Tecnológico Nacional de México
Instituto Tecnológico de Tijuana

Subdirección Académica
Departamento de Sistemas y Computación

Semestre:
Febrero – Junio 2021

Carrera:
Ingeniería en Tecnologías de la Información y Comunicaciones

Materia y serie:
Minería de datos BDD-1703TI9A

Unidad a evaluar: Unidad IV

Nombre de la Tarea:
Práctica Evaluatoria - Unidad 3

Nombre del Alumno:
Hernández Negrete Juan Carlos 16212021
Sifuentes Martinez Manuel Javier 17212934

Nombre del docente:
José Christian Romero Hernández

Analysis of data visualization in the Naive Bayes model

Change directory

By default, R Studio comes with a path that is probably not the one used to store the elements to be used, so you must use getwd to obtain the current directory and setwd to assign a new one.

```
getwd ()
setwd ("../Desktop/8th subjects / Data mining / Practices / U4 /")
getwd ()
```

Import and section dataset

Columns 1-2, 3-4 and 1-4 will be compared, for what the columns are specified and in the last line a vector is used to imply that we only want column 1 and 4.

```
dataset = read.csv ('iris.csv')
dt = dataset [1:2]
dt2 = dataset [3:4]
dt3 = dataset [c (1,4)]
```

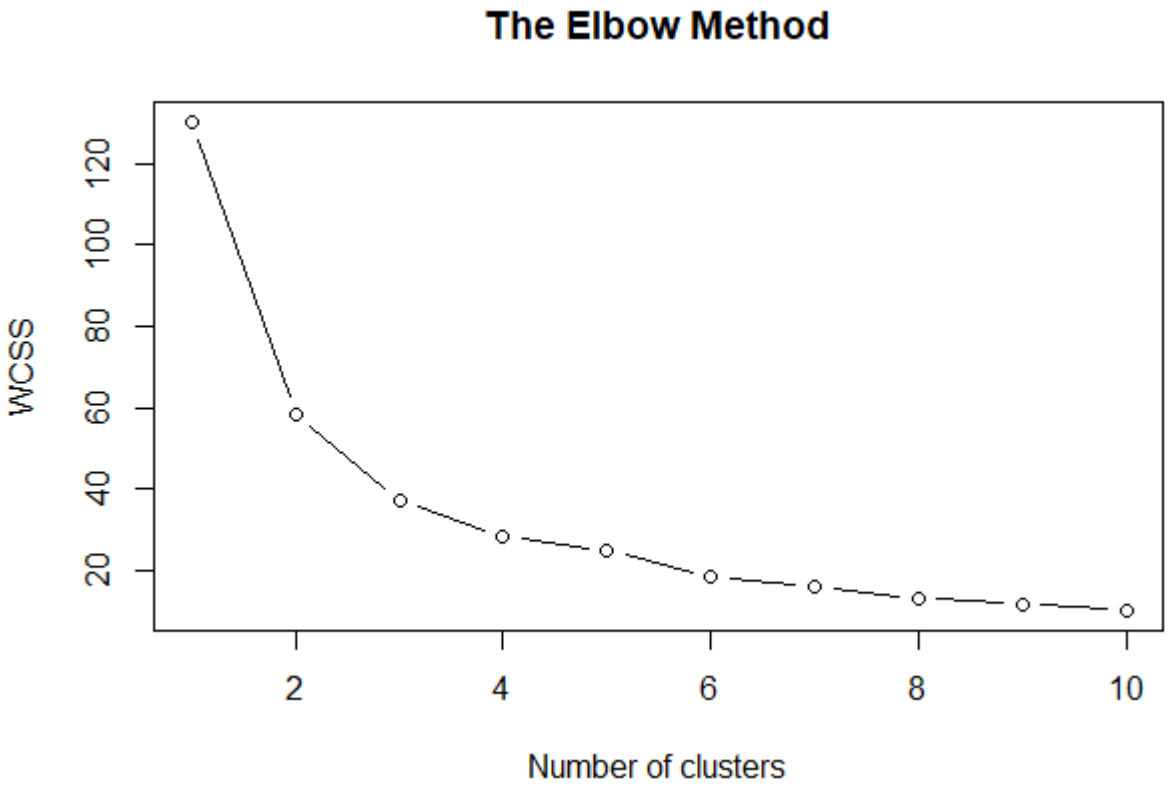
Using the elbow method to find the optimal number of groups

The elbow method is used to find out how many clusters are ideal to give a good representation in Regarding the K means function, and so that the code with the variables in which the columns are stored is not repeated many times, a function is used, and in this way it is only called and the desired variable is sent to it .

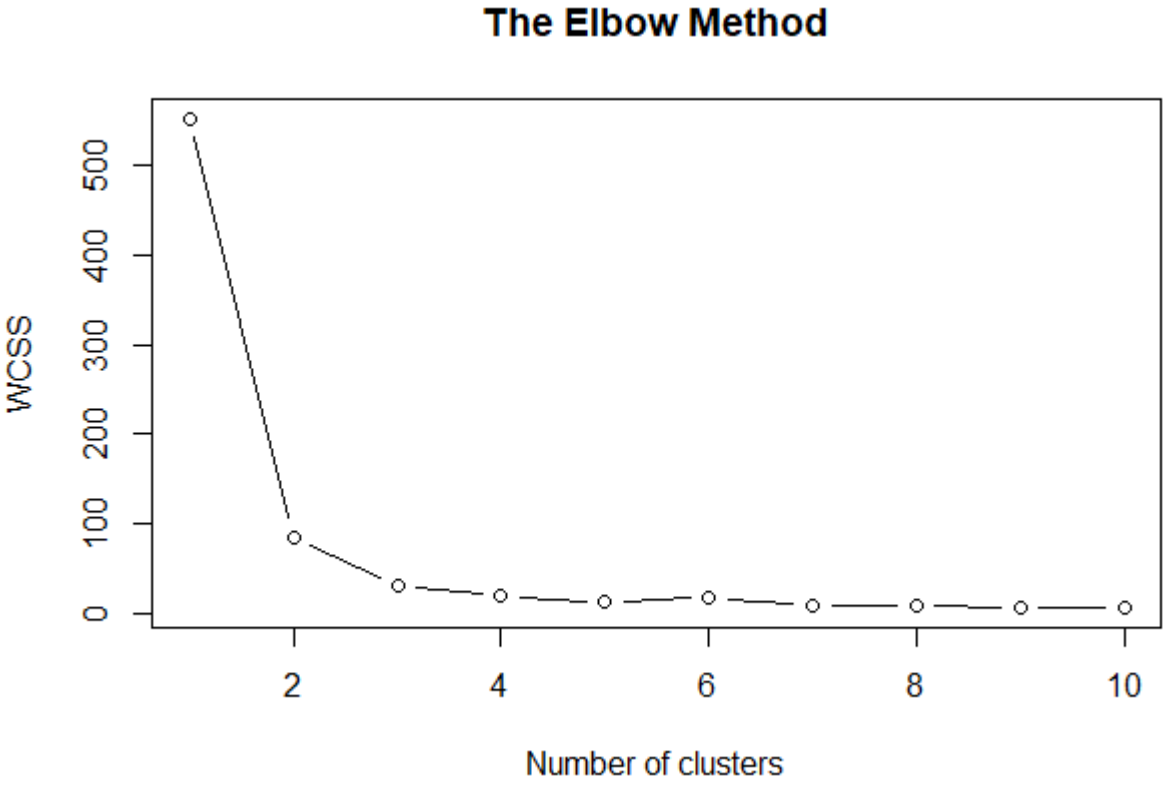
```
TEM <- function(dataset) {
  set.seed (6)
  wcss = vector ()
  for (i in 1:10) wcss [i] = sum (kmeans (dataset, i) $ withinss)
  plot (1:10,
        wcss,
        type = 'b',
        main = paste ('The Elbow Method'),
        xlab = 'Number of clusters',
        ylab = 'WCSS')
}

TEM (dt)
TEM (dt2)
TEM (dt3)
```

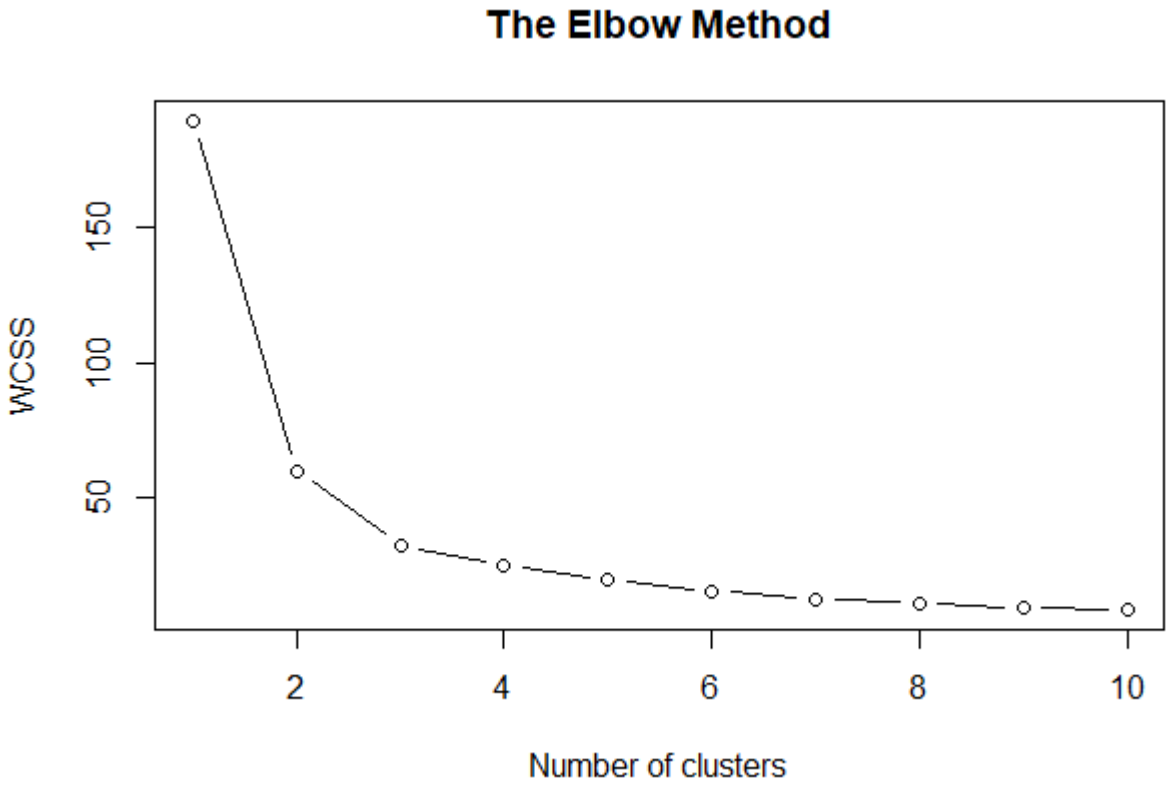
This is the representation of the number of ideal clusters for columns 1 and 2. The way this number of clusters is determined is the point on the graph before the plotted line begins to normalize. In this case it would be number 5.



In this graph it is clearly found that the number of ideal clusters for K means is only 2.



The point before the line begins to normalize is at 4, so this is the number to be used for K means.



Fit K-Means to the data set

With the help of K means, and specifying the clusters, the groups will be drawn with respect to the data. As in a previous step, a function is used to avoid code repetition.

```
Clusters <- function(dataset, cnt) {
  set.seed (29)
  kmeans = kmeans (x = dataset, centers = cnt)
  y_kmeans = kmeans $ cluster
}

ykmeans <- Clusters (dt, 5)
ykmeans2 <- Clusters (dt2, 2)
ykmeans3 <- Clusters (dt3, 5)
```

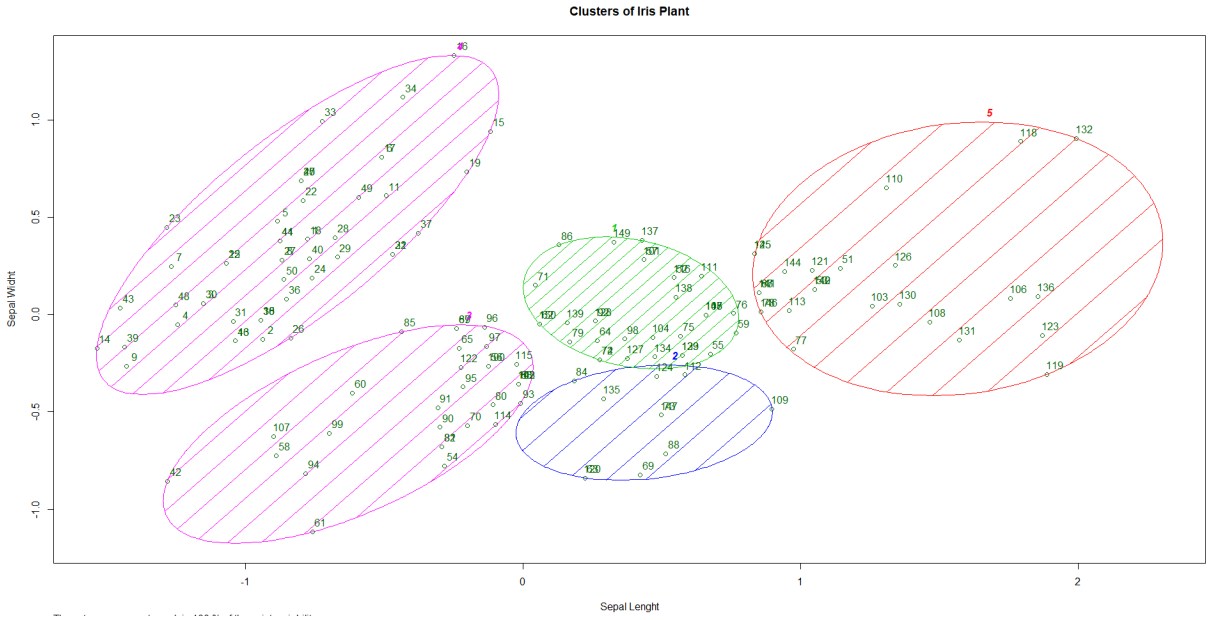
K-means is an unsupervised classification algorithm (clustering) that groups objects into k groups based on their characteristics. Grouping is done by minimizing the sum of distances between each object and the centroid of its group or cluster. Quadratic distance is often used. With this in mind, based on the analysis carried out so far, with the elbow method and the k means function, the different existing groups will be graphed in the selected columns with the help of clusplot.

Viewing the clusters

The cluster library must first be loaded into memory to be able to graph the results. You must specify the data source, the variable where the data of the k means function was stored and then specify the values of the graph display.

```
library(cluster)
clusplot (dt,
          ykmeans,
          lines = 0,
          shade = TRUE,
          color = TRUE,
          labels = 2,
          plotchar = FALSE,
          span = TRUE,
          main = paste ('Clusters of Iris Plant'),
          xlab = 'Sepal Lenght ',
          ylab = ' Sepal Width ')
```

It can be seen that there are 5 different groups in columns 1 and 2. The blue and green ovals have a slight intersection, this means that there is some noise between these 2.

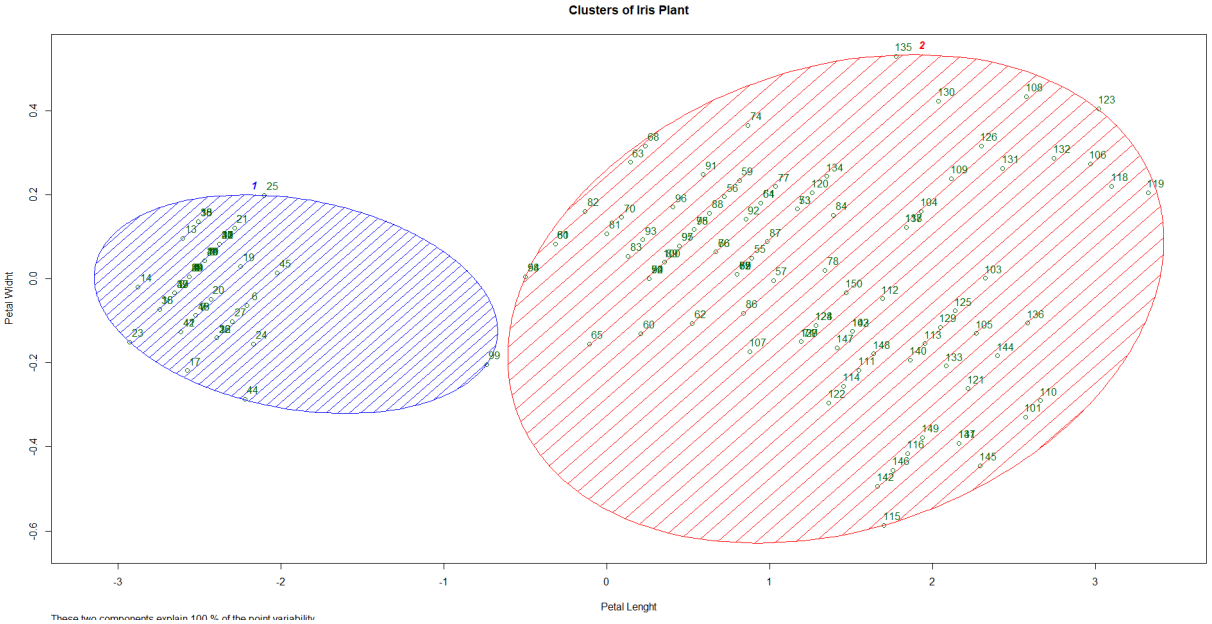


The creation of the graph of columns 3-4 is done in the same way, only changing the data source and the variable of k means.

```
clusplot (dt2,
          ykmeans2,
          lines = 0,
          shade = TRUE,
          color = TRUE,
          labels = 2,
```

```
        plotchar = FALSE,
        span = TRUE,
        main = paste ('Clusters of Iris Plant'),
        xlab = 'Petal Lenght',
        ylab = 'Petal Width')
```

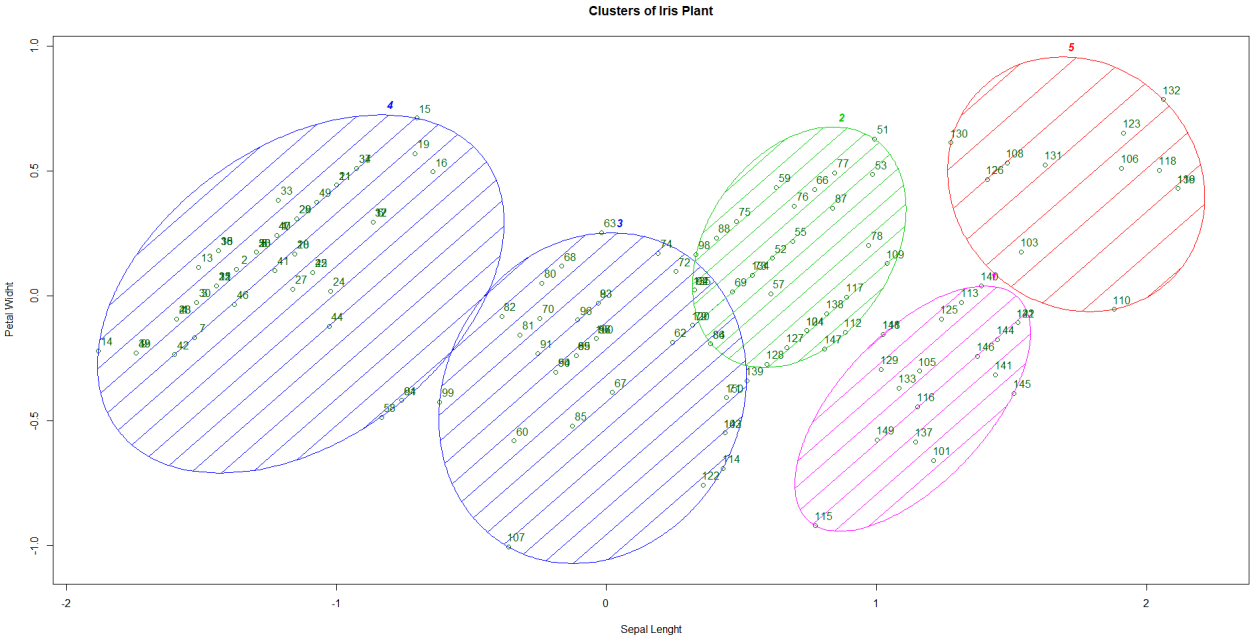
From the result obtained from the elbow method, it can be seen that only with 2 clusters it is possible to represent in a good way types of data sets existing in columns 3 and 4, without generating noise between these 2.



The steps of creating the two previous graphs are followed, changing the dataset and the variable where the k means data is stored.

```
clusplot (dt3,
        ykmeans3,
        lines = 0,
        shade = TRUE,
        color = TRUE,
        labels = 2,
        plotchar = FALSE,
        span = TRUE,
        main = paste ('Clusters of Iris Plant'),
        xlab = 'Sepal Lenght',
        ylab = 'Petal Width')
```

Finally, there is the representation of the different groups of data existing between column 1 and 4. It can be seen that there is some noise between the groups represented.



YouTube video link: https://youtu.be/ezZyQgga3po
GitHub: https://github.com/JuanCarlos-Negrete/Data-Mining/tree/Unit_4/Unit_4/Evaluation