Exploratory Analysis and Predictive Modeling of Alzheimer's Disease Diagnosis

MODELLING EXPERIMENTAL AND OBSERVATIONAL DATA

2211466 | MA335

ABSTRACT

This report explores the relationship between various characteristics and the diagnosis of Alzheimer's disease. Using a dataset containing demographic factors, cognitive assessment scores, and socio-economic status, we conducted descriptive statistics, implemented clustering algorithms, fitted a logistic regression model, and performed feature selection. Through these analyses, we aimed to gain insights into factors influencing Alzheimer's disease diagnosis and identify predictive markers. The findings contribute to understanding the disease's etiology, facilitating early detection, and informing targeted intervention strategies.

Table of Contents

Word Count: 1816

1. INTRODUCTION

The aim of this report is to investigate the relationship between various characteristics of individuals and their diagnosis as either Alzheimer's (Demented) or not (Nondemented). By analyzing a dataset containing demographic, cognitive, and socio-economic factors, we aim to understand the factors associated with Alzheimer's disease and its diagnosis. The report utilizes descriptive statistics, clustering, logistic regression, and feature selection techniques to gain insights into the relationship between these characteristics and the diagnosis. The findings will contribute to our understanding of the disease and inform potential interventions.

2. Pre-liminary Analysis

In the preliminary analysis, started by reading the dataset "project data.csv" and checked the data types of all variables to ensure they are correctly recognized. M/F converted it into a factor variable using the as.factor() function.However, since we need to perform numeric analysis, further converted the "M/F" variable into numeric values using the as.numeric() function. This allowed to assign numerical values to the categories "Female" and "Male" (e.g "Female" is 1 and "Male" is 2).

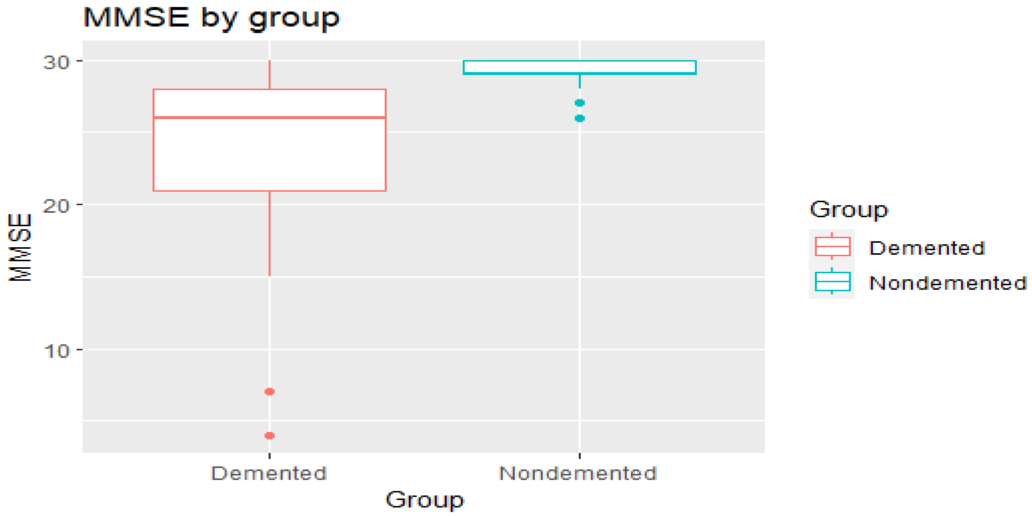
Next, addressed the "Converted" category in the "Group" variable. The "Converted" category represents individuals who did not have symptoms of Alzheimer's disease but showed signs of Alzheimer's during the diagnosis. Since current analysis focuses on the relationship between "Demented" and "Nondemented" categories, decided to remove the rows where the "Group" value was "Converted".Prior to removing the "Converted" rows, the dataset had a total of 373 observations. After removing these rows, left with 363 observations. To ensure a more robust analysis, also removed any rows with missing values using the na.omit() function. This resulted in a final dataset of 317 observations for further analysis. These preliminary steps were necessary to clean and prepare the dataset for subsequent analysis.

3. ANALYSIS

3.1 Descriptive Statistics

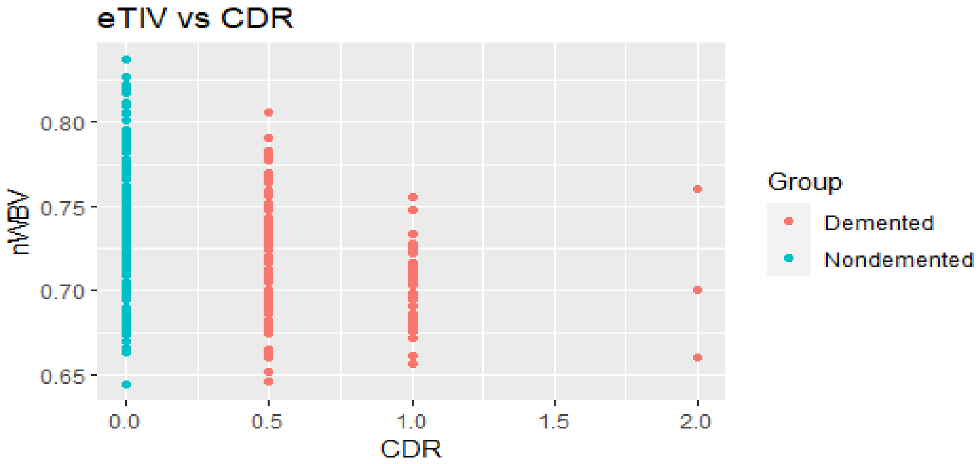
Graphical representation

1.Box plot



The Box plot demonstrates that the "Nondemented" group exhibits a higher average MMSE (Mini Mental State Examination) score compared to the "Demented" group. This suggests a potential association between higher MMSE scores and a lower likelihood of being diagnosed with dementia.

2.Scatter plot



The scatter plot reveals a distinct pattern: individuals classified as Nondemented have a CDR score of 0 and relatively higher values of nWBV (normalized whole brain volume) around 0.83. As the CDR score increases from 0 to 1, representing the progression of dementia, there is a noticeable decrease in nWBV. This trend indicates that individuals with higher CDR scores, particularly in the Demented category, exhibit lower nWBV values, reflecting significant brain volume loss associated with the severity of dementia.

Numerical Representations

Below table is constructed from summary of Demented category and non-Demented category. And the category is filtered using filter () function.

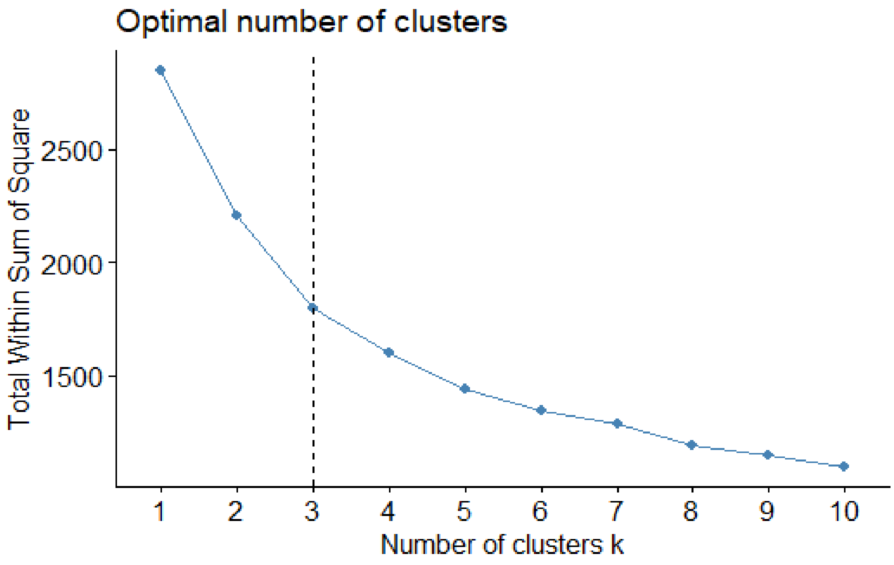
Col Names	Demented		Non-Demented	
Numeric	Mean	Median	Mean	Median
M.F	1.60	2	1.32	1
Age	76.20	76	77.06	77
EDUC	13.80	14	15.14	16
SES	2.77	3	2.39	2
MMSE	24.32	26	29.23	29
CDR	0.67	0.5	0.01	0
eTIV	1490.70	1477	1495.50	1474.5
nWBV	0.72	0.71	0.74	0.74
ASF	1.19	1.19	1.19	1.19

A comparison of the descriptive statistics between the "Non-Demented" and "Demented " datasets reveals important insights about the two groups. The "Non-Demented " group, on average, consists of individuals who are slightly older, have higher levels of education, and exhibit better cognitive function. Their Mini Mental State Examination (MMSE) scores are higher, indicating superior cognitive performance compared to the "Demented " group. Moreover, the "Non_Demented_val" group has a higher proportion of individuals with Clinical Dementia Rating (CDR) scores of 0, indicating no dementia.

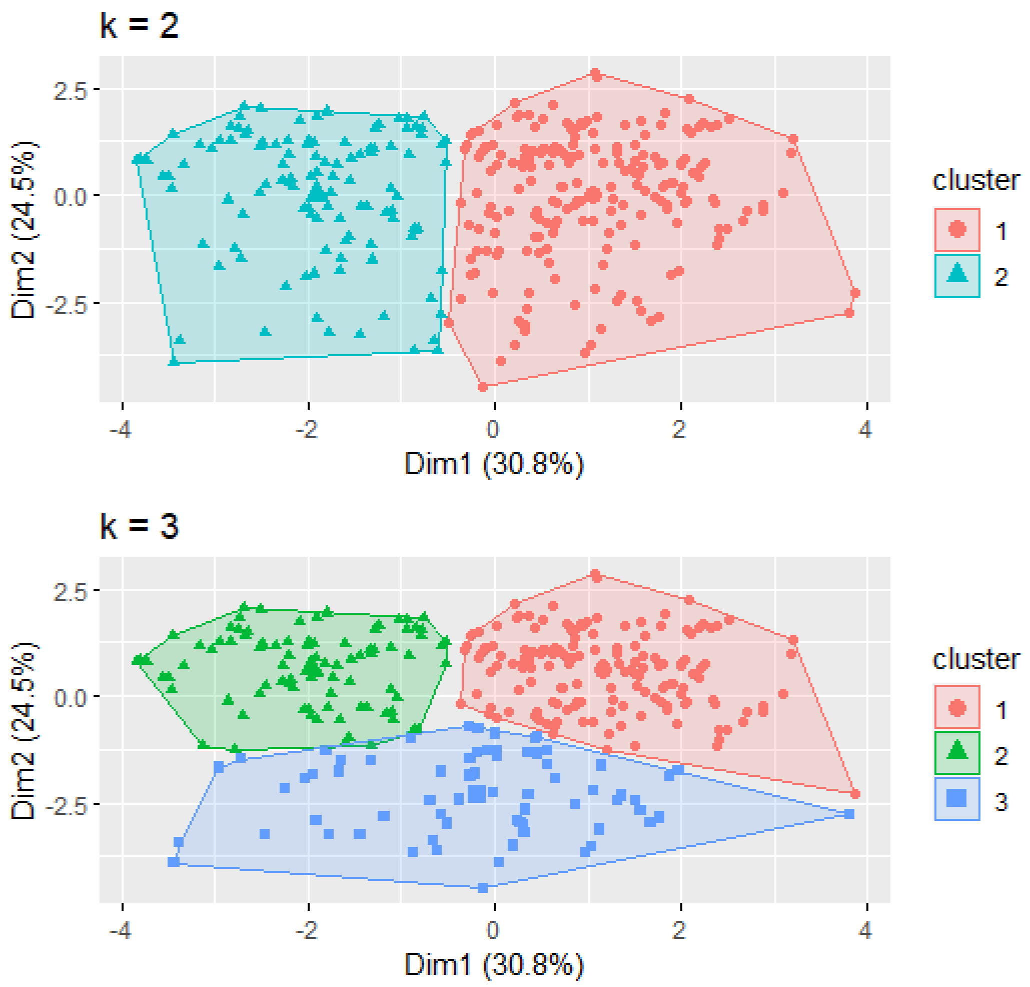
In terms of brain measures, the "Non-Demented " group tends to have slightly larger normalized whole brain volumes (nWBV) and estimated total intracranial volumes (eTIV) compared to the "Demented " group.These findings suggest that age, education, cognitive function, and brain volume may play crucial roles in distinguishing between the non-demented and demented groups. It highlights the potential significance of these factors in understanding the risk factors and characteristics associated with dementia.

3.2 Clustering Algorithm

The clustering algorithm used in the provided code is the K-means clustering algorithm. Here is a brief explanation of the algorithm and the steps involved: Data Preparation: Only numeric variables from the dataset are selected for clustering. This ensures that only numerical features are considered in the clustering process. The selected numeric variables are then scaled to have similar scale units, which is a common preprocessing step in clustering. Determining the Optimal Number of Clusters: The "fviz_nbclust" function is used to determine the optimal number of clusters based on the within-cluster sum of squares (WSS) method. The WSS measures the variability within each cluster, and the optimal number of clusters is typically chosen where the rate of decrease in WSS becomes less pronounced. In this case, the plot suggests that three clusters might be an appropriate choice.



Clustering: The "kmeans" function is used to perform the K-means clustering. The algorithm partitions the data into the specified number of clusters (in this case, 3 and 2) by minimizing the sum of squared distances between the data points and their cluster centroids. The "nstart" parameter controls the number of random initial configurations the algorithm tries to avoid suboptimal solutions. Visualization: The "fviz_cluster" function is used to visualize the clustering results. It creates a scatter plot where each data point is colored according to its assigned cluster. The resulting plots show the data points and the cluster centroids.



Evaluation: It can be seen that when k=2, there is no overlap between the clusters, but when k=3 one or two points of cluster 3 overlaps with cluster 2 in borderline. As overlaps are in border line it can be ignored. When k=2, points are not properly classified as Demented or Non-Demented based on Group values as shown in table 1 inside Appendix. But when k=3, clustering is better. Hence let's analyze Cluster when k=3 as it was also mentioned in elbow method too. Analysis and Interpretation: The "aggregate" function is used to calculate the mean values of the variables within each cluster. This provides insights into the characteristics and profiles of each cluster. The cluster means indicate distinct characteristics of each cluster. Cluster 1 is associated with older individuals, higher cognitive impairment, and lower education and socioeconomic status. Cluster 2 represents a more balanced profile with average values across most variables. Cluster 3 consists of individuals with higher education, larger brain volume, and average cognitive performance. The within-cluster sum of squares (WCSS) suggests moderate compactness within each cluster. These insights provide a glimpse into the demographic, cognitive, and brain-related differences among the clusters.

3.3 Logistic Regression

In this task, logistic regression models were fitted to predict the Group variable in the dataset. The data was split into training and test sets, with 70% of the data used for training and the remaining 30% for testing. Several logistic regression models were trained using different sets of predictor variables. Model1 included all predictors identified as important by the Boruta method, resulting in an accuracy of 0.989. Model2 only considered the CDR variable, which was identified as the most important predictor by Recursive Feature Elimination (RFE) with cross-validation, achieving an accuracy of 1. Model3 included a selected set of predictors determined through stepwise forward selection, achieving an accuracy of 0.957. Model4 focused on the intercept and a subset of predictors, yielding an accuracy of 0.98. Model5 considered a set of least important predictors identified by the Boruta method, resulting in an accuracy of 0.62. Model6 incorporated the most important predictors identified by the Boruta method, resulting in an accuracy of 1.

The models’ predictions were evaluated by calculating probabilities and classifying the test data based on a threshold of 0.5. The accuracy of the predicted classes was determined by comparing them with the actual Group values in the test data. In summary, logistic regression models were used to predict the Group variable based on different sets of predictor variables. The models achieved varying levels of accuracy, with the highest accuracy obtained when considering the most important predictors identified by the Boruta method and RFE. The accuracy results provide insights into the effectiveness of the different models in predicting the Group variable.

3.4 Feature Selection

Method1

In wrapper variable selection methods, two steps were followed to identify important features for predicting the Group variable.First, a model with only an intercept was constructed (model1) using the lm function in R. The purpose of this model was to assess the significance of individual predictors. Then, a step-forward selection procedure (step1_For) was applied to gradually add predictors to the model based on their significance. The scope of predictor variables included M.F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, and ASF.The first set, identified by the model with the intercept-only (model1), consisted of CDR, M.F, EDUC, Age, ASF, and nWBV. These predictors demonstrated significant contributions to the model's predictive power. The second set, confirmed by the model with all variables (model2), included CDR, M.F, eTIV, and EDUC. These variables were also found to be significant for predicting the Group variable.

Method2

The Boruta analysis (boruta1) was performed on the Group_numeric~. formula, considering all variables in the data_new dataset. The final decision from Boruta was obtained, and variables marked as "Confirmed" were considered important for predicting the Group variable.

Model1 included all predictors, as Boruta confirmed their importance. Model5 was constructed using the variables identified as least important by Boruta, namely SES, Age, and M.F. Model6 was developed using the predictors considered most important by Boruta, which were CDR, MMSE, nWBV, and ASF.The Boruta analysis provided insights into the relative importance of predictors for predicting the Group variable.

Method3

RFE with Cross Validation, was used to identify the top predictor for the Group variable. The analysis revealed that CDR was identified as the most important predictor. As a result, Model2 was constructed, specifically considering the CDR variable. RFE with Cross Validation provided valuable insights into the significance of predictors, ultimately contributing to the development of a predictive model focused on the key factor of CDR in relation to the Group variable.

4. DISCUSSION

The analysis aimed to investigate the relationship between various characteristics and the diagnosis of Alzheimer's disease, specifically distinguishing between individuals diagnosed as "Demented" or "Non-Demented." Key findings include differences in age, education, cognitive function, and brain measures between the two groups. Clustering analysis revealed distinct subgroups within the dataset, and logistic regression models effectively predicted the diagnosis based on selected predictors.

5. CONCLUSION

This study provides insights into the relationship between demographic, cognitive, and brain-related characteristics and the diagnosis of Alzheimer's disease. The identified differences in age, education, cognition, and brain measures between the "Demented" and "Non-Demented" groups suggest their potential significance in distinguishing between the two. The clustering analysis and logistic regression models contribute to understanding the factors influencing the diagnosis of Alzheimer's disease. These findings have implications for early detection and tailored interventions targeting Alzheimer's disease.

6. APPENDIX

```
library(dplyr);library(caret);library(MASS);library(prettyR);library(ggplot2);
library(cluster);library(factoextra);library('faraway');library('ISLR');library('corrplot')
library(Boruta) ; #Read the dataset for analysis; data <- read.csv("project data.csv")
#Check the data types of all variables ; glimpse(data)
#Pre-Analysis
#As M/F is Chr converting it into numeric ; data$M.F <- as.factor(data$M.F)
data$M.F <- as.numeric(data$M.F); #Before removing totally 373 observations ;
#Female 1 Male 2 ; #Remove rows , where Group value is Converted and missing values.
data<- data[data$Group != "Converted",] ; #After removing totally 363 observations , So
10 rows had value Converted. ; #Remove missing values; data_new <- na.omit(data) ;
#After removing missing values 317 observations
#Task 1 - Descriptive Statistics  ;#Numerical representations
Demented_val <- data_new %>% filter(data_new$Group=="Demented")
Non_Demented_val <- data_new %>% filter(data_new$Group=="Nondemented")
summary(Demented_val) ; summary(Non_Demented_val)
# Graphical representations ; #Boxplot ;ggplot(data_new , aes(x=Group, y=
MMSE,color=Group)) + geom_boxplot() + labs(title = "MMSE by group")
#Scatterplot; ggplot(data_new , aes(x=CDR,y=nWBV, color=Group)) + geom_point() +
labs(title= "eTIV vs nWBV")
#Task 2 K-means Clustering Algorithm
#Only numeric variables must be used while clustering
data_numeric <- select(data_new, where(is.numeric))
#scaled data ensures all variables in similar scale units. ; data1 <- scale(data_numeric)
#Determining the optimal number of clusters
fviz_nbclust(data1, kmeans, method = "wss")+ geom_vline(xintercept = 3, linetype = 2)
#Clustering
#Using set.seed(), So that everytime same points are used for clustering. ; set.seed(123)
kmeans3 <- kmeans(data1, centers = 3, nstart = 20)  ; kmeans3
kmeans2 <- kmeans(data1, centers = 2, nstart = 20)
```

```
#To visualise the results the fviz_cluster function can be used:
f1 <- fviz_cluster(kmeans2, geom = "point", data = data1) + ggtitle("k = 2")
f2 <- fviz_cluster(kmeans3, geom = "point", data = data1) + ggtitle("k = 3")
#Display Cluster plots ; library(gridExtra) ; grid.arrange(f1, f2, nrow = 2)
#To check classification ; table(kmeans3$cluster,data_new$Group)
table(kmeans2$cluster,data_new$Group)
#Use the aggregate() function to find the mean of the variables in each cluster:
aggregate(data1, by=list(cluster=kmeans3$cluster), mean)
#Task4 Fit logistic regression model  ; data_new$Group <- as.factor(data_new$Group)
#data_new$Group <- as.numeric(data_new$Group)
# Split the data into training and test set ; set.seed(123)
training.samples <- data_new$Group %>% createDataPartition(p = 0.7, list = FALSE)
train.data  <- data_new[training.samples, ] ; test.data <- data_new[-training.samples, ]
#Training the model
#Model1 with accuracy 0.989 based on all predictor confirmed by Boruta method
model1 <- glm( Group ~ ., data = train.data, family = binomial)
#Model2 with accuracy 1 based on CDR as confirmed by RFE with Cross validation
model2 <- glm( Group ~ CDR, data = train.data, family = binomial)
#Model3 with accuracy 0.957 based on Step forward model of all variables
model3 <-glm(Group~ CDR +  M.F + EDUC + Age + ASF +
nWBV,data=train.data,family=binomial)
#Model4 with accuracy 0.98 based on Step forward model of only intercept
model4 <-glm(Group~ CDR +  M.F + eTIV + EDUC ,data=train.data,family=binomial)
#Model5 with accuracy 0.62 based on least important predictors confirmed by Boruta
method ; model5<-glm(Group~ Age + SES + M.F,data=train.data,family=binomial)
#Model6 with accuracy 1 based on most important predictors confirmed by Boruta method
model6 <-glm(Group~ CDR + MMSE + nWBV + ASF,data=train.data,family=binomial)
#Prediction by model ; probabilities <- model1 %>% predict(test.data, type = "response")
contrasts(test.data$Group)
predicted.classes <- ifelse(probabilities > 0.5, "Nondemented", "Demented")
```

```
#Accuracy of the model ; mean(predicted.classes == test.data$Group)
#Task5 Feature Selection method
#To find most important features. ; #Method 1
#Wrapper variable selection methods
Group_factor <-as.factor(data_new$Group) ; Group_numeric <-as.numeric(Group_factor)
#Model with intercept only ; model1<-lm(Group_numeric~1,data=data_new[,-1])
#Step forward ; step1_For <-step(model1,scope = ~ M.F + Age + EDUC+SES + MMSE +
CDR + eTIV + nWBV + ASF,method='forward')
#Model with all variables ; model2<-lm(Group_numeric~.,data=data_new[,-1])
#Step forward ; step1_For_all <-step(model2,method="forward")
#Based on the output of this method Model3 and Model4  were executed.
#CDR +  M.F + EDUC + Age + ASF + nWBV was confirmed by Model with intercept
only
#CDR +  M.F + eTIV + EDUC was confirmed by Model with all variables
#Method 2 : Boruta ; boruta1 <- Boruta(Group_numeric~., data=data_new, doTrace=1)
decision<-boruta1$finalDecision ; signif <- decision[boruta1$finalDecision %in%
c("Confirmed")] ; print(signif)
plot(boruta1, xlab="", main="Variable Importance") ; attStats(boruta1)
#Based on the output of this method Model1 Model5 , Model6  were executed.
#Boruta also confirmed that all predictors are good(Model1)
#Boruta output suggests CDR,MMSE,nwBV & ASF as Most important
predictors(Model6).
#Boruta output suggests SES,Age & M.F as Least important predictors(Model5).
#Method 3 : RFE with Cross validation
# Define the control parameters for RFE control <- rfeControl(functions = rfFuncs, method
= "cv", number = 5)
# Perform feature selection using RFE
rfe_results <- rfe(x = data_new[, -1], y = data_new$Group, sizes = c(1:ncol(data_new) -
1), rfeControl = control) ; # Print the results ; print(rfe_results)
#As this method says CDR is top predictor Model2 was executed.
```