroides de los clusters encontrados los siguientes:

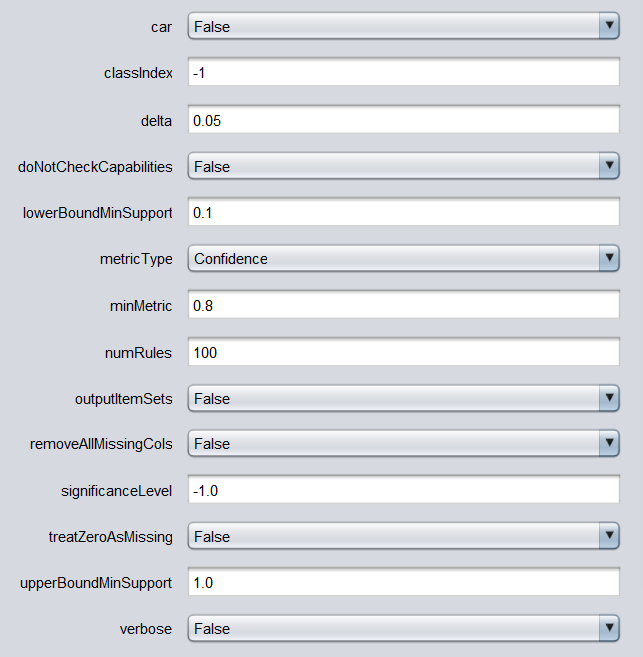


|  |  |  |
| --- | --- | --- |
| *Cluster #* | *Nombre* | *Descripción* |
| 0 | Hombres mayores solo con buena presión | Hombres mayores con presión sanguínea baja, colesterol alto, niveles de sodio y potasio altos, con medicamento durgY |
| 1 | Mujeres mayores hipertensas | Mujeres mayores con presión sanguínea alta, colesterol normal, niveles de sodio y potasio altos, con medicamento drugX |
| 2 | Hombres con colesterol alto | Hombres de 30 años con presión sanguínea baja, colesterol alto, niveles de sodio y potasio bajos, con medicamento drugY |
| 3 | Hombres cuarentones hiper-graves | Hombres de 40 años con presión sanguínea alta, colesterol alto, niveles de sodio y potasio altos, con medicamento drugA |
| 4 | Hombres cuatentones que cuidan su presión | Hombres de 40 años con presión sanguínea normal, colesterol alto, niveles de sodio y potasio altos, con medicamento drugX |
| 5 | Mujeres cuarentonas hipertensas | Mujeres cerca de los 40 años con presión sanguínea alta, colesterol normal, niveles de sodio alto y potasio bajo, con medicamento drugY |

## Asociación

### d.(0.5) Crear reglas de asociación que permitan determinar patrones en los datos.

Se establece la mínima confianza en 80% y que se creen máximo 100 reglas

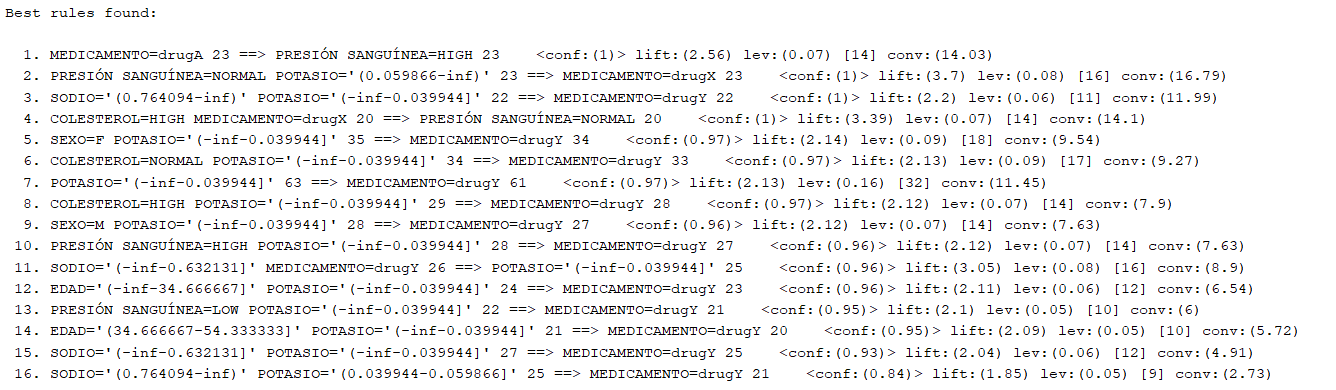


### e.(0.5) Explicar las 3 reglas más relevantes.

Las tres primeras reglas son:

1. *MEDICAMENTO=drugA ==> PRESIÓN SANGUÍNEA=HIGH*: Las personas a las cuales se le receta el medicamento *drugA* son las que tienen presión sanguínea alta.
2. *PRESIÓN SANGUÍNEA=NORMAL POTASIO=(0.0598,inf)==> MEDICAMENTO=drugX*: Las personas cuya presión saguínea es normal y presentan alto nivel de potasio (ya que es el último rango), se les receta el medicamento *drugX*.
3. *SODIO=(0.7640,inf) POTASIO=(-inf,0.0399) ==> MEDICAMENTO=drugY*: Las personas que tienen altos niveles de sodio (ya que es el último rango) y bajos niveles de potasio (están en el primer rango), tienen como receta el medicamento *drugY*.

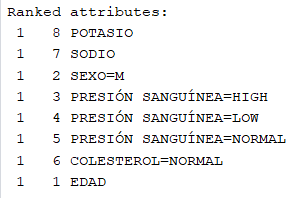
Todas las reglas que se encontraron son las siguientes:



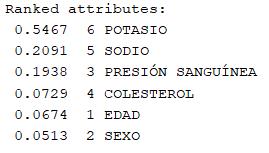
## Selección de Factores

### f.(2.5) Cuáles son las 3 variables que más influyen sobre la variable objetivo de los datos? Aplicar 5 técnicas diferentes.

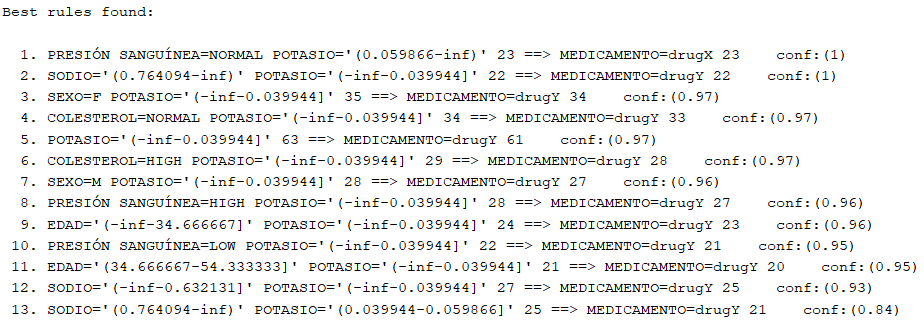
Se hace el análisis por medio de *Componentes principales*, poniendo en *true* el parametro t*ransformBackToOriginal* y dan el siguiente resultado



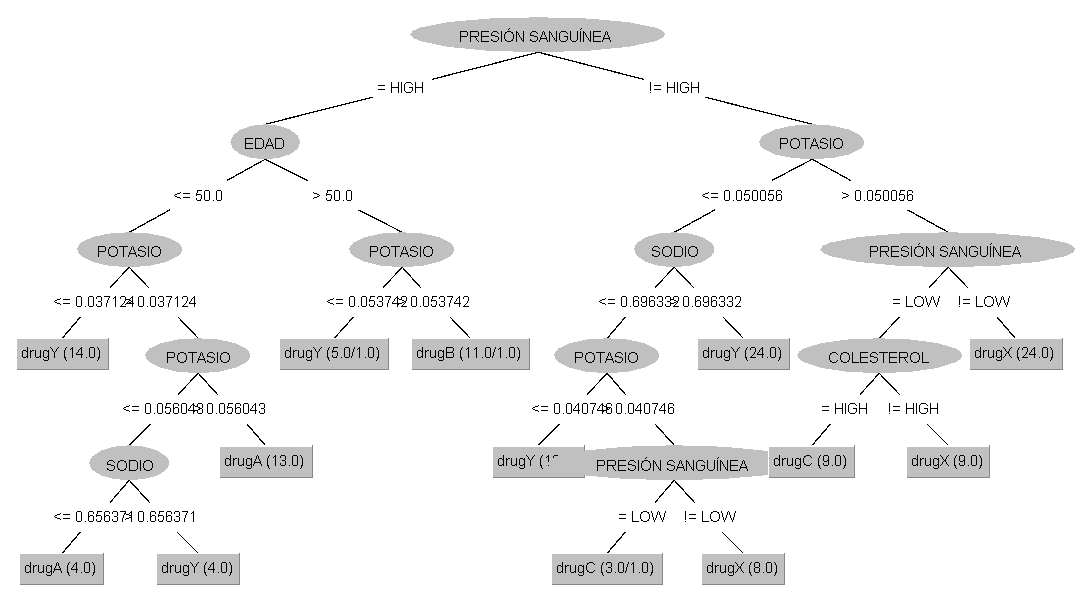
Luego con por medio de las *Correlaciones*



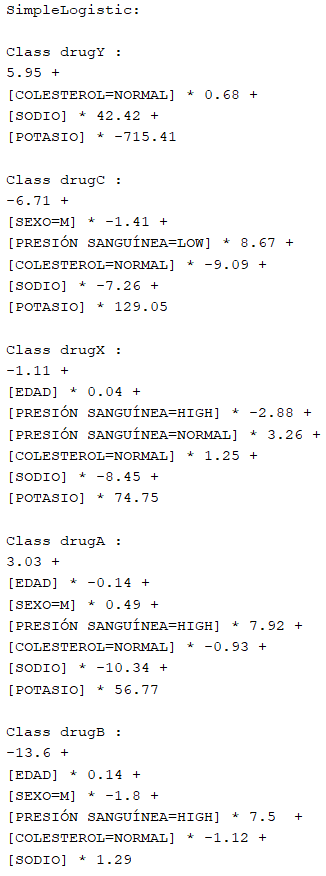
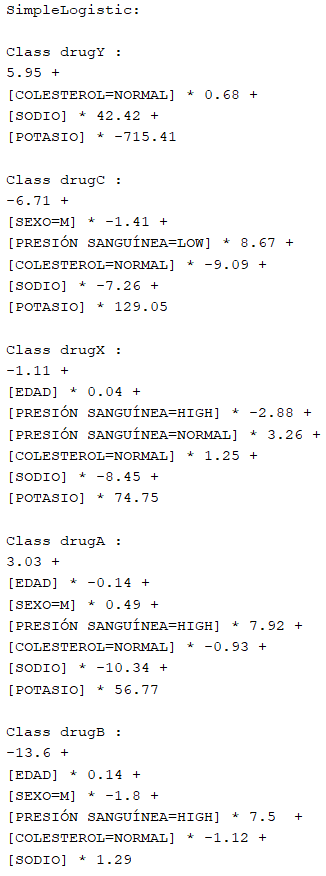
También se realiza por medio de *Apriori* configurando el parametro *car* en *true* y especificando el indice de la variable objetivo *(classIndex)* y se obtienen las siguientes reglas



Además, se realiza por un *árbol de decisiones* y nos da el siguiente grafico



Por último, se realiza la *regresión logística* dando las siguientes ecuaciones:



Según las 5 técnicas se concluye que las 3 variables más relevantes son: ***POTASIO***, ***SODIO*** y ***PRESIÓN SANGUÍNEA***