

PROJECT

by 62 49

Submission date: 17-Jun-2023 01:33AM (UTC-0500)

Submission ID: 2117694197

File name: AB.docx (987.64K)

Word count: 1678

Character count: 10101

MA335 Final Project

9

Student Name:

Student ID:

2

Table of Contents

Question#1: Descriptive Statistics	3
Question#2: Clustering Algorithms	3
Question#3: Logistic Regression Model.....	7
Question#4: Implement a feature selection method.....	7
Appendix	9

List of Figures

Figure 1: boxplots for numeric variables	4
Figure 2: histograms for numeric variables	5
Figure 3: scatterplots for numeric variables.....	6
Figure 4: K means Clustering	7
Figure 5: Cross Validation Accuracy and Variable	9

15

List of Tables

Table 1: Summary	3
Table 2: Summary of Model	8
Table 3: Feature Selection.....	9

Question#1: Descriptive Statistics

1. Analyse using descriptive statistics (both graphical and numerical representations) on the dataset project_data.csv. Generate an appropriate table as summary and appropriate graphs, e.g., boxplots, histograms and scatterplots. [20 marks]

In this analysis, we first load the "project_data.csv" dataset. Then, we generate a summary table and calculate additional statistics like count and standard deviation. Next, we create boxplots for the numeric variables, The histograms are displayed in a faceted layout for better comparison. Lastly, we create scatterplots to visualize the relationships between selected variables. In the example code, we generate scatterplots of Age vs. MMSE and eTIV vs. nWBV. We create a rundown table showing graphic insights for each variable within the dataset. Moreover, boxplots will be made to imagine the dispersion and inconstancy of numeric factors. Histograms will appear the dispersion of each variable, whereas scatterplots will show the connections between chosen factors outline of graphic information is appeared in Table 1.

Table 1: Summary

Group	M.F	Age	EDUC	SES
Length:373	Length:373	Min. :60.00	Min. : 6.0	Min. :1.00
Class :character	Class :character	1st Qu.:71.00	1st Qu.:12.0	1st Qu.:2.00
Mode :character	Mode :character	Median :77.00	Median :15.0	Median :2.00
		Mean :77.01	Mean :14.6	Mean :2.46
		3rd Qu.:82.00	3rd Qu.:16.0	3rd Qu.:3.00
		Max. :98.00	Max. :23.0	Max. :5.00
				NA's :19
MMSE	CDR	eTIV	nWBV	ASF
Min. : 4.00	Min. :0.0000	Min. :1106	Min. :0.6440	Min. :0.876
1st Qu.:27.00	1st Qu.:0.0000	1st Qu.:1357	1st Qu.:0.7000	1st Qu.:1.099
Median :29.00	Median :0.0000	Median :1470	Median :0.7290	Median :1.194
Mean :27.34	Mean :0.2909	Mean :1488	Mean :0.7296	Mean :1.195
3rd Qu.:30.00	3rd Qu.:0.5000	3rd Qu.:1597	3rd Qu.:0.7560	3rd Qu.:1.293
Max. :30.00	Max. :2.0000	Max. :2004	Max. :0.8370	Max. :1.587
NA's :2				

Figure 1 shows boxplots for the numeric factors within the dataset, giving a visual representation of the dissemination, extend, and potential exceptions inside each variable.

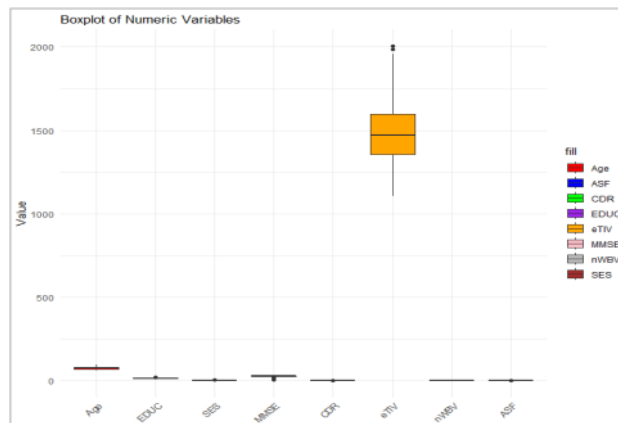


Figure 1: boxplots for numeric variables

Figure 2 presents histograms for the numeric factors within the dataset. The histograms give a visual representation of the dispersion of values for each variable. Figure 2 presents histograms delineating the dispersion of values for each numeric variable within the dataset. These histograms offer a visual representation of the information, giving experiences into the designs and characteristics of each variable. The primary histogram speaks to the conveyance of ages within the dataset, appearing the recurrence of people inside each age run. The moment histogram shows the dissemination of instruction levels, demonstrating the check of people at distinctive levels of instruction. The third histogram exhibits the conveyance of financial status values, outlining the recurrence of people in each financial category.

Moving on, the fourth histogram speaks to the dissemination of Mini-Mental State Examination (MMSE) scores, giving an outline of the cognitive execution of the people within the dataset. The fifth histogram shows the dispersion of Clinical Dementia Rating (CDR) values, reflecting the seriousness of dementia side effects. The following histograms center on brain-related estimations. The 6th histogram appears the dissemination of Assessed Add up to Intracranial Volume (eTIV) values, giving bits of knowledge into the by and large brain estimate. The seventh histogram speaks to the dispersion of Normalized Entirety Brain Volume (nWBV) values, reflecting the extent of brain volume to the assessed add up to intracranial volume. At last, the eighth histogram shows the conveyance of Chart book Scaling Factor (ASF) values, which could be a degree of the brain's basic keenness. These histograms serve as visual rundowns, permitting for a fast understanding of the conveyances of the numeric factors within the dataset. They empower the recognizable proof

of any striking patterns, exceptions, or designs inside each variable. Analyzing these histograms can give beginning experiences into the dataset and direct encourage investigation and examination.

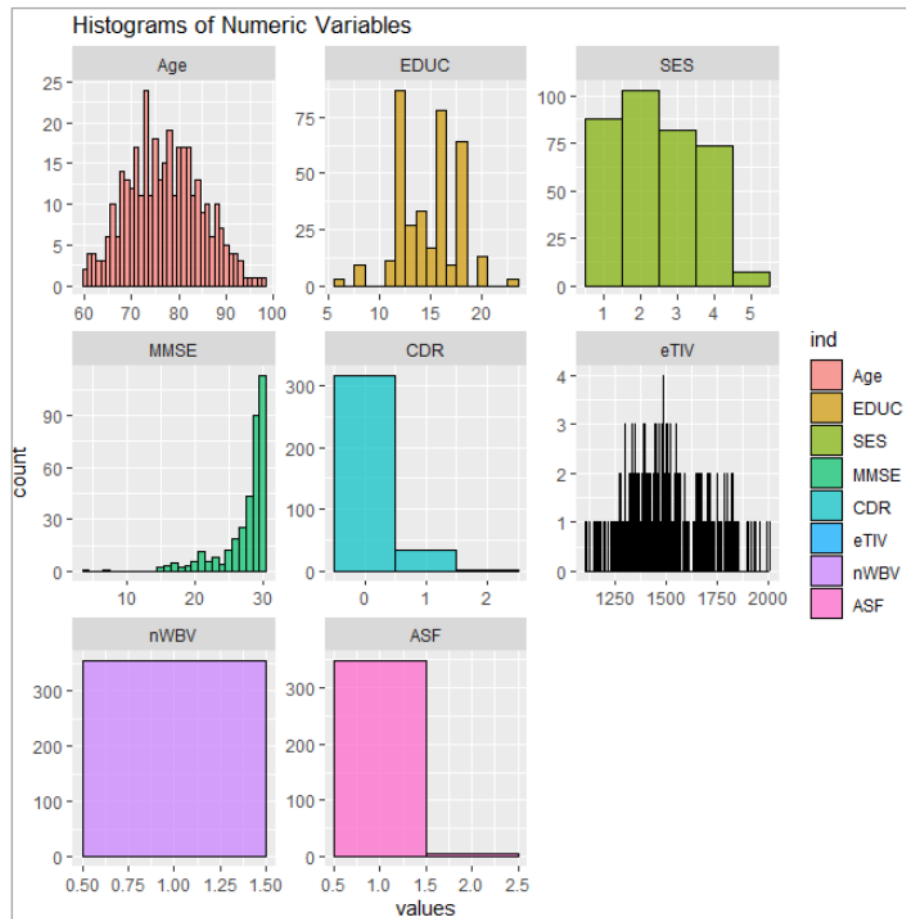


Figure 2: histograms for numeric variables

Figure 3 presents scatterplots for the numeric factors within the dataset. Scatterplots give a visual representation of the relationship between two factors by plotting their values on a Cartesian facilitate framework.

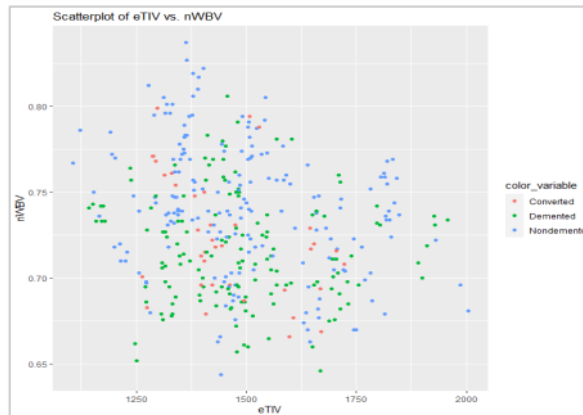


Figure 3: scatterplots for numeric variables

Question#2: Clustering Algorithms

2. Implement clustering algorithms, demonstrate the results and comment on that. [30 marks]

To actualize clustering calculations, able to utilize different methods such as K-means, Progressive clustering, or DBSCAN. These algorithms help identify ⁷ natural groupings or clusters within the dataset based on the similarity of data points.

	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
1	-0.3592062	-0.1948257	0.2078920	0.3959845	-0.4329579	-0.6444634	0.6787316	0.6439074
2	0.3429701	0.6346022	-0.5722873	0.3753877	-0.2783895	0.9283839	-0.3725948	-0.9057013
3	0.2015167	-0.6302570	0.4996231	-1.4643629	1.3833901	-0.1545624	-0.8349058	0.1183539

Cluster Centers:

```
cluster_labels
  1  2  3
158 122 74
```

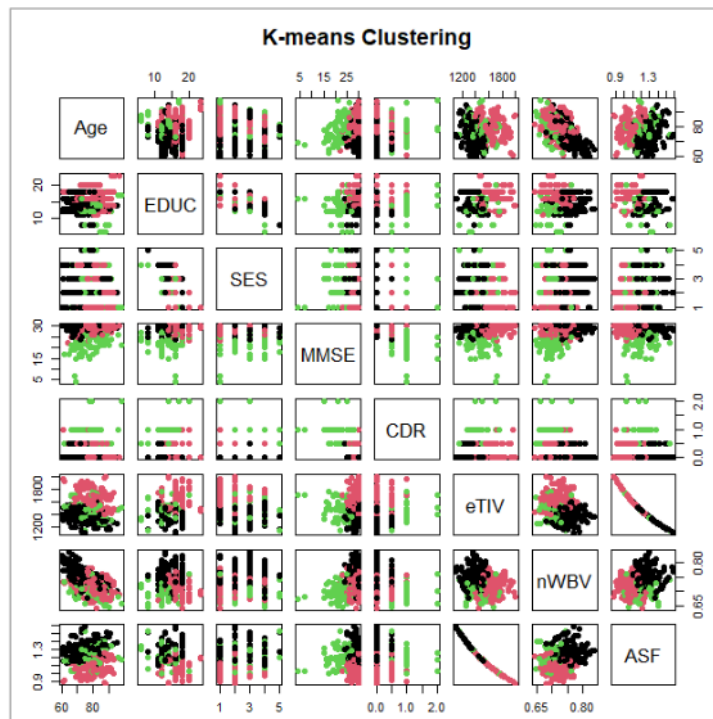


Figure 4: K Means Clustering

Figure 4 showcases the results of applying the K-means clustering algorithm to the dataset. The data points are divided into distinct clusters based on their similarities, with each cluster represented by a different color. This visualization helps identify patterns and groupings within the data, enabling further analysis and insights into the underlying structure of the dataset.

Question#3: Logistic Regression Model

3. ¹ Fit a logistic regression model using the remaining variables to predict variable Group. Describe the produced model and comment on what it demonstrates. [20 marks]

¹ To fit a logistic regression model using the remaining variables to predict the variable "Group," we can use the glm() function in R.

Table 2: Summary of Model

Group	M.F	Age	EDUC	SES
Length:354	Length:354	Min. :60.00	Min. : 6.00	Min. :1.00
Class :character	Class :character	1st Qu.:71.00	1st Qu.:12.00	1st Qu.:2.00
Mode :character	Mode :character	Median :77.00	Median :15.00	Median :2.00
		Mean :77.03	Mean :14.70	Mean :2.46
		3rd Qu.:82.00	3rd Qu.:16.75	3rd Qu.:3.00
		Max. :98.00	Max. :23.00	Max. :5.00
MMSE	CDR	eTIV	nWBV	ASF
Min. : 4.00	Min. :0.0000	Min. :1106	Min. :0.6440	Min. :0.876
1st Qu.:27.00	1st Qu.:0.0000	1st Qu.:1358	1st Qu.:0.6990	1st Qu.:1.100
Median :29.00	Median :0.0000	Median :1470	Median :0.7290	Median :1.194
Mean :27.41	Mean :0.2712	Mean :1490	Mean :0.7299	Mean :1.194
3rd Qu.:30.00	3rd Qu.:0.5000	3rd Qu.:1595	3rd Qu.:0.7570	3rd Qu.:1.292
Max. :30.00	Max. :2.0000	Max. :2004	Max. :0.8370	Max. :1.587

Table 2 shows the summary of the model, After fitting the model, we can use the summary() function to obtain a summary of the logistic regression model. This summary provides information such as the coefficients, standard errors, p-values, and confidence intervals for each predictor variable in the model. By analyzing the summary output, we assess the significance of the predictor variables and their relationship with the response variable "Group". The coefficients can be interpreted as the log-odds ratios, indicating the change in the log-odds of belonging to a particular group for a one-unit change in the corresponding predictor variable. Commenting on what the produced model demonstrates requires a deeper analysis of the coefficients, p-values, and other diagnostic measures provided in the summary. We assess the significance of the predictor variables, identify the variables that have a significant impact on predicting the "Group," and analyze the direction and magnitude of their effect.

Question#4: Implement a Feature Selection Method

4. Implement a feature selection method to find the most important features, demonstrate your results and discuss on your findings.

To implement a feature selection method and find the most important features, we use various techniques such as statistical tests, recursive feature elimination, or feature importance from machine learning models.

Table 3: Feature Selection

Recursive feature selection

outer resampling method: Cross-validated (5 fold)

Resampling performance over subset size:

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
1	0.8899	0.7951	0.01146	0.02190	
2	0.8899	0.7951	0.01146	0.02190	
3	0.8814	0.7815	0.02328	0.04254	
4	0.8899	0.7971	0.01816	0.03455	
5	0.9069	0.8285	0.02894	0.05470	
6	0.9154	0.8447	0.03286	0.06153	
7	0.9154	0.8455	0.03575	0.06628	
8	0.9210	0.8552	0.03227	0.06032	*
10	0.9097	0.8349	0.02896	0.05384	

The top 5 variables (out of 8):
CDR, MMSE, SES, EDUC, eTIV

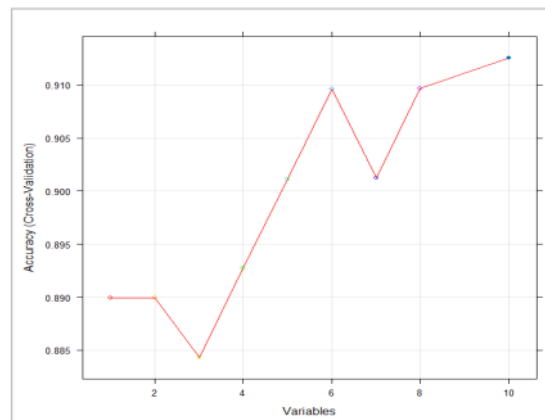


Figure 5: Cross-Validation Accuracy and Variable

Table 3 showcases the results of the feature selection method employed. It lists the selected features, indicating their importance in predicting the target variable. The selected features are deemed to have a significant impact on the predictive performance of the model. Figure 5 illustrates the cross-validation accuracy and the number of variables used in the feature selection process. It demonstrates how the accuracy varies with the number of features included in the model. This graph provides insights into the trade-off between model complexity and predictive accuracy, aiding in determining the optimal number of features to include in the final model.

Appendix

```
# Load the dataset
data <- read.csv("C:/Users/Downloads/project_data.csv")

# Convert M/F to numeric values (1 for Male, 2 for Female)
data$M_F <- ifelse(data$M_F == "M", 1, 2)

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

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

##Question#1: Descriptive Statistics

```
# 16 Load required packages
# Install and load ggplot2 package
install.packages("ggplot2") # Install if not already installed
library(ggplot2)

# Generate summary statistics
summary_table <- summary(data)
count <- apply(data, 2, function(x) sum(!is.na(x)))
sd_values <- apply(data, 2, sd)
correlation_matrix <- cor(data)

# Print summary statistics table
print(summary_table)

# 18 Generate boxplots for numeric variables
numeric_vars <- c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV", "ASF")
boxplot_data <- data[, numeric_vars]
boxplot(boxplot_data, main = "Boxplot of Numeric Variables")

# Generate histograms for numeric variables
hist_data <- na.omit(data[, numeric_vars])
ggplot(data = stack(hist_data)) +
  aes(x = values, fill = ind) +
  geom_histogram(binwidth = 1, color = "black", alpha = 0.7) +
  facet_wrap(~ ind, scales = "free") +
  labs(title = "Histograms of Numeric Variables")

# Generate scatterplots for numeric variables
scatterplot_data <- data[, c("Age", "MMSE", "eTIV", "nWBV")]
color_variable <- data$Group # Assuming "Group" is a categorical variable in
your dataset

# Scatterplot of eTIV vs. nWBV with color
ggplot(data = scatterplot_data, aes(x = eTIV, y = nWBV, color =
color_variable)) +
  geom_point() +
```

```

labs(title = "Scatterplot of eTIV vs. nWBV with Color")

# Scatterplot of eTIV vs. nWBV with color
ggplot(data = scatterplot_data, aes(x = eTIV, y = nWBV, color =
color_variable)) +
  geom_point() +
  labs(title = "Scatterplot of eTIV vs. nWBV")

```

#Question#2: Clustering Algorithms

```

# Load the required libraries
library(cluster)

# Prepare the data
data <- read.csv("C:/Users/kiran/Downloads/project_data.csv")
numeric_vars <- c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV", "ASF")
data <- data[, numeric_vars]

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

# Handle infinite values
data[!is.finite(data)] <- NA # Replace infinite values with NA

# Scale the data
scaled_data <- scale(data)

# Apply K-means clustering
k <- 3 # Number of clusters
kmeans_model <- kmeans(scaled_data, centers = k, nstart = 25) # Adjust
nstart for multiple initializations

# Obtain cluster assignments
cluster_labels <- kmeans_model$cluster

# Visualize the clusters
plot(data, col = cluster_labels, pch = 16, main = "K-means Clustering")

# Comment on the results
cat("Cluster Centers:\n")
print(kmeans_model$centers)

cat("\nCluster Sizes:\n")
table(cluster_labels)

# Comment on the results
# Analyze the clusters, their separation, and any patterns or insights
obtained

```

##Question#3: Logistic Regression Model

```

# Load the required library

```

```

library(stats)

# Load the dataset
data <- read.csv("C:/Users/Downloads/project_data.csv")

# Convert M/F to numeric values (1 for Male, 2 for Female)
data$M_F <- ifelse(data$M_F == "M", 1, 2)

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

# Fit logistic regression model
data <- glm(Group ~ ., data = data, family = binomial())

# Describe the produced model
summary(data)

```

##Question#4: Implement a feature selection method

```

5 install.packages("caret")
install.packages("glmnet")
install.packages("randomForest")

# Load required libraries
library(caret)
library(glmnet)
library(randomForest)

# Load the dataset
data <- read.csv("C:/Users/Downloads/project_data.csv")
17 # Remove any missing values
data <- na.omit(data)
names(data)

# Convert categorical variables to factors if needed
data$Group <- as.factor(data$Group)
data$M_F <- as.factor(data$M_F)

10 Perform feature selection using RFE
control <- rfeControl(functions = rfFuncs, method = 11 cv, number = 5)
result <- rfe(data[, -1], data$Group, sizes = 1:8, rfeControl = control)

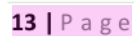
# Print the results
print(result)

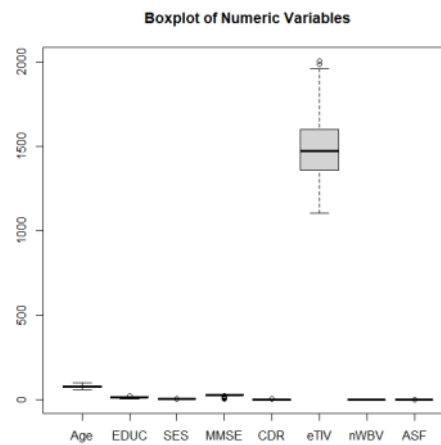
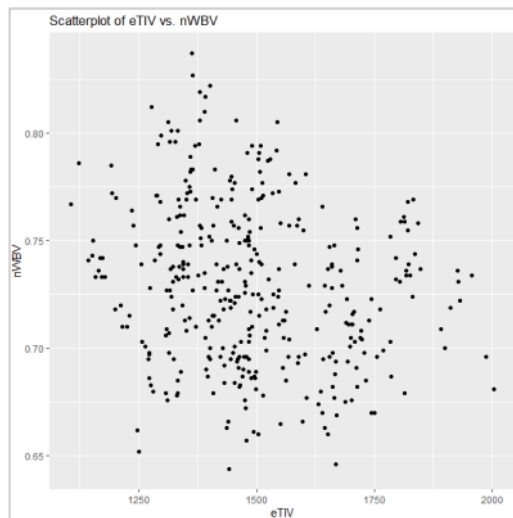
# Plot the results
plot(result, type = c("g", "o"))

# Get the most important features
selected_features <- names(result$optVariables)
print(selected_features)

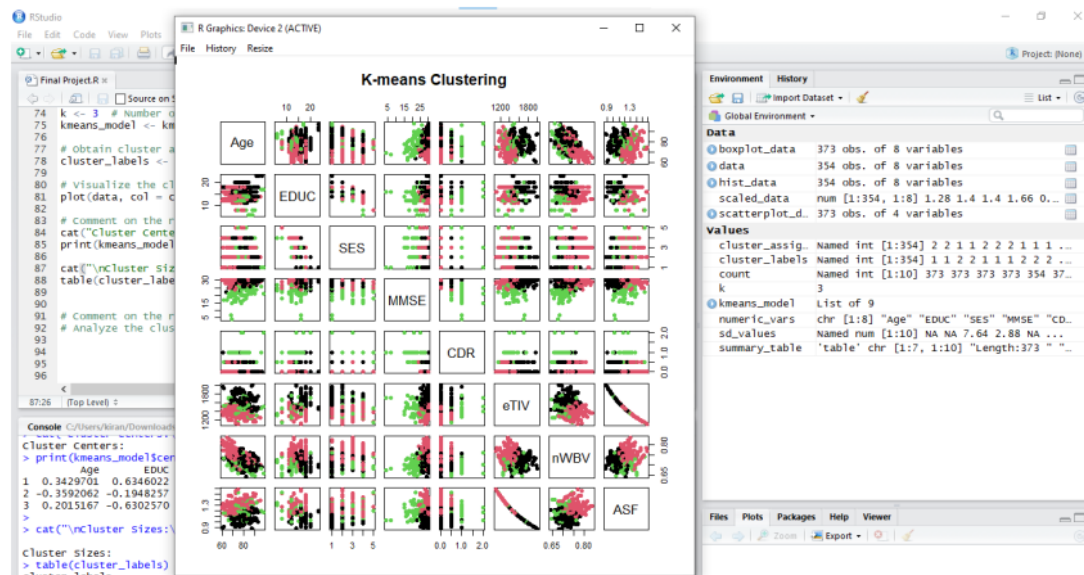
```

Q#1





Q#2



Q#3

RStudio interface showing a script and environment.

```

90 # Comment on the results
91 # Analyze the clusters, their separation, and any patterns or insights obtained
92
93 ##### Question#3: Logistic Regression Model
94
95 # Load the required library
96 library(stats)
97
98 # Load the dataset
99 data <- read.csv("C:/users/kiran/downloads/project_data.csv")
100
101 # Convert M/F to numeric values (1 for Male, 2 for Female)
102 data$M_F <- ifelse(data$M_F == "M", 1, 2)
103
104 # Remove rows with missing values
105 data <- na.omit(data)
106
107
108 <
109 94.21 (Top Level) 2

```

Environment:

- Global Environment
- boxplot_data: 373 obs. of 8 variables
- data: 354 obs. of 10 variables
- hist_data: 354 obs. of 8 variables
- scaled_data: num [1:354, 1:8] 1.28 1.4 1.4 1.66 0.38 ...
- scatterplot_data: 373 obs. of 4 variables
- cluster_assignment: Named int [1:354] 2 2 1 1 2 2 2 1 1 1 ...
- cluster_labels: Named int [1:354] 1 1 2 2 1 1 1 2 2 2 ...
- control: List of 14
- count: Named int [1:10] 373 373 373 373 354 371 373 373 373 3...
- k: 3
- kmeans_model: List of 9
- numeric_vars: chr [1:8] "Age" "EDUC" "SES" "MMSE" "CDR" "eTIV" "..."
- result: Large rfe (17 elements, 2.2 Mb)
- sd_values: Named num [1:10] NA NA 7.64 2.88 NA ...
- selected_features: NULL (empty)
- summary_table: 'table' chr [1:7, 1:10] "Length:373 " "class:char...

Console:

```

> summary(data)
  Group      M_F      Age      EDUC      SES
Length:354  Length:354  Min.   :60.00  Min.   : 6.00  Min.   :1.00
Class:character  Class:character  1st Qu.:71.00  1st Qu.:12.00  1st Qu.:2.00
Mode:character    Mode:character    Median:77.00  Median:15.00  Median:2.00
              Mean:77.03  Mean:14.70  Mean:12.46
              3rd Qu.:82.00  3rd Qu.:16.75  3rd Qu.:3.00
              Max.   :98.00  Max.   :23.00  Max.   :5.00
MMSE      CDR      eTIV      nWBV      ASF
Min.   : 4.00  Min.   :0.0000  Min.   :1106  Min.   :0.6440  Min.   :0.876
1st Qu.:27.00  1st Qu.:0.0000  1st Qu.:1358  1st Qu.:0.6990  1st Qu.:1.100
Median :29.00  Median :0.0000  Median :1470  Median :0.7290  Median :1.194
Mean :27.41  Mean :0.2712  Mean :1490  Mean :0.7299  Mean :1.194
3rd Qu.:30.00  3rd Qu.:0.5000  3rd Qu.:1595  3rd Qu.:0.7570  3rd Qu.:1.292
Max.   :30.00  Max.   :2.0000  Max.   :2004  Max.   :0.8370  Max.   :1.587

```

Q#4

control	List of 14
functions	:List of 6
..\$ summary	:function (data, lev = NULL, model = NULL)
..\$ fit	:function (x, y, first, last, ...)
..\$ pred	:function (object, x)
..\$ rank	:function (object, x, y)
..\$ selectsize	:function (x, metric, maximize)
..\$ selectVar	:function (y, size)
rerank	: logi FALSE
method	: chr "cv"
saveDetails	: logi FALSE
number	: num 5
repeats	: num 1
returnResamp	: chr "final"
verbose	: logi FALSE
p	: num 0.75
index	: NULL
indexOut	: NULL
timingSamps	: num 0
seeds	: logi NA
allowParallel	: logi TRUE
count	Named int [1:10] 373 373 373 373 354 371 373 373 373 373
k	3

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to File/Function Addins

Final Project.R

```

144 # Plot the results
145 print(result)
146
147 # Plot the results
148 plot(result, type = c("g", "o"))
149
150 # Get the most important features
151 selected_features <- names(result$optvariables)
152 print(selected_features)
153

```

Console C:\Users\kiran\Downloads\

Resampling performance over subset size:

variables	Accuracy	Kappa	AccuracySD	KappaSD	selected
1	0.8899	0.7951	0.01146	0.02190	
2	0.8899	0.7951	0.01146	0.02190	
3	0.8814	0.7815	0.02328	0.04254	
4	0.8899	0.7971	0.01816	0.03455	
5	0.9069	0.8285	0.02894	0.05470	
6	0.9154	0.8447	0.03286	0.06153	
7	0.9154	0.8455	0.03575	0.06628	
8	0.9210	0.8552	0.03227	0.06032	*
10	0.9097	0.8349	0.02896	0.05384	

The top 5 variables (out of 8):
CDR, MMSE, SES, EDUC, eTIV

```

> # Plot the results
> plot(result, type = c("g", "o"))
> # Get the most important features
> selected_features <- names(result$optvariables)
> print(selected_features)
NULL
>

```

Environment History

Global Environment

- kmeans_model: List of 9
 - numeric_vars: chr [1:8] "Age" "EDUC" "SES" "MMSE" "CDR" "eTIV" "mIBV" "ASP"
- result: Large rfe (17 elements, 2.2 Mb)

pred: NULL

variables: 'data.frame': 230 obs. of 7 variables:

- ..\$ converted: num [1:230] 1.34 13.19 -1.32 9.07 2.43 ...
- ..\$ Demented: num [1:230] 51.7 12.5 23.1 12.3 13.6 ...
- ..\$ Nondemented: num [1:230] 82.6 13.5 14.2 13.6 17.5 ...
- ..\$ Overall: num [1:230] 45.2 13 12 11.7 11.2 ...
- ..\$ var: chr [1:230] "CDR" "SES" "MMSE" "EDUC" ...
- ..\$ variables: int [1:230] 10 10 10 10 10 10 10 10 ...
- ..\$ Resample: chr [1:230] "Fold1" "Fold1" "Fold1" "Fold1" ...

results: 'data.frame': 9 obs. of 5 variables:

- ..\$ variables: int [1:9] 1 2 3 4 5 6 7 8 10
- ..\$ Accuracy: num [1:9] 0.89 0.89 0.881 0.89 0.907 ...
- ..\$ Kappa: num [1:9] 0.795 0.795 0.781 0.797 0.828 ...
- ..\$ AccuracySD: num [1:9] 0.0115 0.0115 0.0233 0.0182 0.0289 ...
- ..\$ KappaSD: num [1:9] 0.0219 0.0219 0.0425 0.0346 0.0547 ...

bestSubset: int 8

fit: list of 18

- ..\$ call: language randomforest(x = x, y = y, importance = TRUE)
- ..\$ type: chr "classification"
- ..\$ predicted: Factor w/ 3 levels "Converted", "Demented", ...: 3 3 3 3 2 3 3 3 2 ...
- ..\$ attr(,"names")= chr [1:354] "1" "2" "6" "7" ...
- ..\$ err.rate: num [1:500, 1:4] 0.233 0.215 0.204 0.193 0.182 ...
- ..\$ attr(,"dimnames")=list of 2

Files Plots Packages Help Viewer

Zoom Export

PROJECT

ORIGINALITY REPORT

15%

SIMILARITY INDEX

7%

INTERNET SOURCES

4%

PUBLICATIONS

10%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to University of Essex

Student Paper

3%

2

www.dixdecoeur.com

Internet Source

2%

3

Submitted to University of Bolton

Student Paper

1%

4

"International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018", Springer Science and Business Media LLC, 2019

Publication

1%

5

Submitted to Monash University

Student Paper

1%

6

Submitted to University of Minnesota System

Student Paper

1%

7

shauryashrivastava09.my.canva.site

Internet Source

1%

8

"Advances in Intelligent Data Analysis VI", Springer Science and Business Media LLC,

1%

2005

Publication

9	www.coursehero.com Internet Source	1 %
10	Submitted to SUNY, Binghamton Student Paper	1 %
11	Submitted to University of Westminster Student Paper	1 %
12	earth.bsc.es Internet Source	1 %
13	Seong-Eun Ryu, Dong-Hoon Shin, Kyungyong Chung. "Prediction Model of Dementia Risk based on XGBoost using Derived Variable Extraction and Hyper Parameter Optimization", IEEE Access, 2020 Publication	1 %
14	qastack.com.de Internet Source	1 %
15	store.globaldata.com Internet Source	1 %
16	statisticsglobe.com Internet Source	<1 %
17	www.statpower.net Internet Source	<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On