Investigation of The Relationship Among Selected Characteristics And The Diagnosis of Alzheimer.

Abstract

Alzheimer's disease is the main cause of dementia. [1] The report analyses the relationship of given characteristics with the diagnosis of Alzheimer's disease. The results show that the CDR, M.F, eTIV and EDUC are good features to explain or predict the demented and nondemented groups considered.

Contents

1 Introduction

Alzheimer's disease is the main cause of dementia. [1] The term Dementia is used to describe a category of disease characterized by a decline in cognitive function that impairs daily function.[2] The aim of this report is to investigate the relationship between the given characteristics of Alzheimer and its diagnosis (Demented) or not (Demented). The characteristics include age, gender (M.F), year of education (EDUC), socioeconomic status (SES), mini mental state examination (MMSE), clinical dementia rating (CDR), estimated total intracranial volume (eTIV), normalize whole brain volume (nWBV) and atlas scaling factor (ASF).

2 Preliminary Analysis

The given data set "project data.csv" was loaded and previewed. The M.F variable, which had M and F values, was converted into numeric values with $M = 1$ and $F = 0$. The rows where the Group variable has values "Converted" were removed. Missing values were also removed from the data set.

3 Analysis and Discussion

3.1 Descriptive Analysis

The Table 1 below shows the statistical summary of the Alzheimer's characteristics given in the data set. The Age range from 60 years to 98 years shows that the data was obtained from elderly people with an average age of 76.72. The highest (EDUC) of the individuals in the data set is 23 years and the minimum 6 years. Their (SES) vary from 1 the lowest to 5 the highest. The (MMSE) which is a measure of cognitive function shows that half of the individuals have MMSE score of 29 or less. The highest clinical dementia rating (CDR) is 2 while the minimum is 0. The average (eTIV) is 1494 and the values vary by 179.72 units. The (nWBV) has an average of 0.73 and half

of the individuals have nWBV of 0.73 or lower. The maximum (ASF) is 1.59 and the minimum is 0.88.

Table 1: Statistical Summary Of The Alzheimer's Characteristics

Variables	Mean	Standard Deviation	Minimum	Maximum	Median
Age	76.72	7.81	60.00	98.00	76.00
EDUC	14.62	2.93	6.00	23.00	15.00
SES	2.55	1.12	1.00	5.00	2.00
MMSE	27.26	3.86	4.00	30.00	29.00
CDR	0.27	0.38	0.00	2.00	0.00
eTIV	1494.00	179.72	1106.00	2004.00	1476.00
nWBV	0.73	0.038	0.64	0.84	0.73
ASF	1.19	0.14	0.88	1.59	1.19

The figure 1 below is a box plot that shows the demented females are within the age of 66 to 87 years. Half of them are younger than 75 years and there are outliers above age 95 years. The nondemented females are more spread within the ages 60 to 97 years and 50% of them are younger than 77 years. The demented males are spread over the age of 61 years to 92 years. Half of them are younger than 76 years. The nondemented males are within 60 to 92 years. 50% of them are younger than 77 years.

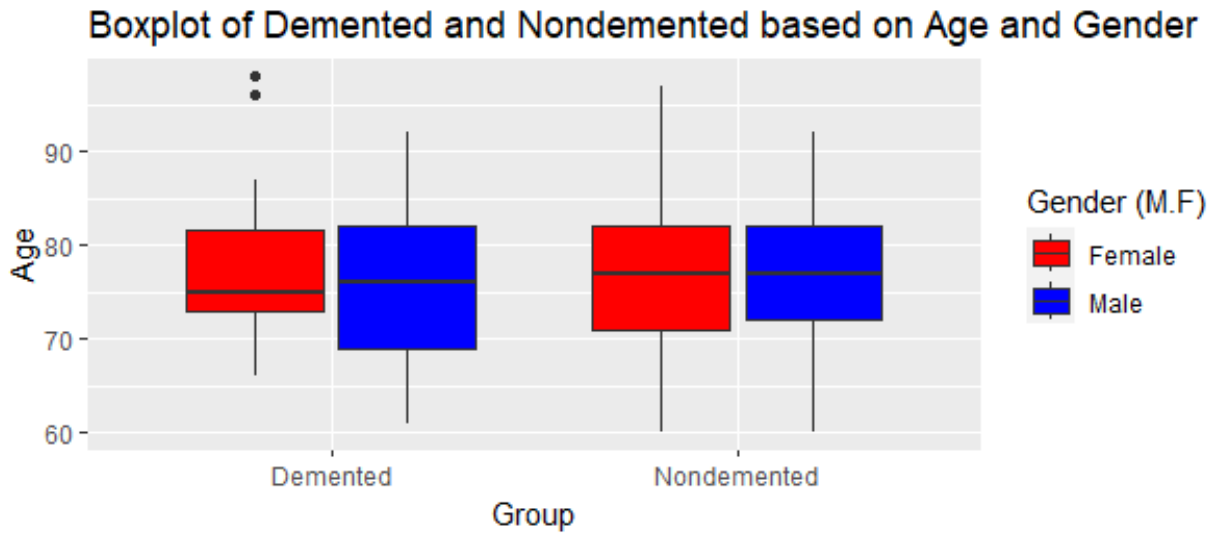

Figure 1: Box plot of the Demented and Nondemented Groups by Gender

There are 180 females in the dataset and 137 males. This indicates that the females are more

than the males. The bar chart also shows that 51 females are demented (28.33% of the total females) while 129 females are nondemented (71.67% of the total females). 76 males are demented (55.47% of the total males) while 61 males are nondemented (44.53% of the total males). The proportions of the female and male groups above suggests that dementia is more prevalent in males with a higher proportion of demented males being 55.47% against 28.33% for demented females. It also suggests that the proportion of nondemented individuals is higher among females with 71.67% for females against 44.53% for nondemented males.

3.2 Clustering Algorithms

Figure 2, indicated that the optimal k value is 3 based on the point where the elbow joint is observed for the K-means clustering algorithm implemented.

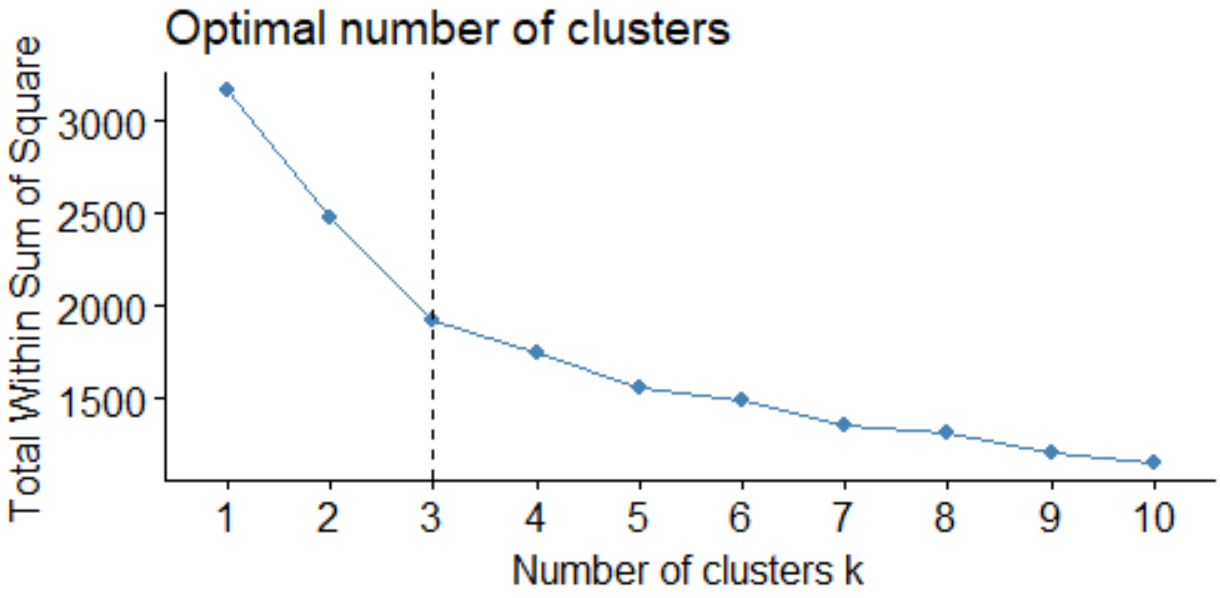

Figure 2: Optimal number of clusters

Figure 3 below shows three clusters that have sizes of 102, 138 and 77 observations in each cluster respectively. Although the data set was prepared with two groups (demented and nondemented), the cluster algorithms has grouped the variables (characteristics) into 3 with similar patterns.

Table 2 below provides details on the centres which is the means of each variable present in

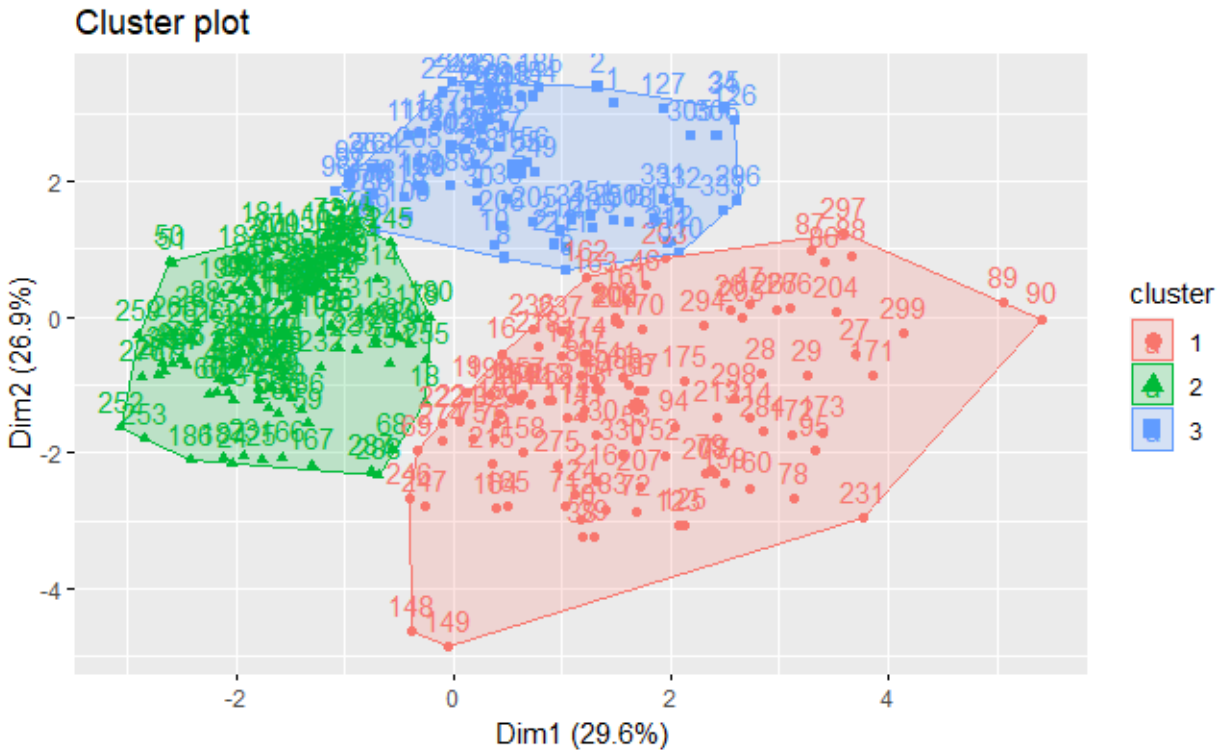

Figure 3: Cluster Plot

each cluster.

Table 2: Cluster Means

	Group	M.F	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
1	1.22	0.26	-0.01	-0.54	0.43	-0.97	1.13	-0.22	-0.52	0.18
2	-0.73	-0.68	-0.18	-0.02	0.00	0.50	-0.66	-0.55	0.53	0.53
3	-0.31	0.88	0.33	0.76	-0.58	0.38	-0.32	1.28	-0.26	-1.20

The values of (CDR) and (MMSE) are strong indicators to tell which group the clusters belong to based on domain knowledge. The CDR is highest in cluster 1 (1.13) with the lowest MMSE (-0.97) which is a measure of cognitive function. This suggests cluster 1 to be the demented group. The least CDR value (-0.66) and highest MMSE value (0.50) suggests that cluster 2 is the nondemented group while cluster 3 with CDR value of -0.32 and MMSE value of 0.38 can be taken as the group with tendencies of becoming demented. But this is subject to further analysis as the mean values of the other characteristics can not be used solely to determine the groups.

3.3 Feature Selection

The backward and forward method were used to select the important features for the model. The backward method gave the b_$model$ as y$_{G}roup\ M.F + Age + EDUC + CDR + nWBV + ASF$ while the forward method gave the f_$model$ as y$_{G}roup\ CDR + M.F + eTIV + EDUC$. The AIC values from the forward method were lower than the AIC values from the backward method. The f_$model$ has less features which might likely make it less complex than the b_$model$. Analysis of variance was carried out on both models. Based on the p-value (0.1587), we fail to reject the null hypothesis. The difference between the two models is not statistically significant at a 0.05 significant level. Therefore, f_$model$ with the fewer features was selected.

3.4 Logistic Regression

A five-fold cross validation was used split the data into 5 subsets. The (glm), (lda) and (knn) classifications were used to fit the logistic regression model (f_$model$) obtained above. The lda, glm and knn results showed accuracy of 0.9937, 0.9937 and 1 respectively. This indicates that the f_$model$ performed very well in the proportion of predictions. Based on the Kappa values of 0.9869, 0.9869 and 1 for the lda, glm and knn models respectively, there is a high level of agreement between the predicted and actual values of each model. The confusion matrix for each of them also show good predictions.

4 Conclusion

The analysis on the data shows that of all the given characteristics associated with the diagnosis of Alzheimer's disease, the (CDR), (M.F), (eTIV) and (EDUC) are sufficient predictors for the response variable (Group) which include demented and nondemented.

References

[1] Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E Teunissen, Jeffrey Cummings, and Wiesje M van der Flier. Alzheimer's disease. *The Lancet*, 397(10284):1577–1590, 2021.

[2] Temitope Ayodele, Ekaterina Rogaeva, Jiji T Kurup, Gary Beecham, and Christiane Reitz. Early-onset alzheimer's disease: what is missing in research? *Current neurology and neuroscience reports*, 21:1–10, 2021.

5 Appendix

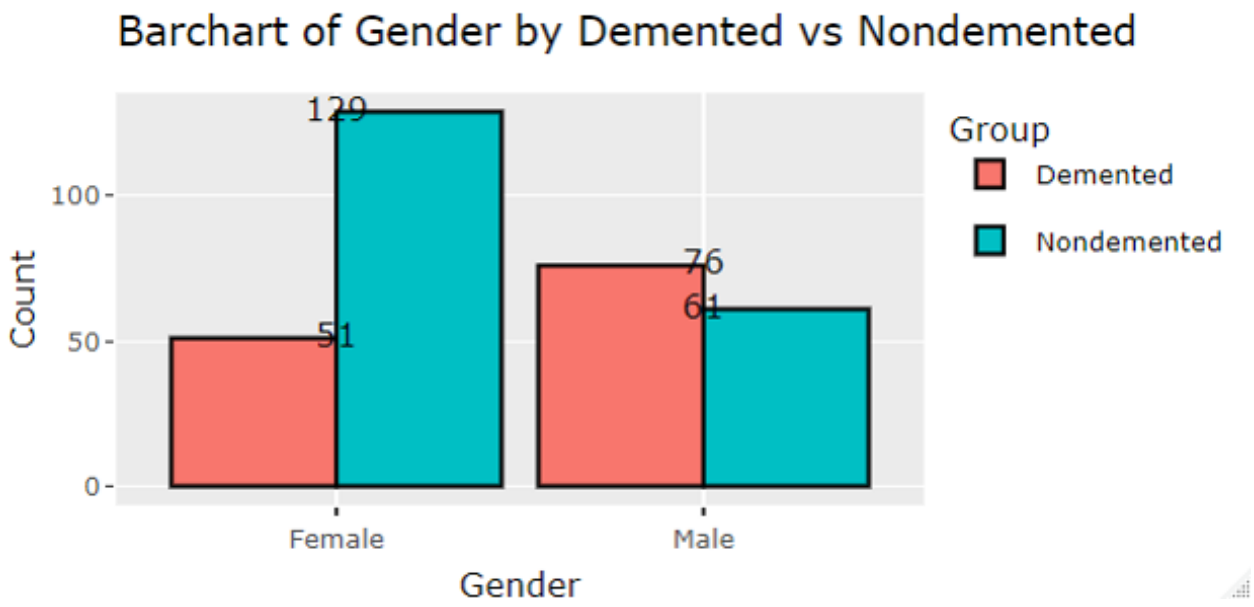

Figure 4: Bar chart of Gender by Demented and Nondemented Groups

```r
# Load libraries
library(dplyr)
library(ggplot2)
library(plotly)
library(tidyverse)
library(cluster)
library(factoextra)
library(gridExtra)
library(caret)
# ============================================================================
# Load data
Alzheimer <- read.csv('project data.csv')
head(Alzheimer)
str(Alzheimer)
```

```r
# Preliminary Analysis
# Convert M/F into numeric values
Alzheimer$M.F <- ifelse(Alzheimer$M.F == 'M', 1,
                        ifelse(Alzheimer$M.F == 'F', 0, NA))

# Confirm the conversion
head(Alzheimer)

# Remove rows with Group = 'Converted'
Alzheimer <- Alzheimer %>%
  filter(Group != 'Converted')

# Remove missing values
Alzheimer <- na.omit(Alzheimer)

# Analysis
# Generate summary of Alzheimer
summary(Alzheimer)
```

```r
# ============================================================================
# Select all numeric variables
attach(Alzheimer)
numeric_vars <- c('Age', 'EDUC', 'SES', 'MMSE', 'CDR', 'eTIV', 'nWBV', 'ASF')
numeric_vars

# Find standard deviation of variables
sds <- apply(Alzheimer[, numeric_vars], 2, sd)
print(sds)

# Create appropriate plots
ggplot(Alzheimer,
       aes(x = Group, y = Age, fill = as.factor(M.F))) +
  geom_boxplot() +
  labs(x = 'Group', y = 'Age',
       title = paste('Boxplot of Demented and Nondemented based on Age and Gender'),
       fill = 'Gender (M.F)') +
  scale_fill_manual(
    values = c("0" = "tomato1", "1" = "lightseagreen" ),
```

```
    labels = c("0" = "Female", "1" = "Male"))
```

```
# Convert M.F to factor
Alzheimer$M.F <- as.factor(ifelse(M.F == 1, 'Male', 'Female'))

gender_G <- ggplot(Alzheimer,
                   aes(x = M.F,
                       fill = Group)) +
  geom_bar(position = 'dodge', color = 'black') +
  geom_text(aes(label = ..count..), stat = 'count', vjust = 0.5, colour = 'black') +
  labs(x = 'Gender', y = 'Frequency',
       title = paste('Barchart of Gender by Demented vs Nondemented'))

ggplotly(gender_G)
```

```
# ========================================================================
# Convert Group variable to numeric values
Alzheimer$Group <- ifelse(Alzheimer$Group == 'Demented', 1,
                          ifelse(Alzheimer$Group == 'Nondemented', 0, NA))

# Convert M/F into numeric values
Alzheimer$M.F <- ifelse(Alzheimer$M.F == 'Male', 1,
                        ifelse(Alzheimer$M.F == 'Female', 0, NA))
head(Alzheimer)

# Similarity measure
distance.Euclidean <- get_dist(Alzheimer)
fviz_dist(distance.Euclidean,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

distance.corr <- get_dist(Alzheimer, stand = TRUE, method = "pearson")
fviz_dist(distance.corr,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

# Standardize features
scaled_A_vars <- scale(Alzheimer)

# Determining the optimal number of clusters
fviz_nbclust(scaled_A_vars, kmeans, method = "wss")+
  geom_vline(xintercept = 3, linetype = 2)

# K-Means Clustering
set.seed(123)
kmeans2 <- kmeans(scaled_A_vars, centers = 2, nstart = 20)
kmeans3 <- kmeans(scaled_A_vars, centers = 3, nstart = 20)
kmeans4 <- kmeans(scaled_A_vars, centers = 4, nstart = 20)
kmeans3

# To visualise the results the fviz_cluster function can be used:
fviz_cluster(kmeans2, data = scaled_A_vars, stand = FALSE)
fviz_cluster(kmeans3, data = scaled_A_vars, stand = FALSE)
fviz_cluster(kmeans4, data = scaled_A_vars, stand = FALSE)
```

```r
f1 <- fviz_cluster(kmeans2,
                   geom = "point", data = scaled_A_vars) + ggtitle("k = 2")
f2 <- fviz_cluster(kmeans3,
                   geom = "point", data = scaled_A_vars) + ggtitle("k = 3")
f3 <- fviz_cluster(kmeans4,
                   geom = "point", data = scaled_A_vars) + ggtitle("k = 4")
grid.arrange(f1, f2, f3, nrow = 2)
```

```r
# ========================================================================
# Implement feature selection on the data set
attach(Alzheimer)
y_Group <- as.numeric(Alzheimer[,1])
X <- Alzheimer[,2:10]

model1 <- glm(y_Group~.,data=X)
summary(model1)

step1 <- step(model1,method="backward")
summary(step1)
# y_Group ~ M.F + Age + EDUC + CDR + nWBV + ASF (Features selected)

model2 <- lm(y_Group~1,data=X)
step2 <- step(model2,
              scope=~ M.F + Age + EDUC + SES + MMSE + CDR+ eTIV + nWBV + ASF,
            method="forward")
summary(step2)
# y_Group ~ CDR + M.F + eTIV + EDUC (Features selected)
b_model <- lm(y_Group ~ M.F + Age + EDUC + CDR + nWBV + ASF)
summary(b_model)

f_model <- lm(y_Group ~ CDR + M.F + eTIV + EDUC)
summary(f_model)

anova(b_model, f_model)
```

```r
# ========================================================================
# Convert Group variable to factor
Alzheimer$Group <- as.factor(Alzheimer$Group)

# Cross Validation (CV)
# For 5-fold CV
trControl <- trainControl(method = "cv", number = 5)

#lda
lda.fit <- train(Group ~ CDR + M.F + eTIV + EDUC,
                 method = "lda",
                 trControl = trControl,
                 metric = "Accuracy",
                 data = Alzheimer)

lda.pred <- predict(lda.fit,Alzheimer)
t1 <- table(lda.pred, Alzheimer$Group)
confusionMatrix(t1)
```

```r
# ========================================================================
#glm
glm.fit <- train(Group ~ CDR + M.F + eTIV + EDUC,
                 method = "glm",
                 trControl = trControl,
                 metric = "Accuracy",
                 data = Alzheimer)

glm.pred <- predict(glm.fit,Alzheimer)
t2 <- table(glm.pred, Alzheimer$Group)
confusionMatrix(t2)
# ========================================================================
#knn
knn.fit <- train(Group ~ CDR + M.F + eTIV + EDUC,
                 method = "knn",
                 tuneGrid = expand.grid(k = 1:10),
                 trControl = trControl,
                 metric = "Accuracy",
                 data = Alzheimer)

knn.pred <- predict(knn.fit,Alzheimer)
t4 <- table(knn.pred, Alzheimer$Group)
confusionMatrix(t4)
```