”Statistical Analysis and Machine Learning Approaches in Alzheimer’s Diagnosis Prediction and Characteristics”

Soliu Kafayat Adewumi (Ks22317)

June 21 2023

Contents

Abstract

The objective of this report is to decipher the association between various patient attributes and the diagnosis of Alzheimer's disease by utilizing sophisticated statistical methods and machine learning approaches. Through a comprehensive examination of a dataset encompassing patient data, we will undertake descriptive statistical analysis, deploy clustering algorithms, fit a logistic regression model, and execute feature selection to pinpoint significant indicators of Alzheimer's diagnosis. The findings present crucial understanding about the determinants influencing the onset of Alzheimer's disease and illustrate the effectiveness of these analytical techniques in the realm of predictive healthcare.

0.1 Introduction

Alzheimer's disease, a particularly challenging type of dementia, has garnered considerable research interest due to its rising incidence rates and profound effects on public health. Accurate diagnosis and comprehensive understanding of this disease are critical for the formulation of efficacious treatment and prevention strategies. This report deploys the tools of data science to deeply analyze a dataset composed of various patient attributes pertinent to Alzheimer's disease.

The dataset under investigation comprises several factors, including the patient's age, gender, years of education, socioeconomic standing, Mini Mental State Examination (MMSE) score, Clinical Dementia Rating (CDR), estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and the Atlas Scaling Factor (ASF). The goal variable, Group, classifies patients into Demented (Alzheimer's), Nondemented, or Other categories.

Our analysis commences with a thorough descriptive examination, incorporating both numerical and graphical data representations. Subsequently, we will employ clustering algorithms to unearth potential patterns and associations within the data. This will be followed by the application of a logistic regression model, utilizing the remaining variables to predict the Group variable. Finally, a feature selection method will be utilized to identify the most impactful predictors in the dataset.

This study aims to shed light on the contributing factors to Alzheimer's disease and highlight the utility of machine learning in the domain of healthcare data analysis. The insights gathered could serve as invaluable input for future research initiatives and the development of therapeutic approaches in the management of Alzheimer's disease.

0.2 Preliminary Analysis

- Pre -processing:
  - Filtering the data to only include rows where `Group` variable is either Nondemented or Demented : The `Group` variable is done using the
  - Converting `M/F` into Numeric Values: The `M.F` column is converted into numeric values using the `ifelse` function. "M" is converted to 1, and "F" is converted to 0.
  - Removing Rows with `Group = "Converted"`: Rows with the value "Converted" in the `Group` column are removed from the dataset.
  - Removing Missing Values: Rows with missing values are removed using the `na.omit` function.

0.3 Analysis and Discussion

0.3.1 Graphical Representation

The boxplot visualizations depicted the distribution of all variables.The boxplot for age showed the range of ages, with the median age around 76 years.The box plot of Atlas Scaling Factor has the median of 1.2. The boxplot for eTIV shows the intracranial volume tha has a median volume around 1500.The boxplot for education showed the spread of education levels, with the median around 15 years of education

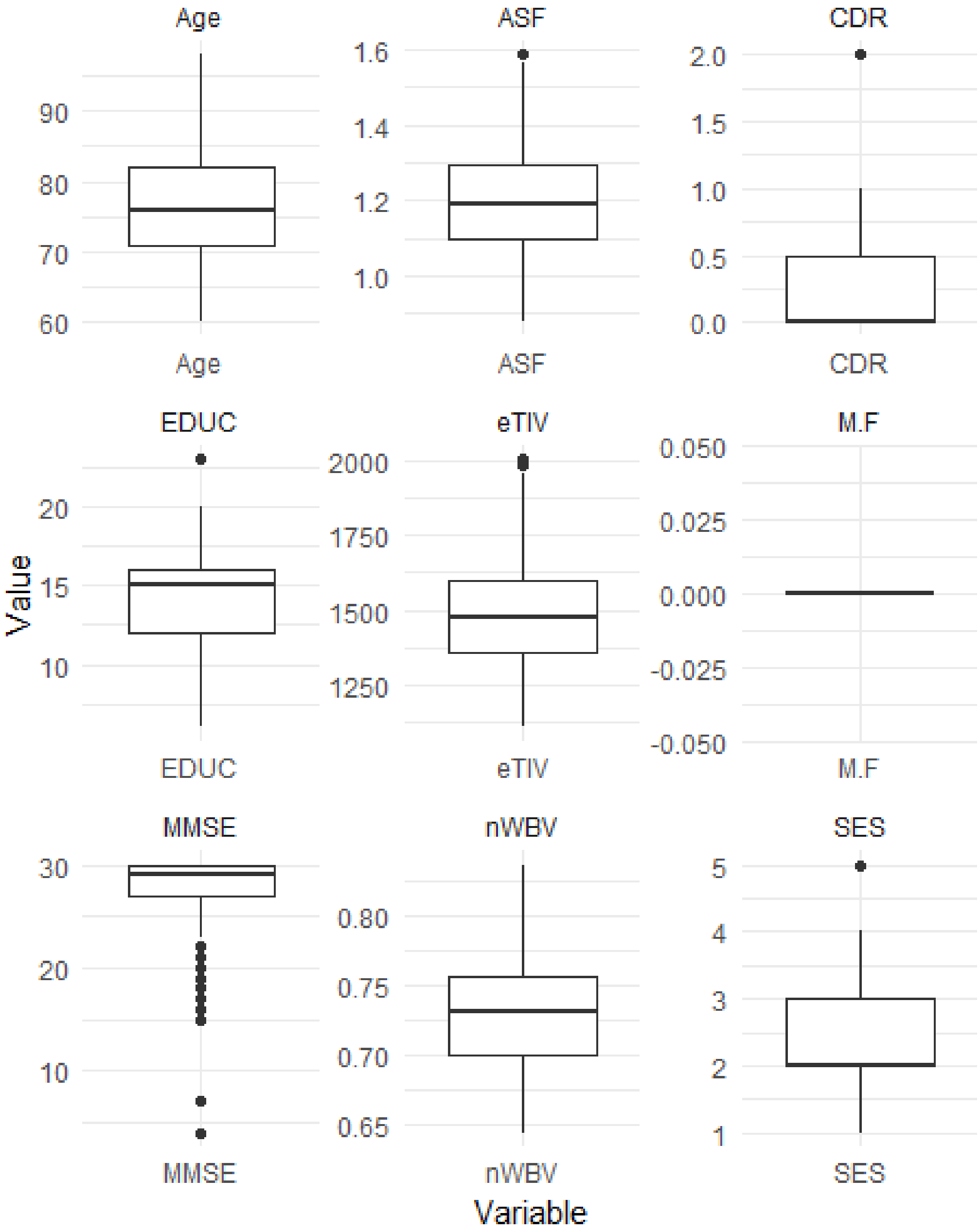

Figure 1: Boxplot Of All Variables

Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Group	0.0000	0.0000	0.0000	0.4322	1.0000	1.0000
M.F	0.0000	0.0000	0.0000	0.4322	1.0000	1.0000
Age	60.00	71.00	76.00	76.72	82.00	98.00
EDUC	6.00	12.00	15.00	14.62	16.00	23.00
SES	1.000	2.000	2.000	2.546	3.000	5.000
MMSE	4.00	27.00	29.00	27.26	30.00	30.00
CDR	0.0000	0.0000	0.0000	0.2729	0.5000	2.0000
eTIV	1106	1358	1476	1494	1599	2004
nWBV	0.6440	0.7000	0.7320	0.7306	0.7570	0.8370
ASF	0.876	1.098	1.189	1.192	1.293	1.587

Table 1: Statiscal overview of Alzheimers disease.

0.3.2 Summary Table

Table 1 above provide a statistical overview of the dataset columns (variables). The quartile values (1st Qu., Median, and 3rd Qu.) signify the 25th, 50th (also the median), and 75th percentiles of the data, respectively. For instance, in terms of 'Age', 25% of individuals are 71 years old or younger, half are 76 or younger,and 75% are 82 years old or younger.

0.3.3 Clustering

The dataset was preprocessed and analyzed using k-means clustering, creating two distinct clusters. The cluster centers, which represent the 'typical' observation within each cluster, varied notably across all variables. The extent of variation differed for each variable, with some showing a broad range and others a narrower one. This suggests that the clusters represent distinct groups within the data and these groups differ along multiple dimensions such as age, education, and various brain measures.

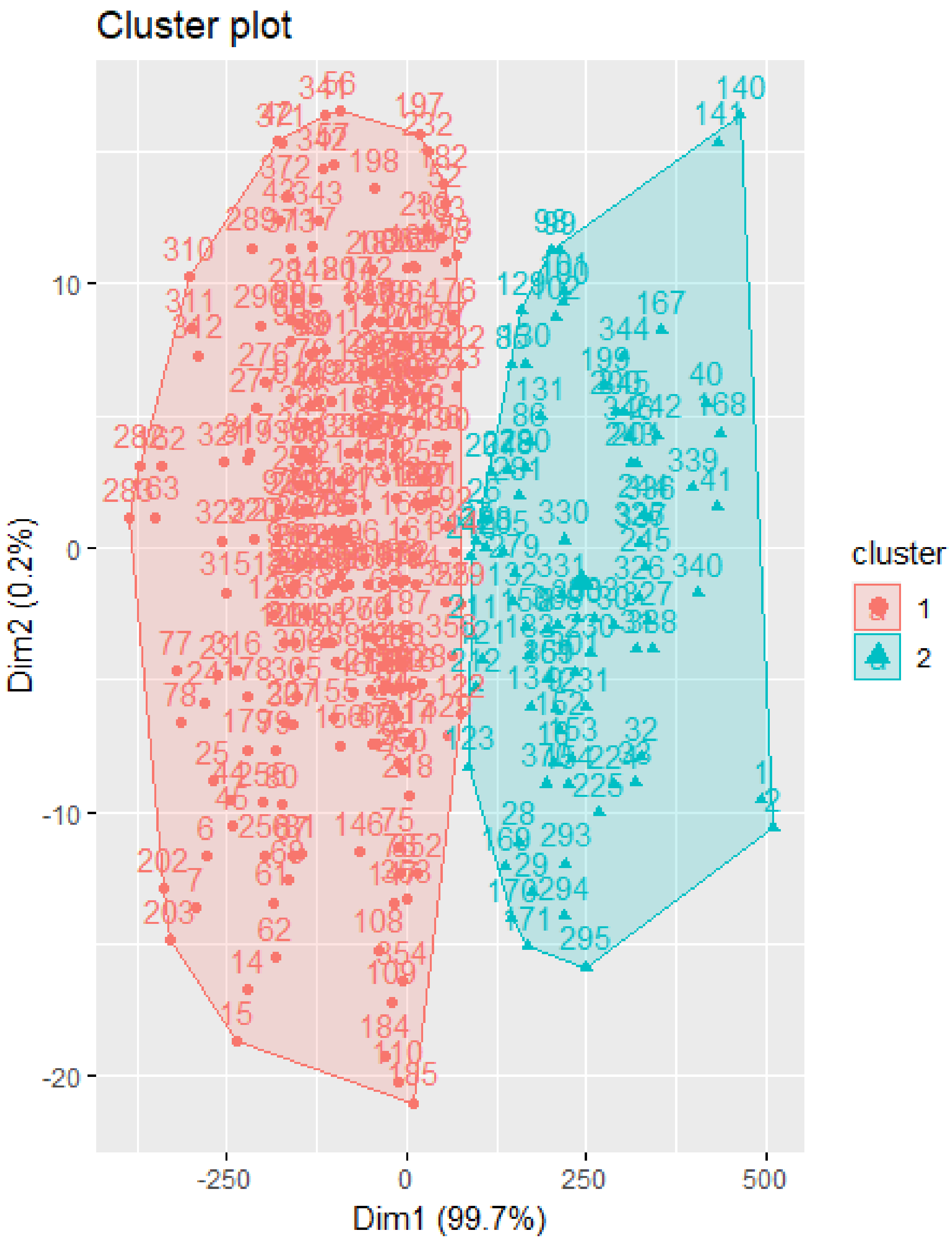

Figure 2: Cluster Plot

0.4 Logistic Regression Analysis

Using generalized linear model (GLM) to predict the outcome of the binary variable 'Group' based on several independent variables (Age, EDUC, MMSE, SES, CDR, eTIV, nWBV, ASF, M.F). This particular GLM is using the logit link function, making it a logistic regression model. The estimates and standard errors for some of the coefficients are extremely large (e.g., for CDR, nWBV, and ASF),suggesting a high degree of uncertainty around those estimates.This might be due to unstable estimates of the model parameters,which could occur if there is multicollinearity,or if the model is overfitting the data,The p-values are almost 1 and none of the predictors appear to be statistically significant,implying that none of these predictors have a statistically significant effect on the outcome 'Group' at usual significance levels.The null and residual deviances show that this model explains the data almost perfectly, which, combined with the earlier issues, could suggest that the model might be overfitting the data.

0.4.1 Feature Selection Method

The procedure of backward and forward elimination was utilized to ascertain the most influential features for predicting the "Group" variable, a diagnostic marker for Alzheimer's disease (AD). This analysis commenced with a comprehensive set of variables: M.F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, and ASF. The evaluation metric employed was the AIC (Akaike Information Criterion). The outcome of the feature selection method indicated that several variables, namely ASF, M.F, EDUC, eTIV, SES, MMSE, Age, and nWBV were sequentially eliminated. The ultimate model constituted the variables Age, SES, CDR, eTIV, and nWBV. The summary of the selected model provides the estimated coefficients for each included variable.

0.5 Conclusion And Recommendations

To sum up, the analysis of the Alzheimer's disease data, pertaining to both demented and non-demented individuals,led to several key findings.Primarily,the data was properly scaled to facilitate precise comparisons and in-depth exploration of the variables.The K-means clustering algorithm,by focusing on the distinct features within each group,effectively distinguished separate categories within the dataset.These insights contribute to our understanding of the elements influencing Alzheimer's disease. They may steer forthcoming research and endeavors in predictive modeling related to this condition.

References

Alzheimer's Association.2021 Alzheimer's Disease Facts and Figures. Frisoni, G. B., Boccardi, M., Barkhof, F., Blennow, K., Cappa, S., Chiotis, K.2017. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. The Lancet Neurology, 16(8)

'

Appendix:Functions

,

Untitled

Soliu Kafayat

2023-06-19

```{r, eval=FALSE, warning=FALSE}

# Load required packages

library(ggplot2)

library(dplyr)

library(tidyr)

library(skimr)

library(cluster)

library(factoextra)

library(caret)

library(summarytools)

library(randomForest)

library(Boruta)

```
```{r, eval=FALSE,
warning=FALSE}

#Loading of dataset
data <- read.csv('project data.CSV')
head(data)
str(data)

#Pre-Processing

#combination of nondemented and demented
data <- data[data$Group %in% c("Nondemented", "Demented"), ]

# Convert M/F into numeric values
data$M.F <- ifelse(data$M.F == "M", 1, 0)

# Remove rows with Group = "Converted"
data <- data[data$Group != "Converted", ]

# Remove rows with missing values
data <- na.omit(data)
```

```{r, eval=FALSE, warning=FALSE}

#No.1
```

Descriptive statistics summary

summary_table <- skim(data) summary(data)

Graphical representations

Boxplots

boxplot_data <- data %>% gather(key = “Variable”, value = “Value”, -Group)

ggplot(boxplot_data, aes(x = Variable, y = Value)) + geom_boxplot() + facet_wrap(~ Variable, scales = “free”) + theme_minimal()

Histograms of all variables

par(mfrow = c(3, 3)) for (i in 3:10) { hist(data[, i], main = ““, col =”blue”, xlab = names(data)[i]) }

~~~
```{r, eval=FALSE,
warning=FALSE}

#No2.

# Select relevant variables for clustering
#clustering_data <- data[, c("Age", "EDUC", "MMSE", "nWBV")]

data$Group <- ifelse(data$Group == "Nondemented", 0, 1)
data_2_zscore <- scale(data[,2:10], center = TRUE, scale = FALSE)

# Perform k-means clustering
set.seed(123)  # For reproducibility
kmeans_model <- kmeans(data_2_zscore, centers = 2, nstart = 10)

# Cluster assignments
cluster_labels <- kmeans_model$cluster
summary(cluster_labels)

# Cluster centroids
cluster_centers <- kmeans_model$centers
summary(cluster_centers)

# Visualize cluster assignments
fviz_cluster(kmeans_model, data = data_2_zscore,stand = FALSE)
~~~

```{r, eval=FALSE, warning=FALSE}

#NO3.

# Convert categorical variables to dummy variables
```

data <- cbind(data, model.matrix(~ Group - 1, data)) data$Group < -factor(data$Group)

Fit logistic regression modelction

model <- glm(Group ~ Age + EDUC + MMSE +SES + CDR + eTIV + nWBV + ASF +M.F, data = data, family = binomial) summary(model)

feature selection using backward sele

step2 <- step(model, method = “backward”) summary(step2)

feature selection using forward selection

model1 <- glm(Group ~ 1, data = data, family = binomial) step1<-step(model1,scope = ~ Group + M.F + Age + EDUC + SES + MMSE + CDR + eTIV + nWBV + ASF, method=‘forward’) summary(step1)

~~~
```{r, eval=FALSE,
warning=FALSE}

#No4

#split data

set.seed(123)
trainingRowIndex <- sample(1:nrow(data), 0.8*nrow(data))  # row indices for 80% training data
trainingData <- data[trainingRowIndex, ]  # model training data
testData  <- data[-trainingRowIndex, ]

y_test<- testData$Group
testData<- testData[,-1]

# dim(testData)
# dim(trainingData)

#Making predictions

y_test<- ifelse(y_test== 1, "Demented", "Nondemented")
glm.probs <- predict(model,newdata = testData,type="response") #Pr(Y=1|X)
confusion_matrix<- table(y_test, glm.probs>0.5)
confusion_matrix

# Running the Boruta algorithm on the training data
set.seed(123)  # For reproducibility
boruta_output <- Boruta(Group ~ ., data = trainingData, doTrace = 0)

# Printing the results
print(boruta_output)

# Plotting the results
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance Plot")

# Get significant variables
significant_vars <- getSelectedAttributes(boruta_output, withTentative = F)
print(significant_vars)
~~~