MA335 Final Project

Analysis of the Relationship Between the Various Characteristics of Alzheimer's Disease and the Diagnosis

Reethu Mathew

Reg. No: 2213186

June 20, 2023

Abstract

This report presents an analysis of a dataset that includes various characteristics of Alzheimer's disease. The aim of the analysis is to investigate the relationship between these characteristics and the diagnosis of Alzheimer's (demented) or non-Alzheimer's (non-demented) cases. The analysis is conducted using R on the 'project data.csv' file containing the dataset under investigation.

This report provides a comprehensive analysis of the dataset, exploring descriptive statistics, clustering, logistic regression, and feature selection. The findings contribute to a better understanding of the relationships between the characteristics of Alzheimer's disease and its diagnosis, potentially facilitating improved diagnosis and treatment strategies.

Contents

Word Count.....1996

1 Introduction

Dementia results from a wide range of diseases, brain disorders, and traumas [3]. Alzheimer's disease is the leading cause of dementia and one of the greatest health-care challenges of the 21st century [5]. It is a neurodegenerative illness that often develops gradually and gets worse over time [1]. It may impair one's memory, ability to think clearly, and other mental capacities [2]. Dementia, which affects more than 55 million people globally [3], imposes a tremendous burden on patients, families, and healthcare systems. Understanding the association between different characteristics and the diagnosis of Alzheimer's disease is critical for early detection, intervention, and management of the condition.

Data science and statistical analysis methods have been growing in significance in recent years for detecting intricate relationships and trends in huge datasets. Researchers have made great progress in identifying potential risk factors, establishing diagnostic tools, and enhancing treatment techniques by applying these methods to Alzheimer's disease datasets. The objective of the current study is to examine the association between various characteristics and the diagnosis of Alzheimer's disease using advanced statistical methods.

The dataset used for this analysis includes variables such as age, gender, education, socioeconomic status, mini-mental state examination (MMSE) scores, clinical dementia ratings (CDR), estimated total intracranial volume (eTIV), normalised whole brain volume (nWBV), and atlas scaling factor (ASF) [4]. To carry out this analysis, we will use R, a widely known statistical programming language, and a variety of approaches such as descriptive statistics, clustering algorithms, logistic regression, and feature selection methods. Such approaches will offer valuable insights into the dataset while encouraging the interpretation of the underlying relationships and patterns.

2 Preliminary Analysis

2.1 Data Pre-processing

Before using R or another data analysis programme to do statistical analysis, data preprocessing is essential. It involves cleaning, trans-

forming, and organising the raw data to ensure its quality, compatibility, and suitability for analysis. It assists with handling missing values, which may lead to results that are inaccurate or biased, and it finds and fixes errors, outliers, or inconsistencies that may affect the accuracy of statistical analysis. Transformation techniques such as normalisation and categorical variable encoding enable variables to fulfil specific statistical analysis needs.

The first step after loading the dataset into R Studio is to convert the variable "M.F" into a numeric type because it contains categorical values for the test participants' gender in character type. This will allow the variable to be useful for calculations. The rows where the "Group" variable is "Converted" are then eliminated because we are only interested in situations in which the subject is either demented or not. Finally, the most crucial phase of eliminating missing values concludes the data pre-processing for this dataset to avoid the possible generation of misleading results or the hindrance of some calculations that might not be designed to deal with missing values. The row names are adjusted after removing the missing values. The response variable Group is also converted to factor type to help with the analyses. This clean and transformed data is then subjected to descriptive statistics to get an idea of the type of data we will be analysing.

2.2 Descriptive Statistics

Descriptive statistics are used to summarise and define the key features of a dataset. It supports data exploration, visualisation, interpretation, and quality assessment while offering insightful information for later statistical analysis and decision-making. Table 1 shows the variables in the dataset.

The numerical variables in the dataset are summarized in Table 2 which is attached in Appendix B. This table provides us an idea about the range and distribution of the numerical variables. The histogram in Figure 5 in Appendix C illustrates how Age is distributed throughout the Demented and Nondemented categories in the dataset, providing insight into the shape, central tendency, and spread of the dataset. The dissimilarities in the Demented and Nondemented categories in the distribution of the Age variable are exhibited in the plot. We can see that the range of the maxi-

Variables	Description
Group	Group of the diagnosis (Nondemented or Demented)
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
ASF	Atlas scaling factor

Table 1: Variables in the dataset

mum frequency for the Demented group is between 70 and 75, while the range for the Nondemented group is between 80 and 85. Even though the Demented group contains several outliers, the majority of the data falls within the range of 65 to 90.

Both directly and indirectly, dementia has a substantial effect on women. Women provide 70% of dementia patients' care hours while simultaneously having greater disability-adjusted life years and mortality rates than men [3]. The pie chart (Figure 6 from Appendix C) displays the proportional representation of gender data that was gathered in the dataset, showing that there are more women than men, which lends credibility to the aforementioned statement. Finally, Figure 3 shows the correlation plot between the variables that are related to the brain (MMSE, CDR, eTIV, nWBV, and ASF), which is also included in Appendix C.

3 Analysis

For the analysis, we will employ clustering algorithms, logistic regression, and feature selection. Clustering algorithms are a set of unsupervised machine learning techniques used to detect groups or clusters within a dataset. They can assist in the identification of distinct patient profiles or the grouping of people who share common characteristics, which can be helpful in figuring out the many subtypes of Alzheimer's disease. Logistic regression, a statistical modelling tool, is used to predict categorical outcomes. Logistic regression can be used to identify the relationship between the char-

acteristic variables in the dataset and the diagnosis of Alzheimer's disease, facilitating the prediction and understanding the risk factors associated with the disease. The idea of feature selection focuses on determining the most crucial elements that significantly influence the predictive model. Feature selection reduces the dimensionality of the dataset by focusing on the most relevant factors, enhancing model performance, interpretability, and computational efficiency.

3.1 Clustering Algorithms

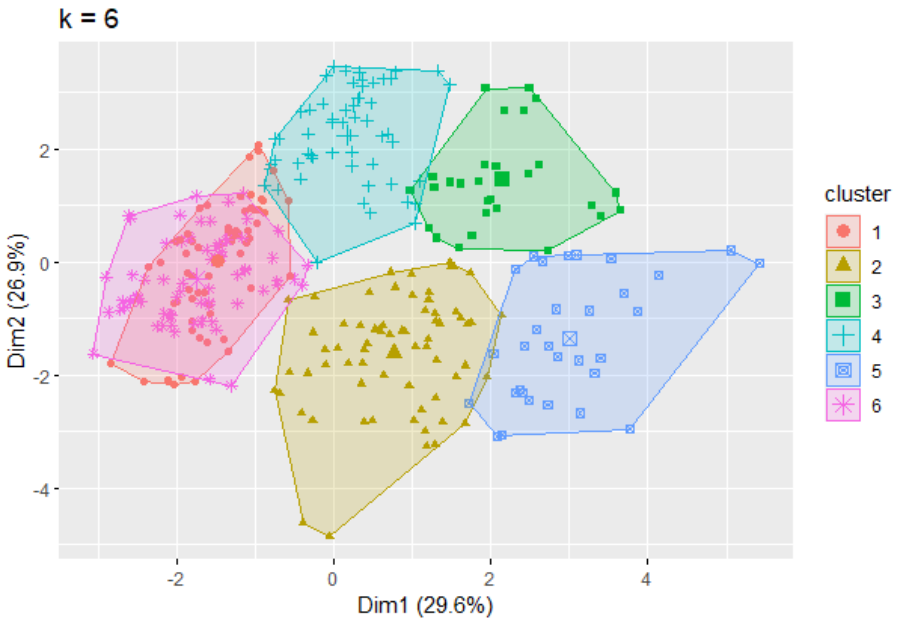

Figure 1: K-Means Clustering for k = 6

We will deploy two different types of clustering algorithms for this dataset: K-Means clustering and Hierarchical clustering. K-Means clustering is a partition-based approach that attempts to divide the dataset into a specified number of clusters, represented by centroids. Although it assumes spherical clusters and requires the user to define the number of clusters, it is computationally efficient. On the other hand, hierarchical clustering is more versatile in handling different shapes and sizes and constructs a hierarchy of clusters without assuming the number of clusters, yet it can be

computationally expensive for large datasets.

To produce the clustering plot for K-Means clustering, we initially assume the k value to be 2. Following that, the k value is raised, and the corresponding clusters are generated. Figure 8 in Appendix C shows the clusters for k = 2, 3, 4, and 5. The optimal number of clusters can be determined from Appendix C figure 9 based on the clusters generated for each value of k. We can confirm that k = 6 is the elbow point, which is the point of inflection where the within-cluster sum of squares (WSS) begins to level off, and thus the optimal k value for this data set is 6. Figure 1 presents the clusters for the value of k = 6.

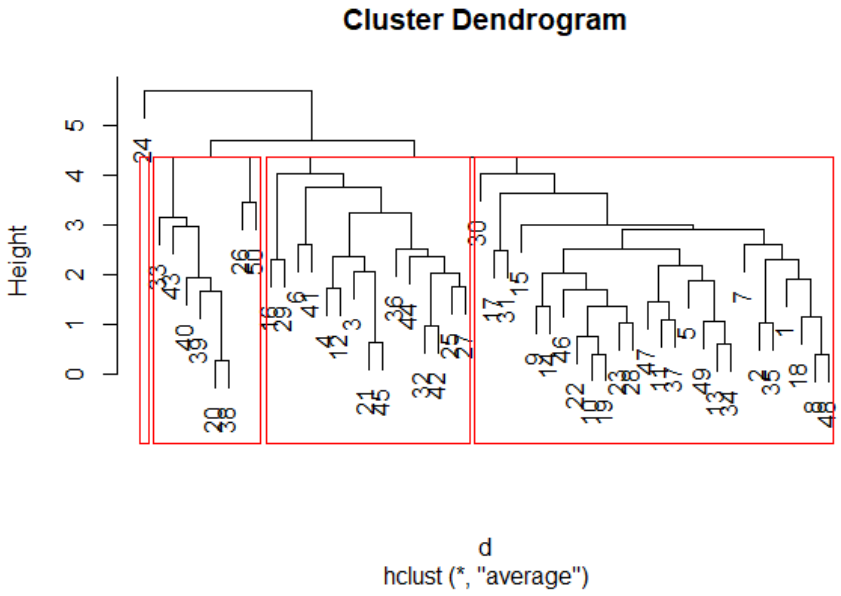

Figure 2: The Cluster Dendrogram with Average Linkage

We can create hierarchical clustering using the Complete, Average, Single, and Centroid linkage methods. We are using clustering with average linkage for this analysis because it tends to result in clusters with more evenly distributed sizes and can tolerate certain outliers. We will also generate a random sample of size 50 because hierarchical clustering can be too crowded for our dataset. The cluster dendrogram for average linkage is illustrated in Figure 2. Figure 10 in Appendix C shows the four different types of hierarchical clus-

tering methods. Finally, figure 11 in Appendix C visualises the clusters in relation to the categorical variable Group. The four clusters in the hierarchy are represented by the colours black, red, green, and blue, with the numbers 1 and 2 standing for the Nondemented and Demented groups, respectively.

3.2 Logistic Regression

We are splitting the dataset into training and test datasets to avoid overfitting in the ratio 80:20. The train set is fed into the logistic regression model with all the predictor variables. The summary of the logistic regression model is attached in Figure 3.

```
Call:
glm(formula = Group ~ M.F + Age + EDUC + SES + MMSE + CDR + eTIV +
    nWBV + ASF, family = "binomial", data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.546e+03  1.348e+06   0.002    0.998
M.F         -4.321e+01  2.907e+04  -0.001    0.999
Age         -6.353e+00  2.838e+03  -0.002    0.998
EDUC         3.912e+00  3.789e+03   0.001    0.999
SES         -1.272e+01  1.195e+04  -0.001    0.999
MMSE        -7.366e+00  7.791e+03  -0.001    0.999
CDR          2.991e+02  7.385e+04   0.004    0.997
eTIV        -5.346e-01  4.606e+02  -0.001    0.999
nWBV        -8.530e+02  7.146e+05  -0.001    0.999
ASF         -3.974e+02  6.695e+05  -0.001    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.3683e+02  on 253  degrees of freedom
Residual deviance: 4.3538e-08  on 244  degrees of freedom
AIC: 20

Number of Fisher Scoring iterations: 25
```

Figure 3: Summary of Logistic Regression Model

The response variable in the test dataset is then predicted using this logistic regression model. This logistic model's accuracy is 23.81%. With the goal of constructing a better model, we can now proceed to feature selection methods.

3.3 Feature Selection

The three strategies of forward, backward, and Boruta are considered for feature selection on the current model. Both the forward and backward models have resulted in the same model, which is shown below:

$$Group \sim Age + MMSE + CDR + eTIV$$

Summaries of both the forward and backward models are attached in Appendix C as figures 12 and 13. The logistic regression performed on this model yields an accuracy of 25.4%. A summary of the modified logistic regression model is shown in Figure 14, Appendix C. Despite the minor increase in accuracy, this model is definitely an improvement over the full model. It is crucial to keep in mind that these accuracy figures are specific to the train and test datasets and may vary with other train and test datasets.

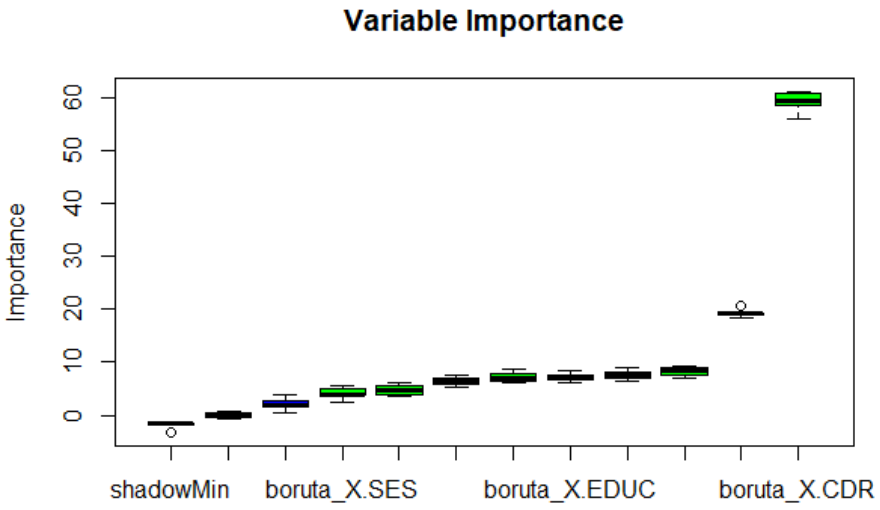

Figure 4: Variable Importance according to Boruta Feature Selection

Boruta feature selection uses a random forest approach to classify items based on their relevance. Figure 4 depicts a boxplot of the variables based on their relevance using the Boruta approach. We

can see that the variable CDR has the most relevance out of the predictor variables. Table 3 in Appendix B contains the attribute statuses of the Boruta operation done on the dataset.

4 Discussion

The analysis executed throughout this report is designed to evaluate the association between several characteristics of Alzheimer's disease and its diagnosis. The analysis executed throughout this report is designed to evaluate the association between several characteristics of Alzheimer's disease and its diagnosis. The results of the investigation reveal patterns in the dataset and offer various possibilities for identifying Alzheimer's disease. The identification of different patient profiles by clustering can be helpful in understanding the heterogeneity of the disease and generating personalised treatments. Logistic regression analysis and feature selection lead to the discovery of key variables for diagnosis prediction which are Age, MMSE, CDR, and eTIV. These findings emphasise the role of cognitive and brain-related criteria in detecting Alzheimer's disease. However, it is important to be aware of the limitations of this analysis. Due to its small number of variables, the dataset utilised for the analysis may not fully reflect the complexity of Alzheimer's disease. Additionally, the accuracy achieved by the logistic regression model was relatively low, indicating the need for further refinement.

5 Conclusion

In conclusion, this analysis delivers a comprehensive statistical analysis of a dataset containing the characteristics of Alzheimer's disease. The findings point out the significance of age, MMSE, CDR, and eTIV as key variables for diagnosing the disease. Different patient profiles were uncovered by clustering algorithms, demonstrating the variety of Alzheimer's. Logistic regression and feature selection methods enhance the prediction model by identifying potential risk factors and important predictors. These findings provide valuable insights for healthcare professionals and researchers, facilitating improved detection, intervention, and management strategies for Alzheimer's disease. However, further research is required to ex-

plore additional variables and enhance model accuracy.

Overall, this analysis contributes to Alzheimer's disease research by utilising advanced statistical techniques and offering valuable insights into the dataset. The results could help in the development of more efficient diagnostic and therapeutic strategies, ultimately benefiting individuals affected by Alzheimer's disease.

References

[1] *Alzheimer's disease - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Alzheimer\%27s_disease\#cite_note-WHO2020-2. [Accessed 19-Jun-2023].

[2] *Alzheimer's disease — nhs.uk*. https://www.nhs.uk/conditions/alzheimers-disease/. [Accessed 19-Jun-2023].

[3] *Dementia - World Health Organization*. Retrieved June 19, 2023. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/dementia.

[4] Afreen Khan and Swaleha Zubair. "Longitudinal Magnetic Resonance Imaging as a Potential Correlate in the Diagnosis of Alzheimer Disease: Exploratory Data Analysis". In: *JMIR Biomedical Engineering* 5 (Apr. 2020), pp. 1–13.

[5] Philip Scheltens et al. "Alzheimer's disease". In: *The Lancet* 388.10043 (2016), pp. 505–517. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(15)01124-1. URL: https://www.sciencedirect.com/science/article/pii/S0140673615011241.

A R Code

Data Pre-processing

```r
# Setting the working directory to the file location
setwd("D:/DATA SCIENCE AND ITS APPLICATIONS/MA335")
# loading the data
base_data <- read.csv("project data.csv")
# Converting the M.F column to numerical values
base_data$M.F <- as.numeric(factor(base_data$M.F,
levels = c("M", "F")))
# Removing rows where Group value is Converted
data <- base_data[!base_data$Group == "Converted",]
# Removing missing values from the data set
clean_data <- na.omit(data)
# Correcting the row names
rownames(clean_data) <- NULL
# Converting the Group to factor type
clean_data$Group <- factor(clean_data$Group,
                           levels = c("Nondemented",
                                      "Demented"))
```

Descriptive Statistics

```r
# Summary table
summary__data <- cbind(summary(clean_data$Age),
                       summary(clean_data$EDUC),
                       summary(clean_data$MMSE),
                       summary(clean_data$CDR),
                       summary(clean_data$eTIV),
                       summary(clean_data$nWBV),
                       summary(clean_data$ASF))
colnames(summary__data) <- c("Age", "EUDC", "MMSE",
                             "CDR", "eTIV", "nWBV",
                             "ASF")
summary__data
# Histogram of Age
library(ggplot2)
ggplot(clean_data, aes(x = Age, fill = Group)) +
```

```
  geom_histogram(binwidth = 3) +
  labs(x = "Age", y = "Frequency") +
  facet_wrap(~Group, ncol = 1)
# Pie chart
gender <- table(clean_data$M.F)
pie_labels <- c("Male", "Female")
pie(gender, labels = pie_labels,
    main = "Gender Distribution")
# Correlation plot
cor_matrix <- cor(clean_data[, c("MMSE",
                                 "CDR", "eTIV",
                                 "nWBV", "ASF")])
library(corrplot)
corrplot(cor_matrix, method = "color", type = "lower")
```

Clustering Algorithms

K-Means Clustering

```
library(factoextra)
library(gridExtra)
cluster_data <- clean_data
cluster_data$Group <- as.numeric(cluster_data$Group)
cluster_data <- scale(cluster_data)
set.seed(12345)
kmeans2 <- kmeans(cluster_data, centers = 2,
                  nstart = 20)
kmeans3 <- kmeans(cluster_data, centers = 3,
                  nstart = 20)
kmeans4 <- kmeans(cluster_data, centers = 4,
                  nstart = 20)
kmeans5 <- kmeans(cluster_data, centers = 5,
                  nstart = 20)
f1 <- fviz_cluster(kmeans2, geom = "point",
                   data = cluster_data) +
  ggtitle("k = 2")
f2 <- fviz_cluster(kmeans3, geom = "point",
                   data = cluster_data) +
  ggtitle("k = 3")
```

```
f3 <- fviz_cluster(kmeans4, geom = "point",
                   data = cluster_data) +
  ggtitle("k = 4")
f4 <- fviz_cluster(kmeans5, geom = "point",
                   data = cluster_data) +
  ggtitle("k = 5")
grid.arrange(f1, f2, f3, f4, nrow = 2)
fviz_nbclust(cluster_data, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)
kmeans6 <- kmeans(cluster_data, centers = 6,
                  nstart = 20)
fviz_cluster(kmeans6, geom = "point",
             data = cluster_data) +
  ggtitle("k = 6")
```

Hierarchical Clustering

```
# Creating a sample of 50
set.seed(12345)
cluster_id <- sample(1:dim(cluster_data)[1], 50)
data_hier <- cluster_data[cluster_id,]
# Calculating the distance matrix
d <- dist(data_hier, method = "euclidean")
# average linkage for clustering
fit.average <- hclust(d, method="average")
# the dendrogram for average linkage
plot(fit.average)
groups.fit.average <- cutree(fit.average, k=4)
rect.hclust(fit.average, k=4, border="red")
# Hierarchies for other 3 linkages
fit.complete <- hclust(d, method="complete")
fit.single <- hclust(d, method="single")
fit.centroid <- hclust(d, method="centroid")
# the dendrograms
par(mfrow = c(2,2))
# average
plot(fit.average)
groups.fit.average <- cutree(fit.average, k=4)
rect.hclust(fit.average, k=4, border="red")
```

```
# complete
plot(fit.complete)
groups.fit.complete <- cutree(fit.complete, k=4)
rect.hclust(fit.complete, k=4, border="red")
# Single
plot(fit.single)
groups.fit.single <- cutree(fit.single, k=4)
rect.hclust(fit.single, k=4, border="red")
# Centroid
plot(fit.centroid)
groups.fit.centroid <- cutree(fit.centroid, k=4)
rect.hclust(fit.centroid, k=4, border="red")
par(mfrow = c(1,1))
ggplot(data_hier,
       aes(data_hier$MMSE,
           data_hier$nWBV,
           color = as.factor(data_hier$Group))) +
  geom_point(alpha = 0.4, size = 3.5) +
  geom_point(col = groups.fit.average) +
  scale_color_manual(values = c('black', 'red',
                                'green', 'blue')) +
  labs(col = "Group") +
  xlab("MMSE") + ylab("nWBV")
```

Logistic Regression

```
# train/test split
size_index <- round(dim(clean_data)[1]*0.8)
set.seed(526)
size_id <- sample(1:dim(clean_data)[1], size_index)
train = clean_data[size_id,]
test = clean_data[-size_id,]
# log regression
dem_log_reg <- glm(Group ~ M.F + Age + EDUC + SES +
                     MMSE + CDR + eTIV + nWBV + ASF,
                   data = train, family = "binomial")
summary(dem_log_reg)
# Prediction and Accuracy Calculation
test_data <- test[, -1]
```

```
dem_pred_log <- predict(dem_log_reg,
                        newdata = test_data,
                        type = "response")
predictions <- round(dem_pred_log) + 1
out <- rep("Nondemented", 63)
out[predictions>1] = "Demented"
out <- factor(out, levels = c("Nondemented",
                              "Demented"))
acc_out <- sum(as.numeric(out) ==
                 test_data[,1])/nrow(test_data)
accuracy <- round(100*acc_out,2)
accuracy
```

Feature Selection

Forward and Backward Selection

```
forward_model <- step(dem_log_reg,
                      method = 'forward')
summary(forward_model)
backward_model <- step(dem_log_reg,
                       method = 'backward')
summary(backward_model)
# modified model
mod_dem_log_reg <- glm(Group ~ Age + MMSE
                       + CDR + eTIV,
                       data = train,
                       family = "binomial")
summary(mod_dem_log_reg)
# prediction
test_data <- test[, -1]
mod_dem_pred_log <- predict(mod_dem_log_reg,
                            newdata = test_data,
                            type = "response")
mod_predictions <- round(mod_dem_pred_log) + 1
mod_out <- rep("Nondemented", 63)
mod_out[mod_predictions>1] = "Demented"
mod_out <- factor(mod_out, levels =
                    c("Nondemented", "Demented"))
```

```
mod_acc_out <- sum(as.numeric(mod_out) ==
                     test_data[,1])/nrow(test_data)
mod_acc <- round(100*mod_acc_out, 2)
mod_acc
```

Boruta Feature Selection

```
library(Boruta)
boruta_data <- clean_data
boruta_data$Group <-
  as.numeric(boruta_data$Group)
boruta_X <- boruta_data[,2:10]
boruta_y <- boruta_data[,1]
boruta_features <- Boruta(boruta_y~boruta_X,
                          doTrace = 1)
decision<-boruta_features$finalDecision
signif <-
  decision[boruta_features$finalDecision
           %in% c("Confirmed")]
plot(boruta_features,
     xlab="", main="Variable Importance")
attStats(boruta_features)
```

B Additional Tables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	60.00000	71.00000	76.00000	76.71609	82.00000	98.00000
EUDC	6.00000	12.00000	15.00000	14.61514	16.00000	23.00000
MMSE	4.00000	27.00000	29.00000	27.26183	30.00000	30.00000
CDR	0.0000000	0.0000000	0.0000000	0.2728707	0.5000000	2.0000000
eTIV	1106.000	1358.000	1476.000	1493.577	1599.000	2004.000
nWBV	0.6440000	0.7000000	0.7320000	0.7305962	0.7570000	0.8370000
ASF	0.876000	1.098000	1.189000	1.191606	1.293000	1.587000

Table 2: Summary Table

	meanImp	medianImp	minImp	maxImp	normHits	decision
M.F	6.414914	6.524928	5.297504	7.459075	1	Confirmed
Age	4.750888	4.750785	3.442161	6.226710	1	Confirmed
EDUC	7.139083	6.991844	6.170027	8.335614	1	Confirmed
SES	4.131349	3.974328	2.358848	5.584025	1	Confirmed
MMSE	19.202150	19.105658	18.503980	20.615323	1	Confirmed
CDR	59.368130	59.384092	56.076531	61.053958	1	Confirmed
eTIV	7.194745	6.953529	6.080945	8.828396	1	Confirmed
nWBV	8.269627	8.418507	6.873248	9.288482	1	Confirmed
ASF	7.559969	7.582974	6.347123	8.883467	1	Confirmed

Table 3: Boruta Attribute Table

C Additional Figures

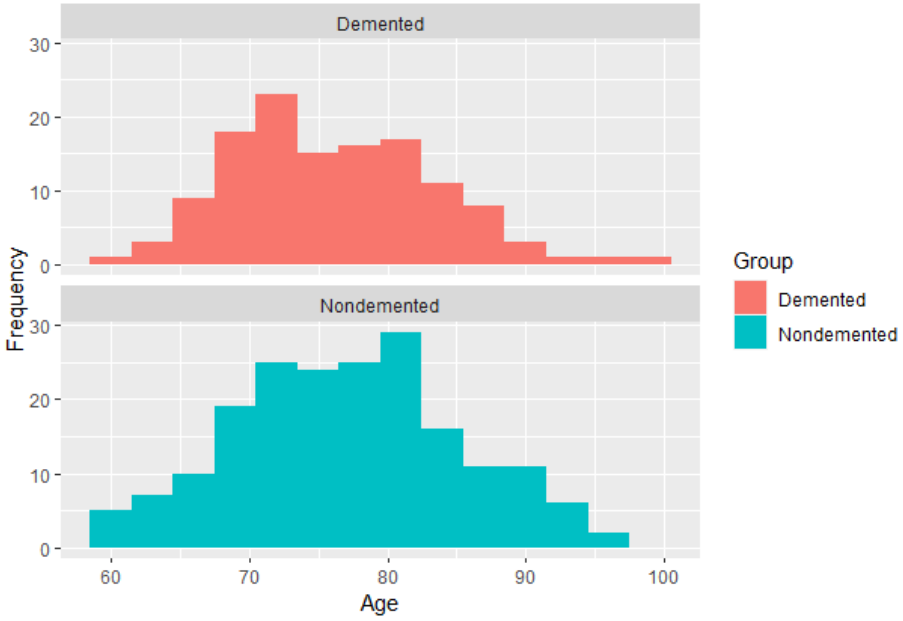

Figure 5: Distribution of Age among the two levels in Group

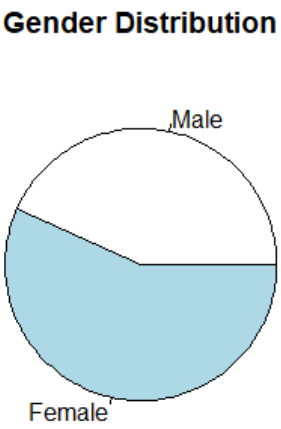

Figure 6: Proportional Distribution of Gender

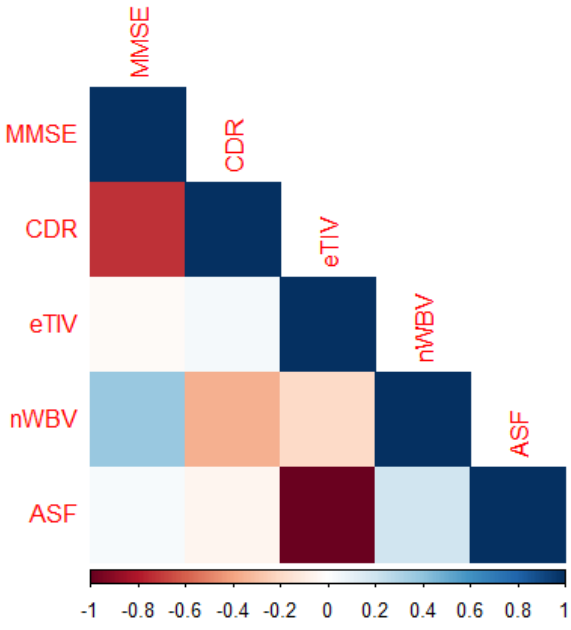

Figure 7: Correlation Plot

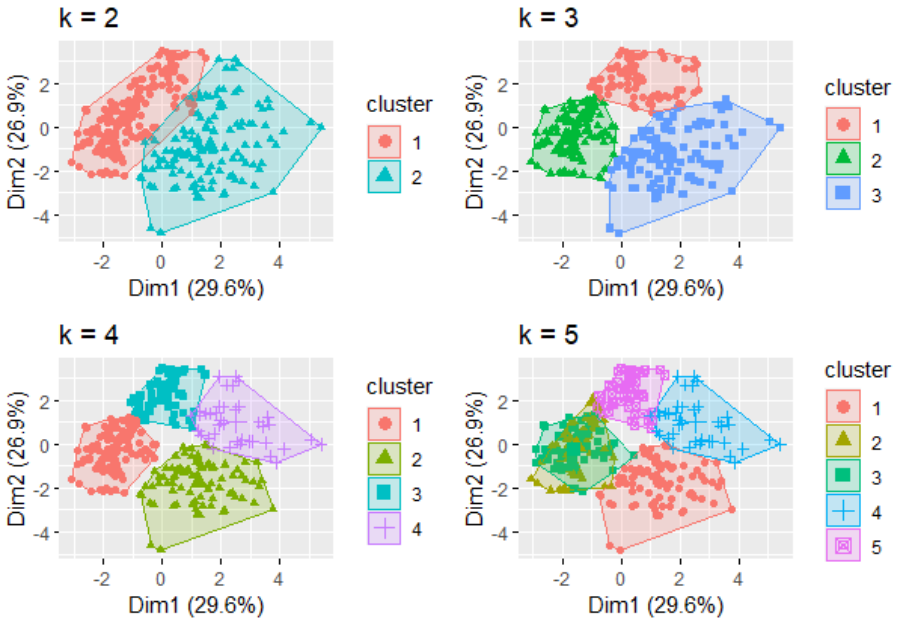

Figure 8: K-Means Clustering for k = 2, 3, 4 & 5

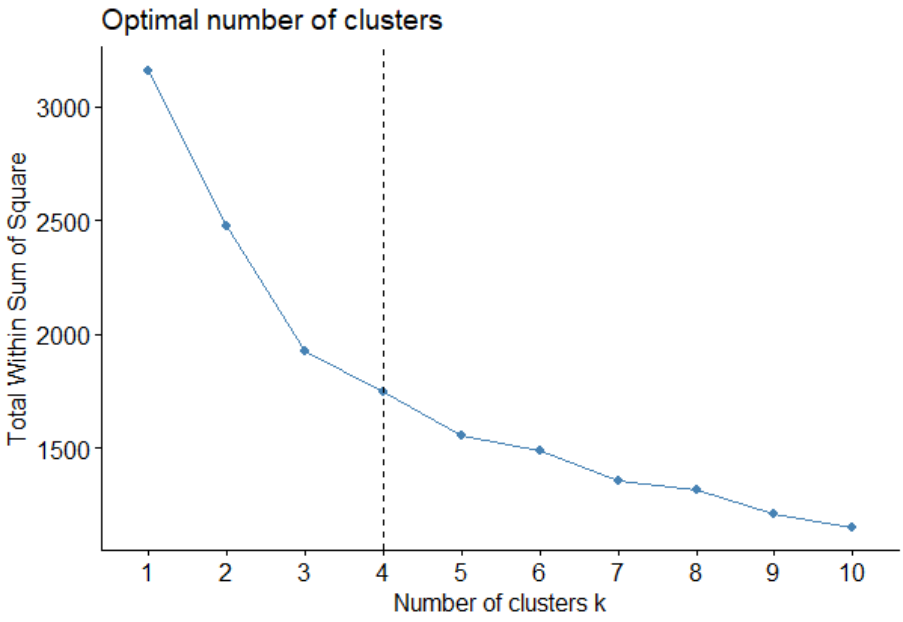

Figure 9: Optimal Number of Clusters

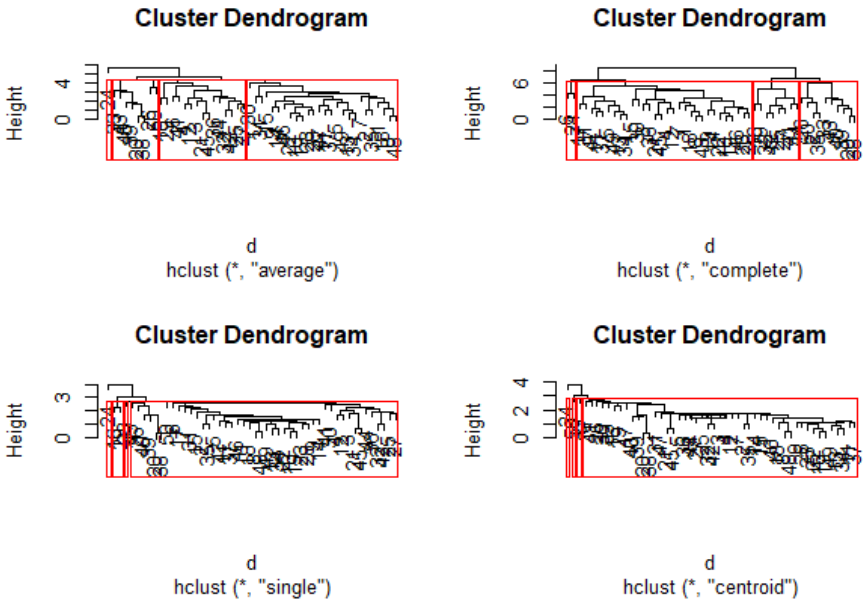

Figure 10: Hierarchical Clustering for Different Linkages

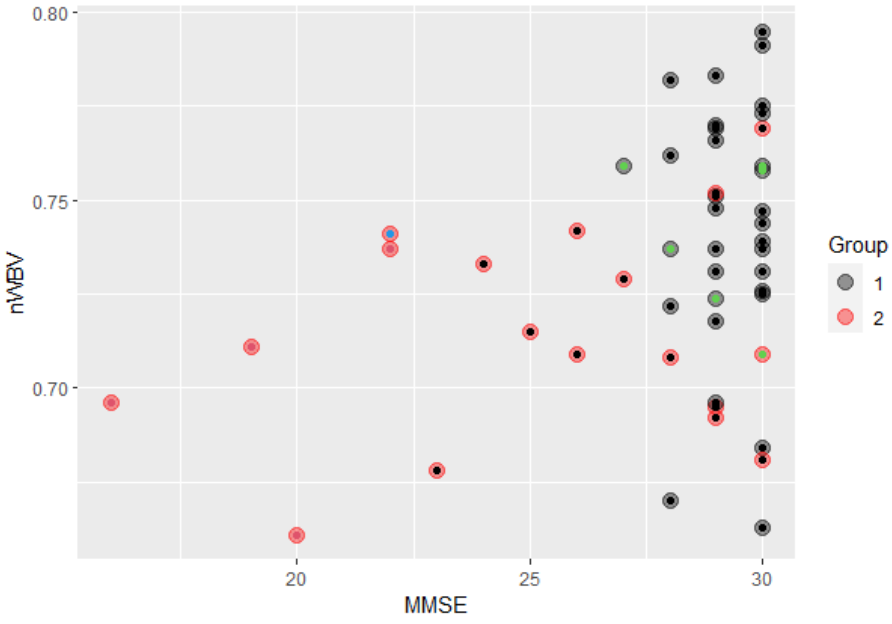

Figure 11: Scatterplot of clusters in relation with Group

```
Call:
glm(formula = Group ~ Age + MMSE + CDR + eTIV, family = "binomial",
    data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.922e+03  2.251e+05   0.009    0.993
Age         -1.418e+01  1.635e+03  -0.009    0.993
MMSE        -2.297e+01  3.037e+03  -0.008    0.994
CDR          8.068e+02  9.007e+04   0.009    0.993
eTIV        -2.936e-01  3.559e+01  -0.008    0.993

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.3683e+02  on 253  degrees of freedom
Residual deviance: 2.0388e-07  on 249  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 25
```

Figure 12: Summary of Forward Feature Selection

```
Call:
glm(formula = Group ~ Age + MMSE + CDR + eTIV, family = "binomial",
    data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.922e+03  2.251e+05   0.009    0.993
Age         -1.418e+01  1.635e+03  -0.009    0.993
MMSE        -2.297e+01  3.037e+03  -0.008    0.994
CDR          8.068e+02  9.007e+04   0.009    0.993
eTIV        -2.936e-01  3.559e+01  -0.008    0.993

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.3683e+02  on 253  degrees of freedom
Residual deviance: 2.0388e-07  on 249  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 25
```

Figure 13: Summary of Backward Feature Selection

```
Call:
glm(formula = Group ~ Age + MMSE + CDR + eTIV, family = "binomial",
    data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.922e+03  2.251e+05   0.009    0.993
Age         -1.418e+01  1.635e+03  -0.009    0.993
MMSE        -2.297e+01  3.037e+03  -0.008    0.994
CDR          8.068e+02  9.007e+04   0.009    0.993
eTIV        -2.936e-01  3.559e+01  -0.008    0.993

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.3683e+02  on 253  degrees of freedom
Residual deviance: 2.0388e-07  on 249  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 25
```

Figure 14: Summary of the Modified Logistic Regression Model