2023

Exploratory analysis

ALZHEIMER DATASET ANALYSIS

SNEKCATO

ABSTRACT

This Final Project focuses on finding a model able to predict if a person is more likely to have Alzheimer's based on a dataset with many characteristics of patients divided into two groups: demented and not demented. In order to do this, certain tasks will be completed using R: Descriptive analytics, k-means clustering, hierarchical clustering, logistic regression, and feature selection.

Contents

Introduction

Alzheimer's is a type of dementia suffered mostly by seniors that impairs memory and cognitive skills. In order to understand the probable causes of this disease and promote a prevention culture, it is important to perform continuous research.

In this task, the R programming language is used to analyse the dataset "project data.csv" to predict what features are associated with Alzheimer’s. To preprocess the data, the missing values and the "Converted" group have been removed, and the categorical variables have been transformed into numerical variables.

To answer this question, descriptive analytics, k-means clustering, hierarchical clustering, logistic regression, and feature selection will be performed.

1.Descriptive analytics

In the dataset, we can observe that all subjects are senior citizens, with an average age of 76.72 years old, of which 180 are women and 137 are men.

After running a correlation, we can observe that dgroup (0 for non-demented and 1 for demented) is negatively correlated to the normalised whole brain volume (-0.33) and Education years (-0.22). Also, we can note that it is positively correlated to sex (0.27). There is a high correlation between dgroup and the Mini Mental State Examination (-0.73), but we will ignore it since they are correlated because they represent the same diagnosis. Another interesting finding is that normalised brain volume is highly negatively correlated with age and positively correlated with sex.

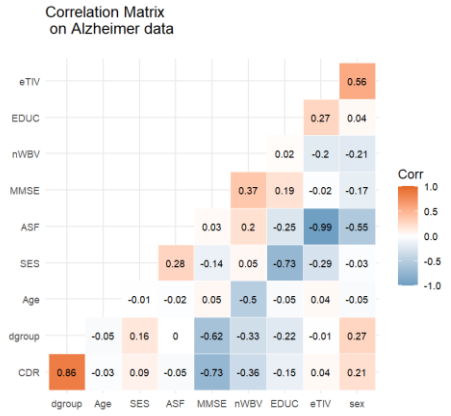


Figure 1

In our dataset, the proportion of men with Alzheimer's (55.47%) almost doubles the proportion of women with Alzheimer's (28.33%).

When comparing demented and non-demented subjects, In figure 2 we can observe that Demented subjects tend to have lower brain volumes than non-demented patients.Also, in Figure 3, it can be observed that those in the Demented group tend to have fewer years of education than the non-demented group.

After this brief descriptive analysis, it makes sense to focus on the variables sex, normalised brain volume, and years of education to create models to successfully predict Alzheimer's.

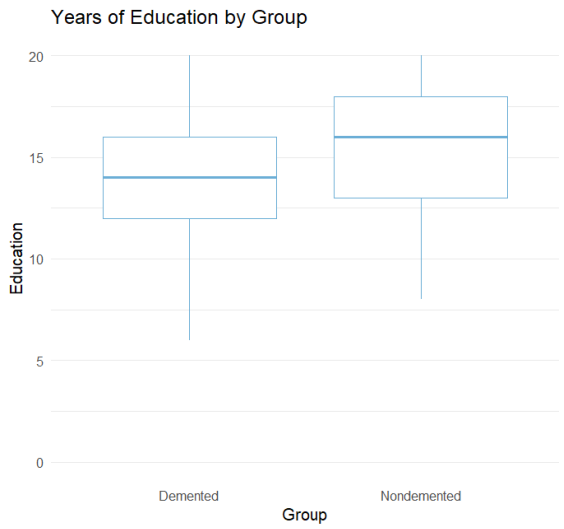


Figure 2

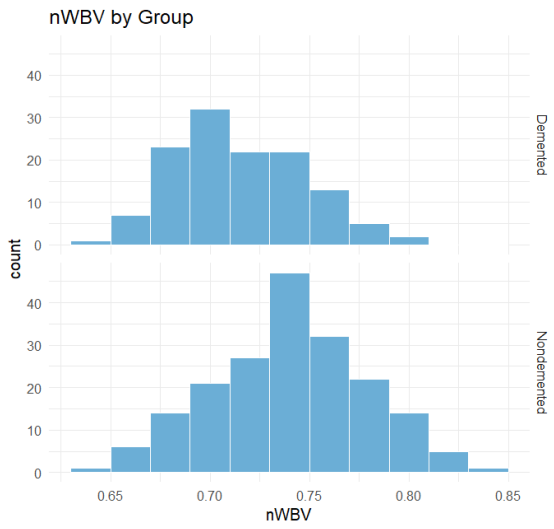


Figure 3

2.Clustering Algorithms

Since our dataset has different scales, we start by normalising the features before using the k-means clustering algorithm with two centres to find the cluster that would predict non-demented patients and the one that would predict demented patients.

In Table 1, we can see clearly how the clusters were distributed. The first cluster contains the highest values of Age, Education, CDR, eTIV, and sex and the lowest values of SES,MMSE, nWBV, and ASF. While the second cluster contains the highest values on SES, MMSE, nWBV, and ASF and the lowest values on Age, Education, CDR, eTIV, and sex.

Table 1

```
K-means clustering with 2 clusters of sizes 112, 205

Cluster means:
         Age       EDUC        SES        MMSE        CDR       eTIV       nWBV        ASF        sex
1  0.2159746  0.5372137 -0.4779747 -0.16724170  0.2088427  1.0327797 -0.4850180 -1.0001668  0.8565072
2 -0.1179959 -0.2935021  0.2611374  0.09137108 -0.1140994 -0.5642504  0.2649854  0.5464326 -0.4679454
```

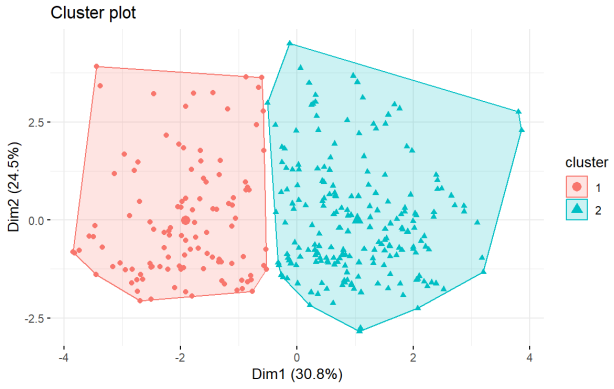


Figure 4

In Figure 4, we can see two defined clusters where the first principal component (Dim1) accounts for 30.8% of the total variance while the second principal component explains 24.5% of the total variation. In total, both components explain 55.3% of the variation in the dataset. Cluster 1 represents the non-demented group, while Cluster 2 represents the demented group.

Table 2.1

```
> table(Alz.km$cluster,diagnosis)
   diagnosis
      0   1
  1  60  52
  2 130  75
```

Table 2.2

```
> table(Alz.hclust.clusters,diagnosis)
                    diagnosis
Alz.hclust.clusters   0   1
                  1  67  76
                  2 123  51
```

To check how accurate the clusters are when using a k-means algorithm, they are compared to the actual diagnosis. In table 2.1, we can observe that cluster 1, which should contain only non-demented patients, in reality contains 60 non-demented and 52 demented patients. Cluster 2 should be composed of demented patients; however, it contains 130 non-demented and 75 demented patients.

When using a hierarchical clustering model, we can observe in Table 2.2 that cluster one, which should contain non-demented patients, contains 67 non-demented and 76 demented patients. Cluster 2 contains 123 non-demented patients and 51 demented patients.

We can easily see that in this case, the k-means model is better at detecting demented patients in the dataset than the hierarchical cluster model. However, it is necessary to reduce the number of components to improve both models.

3.Logistic Regression Model

The Logistic Regression Model was based on the previous analysis. That is the reason why Equation 1 only uses sex, education, and normalised brain mass to explain groups. The data was divided into two parts: 70% was used for training and 30% for testing.

When sex increases by one unit, meaning that when the sex is male, the log odds of being in the demented group increase by 1.08. One unit increase in Education reduces the log odds of being demented by 0.19. In addition, one unit increase in normalised brain volume reduces the log odds of having Alzheimer's by 18.54. In this model, it is easy to see that all explanatory variables are significant at 0.001.

The difference between Null deviance and Residual deviance is quite high, meaning that the model is a good fit.

After running the probabilities of the predictions and calculating the accuracy, we found that the model was correct 67% of the time.

Equation 1

```
Call:
glm(formula = dgroup ~ sex + EDUC + nWBV, family = binomial,
    data = Alzheimer_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0396  -0.8800  -0.5305   0.9773   2.0151

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.43207    2.85978   5.396 6.80e-08 ***
sex1          1.08224    0.26285   4.117 3.83e-05 ***
EDUC         -0.19474    0.04691  -4.151 3.30e-05 ***
nWBV        -18.54555    3.71935  -4.986 6.16e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 426.85  on 316  degrees of freedom
Residual deviance: 356.03  on 313  degrees of freedom
AIC: 364.03

Number of Fisher Scoring iterations: 4
```

```
> accuracy
[1] 0.6736842
```

4.Feature Selection: Principal Component Analysis

When running a principal component analysis to reduce the dimension of the dataset as much as possible while retaining its variance, we can observe that PC1 and PC2 explain most of the variability compared to the other components. In Table 3, we can easily see that PC1 explains 30.84% of variability and PC2 explains 24.1% of variability, making a total of 34.9%.

The first Principal component shows high positive values in CDR, eTIV, and sex and negative values for MMSE, nWBV, and ASF, showing us that males with a high Clinical Dementia Rating and low normalised brain volume are more likely to be part of the demented group and vice versa.

The second principal component has high positive values in Education, MMSE, eTIV, and sex and negative values for SES, CDR, and ASF. In other words, male patients with low-socioeconomic status are more likely to be part of the non-demented group, and vice versa.

Table 3

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.6661 1.4853 1.1718 1.1152 0.71520 0.59362 0.52038 0.50568 0.10419
Proportion of Variance 0.3084 0.2451 0.1526 0.1382 0.05683 0.03915 0.03009 0.02841 0.00121
Cumulative Proportion  0.3084 0.5535 0.7061 0.8443 0.90114 0.94029 0.97038 0.99879 1.00000
```

	Comp.1	Comp.2
Age	0.056213164	0.021004028
EDUC	-0.062290448	0.400470025
SES	0.009894385	-0.393892355
MMSE	-0.434811450	0.228123697
CDR	0.468576021	-0.226866552
eTIV	0.272655975	0.483359548
nWBV	-0.359531030	-0.008963267
ASF	-0.277050604	-0.475250310
sex	0.327493079	0.239303226
dgroup	0.443424177	-0.252342755

In figure 5, the minimum number of principal components needed to explain 80% of the variability is 4. If a biplot is presented (Appendix), Education, sex, and normalised brain volume are located far away from the origin; hence, they are well represented compared to age.

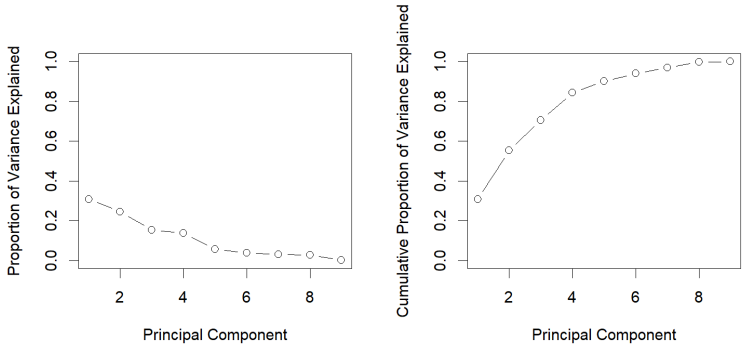


Figure 5

Conclusion

In conclusion, the exploratory analysis showed that men are more likely to suffer from Alzheimer's than women. Also, Demented subjects tend to have lower brain volumes and years of education than non-demented patients.

The k-means clustering algorithm is better at detecting demented patients in the dataset than the hierarchical cluster model. However, it is necessary to reduce the number of components to improve both models. After running a logistic regression model, it was found that being a man increased the log odds of being in the demented group by 1.08 and that one unit increase in normalised brain volume reduced the log odds of having Alzheimer's by 18.54. The Principal component analysis in PCA1 found that males with a high Clinical Dementia Rating and low normalised brain volume are more likely to be part of the demented group, and in PCA2, male patients with low socioeconomic status are more likely to be part of the non-demented group.

References

Alboukadel. “K-Means Clustering Visualization in R: Step by Step Guide.” *Datanovia.Com*, 2 June 2020, www.datanovia.com/en/blog/k-means-clustering-visualization-in-r-step-by-step-guide/.

“Alzheimer’s Disease Fact Sheet.” *Alzheimer’s Disease Fact Sheet*, www.nia.nih.gov/health/alzheimers-disease-fact-sheet#:~:text=Alzheimer’s%20disease%20is%20a%20brain,carry%20out%20the%20simplest%20tasks. Accessed 21 June 2023.

Competitor-Cutter. “Simply Explained Logistic Regression with Example in R.” *Medium*, 29 Dec. 2018, towardsdatascience.com/simply-explained-logistic-regression-with-example-in-r-b919acb1d6b3.

Keita, Zoumana. “Principal Component Analysis (PCA) in R Tutorial.” *DataCamp*, 13 Feb. 2023, www.datacamp.com/tutorial/pca-analysis-r.

Kuhn, Max. “The Caret Package.” *Github*, 27 Mar. 2019, topepo.github.io/caret/index.html.

NABIILAH ARDINI, FAUZIYYAH. “K-Means Clustering and PCA with Visualization.” *Kaggle*, 8 Sept. 2019, www.kaggle.com/code/nabiilahardini/k-means-clustering-and-pca-with-visualization.

Jingjing , Zhang. “MA335 Lecture 7: Feature Selection and Dimensionality Reduction.” *Moodle*, moodle.essex.ac.uk/course/view.php?id=15075&section=15. Accessed 21 June 2023.

Jingjing , Zhang. “MA335 Lab 3a & Lab 3b.” *Moodle*, moodle.essex.ac.uk/course/view.php?id=15075&section=20. Accessed 21 June 2023.

Jingjing , Zhang. “MA335 Lab 4a & 4b.” *Moodle*, moodle.essex.ac.uk/course/view.php?id=15075&section=21. Accessed 21 June 2023.

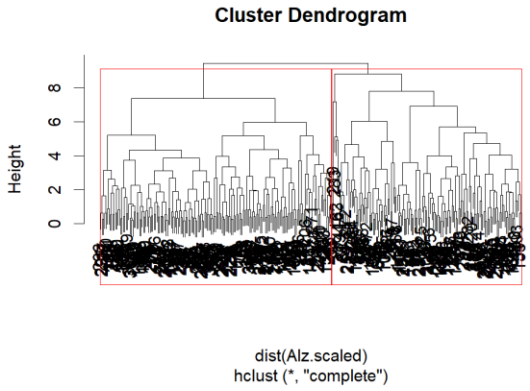
5.Appendix

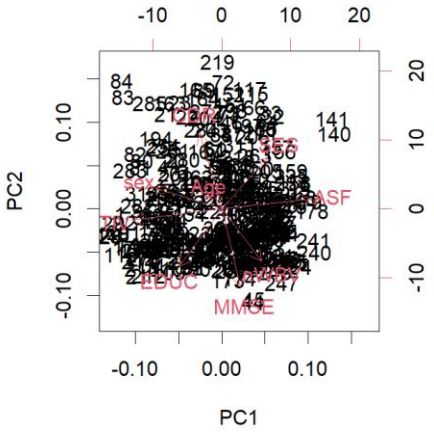
Code:

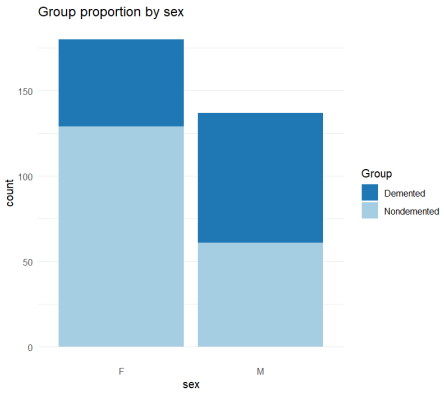
```
Alzheimer_data<-read.csv("C:/Users/Fer_c/OneDrive/Escritorio/Modeling/FINAL PROJECT/project data.csv",head=TRUE)

library(dplyr)

library(tidyr)

library(ggplot2)

library(tidyverse)

library(corrplot)

library(ggcorrplot)

library(plotly)

attach(Alzheimer_data)

library(RColorBrewer)

library(boot)

library(caret)


#Remove all NAs

Alzheimer_data<-Alzheimer_data %>%  na.omit()


#remove rows with Group "Converted"

Alzheimer_data<-Alzheimer_data %>%  filter(Group=='Demented' | Group=='Nondemented')


#Transform M/F into numeric values

Alzheimer_data$sex<- factor(Alzheimer_data$M.F,

          levels=c("F","M"),

          labels=c(0,1))

#Transform socioeconomic status into a factor

Alzheimer_data$SES<- factor(Alzheimer_data$SES,levels = c("1","2","3","4","5"))


#Transform Group into a factor

Alzheimer_data$Group<-as.factor(Alzheimer_data$Group)


#Obtain the names of the columns and a summary of the data

names(Alzheimer_data)

summary(Alzheimer_data)

#####1.DESCRIPTIVE STATISTICS############################

attach(Alzheimer_data)

#bar chart and table of Group by sex

tab_gen<-table(M.F,Group)

prop_group<-prop.table(tab_gen,1)

perc_group<-round((100*prop_group),2)

perc_group

ggplot(Alzheimer_data,aes(x=M.F,fill=Group))+

 geom_bar()+scale_fill_brewer(palette="Paired",direction=-1)+ theme_minimal()+

 theme(panel.grid.major.x = element_blank())+

 labs(title = "Group proportion by sex",x = "sex")


#Plot of Group by sex and Socioeconomic status

tab_gen<-table(SES,Group)

prop_group<-prop.table(tab_gen,1)

perc_group<-round((100*prop_group),2)

perc_group

ggplot(Alzheimer_data,aes(x=M.F,fill=Group))+

 geom_bar()+scale_fill_brewer(palette="Paired",direction=-1)+ theme_minimal()+

 theme(panel.grid.major.x = element_blank())+

 labs(title = "Group proportion by sex and SES",x = "sex")+

 facet_wrap(~SES)


#Plot of Group by SES

ggplot(Alzheimer_data,aes(x=SES,fill=Group))+

 geom_bar()+scale_fill_brewer(palette="Paired",direction=-1)+ theme_minimal()+

 theme(panel.grid.major.x = element_blank())+

 labs(title = "Group proportion by SES",x = "Socioeconomic Status")

ggplot(Alzheimer_data, aes(x = reorder(SES,MMSE, .fun='median'), y = MMSE)) +

 geom_boxplot(color = "#6baed6") +

 theme_minimal() +

 theme(panel.grid.major.x = element_blank()) +

 labs(title = "Mini mental state Examination score by SES",x = "Socioeconomic status", y = "MMSE score")+

 ylim(0, 35)


#Plot of group by AGE

summary(Alzheimer_data$Age)

ggplot(Alzheimer_data,aes(x=Age))+geom_histogram(binwidth=4,fill="#6baed6", color="white")+

 theme_minimal()+labs(title = "Age Distribution",x = "Age")


#Plot of group by Education
```

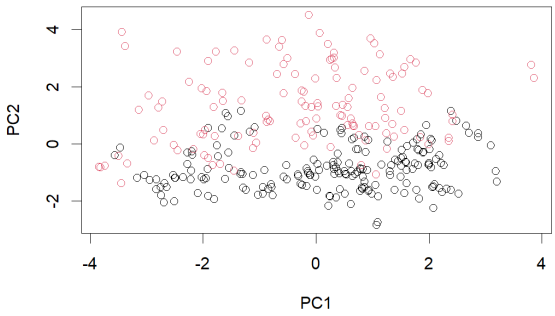
```
ggplot(Alzheimer_data, aes(x = reorder(Group,EDUC, .fun='median'), y = EDUC)) +
 geom_boxplot(color = "#6baed6") +
 theme_minimal() +
 theme(panel.grid.major.x = element_blank()) +
 labs(title = "Years of Education by Group",x = "Group", y = "Education")+
 ylim(0, 20)
#Plot of MMSE by group
ggplot(Alzheimer_data, aes(MMSE))+
 geom_histogram(binwidth=2,fill="#6baed6",
color="white")+facet_grid(rows=vars(Group))+
 theme_minimal()+
 labs(title = "MMSE by Group")
#Plot of nwBv by group
ggplot(Alzheimer_data, aes(nWBV))+
 geom_histogram(binwidth=0.02,fill="#6baed6",
color="white")+facet_grid(rows=vars(Group))+
 theme_minimal()+
 labs(title = "nWBV by Group")

#CORRPLOT
Alzheimer_data$dgroup<- factor(Alzheimer_data$Group,
                    levels=c("Nondemented","Demented"),
                    labels=c(0,1))

Alzcorr.df<-as.data.frame(Alzheimer_data[3:12])
Alzcorr.df <- as.data.frame(apply(Alzcorr.df, 2, as.numeric))
corr.mat <- round(cor(Alzcorr.df), 2)
pval.cor <- cor_pmat(Alzcorr.df)
corr_plot = ggcorrplot(corr.mat,hc.order = TRUE,type = "lower", lab = TRUE,outline.color
= "white",lab_size = 3,colors = c("#6D9EC1", "white", "#E46726"))+
 labs(title = "Correlation Matrix \n on Alzheimer data \n",
     x = "", y = "")+theme(axis.text.x = element_text(angle = 90)) + theme_minimal()
corr_plot
##########################2.CLUSTERING ALGORITHMS#########
#Creates a dataframe from the dataset with only numeric variables
Alz.df<-as.data.frame(Alzheimer_data[3:11])
Alz.df <- as.data.frame(apply(Alz.df, 2, as.numeric))
# Convert variables to numeric

# Set seed
set.seed(1)

# Creates a vector that sets demented as 1 and non demented as 0
diagnosis <- as.numeric(Alzheimer_data$Group == "Demented")

# Means and sd are too different from each other, they need to be scaled
colMeans(Alz.df)
apply(Alz.df,2,sd)

# Scalede data
Alz.scaled<-scale(Alz.df)

# hierarchical clustering model
hclust.Alz<-hclust(dist(Alz.scaled), method="complete")
rect.hclust(hclust.Alz, k=2, border="red")

# Cut tree
Alz.hclust.clusters<-cutree(hclust.Alz,k=2)

# Compare hierarchichal cluster to actual diagnoses
table(Alz.hclust.clusters,diagnosis)

# k-means model
Alz.km<-kmeans(scale(as.matrix(Alz.scaled)),centers=2,nstart=20)
Alz.km
#Visualize kmeans
fviz_cluster(Alz.km, data = Alz.scaled,ggtheme = theme_minimal(),geom = "point")
#determine optimal number of clusters
fviz_nbclust(Alz.scaled, kmeans, method = "wss")+
 geom_vline(xintercept = 2, linetype = 2)
Alzheimer_data$cluster <- as.factor(Alz.km$cluster)

# Compare k-means to actual diagnoses
table(Alz.km$cluster,diagnosis)

# Compare k-means to hierarchical clustering
table(Alz.km$cluster,Alz.hclust.clusters)
Alz.pr.hclust <- hclust(dist(Alz.pr$x[, 1:7]), method = "complete")

#cut model
Alz.pr.hclust.clusters<-cutree(Alz.pr.hclust,k=2)

# Compare to actual diagnoses
```

```
table(diagnosis, Alz.pr.hclust.clusters)
table(diagnosis, Alz.hclust.clusters)

# Compare to k-means and hierarchical
table(diagnosis,Alz.km$cluster)
#########################3.LOGISTIC REGRESSION########
set.seed(1)
# training and testing sets
trainIndex <- createDataPartition(Alzheimer_data$dgroup, p = 0.7, list = FALSE)
trainData <- Alzheimer_data[trainIndex, ]
testData <- Alzheimer_data[-trainIndex, ]

glm.fit<-glm(dgroup~sex+EDUC+nWBV,data=Alzheimer_data,family=binomial)
summary(glm.fit)

# predicted probabilities
glm.probs <- predict(glm.fit, newdata = testData, type = "response")
threshold <- 0.5
glm.predicted <- ifelse(glm.probs > threshold, 1, 0)

# Table of predicted and real diagnosis
table(glm.predicted, testData$dgroup)

# Accuracy of the predictions
accuracy <- mean(glm.predicted == testData$dgroup)
accuracy

#########################4.FEATURE SELECTION#####
# PCA
Alz.pr<-prcomp(Alz.df,scale=TRUE)
summary(Alz.pr)

#PCA with correlation matrix
Alz.pca <- princomp(corr.mat)
Alz.pca$loadings[, 1:2]

# Visualize the biplot
biplot(Alz.pr)

# Scatter plot of PC1 and PC2
plot(Alz.pr$x[, c(1, 2)], col = (diagnosis + 1),
    xlab = "PC1", ylab = "PC2")

# Scatter plot of PC1 and PC2
plot(Alz.pr$x[, c(1, 3)], col = (diagnosis + 1),
    xlab = "PC1", ylab = "PC3")
par(mfrow = c(1, 2))

# Variability of each component
pr.var<-Alz.pr$sdev^2

# Variance explained by each principal component
pve<-pr.var/sum(pr.var)

# Variance plot
plot(pve, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained",
    ylim = c(0, 1), type = "b")

# Cumulative plot
plot(cumsum(pve), xlab = "Principal Component",
    ylab = "Cumulative Proportion of Variance Explained",
    ylim = c(0, 1), type = "b")
```


Cluster Dendrogram
Height
0
2
4
6
8
dist(Alz.scaled)
hclust (*, "complete")


-10
0
10
20
PC2
0.10
0.00
-0.10
PC1
-0.10
0.00
0.10
MMSE


Group proportion by sex
count
150
100
50
0
F
M
sex
Group
Demented
Nondemented


PC2
4
2
0
-2
-4
-2
0
2
4
PC1