
# MA335 Final Project

Name: Rahul Kithalamane Basavaraj

Date: 20/06/2023

2212064

## Contents:



## Word Count: 1537 words

## Abstract:

The given dataset consists of various characteristics of Alzheimer's. The variables included in the dataset are "Group" (Group of the diagnosis (Nondemented, Demented, Other)), "M/F" (Gender), "Age" (Age), "EDUC" (Year of education), "SES" (Socioeconomic Status (1-5, 1-low, 5-high)), "MMSE" (Mini mental state examination), "CDR" (Clinical dementia rating), "eTIV" (Estimated total intracranial volume), "nWBV" (Normalize whole brain volume), and "ASF" (Atlas scaling factor). In summary, the given dataset provides information on various factors related to Alzheimer's diagnosis, including demographic characteristics (gender, age, education, socioeconomic status), cognitive assessments (MMSE, CDR), and neuroimaging measures (eTIV, nWBV, ASF). These data can be further analysed and modelled to the relationship between these variables and the diagnosis of Alzheimer's.

## Introduction:

The dataset contains information related to the diagnosis and characteristics of individuals with Alzheimer's. It is a neurological disorder that affects cognitive functions like memory, behaviour, and thinking. Early detection can help in intervention of that individual with an improved patient care. The variables present in the dataset also provides insights into the diagnosis and characteristics of individuals. These variables include "Group," which represents the diagnosis group (Nondemented, Demented, and Converted), "M/F" indicating the gender of an individual, "Age" representing the age of an individual, "EDUC" indicating the number of years of education completed by the individuals, "SES" representing socio-economic status on a scale from 1 to 5, "MMSE" which stands for Mini Mental State Examination to assesses

cognitive function, "CDR" which stands for Clinical Dementia Rating to assesses the severity of dementia, "eTIV" representing the estimated total intracranial volume, "nWBV" indicating the normalized whole brain volume, and "ASF" which represents the Atlas scaling factor.

Through this dataset, we can get information about demographic characteristics (such as age and gender), educational background, socioeconomic status, cognitive function, and neuroimaging measures. On further statistical analysis and modelling techniques, we can uncover patterns, identify risk factors, and gain insights into the various factors, diagnosis, and complex nature of Alzheimer's.

## Preliminary Analysis:

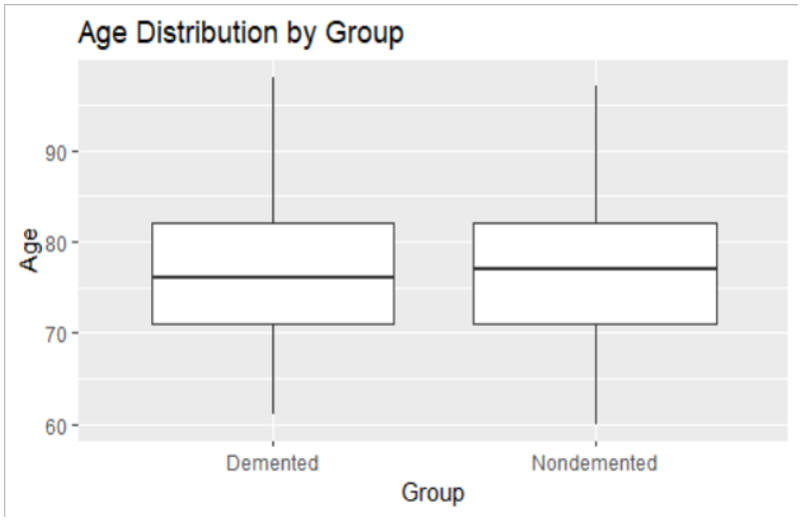
The preliminary analysis shows the length of the individuals in this dataset is 317. The dataset includes groups of individuals classified as Nondemented, Demented and Converted, with some missing values for the SES and MMSE variable. The gender distribution consists of males and females. The age ranges from 60 to 98, with varying levels of education range from 6 to 23. The MMSE scores ranges from 4 to 30, indicating relatively intact cognitive function. The CDR values range from 0 to 2, with some individuals also having a CDR of 0.5 and 1 which indicates varying levels of dementia severity. The eTIV values ranges from 1106 to 2004, while the normalized whole brain volume (nWBV) ranges from 0.644 to 0.837. The ASF values range from 0.876 to 1.587, indicating the scaling factor applied to brain images.

The findings and insights derived from utilizing statistical analysis techniques, machine learning algorithms, and predictive modelling on this dataset can contribute to advancing research in the field of Alzheimer's, enabling healthcare professionals to make informed decisions regarding diagnosis, treatment, and patient management. Additionally, they aid in the early identification of individuals at risk of developing dementia, potentially leading to interventions that could delay or mitigate the progression of the disease.

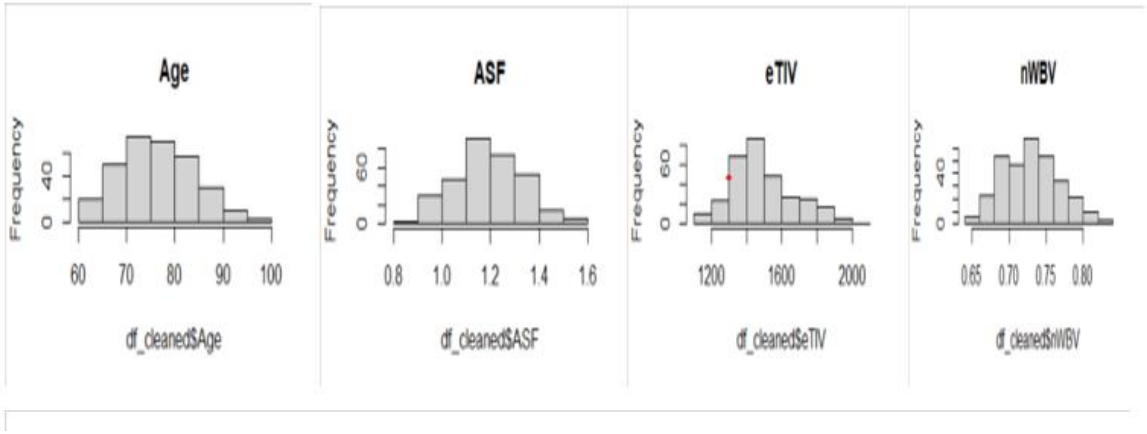
## Analysis:

The boxplot here displays the IQR, median, and outliers. It also provides insights into the spread of each variable present in the dataset. The boxplot for eTIV represents the distribution of estimated intercranial volume. It provides insights into the range and variability of brain sizes among the participants. Here it ranges from 1106 to 2004, the median being 1450. The median age group of individuals nondemented are slightly higher than those who are demented. Looking into the distribution of brain size estimates it indicates majority of brain sizes are present in this region and the average age group for these brain sizes estimates are present between 65 to 80 years old. The histogram chart of nWBV, suggest that the brain volume is more or less equally distributed in this region.

## Boxplot




## Histogram




The intercranial volume in individuals is concentrated in 1200 to 1600 range. The scatterplot analysed below further helps in visualizing the distribution of data points, which in turn identifies the patterns, trends or any kind of correlations between variables. Here they provide valuable insights into the relationships between one variable onto the another.

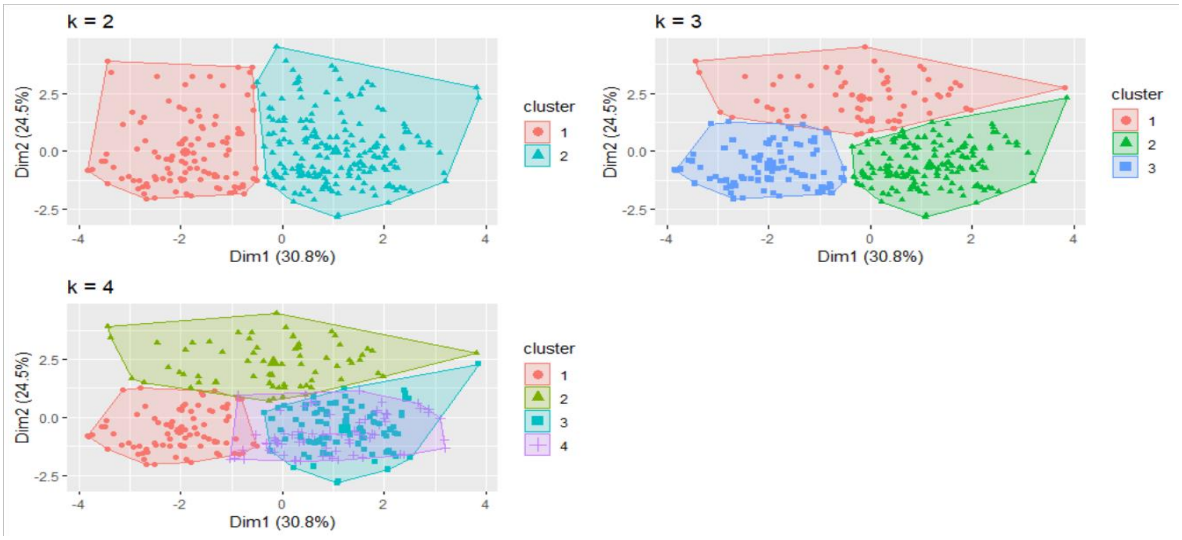
In k-means, clustering is performed for different values of k using algorithms which are very powerful. The total number of clusters are k to 3. We then used the k-means () function to perform K-means clustering on the dataset project data. Here we use the factoextra package for visualizations of the clustering results from fviz_cluster function. Separate scatter plots are used for generating results where the points are coloured based on their assigned clusters. These plots are then arranged in a grid layout for easy comparison.

The k-means algorithm is used to create clusters in a manner in which the variations are minimised by grouping similar data points together. The scatter plot visualizes the clustering results in colour coded format. We can here observe how the algorithm has grouped similar

data points together. This well-separated and distinct clusters in the dataset have successfully identified meaningful patterns or groups after we tried for different values of k and decided that k=3 is the best.

In summary, the K-means clustering results provide an initial grouping of the data points based on their similarity. The scatter plot helps visualize these clusters and provides insights into the structure of the data. Further analysis and interpretation are necessary to determine the practical significance of the clusters and their relevance to the underlying problem. While performing hierarchical clustering we chose average as the best linkage method.

**K-Means**




## Discussion:

The model provides several pieces of information based on the output of logistic regression. The deviance residuals are the differences between the predicted and observed values. Since this value is very small, it indicates a fit model to the data. The coefficients here represent the estimated log-odds of the response variable (Group) based on the predictor variables (Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF, M_F). The null hypothesis is tested using standard error, z-value, and p-value between the predictor and the response variable. If the intercept coefficient and standard error is very large it indicates issues with the model. In this we set the dispersion parameter to 1 for the binomial family. If the predicted value is positive and is greater than 0.5 we assign the label as non demented else demented. The accuracy of the predicted labels comparing them with actual labels we get the resultant mean equal to 1. The null deviance represents deviance only when the intercept (null model) is considered. Here the residual deviance is extremely close to zero which suggest an excellent model fit.

The Akaike Information Criterion (AIC) measures the goodness of fit considering the number of parameters. Lower AIC values indicates a better value between model complexity and fit. The produced model represents a good fit between the predictor variables (Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF, M_F) and the response variable (Group). Since the value of AIC is 20 along with a very small residual deviance, it indicates that the model is able to explain the variability in the data very well.

However, if the large intercept coefficient has a very large standard error, it suggests potential issues of collinearity in the model. It is very important to carefully examine which model we are assuming, check for its multicollinearity among the predictors, and consider potential transformations or variable selection techniques so as to improve the model's performance.

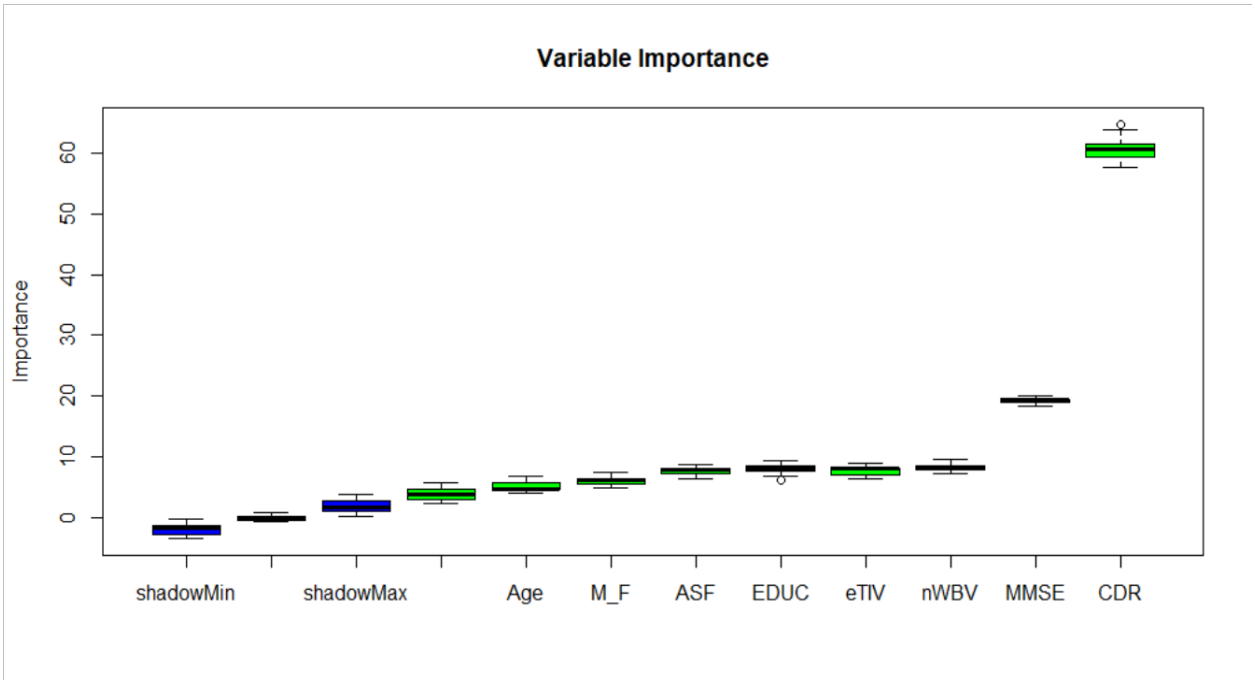
# Conclusion:

The results suggest that all these important features such as age, education, socioeconomic status, MMSE score, CDR score, eTIV, nWBV, ASF, and gender (M/F) have been marked as "Confirmed" by the Boruta algorithm. These are significant in predicting the target variable. Each feature in this dataset can be evaluated based on various measures of importance, such as mean importance, median importance, minimum importance, maximum importance, normalized hits, and decision. The mean importance value suggests that where there are higher values it indicates stronger importance considering the overall assessment. The decision column suggests whether the feature is confirmed or rejected based on importance. Since all the features in the dataset have been confirmed it means that they contribute significantly to predict the target variable.

Based on the consideration of all features in a predictive model, we can get better accuracy and performance in classifying individuals into their respective groups. Analysts, Researchers and healthcare professionals can leverage these findings to gain useful insights into the importance of various factors in diagnosing and understanding different groups within the context of the dataset.

**Table 1:**

| | meanImp | medianImp | minImp | maxImp | normHits | decision |
|---|---|---|---|---|---|---|
| **Age** | 5.060941 | 4.568310 | 4.009511 | 6.765981 | 1 | Confirmed |
| **EDUC** | 7.849720 | 7.950538 | 6.03883 | 9.283043 | 1 | Confirmed |
| **SES** | 3.911519 | 3.662180 | 2.36198 | 5.756468 | 0.9285714 | Confirmed |
| **MMSE** | 19.282879 | 19.379567 | 18.226336 | 20.113587 | 1 | Confirmed |
| **CDR** | 60.754030 | 60.656830 | 57.686746 | 64.697624 | 1 | Confirmed |
| **eTIV** | 7.689028 | 7.963369 | 6.433921 | 8.853022 | 1 | Confirmed |
| **nWBV** | 8.199716 | 8.269870 | 7.146284 | 9.614461 | 1 | Confirmed |
| **ASF** | 7.590078 | 7.805729 | 6.419011 | 8.741099 | 1 | Confirmed |
| **M_F** | 6.085514 | 6.216009 | 4.767434 | 7.425599 | 1 | Confirmed |

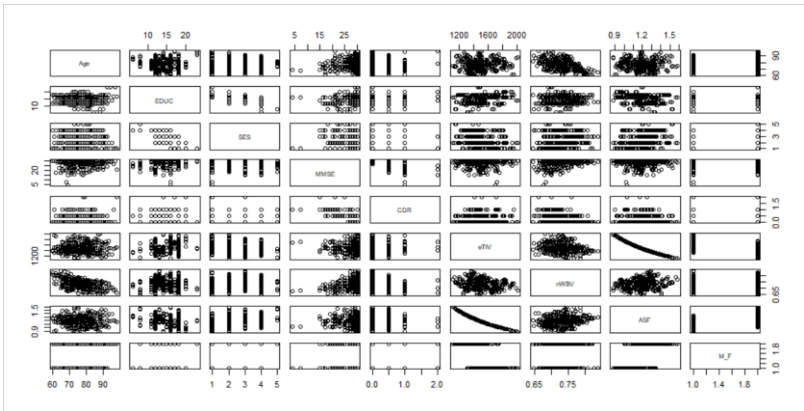



# Appendix

**#1) Descriptive statistics**

```
setwd("C:/Users/Rahul/Desktop")  # set working directory

a <- read.csv("project data.csv",header = T) # reading the data file

df_cleaned <- na.omit(a) #removing the blank rows

df_cleaned$M_F <- ifelse(df_cleaned$M.F == "M", 1, 2)  #"M/F" variable converted to numeric values

df_cleaned <- subset(df_cleaned, Group != "Converted") #removing the rows containing converted in group column

df_cleaned <- subset(df_cleaned, select = -c(M.F)) #removing the M/F variable column

print(df_cleaned) # Print the updated dataset

summary_table <- summary(df_cleaned) #summary of the dataset

print(summary_table)

ggplot(df_cleaned, aes(x = Group, y = Age)) + geom_boxplot() + labs(x = "Group", y = "Age") +

  ggtitle("Age Distribution by Group") #boxplot

par(mfrow = c(1, 1))  # Set the layout to display histograms in a 3x3 grid

hist(df_cleaned$Age, main = "Age") #histogram

hist(df_cleaned$eTIV, main = "eTIV") #histogram
```

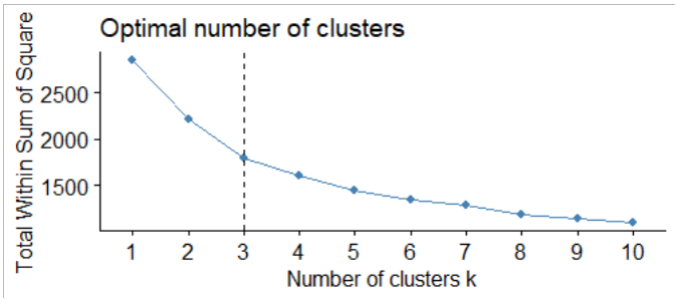
```
hist(df_cleaned$nWBV, main = "nWBV") #histogram
hist(df_cleaned$ASF, main = "ASF") #histogram
pairs(df_cleaned[,-1]) #Scatterplot
dim(df_cleaned) #dimension
```




**# 2)Clustering**

```
# Loading necessary libraries
library(factoextra)
library(ggplot2)
set.seed(123)
Variable= scale(df_cleaned[,-1])
#Apply k-means clustering
kmeans2 <- kmeans(Variable, centers = 2, nstart = 20)
kmeans3 <- kmeans(Variable, centers = 3, nstart = 20)
kmeans4 <- kmeans(Variable, centers = 4, nstart = 20)
kmeans2
str(kmeans2)
fviz_cluster(kmeans2, data = Variable)
fviz_cluster(kmeans3, data = Variable)
fviz_cluster(kmeans4, data = Variable)
# Clustering diagrams
f1 <- fviz_cluster(kmeans2, geom = "point", data = Variable) + ggtitle("k = 2")
```

```
f2 <- fviz_cluster(kmeans3, geom = "point", data = Variable) + ggtitle("k = 3")
f3 <- fviz_cluster(kmeans4, geom = "point", data = Variable) + ggtitle("k = 4")
library(gridExtra)
grid.arrange(f1, f2, f3, nrow = 2)
fviz_nbclust(Variable, kmeans, method = "wss")+
  geom_vline(xintercept = 3, linetype = 2)
```




```
#Start my calculating the distance matrix
d <- dist(Variable, method = "euclidean")
#Apply hierarchical clustering for differnt linkage methods
fit.single <- hclust(d, method="single")
fit.complete <- hclust(d, method="complete")
fit.average <- hclust(d, method="average")
fit.centroid <-hclust(d, method="centroid")
plot(fit.single) # print the dendrogram
groups.fit.single <- cutree(fit.single, k=3) # cut tree into k=4 clusters
# draw dendrogram with red borders around the 4 clusters
rect.hclust(fit.single, k=3, border="red")
#Checking how many observations are in each cluster
table(groups.fit.single)
plot(fit.complete)
groups.fit.complete <- cutree(fit.complete, k=4)
table(groups.fit.complete)
plot(fit.average)
groups.fit.average <- cutree(fit.average, k=4)
```

rect.hclust(fit.average, k=4, border="red")

**# 3) Logistic Regression**

df_cleaned$Group <- as.factor(df_cleaned$Group)

glm.fit<-glm(Group ~., data = df_cleaned,family=binomial) #fitting a logistic regression model

summary(glm.fit) #summary

contrasts(df_cleaned$Group) #encoding

glm.probs <- predict(glm.fit,type="response") #Pr(Y=1|X)

glm.predicted <- rep("Demented",1250) # Initializing labels

glm.predicted[glm.probs>0.5]="Nondemented" # Initializing labels

table(glm.predicted, df_cleaned$Group) #creating contingency tables

mean(glm.predicted==df_cleaned$Group)  #Accuracy of the model

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.196e-04  -2.100e-08   2.100e-08   2.100e-08   1.381e-04

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.245e+03  5.203e+06   0.000    1.000
Age          6.473e+00  7.727e+03   0.001    0.999
EDUC        -3.828e+00  1.205e+04   0.000    1.000
SES          1.505e+01  4.034e+04   0.000    1.000
MMSE         6.396e+00  1.979e+04   0.000    1.000
CDR         -3.304e+02  1.615e+05  -0.002    0.998
eTIV         4.272e-01  1.774e+03   0.000    1.000
nWBV         9.496e+02  2.294e+06   0.000    1.000
ASF          2.403e+02  2.645e+06   0.000    1.000
M_F          4.176e+01  6.482e+04   0.001    0.999
```

**# 4) Feature Selection**

# Loading necessary libraries

library(Boruta)

df_cleaned$Group <- as.factor(df_cleaned$Group)

boruta1 <- Boruta(Group ~., data=df_cleaned, doTrace=1) #Feature selection

decision<-boruta1$finalDecision #Final decision from analysis

signif <- decision[boruta1$finalDecision %in% c("Confirmed")]

print(signif)

plot(boruta1, xlab="", main="Variable Importance") #visualizing using plots

attStats(boruta1) #getting additional statistics