CENG484
Data Mining Assignment 2
Melek Bilgin Tamtürk

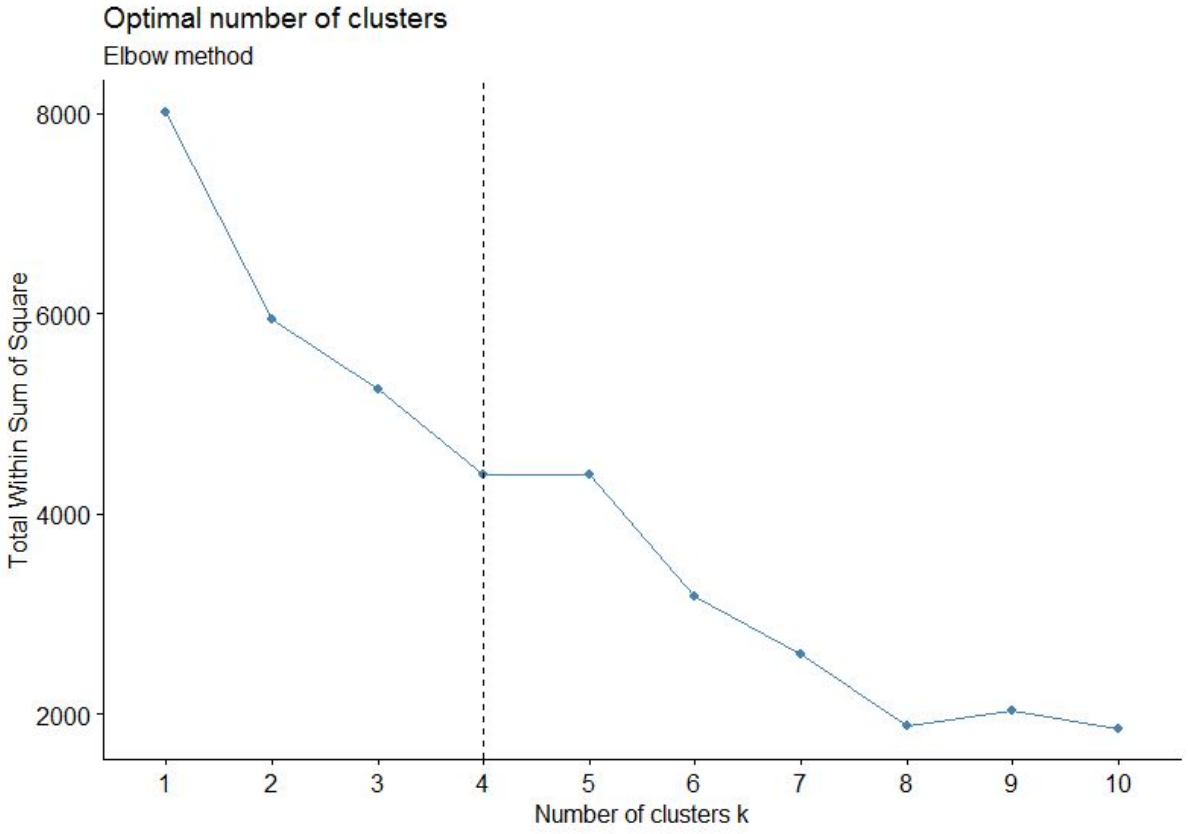
In this assignment, I used RProject since it visualize better than the default interpreter of R. First, I view the Bone Mineral Density dataset to understand the variables better. At first, I prefer to use K-means algorithm which belongs to Clustering method, since it doesn't have test data and validation data I thought I should use unsupervised learning methods.

```
6
7  dataSet<-read.table("spnbmd.csv",header = TRUE, sep =",")
8  dataSet$idnum <- NULL
9  summary(dataSet)
10
11 dataSet$Asian <-ifelse(dataSet$ethnic == "Asian", 1, 0)
12 dataSet$Black <-ifelse(dataSet$ethnic == "Black", 1, 0)
13 dataSet$Hispanic <-ifelse(dataSet$ethnic == "Hispanic", 1, 0)
14 dataSet$White <-ifelse(dataSet$ethnic == "White", 1, 0)
15 dataSet$fem <-ifelse(dataSet$sex == "fem", 1, 0)
16 dataSet$mal <-ifelse(dataSet$sex == "mal", 1, 0)
17
```

First, I uploaded the dataSet to RProject environment. Then, I deleted the idNum attribute as first job to do. Because it doesn't help to cluster the data. Second, I created new columns for my categorical attributes. I converted them into binary attributes.

```
17
18 features <- dataSet
19 features$ethnic <- NULL
20 features$sex <- NULL
21
22 features <- scale(features)
23 # Elbow method
24 fviz_nbclust(features, kmeans, method = "wss") +
25   geom_vline(xintercept = 4, linetype = 2)+
26   labs(subtitle = "Elbow method")
27
```

Then I prepare my feature data set. After that, I use elbow method to decide the number of clusters (k).

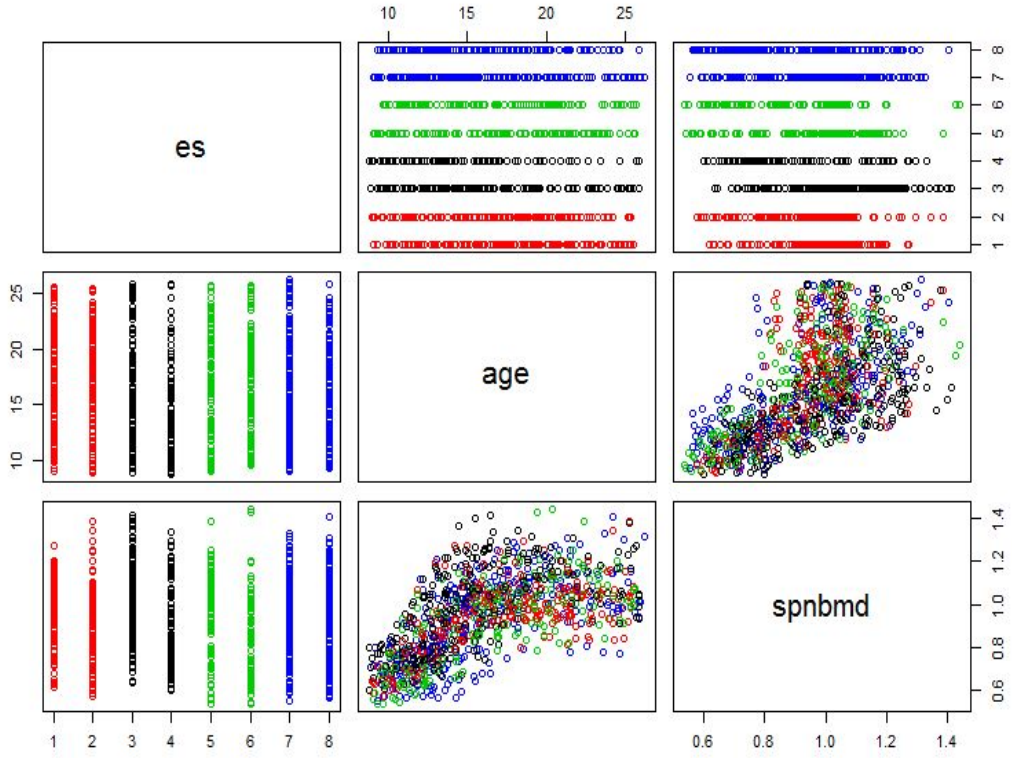


The optimal number of clusters k should be 4 according to this graph.

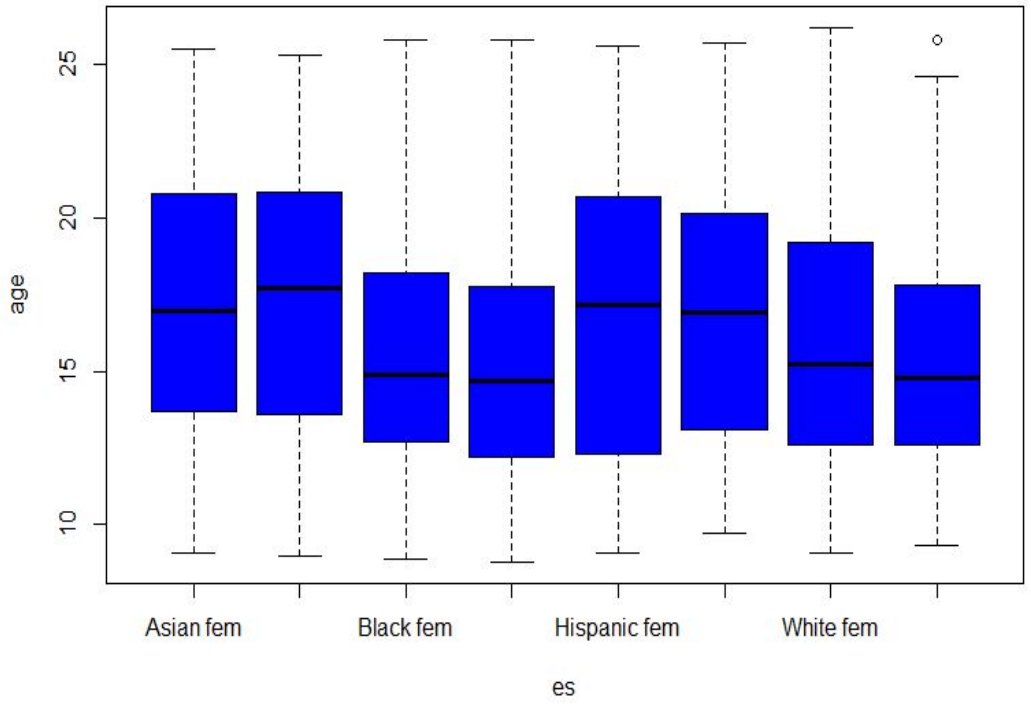
```
27
28  results <- kmeans(features,4)
29  dataSet$es <- paste(dataSet$ethnic,dataSet$sex)
30  dataSet$es <- factor(dataSet$es)
31  table(dataSet$es,results$cluster)
32
33  plot(dataSet[c("es","age","spnbmd")],,col=results$cluster)
34  plot(dataSet[c("es","age")],,col=results$cluster)
35  plot(dataSet[c("es","spnbmd")],col=results$cluster)
36
37
```

I created a new categorical attribute named es where E is for ethnic and S is for sex. Then, I used table method to see if my clusters are reasonable or not.

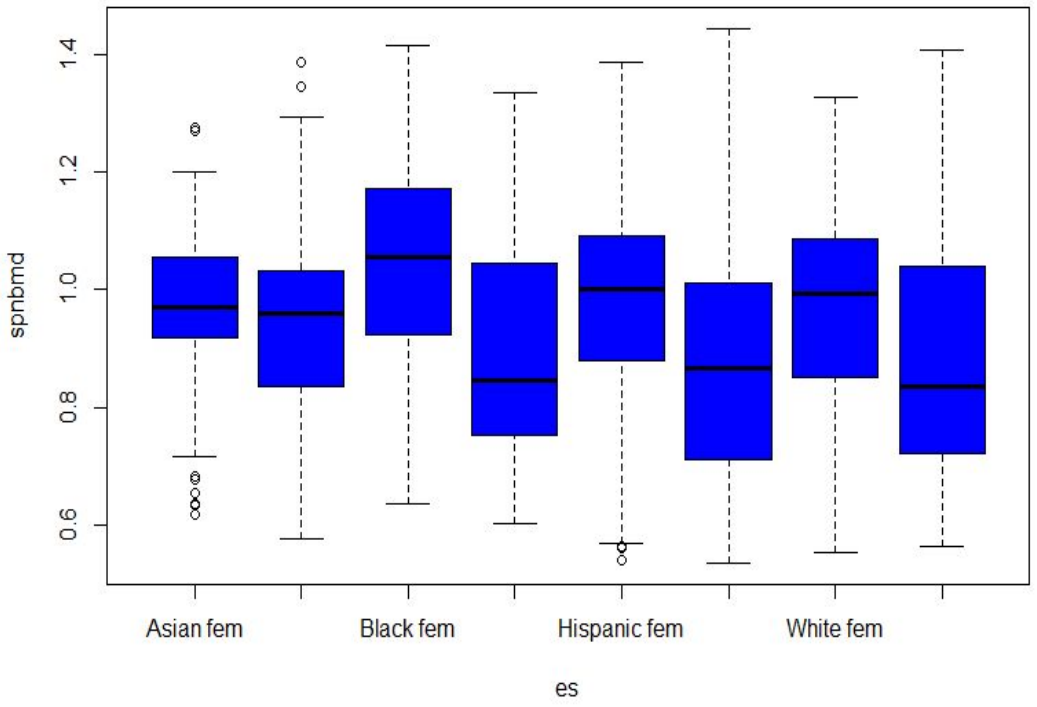
```
              1   2   3   4
Asian fem     0 122   0   0
Asian mal     0 119   0   0
Black fem   149   0   0   0
Black mal    88   0   0   0
Hispanic fem  0   0 110   0
Hispanic mal  0   0  92   0
White fem     0   0   0 166
White mal     0   0   0 157
> |
```



The first plot shows the values of ES, Age and SPNBMD attributes according to the clusters.



The second plot shows the relation between ES and age attributes.



The third plot shows the relationship between ES and spnbmd attributes.

As you se, the graphs are hard to interpret. That's why I try to find a better Then I understand, clustering is better for numerical attributes if you want to visualize these categorical variables and numerical attributes together. It is not the best thing to use clustering for a data set which has two categorical attributes.

Classification is the process of learning a model that elucidate different predetermined classes of data. It is a two-step process, comprised of a **learning** step and a **classification** step. In learning step, a classification model is constructed and classification step the constructed model is used to **prefigure** the class labels for given data.[1]

According to this definition, I thought both sex and ethnic variables can be thought as prefigured class labels. According to someones age and Relative Spinal bone mineral density measurement, we can decide if that person is female or male. Second, We can decide if that person is Asian,White,Hispanic or Black.

After failure of clustering, I used Decision Trees which belongs to Classification. Why I prefer Decision Trees?

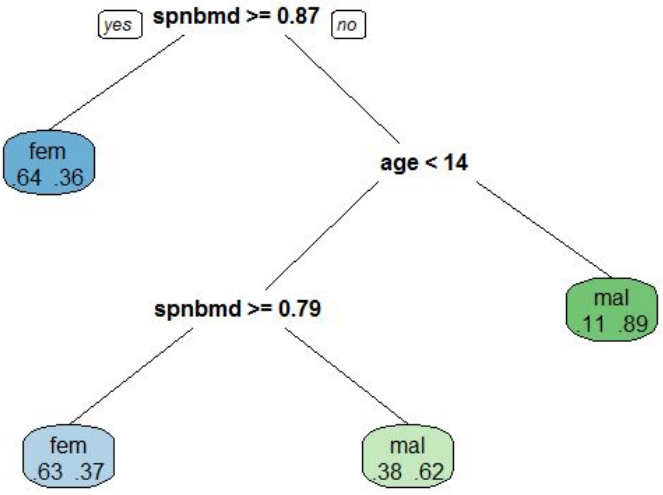
1. It implicitly performs variable screening
2. It doesn't require a lot of effort for data preparation
3. It doesn't require any assumptions between of linearity in the data.
4. It is easy to interpret and explain the results

```
setwd("C:/Users/user/Documents/OKUL/ÖDEVLER/18-19/Data Mining")
install.packages("rpart")
library(rpart)
library(rpart.plot)
dataSet<-read.table("spnbmd.csv",header = TRUE, sep =",")
dataSet$idnum <- NULL #to decrease the dimensionality and unnecessary attributes
dataWithoutEthnic <- dataSet
dataWithoutEthnic$ethnic<-NULL

decisionTree <- rpart(sex~age+spnbmd,data=dataWithoutEthnic)
prp(decisionTree, extra = 4,  box.palette = "auto")
```

I started to implementation with installing package named rpart which helps to plot decision trees. I prefer to use this specific package because it was the most popular and also designable one. I try to use ctree function from party package, but the visualisation of the output tree seemed more complex than the visualisation of otuput tree from prp, that's why I didn't want to use it.

To visualize the relation between only sex, Relative Spinal bone mineral density measurement and age, I create a new data set named dataWithoutEthnic. Then I used rpart function to build the decision tree. After that, I used prp function to visualise the tree. I used prp function rather than rpart.plot because prp was also more understandable and designable than rpart.plot function.

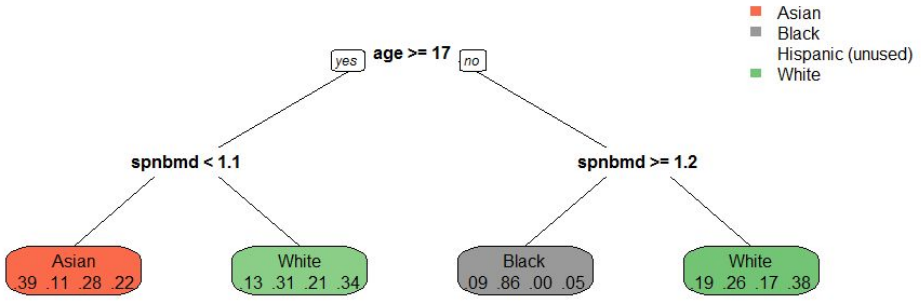


This tree shows how bone density and age is related with sex. According to this tree we can say that, If someone's bone density is equal to or greater than 0.87, then that person is %64 chance of being female. If not, we move to another branch. If someone's spnbmd is less than 0.87 and age is equal to or greater than 14, then that person is %89 chance of being male. If someone's spnbmd is less than 0.87 and equal to or greater than 0.79 and age is less than 14 ,then that person is %63 chance of being female. If someone's spnbmd is less than 0.79 and age is less than 14, then that person is %62 chance of being female.

As a brief result, we can say that mostly female has greater bone density than male independent from age.

```
12  decisionTree2 <- rpart(ethnic~spnbmd+age+sex,data=dataSet)
13  prp(decisionTree2,extra = 4, box.palette = "auto")
```

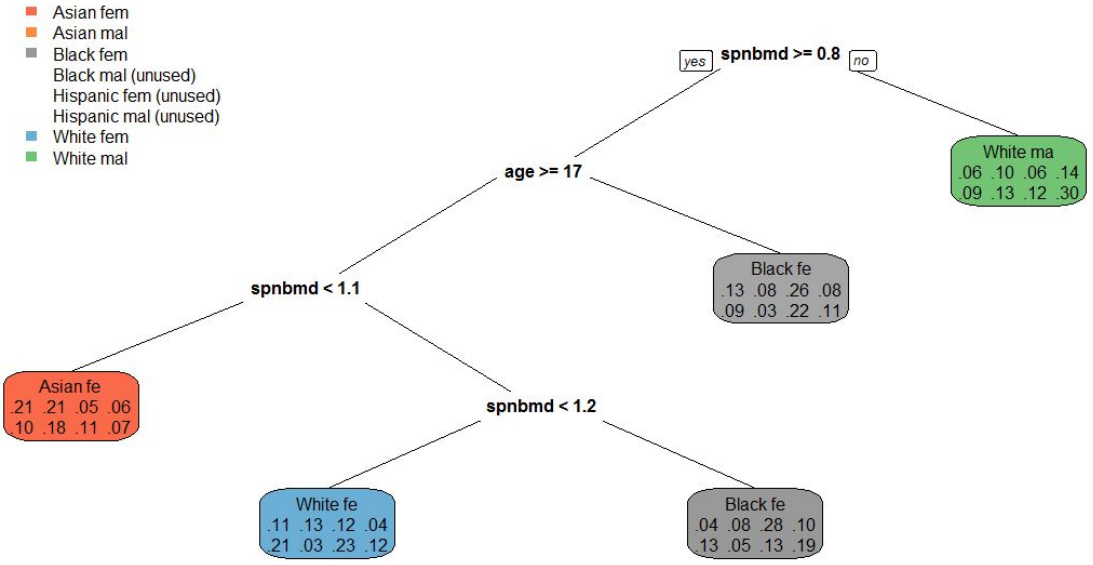
As a second decision tree, I prefer to analyse the relation between ethnic, spnbmd, age and sex.



As seen in the tree, If someone's age is equal to or greater than 17 and spnbmd is less than 1.1, then that person is %39 chance of being Asian. If someone's age is equal to or greater than 17 and spnbmd is equal to or greater than 1.1, that person is %34 chance of being White. If someone's age is less than 17 and spnbmd is equal to or greater than 1.2, then that person is %86 chance of being Black. If someone's age is less than 17 and spnbmd is equal is less than 12, then that person is %38 chance of being White.

```
15  dataSet$es <- paste(dataSet$ethnic,dataSet$sex)
16  dataSet$es <- factor(dataSet$es)
17  dataWithAdditionalClassEs <- dataSet
18  dataWithAdditionalClassEs$ethnic <- NULL
19  dataWithAdditionalClassEs$sex <- NULL
20
21  decisionTree3 <- rpart(es~spnbmd+age,data=dataWithAdditionalClassEs)
22  prp(decisionTree3,extra = 4, box.palette = "auto")
```

Although I can see the ethnicity results from the second tree, I can not decide the relation between etnicity and sex. To understans the relation better, I create another categorical attribute named es where E is for ethnic and S is for sex , for my data set. Then as third decision tree, I prefer to analyse the relation between es, age and spnbmd.



As you can see from the tree, it shows the relation of all attributes.

If someone's spnbmd is less than 0.8 than, that person is %30 chance of being White Male. If someone's spnbmd is greater than or equal to 0.8 and age is less than 17, then that person is %26 chance of being Black female. If someone's spnbmd is greater than or equal to 0.8 and less than 1.1 and age is greater than or equal to 17, then that person is %21 chance of being Asian female or Asian male. If someone's spnbmd is greater than or equal to 1.1 and less than 1.2 and age is greater than or equal to 17, then that person is %23 chance of being White female. If someone's spnbmd is greater than or equal to 1.2 and age is greater than or equal to 17, then that person is %28 chance of being Black female.

If we want to get a brief result, we can definitely say that mostly White Male has the lower bone density, independent from the age.

HOW CAN THE TREE BE ENHANCED?

As you can see from the tree some of the classes seemed as unused(Hispanic female, Hispanic Male, Black male), I think the reason of that is poorness of the data set. ES attribute has 8 different levels and I think the data set is not enough to label all classes because of their probability is low if we compare with the attributes which has less levels. Although we can see the probability of each classes, it is not shown as nodes. So, if we find larger data set, other classes can be shown in the tree.

REFERENCES

[1]Tech Differences. (2019). *Difference Between Classification and Clustering*. [online] Available at: https://techdifferences.com/difference-between-classification-and-clustering.html [Accessed 9 Jan. 2019].