MA335 - FINAL PROJECT

REG NO: 2211560

Musawer Shah
11 JUNE 2023

Contents

Abstract

Alzheimer's is a common disease of the brain where the brain's ability to think and carry out simple tasks is slowly and gradually destroyed. It is a disease most commonly noted in elderly people. The most common symptom of Alzheimer's is loss of memory, uncertain behaviors, poor judgment, and taking longer to complete most menial tasks. This report aims to investigate and present the primary factors associated with the development of Alzheimer's disease, as well as identify the key factor that plays a significant role in demented and non-demented people. In the report, we will be working on a data set that provides certain useful information regarding a person's diverse socio-economic background and education, their Clinical Dementia Rating (CDR), Mini-Mental State Examination (MMSE) results, and various volumes associated with their brain structure. Apart from this, we will do a feature selection method to reduce the number of parameters and find the attributes that have the most significant impact on our data.

1 Introduction

Alzheimer's is a brain disease that slowly and gradually decreases memory and thinking ability, ability to do most common tasks, and difficulty in speaking. It is most common among elderly persons and its risk increase with age. As the "Alzheimer's Disease Fact Sheet", n.d. suggests, approximately 65 million elderly Americans, aged 65 or older, are diagnosed with dementia or Alzheimer's disease, making it the seventh leading cause of death. In this study, I will be analyzing a data set of patients having both demented and non-demented patients. The data set consists of different findings of patients like Clinical Dementia rating (CDR), Mini-mental State Examination (MMSE), their age, education level, economic background, and brain volume ratios. The data set is composed of patients having three stages of Alzheimer's. 1. Demented patients, 2. Non-demented patient and 3. Converted. As per the guidelines, we won't be considering the converted category, and as for the technology is concerned we will be using R-programming language.

2 Methods Used & Findings

In the context of the study, different kinds of statistical methods are implemented ranging from descriptive analysis to clustering algorithms, cleaning and visual-

izing the data, applying appropriate regression models, and feature selection for demonstrating the most important feature variables. Below are the methods and findings performed on the data set.

2.1 Descriptive Analysis

In this section, we will explore key statistical measures to gain insights into our dataset, specifically focusing on demented patients. By examining measures such as mean, median, standard deviation, and variance, we can understand the variations within this particular subgroup. These statistical measures—mean, median, standard deviation, and variance—offer preliminary insights into the variations of attributes among demented patients. Although our analysis is concise due to word constraints, these measures are valuable tools for understanding the dataset and informing decision-making processes. Below is the output of that analysis.

Table 1: Statistical Measures for Demented Patients

Name	Mean	Median	Standard Deviation	Variance	Min	Max
Age	76.2	76	7.35	54.02	61	98
EDUC	13.83	14	3.03	9.16	6	20
MMSE	24.32	26	4.66	21.7	4	30
SES	2.77	3	1.2	1.43	1	5
CDR	0.67	0.5	0.3	0.09	0.5	2
eTIV	1490.7	1477	172.38	29714.29	1143	1957
nWBV	0.72	0.71	0.03	0	0.646	0.806
ASF	1.19	1.19	0.13	0.02	0.897	1.535

The provided table offers valuable insights into the dataset. For example, by examining the age values, we can observe that the majority of patients have an average age of 76. Similarly, the EDUC level and SES values also indicate that most individuals fall within a certain range. Additionally, the standard deviation of the Clinical Dementia Rating (CDR) scores, which is approximately 0.3, signifies the degree of variability or diversity in dementia severity among the patients. This implies that there is a moderate amount of variation in the level of cognitive impairment within the dataset. Furthermore, the minimum and maximum values for each feature provide the lowest and highest ranges, offering a comprehensive understanding of the data distribution. Please refer to Table 3 in Appendix 3 for the summary of non-demented patients.

Below is a box plot for Age for both demented and non-demented groups based on their gender.

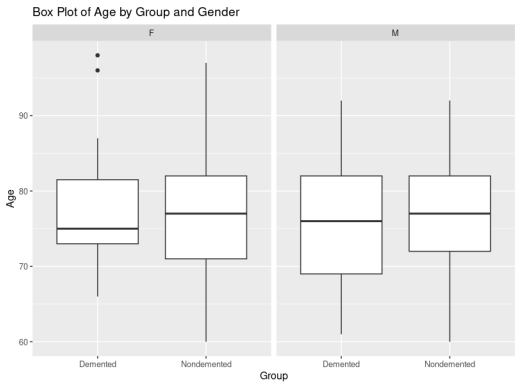

Figure 1: Box plot of Age based on Group & Gender

Based on the provided box plot, we can observe certain patterns in the data. For female demented patients, the majority of data points lie within the age range of approximately 60 to 80. The median age for this group appears to be around 75, suggesting that the central tendency of ages is relatively high. In general, both demented and nondemented individuals show similar median ages, with a value of around 75. The spread of ages within these groups, as represented by the box height (interquartile range), also appears to be relatively consistent. However, it is worth noting that the age distribution for male patients, both demented and non-demented, seems to be more symmetrical or normally distributed compared to the female patients. The box plot suggests a potential positive skewness for the age distribution of female demented patients, with a longer right tail. For the distribution of age please refer to Figure: 5 in Appendix 5.

2.2 Clustering Algorithm

In the context of clustering analysis, we will employ the widely used K-means clustering algorithm. K-means clustering stands as a popular technique for uncovering latent patterns within a dataset. In this particular study, our focus revolves around user profile characteristics, namely age, socioeconomic status, and years of education, as potential factors associated with dementia. By utilizing these specific features, we aim to partition the data into distinct clusters, thus identifying the connection and association between these factors. In this process, our ini-

tial step involves extracting three specific columns (Age, SES, EDUC) from the primary dataset of demented patients. Subsequently, we proceed to scale each column, ensuring that none of them is prioritized due to having larger values. Moving forward, we employ the K-means clustering function, executing it with multiple values of K to determine the most suitable number of clusters. To visualize the outcome, an illustrative image showcasing the resulting clusters obtained from different K values is presented below.

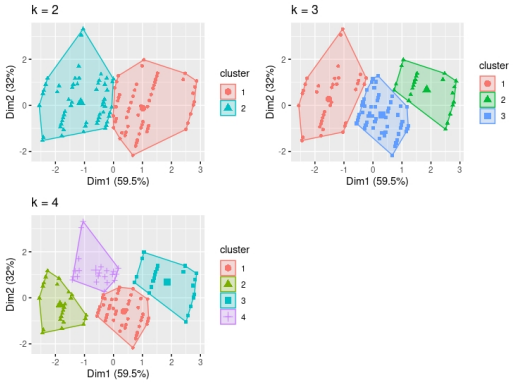

Figure 2: k-Mean Clustering with Different K-Value

In the above image, we use different values of k thus returning a corresponding number of clusters. We can set the value of k to whatever we like but in order to find the optimal number of clusters (k-value) we will use the elbow method to determine the optimal number of k-values.

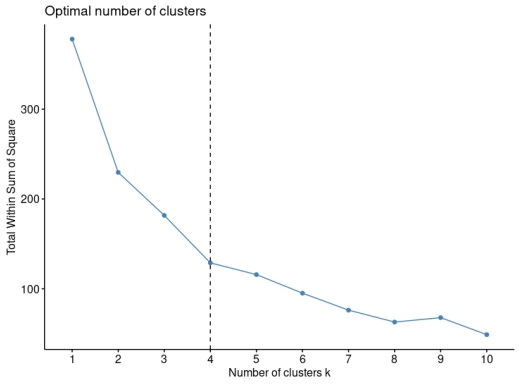

Figure 3: Optimal Clustering with WSS

The provided image illustrates a plot showcasing the relationship between the number of clusters (k) and the Within-Cluster Sum of Squares (WSS) in a K-means clustering analysis. Based on the image, it can be observed that the curve gradually decreases as the number of clusters increases. However, beyond 4 or 5 clusters, the curve begins to flatten, indicating that the reduction in WSS becomes less significant with each additional cluster. This flattening of the curve suggests that increasing the number of clusters beyond 4 may not substantially improve the clustering performance or capture additional meaningful patterns in the data. Therefore, it is reasonable to consider 4 as the optimal value for k in this analysis.

2.3 Feature Selection

Feature selection is the process of selecting the most important and optimal features(independent variables). If for a dataset there are a lot of different features and we want to reduce the number of features we use different kinds of feature selection methods.

In our dataset, despite having a limited number of diverse features, we aim to employ the Boruta feature selection method for our study. Specifically, we seek to identify the significant factors besides CDR (Clinical Dementia Rating) and MMSE (Mini-Mental State Examination), which are widely used for detecting the patient's cognitive state. To determine the most relevant features apart from CDR and MMSE, we exclude them from the dataset and apply the Boruta function. The Importance feature graph displayed below depicts the significance of each feature.

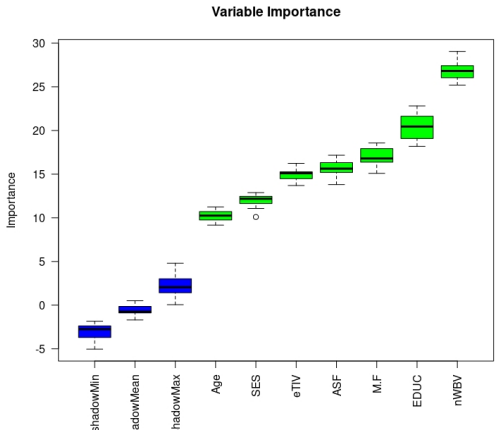

Figure 4: Feature Importance

Based on the provided graph, it can be inferred that, in addition to CDR and MMSE, the next most crucial feature for detecting the demented or non-demented status of patients is the normalized whole brain volume (nWBV). Moreover, the years of education (EDUC) and gender (M.F) also play significant roles in determining the most important feature for distinguishing between patient groups.
Therefore, based on the Boruta feature selection method, it can be concluded that nWBV, EDUC, and M.F are highly influential factors in predicting the cognitive state of patients. These findings highlight the relevance of brain volume, educational background, and gender when considering the classification of individuals into demented and non-demented groups.

2.4 Logistic Regression

Logistic regression is a statistical technique utilized for the classification of categorical target variables. It is specifically designed to predict the probability of an observation belonging to a particular category or class. This method is commonly employed when the outcome of interest is binary, meaning it can only take two distinct values.
To conduct logistic regression, we first select the top five most significant columns from the feature selection process: nWBV, EDUC, M.F, and eTIV. These columns are chosen as our feature variables. Subsequently, the dataset is divided into training and test sets, using a 70% ratio for training and 30% for testing. The training set is used to train the data, employing the five selected variables as features and the "group" column as the target variable. Predictions are then generated for the test set. The predictions provide the probability of each observation belonging to either the Demented or non-demented class. If the probability is equal to or greater than 0.5, the observation is classified as belonging to the Demented class; otherwise, it is assigned to the non-demented class. Based on these predictions, the following outcomes were obtained.

Table 2: Prediction Results

Prediction	Demented	Non-Demented
Demented	89	126
Non-Demented	38	54

The above table summarizes the results of the logistic regression predictions. It presents the number of observations classified as Demented or Non-Demented,

based on the model's predictions compared to the actual class. For example, in the cell where "Prediction" is Demented and "Actual Class" is Demented, there are 89 observations that were correctly predicted as Demented. Similarly, in the cell where "Prediction" is Demented and "Actual Class" is Non-Demented, there are 126 observations that were mistakenly classified as Demented. Upon calculating the accuracy using the provided values, we obtained a result of 0.4511041, indicating that our predictions are correct for approximately 45% of the observations. It is worth noting that the accuracy of the model could have been higher if the datasets were larger or if additional variables such as CDR (Clinical Dementia Rating) and MMSE (Mini-Mental State Examination) were included. Nevertheless, considering the limitations, the model's performance can still be regarded as satisfactory.

3 Conclusion

In conclusion, this report aims to investigate the primary factor associated with the development of Alzheimer's disease and identify the key factor for a patient being demented or non-demented. Various statistical methods were applied to find patterns among datasets and gain valuable insight from them. Different descriptive analysis techniques were done to find variations in data and how the data is distributed. Apart from that, clustering analysis using the K-means algorithm helped uncover latent patterns within the data and how to partition data into their respective clusters. Then some feature selections were done to identify key factors in dementia generation apart from their CDR and MMSE values. And finally, logistic regression was performed using the selected features to classify patients into demented and non-demented groups. The model achieved an accuracy of approximately 45% in predicting the cognitive state based on the provided features.

Overall, this report contributes to the understanding of Alzheimer's disease and highlights the importance of brain volume, education, and gender in predicting cognitive impairment. Further research and analysis with larger datasets and additional variables could improve the accuracy of predictions and provide more comprehensive insights into the disease.

References

Alzheimer's disease fact sheet [Accessed: June 13, 2023]. (n.d.). https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet#:~:text=Alzheimer's%20disease%20is%20a%20brain,first%20appear%20later%20in%20life.

Appendices

A Complete Project Code

Listing 1: Complete Project Code

```
# clear past data
if(!is.null(dev.list())) dev.off()
rm(list = ls())
cat("\014")

# import libraries
library("dplyr")
library("ggplot2")
library("factoextra")
library("gridExtra")
library("Boruta")
library("caret")

# set working directory
setwd("/home/musawer/Documents/data_science/uni/
   observational_modelling/project")

# import data
data <- read.csv("project_data.csv")

#basic structure of data
head(data)
names(data)

# convert some basic columns and remove NA
# remove na
data <- na.omit(data)
# remove converted group
data <- data %>% filter(Group != "Converted")
# convert Group to factor
data$Group <- factor(data$Group)
# convert gender to integer
data$M.F <- as.numeric(factor(data$M.F))
```

```
summary(data)

# some basic visualization and summaries

# summary of demented users
demented_users <- data %>% filter(Group == 'Demented')
summary_of_data <- data.frame();
variable_names <- c("Age","EDUC","MMSE","SES", "CDR","eTIV"
   ,"nWBV","ASF")
for (i in variable_names){
  summary_of_data <- rbind(summary_of_data, c(
    i,
    round(mean(demented_users[,i]), digits = 2),
    round(median(demented_users[,i]), digits = 2),
    round(sd(demented_users[,i]), digits=2),
    round(var(demented_users[,i]), digits=2),
    min(demented_users[,i]),
    max(demented_users[,i])
  ))
}

colnames(summary_of_data) <- c("Name", "Mean", "Median", "
   Standard␣Deviation", "Variance", "Min","Max")
summary_of_data

# summary of non-demented users
non_demented_users <- data %>% filter(Group == 'Nondemented
   ')
summary_of_nondemented_data <- data.frame();
variable_names <- c("Age","EDUC","MMSE","SES", "CDR","eTIV"
   ,"nWBV","ASF")
for (i in variable_names){
  summary_of_nondemented_data <- rbind(summary_of_
     nondemented_data, c(
    i,
    round(mean(non_demented_users[,i]), digits = 2),
    round(median(non_demented_users[,i]), digits = 2),
    round(sd(non_demented_users[,i]), digits=2),
    round(var(non_demented_users[,i]), digits=2),
    min(non_demented_users[,i]),
```

```
    max(non_demented_users[,i])
  ))
}

colnames(summary_of_nondemented_data) <- c("Name", "Mean",
   "Median", "Standard␣Deviation", "Variance", "Min","Max")
summary_of_nondemented_data

# bar plot b/w age and CDR
data %>% group_by(Group) %>%
  select(Age,Group) %>%
  ggplot() +
  geom_bar(aes(x=Age))


ggplot(data, aes(x = Group, y = Age)) +
  geom_boxplot() +
  facet_grid(. ~ M.F) +
  labs(x = "Group", y = "Age") +
  ggtitle("Box␣Plot␣of␣Age␣by␣Group␣and␣Gender")


# performing cluster classification

demented_data <- data %>% filter(Group == "Demented")
# consider only person age,education and status
data1 <- scale(demented_data[,c("Age","SES","EDUC")])

set.seed(123)
kmeans2 <- kmeans(data1, centers = 2, nstart = 20)
kmeans3 <- kmeans(data1, centers = 3, nstart = 20)
kmeans4 <- kmeans(data1, centers = 4, nstart = 20)
kmeans2
str(kmeans2)

fviz_cluster(kmeans2, data = data1)

f1 <- fviz_cluster(kmeans2, geom = "point", data = data1) +
    ggtitle("k␣=␣2")
f2 <- fviz_cluster(kmeans3, geom = "point", data = data1) +
```

```
    ggtitle("k␣=␣3")
f3 <- fviz_cluster(kmeans4, geom = "point", data = data1) +
    ggtitle("k␣=␣4")

grid.arrange(f1, f2, f3, nrow = 2)
print(fviz_nbclust(data1, kmeans, method = "wss")) +
  geom_vline(xintercept = 4, linetype = 2)

# features selection
newDf <- data.frame(data)
# remove CDR & MMSE as they are main factor
newDf <- newDf  %>%
  dplyr::select(-CDR,-MMSE)

boruta1 <- Boruta(Group ~ ., data=newDf, doTrace=1)
decision <- boruta1$finalDecision

signif <- decision[boruta1$finalDecision %in% c("Confirmed"
   )]
print(signif)
plot(boruta1, xlab="", main="Variable␣Importance",las=2)
attStats(boruta1)


# logistic regression
# Select the four attributes with the maximum importance
   values
selected_cols <- c("Group","nWBV", "EDUC", "M.F","ASF","
   eTIV")
selected_data <- data[, selected_cols]
set.seed(123)
trainIndex <- createDataPartition(selected_data$Group,p
   =0.7, list=FALSE)
trainData <- selected_data[trainIndex,]
testData <- selected_data[-trainIndex,]
# formula for logistic regression
formula <- Group ~  nWBV + EDUC+M.F + eTIV

# Fit the logistic regression model
model <- glm(formula, data = trainData, family = binomial)
```

```
# Make predictions on the training data
predictions <- predict(model,testData, type = "response")

pred <- rep("Demented",317)
pred[predictions < 0.5] = 'Nondemented'

table(pred,selected_data$Group)
mean(pred==selected_data$Group)
```

B Tables & Graphs

Table 3: Statistical Measures for Nondemented Patients

Name	Mean	Median	Standard Deviation	Variance	Min	Max
Age	77.06	77	8.1	65.55	60	97
EDUC	15.14	16	2.74	7.52	8	23
MMSE	29.23	29	0.88	0.78	26	30
SES	2.39	2	1.05	1.1	1	5
CDR	0.01	0	0.05	0	0	0.5
eTIV	1495.5	1474.5	184.89	34183.67	1106	2004
nWBV	0.74	0.74	0.04	0	0.644	0.837
ASF	1.19	1.19	0.14	0.02	0.876	1.587

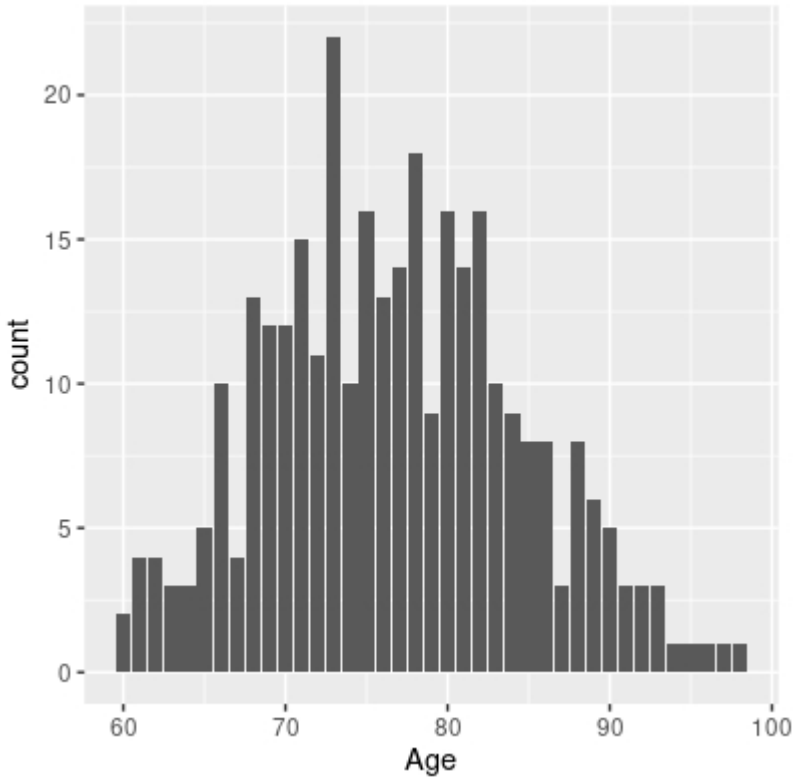

Figure 5: Age Distribution