ANALYSIS ON ALZHEIMER

GEETHANJALI E R

MA335 PROJECT

DATE : 21/06/2023

REGISTRATION NUMBER: 2211420

Abstract

The aim of the project is to find relationship between the characteristics of Alzheimer's disease from the given data set.

Alzheimer's disease is a progressive neurodegenerative disease which affects the brain, leading to decline in memory, thinking skills, and overall cognitive functions. Alzheimer disease develops slowly and worsens over time. The exact cause of Alzheimer is not fully understood, but it is believed to involve a complex interplay of genetics, environmental and lifestyle factors. Here in the data set it's been provided with some test results from which we can understand the severity of dementia, also which cognitive domain is mostly affected by the dementia [mostly the memory is the most affected one than any other].

To find the relationship between these characteristics we will be using R programming for implementing the modelling and experimental techniques, for example linear regressions and logistic regressions.

Some of the characteristics given in the data set includes Mini mental state examination [MMSE], Clinical dementia rating 1 of 2 [CDR], Estimated total intracranial volume [eTIV], Normalize whole brain volume [nWBV], Atlas scaling factor [ASF].

Keywords:

Title : Analyzing various characteristics of demented or non demented

INTRODUCTION:

Alzheimer is disease which affects the brain leads to destruction of memory, thinking skills, and the ability to carry out the daily tasks.

Here for the assignment, we are provided with a data set to be used to investigate the relationship between the characteristics of Alzheimer with the diagnosis. The characteristics given in the data set includes some tests or assessment's result which is used to determine whether a person has dementia or not, Age of the persons, Socioeconomic status, and brain volumes factors. The two tests given in the data are Mini mental state examination [MMSE] and Clinical dementia rating [CDR].

MMSE is an assessment to assess the cognitive impairment and provide a brief of a person's mental state. MMSE consists of questions and tasks to evaluate six cognitive Domains. Each section is scored and the scores are summed to give a total out of 30.

CDR is a tool for assessing the severity of dementia thus provides a structured method for staging dementia based on the persons cognitive and functional abilities. Each domain is rated on a scale 0 to 3. Once scores are assigned for each domain an overall CDR score is determined based on highest score among the domains.

The two brain volume factors are "Normalize whole brain volume" [nwbv] and "Estimated Total Intracranial Volume" [eTIV].

eTIV is a measurement used in neuroimaging studies to estimate the total volume of the brain and other intracranial structures.

Nwbv involves adjusting the measured brain volume to account for individual differences in overall brain size or head size.

PRELIMINARY ANALYSIS :

. Analyse using descriptive statistics(question number 1)

Here lets see small portion of our data set by using the R function head():

R code for this :

head(mydata)

The Output:

```
> head(mydata)
        Group M.F Age EDUC SES MMSE CDR eTIV  nWBV   ASF
1 Nondemented   1  87   14   2   27 0.0 1987 0.696 0.883
2 Nondemented   1  88   14   2   30 0.0 2004 0.681 0.876
6 Nondemented   0  88   18   3   28 0.0 1215 0.710 1.444
7 Nondemented   0  90   18   3   27 0.0 1200 0.718 1.462
8 Nondemented   1  80   12   4   28 0.0 1689 0.712 1.039
9 Nondemented   1  83   12   4   29 0.5 1701 0.711 1.032
>
```

This gives you what all characteristics are there in the data set.

Now we can get summary of the our data using the summary() in R.

The code for this is :

summary_table <- summary(mydata)

summary_table

The output :

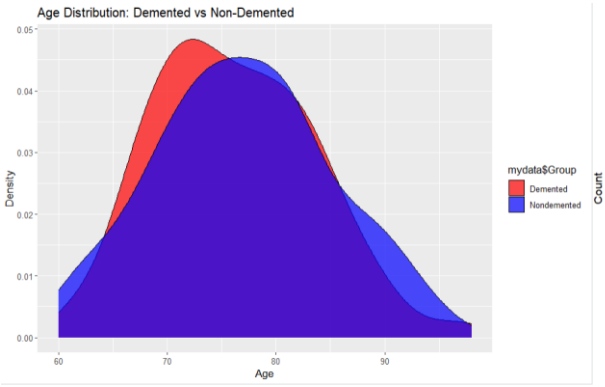
```
> summary_table <- summary(mydata)
> summary_table
        Group          M.F              Age             EDUC            SES             MMSE
 Demented   :127   Min.   :0.0000   Min.   :60.00   Min.   : 6.00   Min.   :1.000   Min.   : 4.00
 Nondemented:190   1st Qu.:0.0000   1st Qu.:71.00   1st Qu.:12.00   1st Qu.:2.000   1st Qu.:27.00
                   Median :0.0000   Median :76.00   Median :15.00   Median :2.000   Median :29.00
                   Mean   :0.4322   Mean   :76.72   Mean   :14.62   Mean   :2.546   Mean   :27.26
                   3rd Qu.:1.0000   3rd Qu.:82.00   3rd Qu.:16.00   3rd Qu.:3.000   3rd Qu.:30.00
                   Max.   :1.0000   Max.   :98.00   Max.   :23.00   Max.   :5.000   Max.   :30.00
      CDR              eTIV           nWBV             ASF
 Min.   :0.0000   Min.   :1106   Min.   :0.6440   Min.   :0.876
 1st Qu.:0.0000   1st Qu.:1358   1st Qu.:0.7000   1st Qu.:1.098
 Median :0.0000   Median :1476   Median :0.7320   Median :1.189
 Mean   :0.2729   Mean   :1494   Mean   :0.7306   Mean   :1.192
 3rd Qu.:0.5000   3rd Qu.:1599   3rd Qu.:0.7570   3rd Qu.:1.293
 Max.   :2.0000   Max.   :2004   Max.   :0.8370   Max.   :1.587
>
```

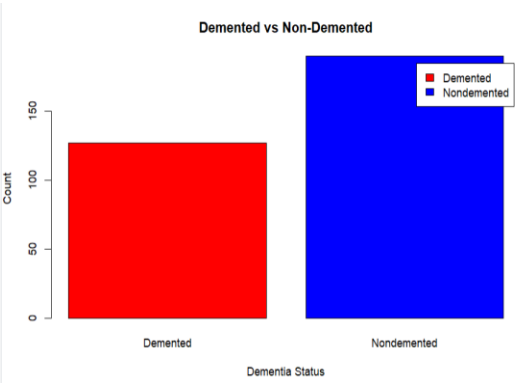
Here we can see that we are getting a summary for each column with the values min, mean, median and max etc. From the table we get to know that there is 127 person's are having dementia and the rest 190 people do not have demetia.

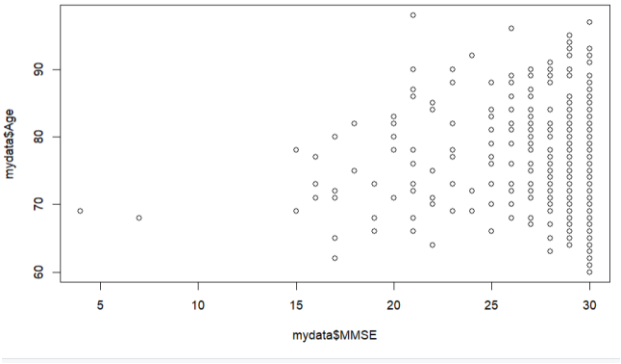
Also this says that the given data belongs to the age group 60 to 98 (assume that as 100).

Also the high mean value of MMSE is says that most of the person are having high scores for that test which suggests to have dementia

Now we can see some Graphical representation the data set:







The first graph is a density plot in which we ploted Age and the total number demented and non demented people in the given age group.

The second graph is simply a bar graph in which the number of demented and non demented is been visualized in which non demented is high.

And the third is a scatter plot which shows the distribution of MMSE along with age.

ANALYSIS:

- Implementing The clustering Algorithms:(question number 2)

Here I have taken the K-means clustering. It is popular unsupervised machine learning algorithm used to partition a dataset into distinct groups or cluster. The main goal of K-means clustering is to group similar data points together while maximizing the dissimilarity between different groups

R code for K-means :

k1 <- kmeans(data2, centers = 3, nstart = 25)

k1

The output:[small portion of the output]

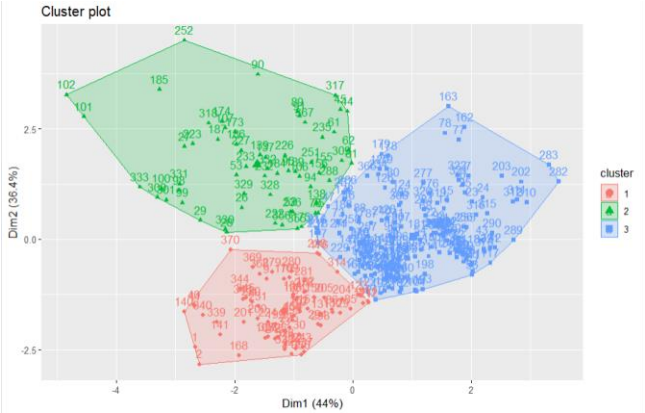
```
> k1 <- kmeans(data2, centers = 3, nstart = 25)
> k1
K-means clustering with 3 clusters of sizes 72, 63, 182

Cluster means:
        MMSE        CDR        eTIV       nWBV        ASF
1  0.4357720 -0.3687781  1.38993430 -0.2299850 -1.2972330
2 -1.6011657  1.4873964  0.03538423 -0.8696635 -0.0909399
3  0.3818563 -0.3689778 -0.56211251  0.3920204  0.5446703
```

In this output we have got the cluster means for all the datapoints from each cluster. Cluster means is the average values of each variable within each cluster.

From the means obtained the first two clusters has higher MMSE values, lower CDR values, significantly higher eTIV values, slightly lower values for both nwBV and ASF . These cluster indicated the values for the characteristics of the non-demented persons. As they show high value for MMSE indicates good scores for cognitive domains and low rate of CDR show less severity.

The third cluster is having a relatively lower MMSE score, higher CDR score, close to average eTIV value. This cluster indicates the characteristics belongs to demented people.

The Graphical representation of the clusters:



- Fit a logistic Regression model: (question number 3)

In statistics, the **Logistic Regression model** is a widely used statistical model which is primarily used for classification purposes. It means that given a set of observations, Logistic Regression algorithm helps us to classify these observations into two or more discrete classes. So, the target variable is discrete in nature.

R Code for logistic regression:

```
logit_model <- glm(mydata$Group ~ mydata$Age + mydata$M.F + mydata$MMSE + mydata$CDR, data = mydata, family = "binomial")
```

R code to get the summary of the model:

```
summary(logit_model)
```

The output:

```
> summary(logit_model)

Call:
glm(formula = mydata$Group ~ mydata$Age + mydata$M.F + mydata$MMSE +
    mydata$CDR, family = "binomial", data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.70840  -0.00001   0.00000   0.00001   2.59963

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.5618  6499.1474   0.000    1.000
mydata$Age     0.1504     0.1457   1.032    0.302
mydata$M.F    17.8766  6854.8051   0.003    0.998
mydata$MMSE    0.4459     0.4235   1.053    0.292
mydata$CDR   -89.7636 18891.9867  -0.005    0.996

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 426.851  on 316  degrees of freedom
Residual deviance:  15.168  on 312  degrees of freedom
AIC: 25.168

Number of Fisher Scoring iterations: 23
```

From the output we can see the Deviance Residuals which gives us an idea about how much the observed response value are deviated from the predicted probabilities .From the min and max values it shows that at least one observation is lower than the observed response value and one observation is higher that the observed response value

The coefficient estimates for mydata$Age, mydata$M.FM, and mydata$MMSE are positive, indicating that an increase in these variables is associated with an increase in the log-odds of belonging to the group. However none of these coefficients are statistically significant as indicated by their large standard errors and p-value >0.05. Therefore, there is no strong evidence to suggest a significant relationship between these variables and the log-odds of belonging to the group.

- Feature selection: (question number 4)

The feature selection method I used here is Wrapper backward selection.

The backward selection is a wrapper method used in feature selection to determine the most relevant subset of features for a predictive model. It starts with an initial set of features and iteratively removes one features at a time until a stopping criterion is met.

The R code :

```
feature_backwards_model <- stepAIC(logit_model, direction = "backward")

summary(feature_backwards_model)
```

The output:

```
> summary(feature_backwards_model)

Call:
glm(formula = mydata$Group ~ mydata$Age + mydata$MMSE + mydata$CDR,
    family = "binomial", data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.76898  -0.03159   0.00000   0.00001   2.71155

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.7843  4018.1094  -0.002    0.999
mydata$Age     0.1955     0.1466   1.333    0.182
mydata$MMSE    0.5354     0.4198   1.275    0.202
mydata$CDR   -54.3123  8036.1367  -0.007    0.995

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 426.851  on 316  degrees of freedom
Residual deviance:  15.862  on 313  degrees of freedom
AIC: 23.862

Number of Fisher Scoring iterations: 22
```

From this output the important features has been selected as the Age, MMSE, CDR

CONCLUSION:

From the analysis we can say that people after 60 are more prone to have dementia.

The age and the results of the two tests are very important factors in determining whether a patient is demented or not.

References:

Moodle labs From MA335

https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial

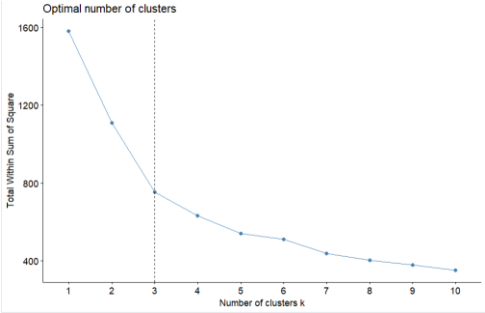
https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

APPENDIX:

R CODES :

```
#the packages relevant to do the analysis
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(ggplot2) #graph ploting
library(MASS) #doing feature selection
#read the data
mydata <- read.csv("C:/Users/HP/Documents/Geethu/project data.csv")
#Appropriate correcting and cleaning of the data
mydata$M.F <- ifelse(mydata$M.F == "M",1,0)
mydata <- mydata[mydata$Group != "Converted",]
mydata <- na.omit(mydata)
mydata$Group <- as.factor(mydata$Group)
#density graph
ggplot(mydata, aes(x = mydata$Age, fill = mydata$Group)) +
geom_density(alpha = 0.7) +
 labs(title = "Age Distribution: Demented vs Non-Demented", x = "Age", y = "Density")
+
 scale_fill_manual(values = c("red", "blue"))
```

```
#bar graph
graph2<-barplot(counts, main = "Demented vs Non-Demented", xlab = "Dementia
Status", ylab = "Count", col = c("red", "blue"), legend = TRUE)
#scatterplot
graph3<-plot(mydata$MMSE,mydata$Age)
#optimal number of cluster
optimal_clusters <- fviz_nbclust(data2, kmeans, method = "wss", k.max = 10)
optimal_clusters + geom_vline(xintercept = 3, linetype = 2)
```




```
data_cluster <- mydata[,6:10] #data to be clusterd
data2<- scale(data_cluster)#before clustering scaling is important
set.seed(123)
```