School Of Science, Technology, Engineering & Mathematics

Report on Exploratory Analysis of Alzheimer's Disease Dataset

MA335 – FINAL PROJECT

Prepared by

Lokesh Palaniraj (Reg no: 2211543)

June 21, 2023

Abstract:

This report aims to perform a detailed analysis of a dataset related to Alzheimer's disease using R. The dataset comprises of many factors like gender, age, year of education, socioeconomic status, mental state, and brain volume estimates which tells whether the patient belongs to demented or nondemented group. To understand the dataset's characteristics, the descriptive statistics with summary measures and graphical representations were presented. Clustering algorithms identify the underlying pattern for better interpretation with the logistic regression predicts the diagnosis of the variables in the dataset. These findings aid in early detection of diseases by understanding the key factors and the potential subgroups associated with it.

Contents:

Word Count: 1300

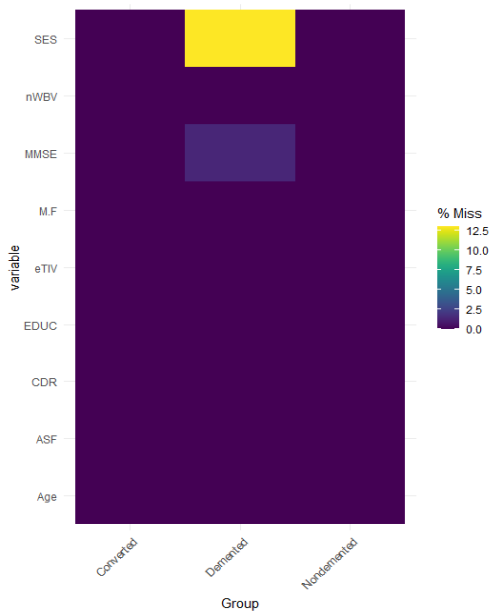
Introduction:

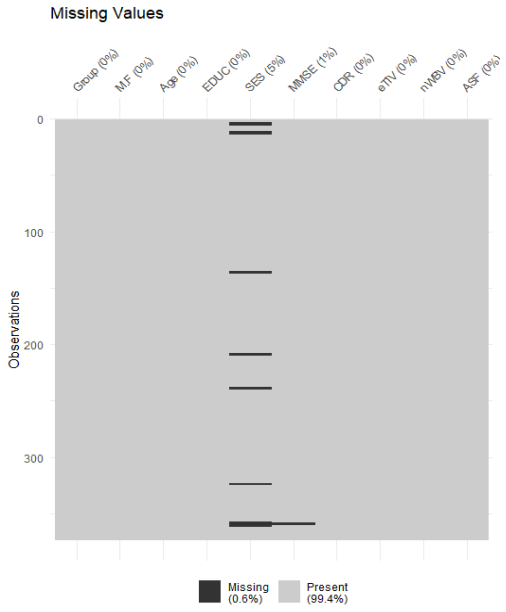
Alzheimer's disease is a brain disorder that slowly depreciates the thinking and memory skills, even the normal day to day activities gradually. This disease is mostly common in elderly people, aged 65 or above. Experts says there may be more than 6 million Americans who may have Alzheimer's. In this report, we analyse a dataset with various factors such as clinical measurements, brain volume estimates, gender, and age to understand the cause of this disease over the population. We also employ descriptive statistics, clustering algorithms and logistic regression model to uncover the underlying patterns and to provide further insights on the

dataset. With the use of this study, the discovery of the Alzheimer's and the treatment plans would be made much sooner for all age groups.

Preliminary Analysis:

The dataset we used in this report provides us with a comprehensive collection of variables related to Alzheimer's disease. Some of the *independent variables* such as Age, Year of Education (EDUC), Socioeconomic status (SES), mini mental state examination (MMSE), and some brain volume estimates like Clinical dementia rating (CDR), Estimated total intracranial volume (eTIV), Normalize whole brain volume (nWBV) and Atlas scaling factor (ASF) and followed by a *dependent* factor which tells whether the patient is demented or non demented, named as 'group'. Before we proceed with the analysis, it is important to check if the data is clean without any missing values. In our preliminary examination, we identified some missing values in some of the variables.





To ensure the quality of the analysis, the missing values should be taken care, so we have removed the rows of the dataset with missing values. This ensures that our analysis is conducted on a whole completed dataset which aids in deriving better insights and meaningful interpretations. In addition to that, the categorical variable group, must be converted to numerical values, as this may affect the accuracy as well. Now that we have performed the preliminary analysis which will establish a solid foundation for our subsequent exploration, it's time to dive deeper into the analysis to provide better insights on the Alzheimer's disease.

Analysis:

Followed by the data cleaning process, we have employed a *descriptive statistic* of the dataset for better understanding of the dataset's characteristics, which includes mean, median, standard deviation, minimum and maximum of all the variables. In our case, the average age of the individuals in the dataset was found to be 76.7 years and with 60 years as the minimum and the 98 as the maximum, indicating relatively a wide range of age groups.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Group	317						
... Demented	127	40%					
... Nondemented	190	60%					
M.F	317	0.43	0.5	0	0	1	1
Age	317	77	7.8	60	71	82	98
EDUC	317	15	2.9	6	12	16	23
SES	317	2.5	1.1	1	2	3	5
MMSE	317	27	3.9	4	27	30	30
CDR	317	0.27	0.38	0	0	0.5	2
eTIV	317	1494	180	1106	1358	1599	2004
nWBV	317	0.73	0.038	0.64	0.7	0.76	0.84
ASF	317	1.2	0.14	0.88	1.1	1.3	1.6

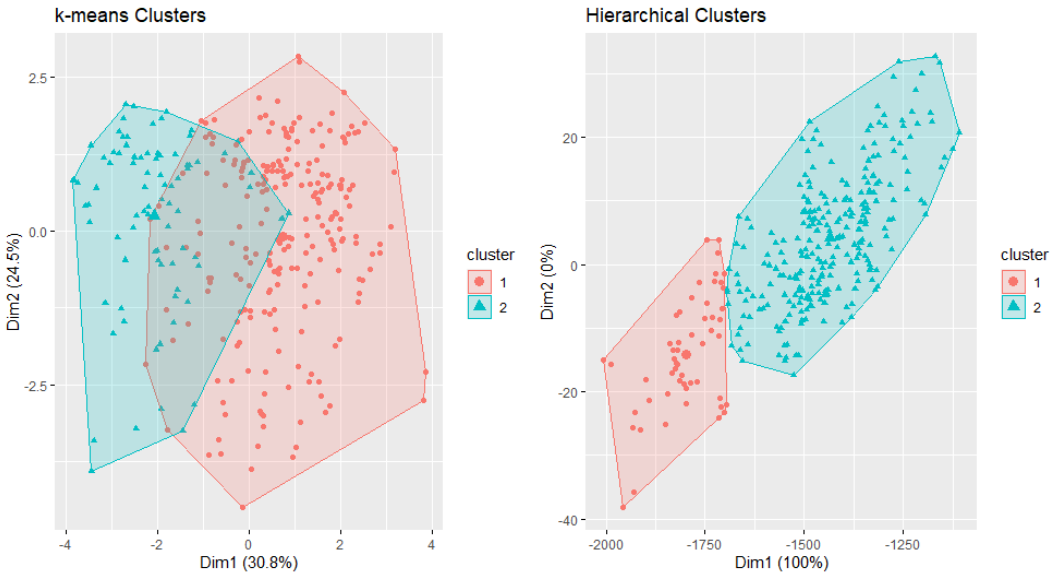
We have also performed correlation between the variables to identify similarities between them and the table has been generated. From the table we can say that the age and the mini mental state examination is positively correlated.

Correlation Table

	Age	EDUC	MMSE	eTIV	nWBV
Age	1.0000000	-0.0453863	0.0460516	0.0370933	-0.4971262
EDUC	-0.0453863	1.0000000	0.1850748	0.2685849	0.0159119
MMSE	0.0460516	0.1850748	1.0000000	-0.0206304	0.3707141
eTIV	0.0370933	0.2685849	-0.0206304	1.0000000	-0.1950752
nWBV	-0.4971262	0.0159119	0.3707141	-0.1950752	1.0000000

To interpret the variables visually, the boxplot for the age has been generated (*refer plots in appendix*) alongside with the histogram of the education from the dataset. The boxplot revealed a symmetrical pattern with no outliers, indicating the balanced age distribution. It is obvious from the histogram, that the cases are at maximum within the range 10-15. The histogram of the socioeconomic status displayed a slightly left-skewed distribution, explaining a higher proportion of individuals with lower socioeconomic status, where the eTIV shows a normal distribution.

We also implemented *clustering methods*, to identify some underlying patterns using **k-means** and **hierarchical** clustering. This aims to identify the similar groups in the dataset which could have inference with the disease progression and can be used to offer personalized treatment methods accordingly.



From the k-means plot, we can interpret that the cluster 1 has a greater percentage of people with *dementia* compared to the 2nd one. On the other hand, in the hierarchical clustering table, the number of individuals diagnosed as *demented* in cluster 2 is significantly higher than in cluster 1. However, the clusters in both approaches appear very different, which is because of the way they initialize the cluster centroids, the assumptions about the cluster size and shape, the number of clusters in k-means, and even with the way of interpretation, where k-means assigns each observation to a single cluster but hierarchical can create nested clusters.

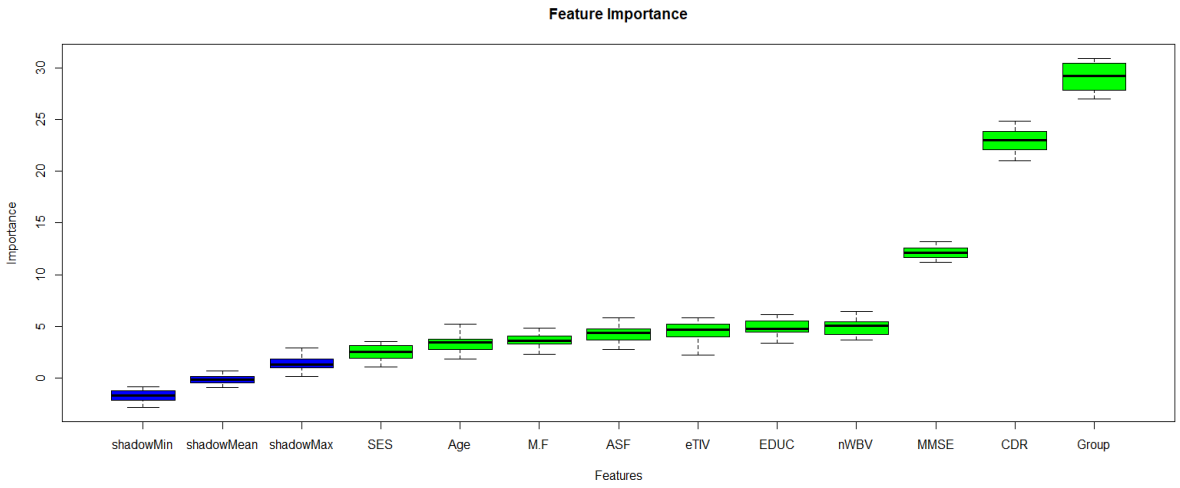
	k-means	Hierarchical
Cluster 1	232	53
Cluster 2	85	264

Extensive Review:

To perform the logistic regression model, initially, a table has been derived from the original one, with the *Group* column as a factor. We split the new dataset into 70% as training and 30% for testing, and we have fitted a *logistic regression model* to the dataset with all the independent variables (M.F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF) with the dependent variable 'group' to predict whether the patient is demented or not. Looking at the

logistic regression outcome with all the parameters, significance levels and the goodness of fit, we got to know that the intercept term, -0.00144, is not statistically significant, suggesting that it doesn't have any importance on the outcome. The AIC value of this model is 20, which indicates the good fit of the model to the data. Finally, the accuracy of the model, i.e., 98.94% explains that the model correctly predicted the variable "Group" nearly 99% of the time. This confirms that the predictor variables are capable of distinguishing the outcome as demented or nondemented, most of the time.

To identify the most important features in this dataset, we have also performed some feature selection methods like forward and Boruta algorithm. These approaches help us to understand better about the variables and their contributions in predicting the final desired outcome. The important features will be evaluated based on their impact on the final prediction, discovering the key factors associated with the disease. The outcome from the Boruta's algorithm has been plotted according to their dominance. From the plot we can infer that the *'group'* has the at most importance of all the other features in the dataset, followed by *'CDR'* and *'MMSE'*.



Additionally, we have performed forward selection with only the intercept term 'Group ~ demented' and by using the specified scope of predictor variables such as M.F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV and ASF, which justifies the importance of 'CDR' whose coefficient value being at the top, i.e., 1.047, indicating that a one-unit increase of *CDR* increases the *Group* variable by 1.047 times, followed by the M.F, eTIV and EDUC. The

multiple R-squared value is 0.7651, implicating that, approximately 76.51% of the variability in the *Group* variable can be explained by the predictor variables.

Conclusion:

This report focused on the statistical analysis of an Alzheimer's disease dataset which encompasses multiple tasks like descriptive statistics, clustering, logistic regression modelling and feature selection, and through these we could be able to uncover lots of valuable insights and its relationship with the Alzheimer's disease. This contributed a comprehensive analysis of the dataset by exploring its characteristics, identifying its potential subgroups, interpreting the predictors of diagnosis and the importance of the features in the dataset by various selection techniques. This analysis would be extremely beneficial in identifying the condition sooner and can provide proactive treatment for the individuals affected by the Alzheimer's disease.

References:

1) National Institute on Aging. (n.d.). Alzheimer's Disease Fact Sheet. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet#:~:text=Alzheimer's%20disease%20is%20a%20brain,first%20appear%20later%20in%20life.
2) Kassambara, A. (2021). fviz_cluster: Visualize Clustering Results. factoextra Package Documentation. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_cluster
3) Xie, Y. (2020). kable. In R Markdown Cookbook. https://bookdown.org/yihui/rmarkdown-cookbook/kable.html

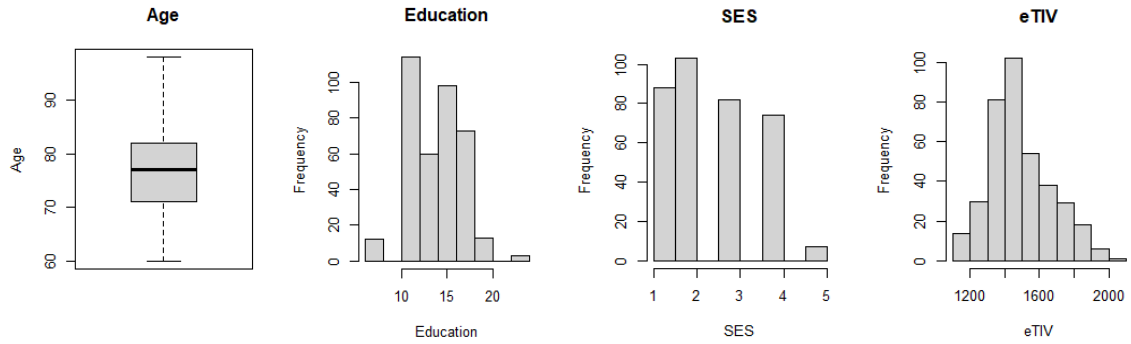
Appendix:

The R-code has been attached below for the reference with properly meaningful comments. The packages that has been used were tidyverse, vtable, cluster, naniar, Boruta, MASS, knitr, kableExtra, factoextra, gridExtra, caTools.

DATA FETCH

setwd("C:/MA335/Final project")

proj_dat <- read.csv("project data.csv")



######## DATA CLEANING ######

Visualization on missing values

vis_miss(proj_dat) + labs(title = "Missing Values")

gg_miss_fct(x = proj_dat, fct = Group)

Converting the m,f values

proj_dat$M.F <- ifelse(proj_dat$M.F == "M", 1, 0)

Removing the rows with group = "converted"

proj_dat <- proj_dat %>%filter(Group != "Converted")

Removing rows with NA

proj_dat <- proj_dat %>%filter(!is.na(SES) & !is.na(MMSE))

############ DESCRIPTIVE ANALYSIS ###########

summary_of_data <- st(proj_dat)

par(mfrow=c(2,2))

b1 <- boxplot(proj_dat$Age, main = "Age", ylab = "Age")

h1 <- hist(proj_dat$EDUC, main = "Education", xlab= "Education")

h2 <- hist(proj_dat$SES, main = "SES", xlab= "SES")

```
h3 <- hist(proj_dat$eTIV, main = "eTIV", xlab= "eTIV")

cor_matrix <- cor(proj_dat[, c("Age", "EDUC", "MMSE", "eTIV", "nWBV")])

print(cor_matrix)

kable(cor_matrix, format = "html", table.attr = "class='table'", caption = "Correlation Table")
%>% kable_styling(full_width = FALSE)

################ CLUSTERING ##################

# Storing the cluster data in a variable, omitting Group

cluster_dat <- subset(proj_dat, select = -c(Group))

cluster_dat1 <- subset(proj_dat, select = -c(Group))

# k-means clustering

kmeans_out <- kmeans(cluster_dat, centers = 2, nstart = 20)

cluster_kmean <- kmeans_out$cluster

c1 <- fviz_cluster(kmeans_out, data = cluster_dat, geom = "point", title = "k-means
Clusters")

table(cluster_kmean)

# Using Hierarchical Clustering

hierarc_out <- hclust(dist(cluster_dat1))

hierarchical_out <- cutree(hierarc_out, k=2)

table(hierarchical_out)

c2 <- fviz_cluster(list(data = cluster_dat1, cluster = hierarchical_out), geom = "point", stand
= FALSE, main = "Hierarchical Clusters")

grid.arrange(c1, c2, nrow = 1)

#######Logistic Regression#############

data1 <- proj_dat

data1$Group <- as.factor(data1$Group)
```

```
set.seed(42)

splitted_dat <- sample.split(data1$Group, SplitRatio = 0.7) # splitting into 70% training 30% testing

train_dat <- data1[splitted_dat, ]

test_dat <- data1[!splitted_dat, ]

logi_model <- glm(Group ~ ., data = train_dat, family = binomial)

summary(logi_model)

preds <- predict(logi_model, newdata = test_dat, type = "response")

predicted_lbls <- ifelse(preds > 0.5, "Nondemented", "Demented")

accuracy <- sum(predicted_lbls == test_dat$Group) / nrow(test_dat)

print(accuracy)

########Feature Selection##############

# Using Boruta

dat <- proj_dat

dat$Group <- ifelse(proj_dat$Group == "Demented", 1, 0)

boruta_out <- Boruta(dat$Group ~., data=dat, doTrace=1)

decision <- boruta_out$finalDecision

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

plot(boruta_out, xlab="Features", main="Feature Importance")

att1 <- attStats(boruta_out)

# Using wrapper - forward

model1 <- lm(Group~1, data = dat)

step1 <- step(model1, scope=~ M.F + Age + EDUC + SES + MMSE + CDR + eTIV + nWBV + ASF, method='forward')

summary(step1)
```