TITLE

"EXPLORING THE RELATIONSHIP BETWEEN CHARACTERISTICS OF ALZHEIMER'S DISEASE AND DIAGNOSIS: A DATA SCIENCE ANALYSIS"

NAME: OBED MAWUKO KWADZO BANINI

STUDENT NO.: 2213880

DATE: 18TH JUNE 2023

COURSE:

MA 335: MODELLING EXPERIMENTAL AND OBSERVATIONAL DATA

ABSTRACT

This study aims to explore the relationship between characteristics of Alzheimer's disease and its diagnosis through a data science analysis. The analysis includes descriptive statistics, clustering algorithms, logistic regression, and feature selection. Descriptive statistics provide insights into the dataset, while clustering algorithms reveal patterns and groupings. Logistic regression helps predict the diagnosis based on the remaining variables, and feature selection identifies the most important features.

TABLE OF CONTENTS

WORD COUNT: 1560

INTRODUCTION

Alzheimer's disease, a complex neurodegenerative disorder, presents a pressing challenge in the field of healthcare, necessitating a comprehensive investigation into its characteristics and diagnosis. The overarching objective is to analyze an extensive dataset encompassing diverse Alzheimer's disease characteristics. By exploring the intricate relationship between these characteristics and the diagnosis, specifically distinguishing between Alzheimer's (Demented) and non-Alzheimer's (Nondemented) cases, this study employs a data-driven approach. Through the application of descriptive statistics, clustering algorithms, logistic regression, and feature selection methods, this analysis seeks to uncover crucial insights.

PRELIMINARY ANALYSIS

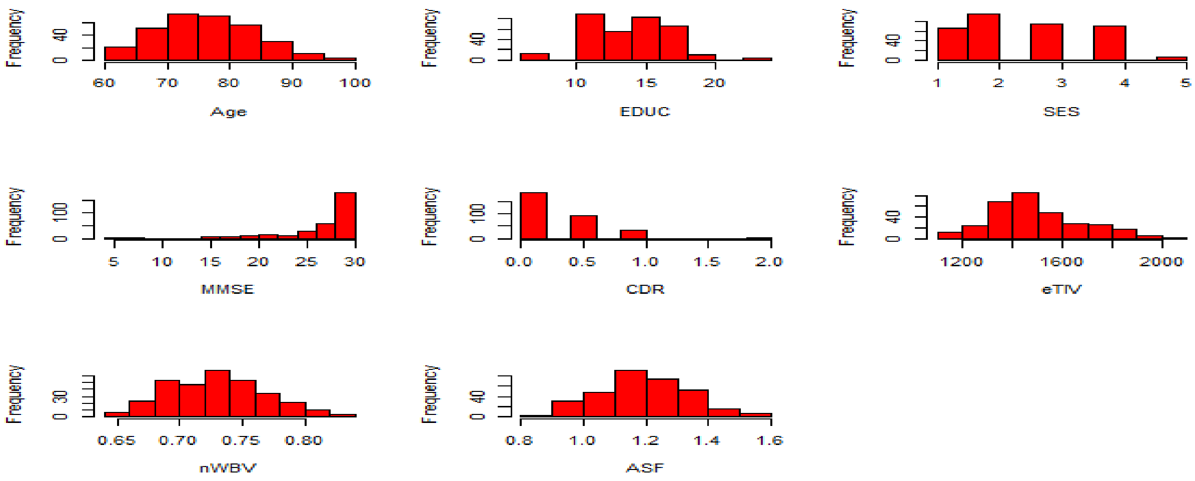
As part of the preliminary analysis and data cleaning procedures, the "M.F" variable was transformed into numeric values. Specifically, a value of 1 was assigned to represent males, while a value of 0 was assigned for females. This conversion was carried out using the ifelse function, ensuring consistency and ease of analysis across the dataset. Additionally, rows where the "Group" variable had the value "Converted" were excluded from further analysis. Furthermore, missing values were removed from the dataset. This data cleaning step aimed to maintain data integrity and ensure reliable analysis by working with complete and valid observations.

ANALYSIS AND DISCUSSION

a. Descriptive statistics

Out of the total population, 127 individuals were diagnosed with Alzheimer's disease (Demented), while 190 individuals did not exhibit signs of the disease (Nondemented).

	Age	Education	SES	MMSE	CDR	eTIV	nWBV	ASF
Min.	60.00	6	1	4	0	1106	0.6440	0.876
1st Qu.	71.00	12	2	27	0	1358	0.7000	1.098
Median	76.00	15	2	29	0	1476	0.7320	1.189
Mean	76.72	14.62	2.546	27.26	0.2729	1494	0.7306	1.192
3rd Qu.	82.00	16	3	30	0.5	1599	0.7570	1.293
Max.	98.00	23	5	30	2	2004	0.8370	1.587



A further analysis the table provides a comprehensive summary of the dataset, focusing on key variables related to Alzheimer's disease.

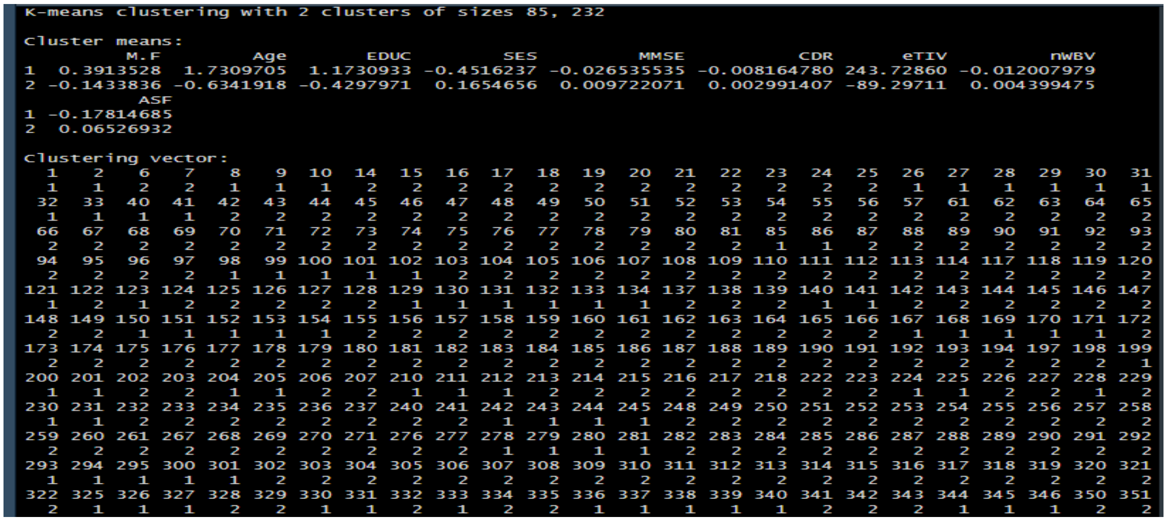
Demographically, the dataset includes both male (137) and female (180) participants. The age range spans from 60 to 98 years, with a median age of 76. The individuals have varying levels of education, with the years of education ranging from 6 to 23 years. The socioeconomic status ranges from 1 to 5, indicating a moderate socioeconomic distribution in the sample.

Cognitive measures are represented by the Mini-Mental State Examination (MMSE) scores, which range from 4 to 30. The dataset indicates a relatively high cognitive performance, as the mean MMSE score is 27.26. Furthermore, the Clinical Dementia Rating (CDR) values, ranging from 0 to 2, demonstrate a predominantly low severity of dementia symptoms, with a mean CDR score of 0.2729.

Brain metrics are captured through the estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and atlas scaling factor (ASF). The eTIV values range from 1106 to 2004, with a mean value of 1494. The nWBV, representing the normalized whole brain volume, ranges from 0.6440 to 0.8370, with a mean of 0.7306. The ASF, an atlas scaling factor, spans from 0.876 to 1.587, with a mean ASF of 1.192.

b. Implementing clustering algorithms.

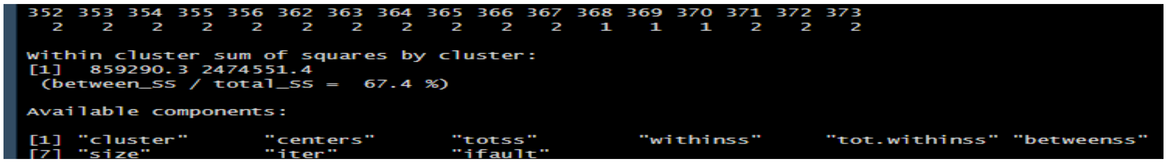
To implement clustering algorithms in the context of analyzing the relationship between characteristics of Alzheimer's disease and diagnosis, we can employ techniques such as Kmeans clustering and hierarchical clustering. These algorithms aim to identify natural groupings or clusters within the dataset based on the available variables. K-means clustering will be used in this project.



```
K-means clustering with 2 clusters of sizes 85, 232

Cluster means:
         M.F        Age       EDUC        SES         MMSE          CDR      eTIV         nWBV
1  0.3913528  1.7309705  1.1730933 -0.4516237 -0.026535535 -0.008164780 243.72860 -0.012007979
2 -0.1433836 -0.6341918 -0.4297971  0.1654656  0.009722071  0.002991407 -89.29711  0.004399475
          ASF
1 -0.17814685
2  0.06526932

Clustering vector:
  1   2   6   7   8   9  10  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31 
  1   1   2   2   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2   1   1   1   1   1   1 
 32  33  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  61  62  63  64  65 
  1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2 
 66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  85  86  87  88  89  90  91  92  93 
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1   1   2   2   2   2   2   2   2 
 94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 117 118 119 120 
  2   2   2   2   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2 
121 122 123 124 125 126 127 128 129 130 131 132 133 134 137 138 139 140 141 142 143 144 145 146 147 
  1   2   1   2   2   2   2   2   1   1   1   1   1   1   2   2   2   1   1   2   2   2   2   2   2 
148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 
  2   2   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2   1   1   1   1   1   2 
173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 197 198 199 
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1 
200 201 202 203 204 205 206 207 210 211 212 213 214 215 216 217 218 222 223 224 225 226 227 228 229 
  1   1   2   2   1   1   2   2   1   1   1   2   2   2   2   2   2   2   2   1   1   2   2   1   2 
230 231 232 233 234 235 236 237 240 241 242 243 244 245 248 249 250 251 252 253 254 255 256 257 258 
  1   1   2   2   2   2   2   2   2   2   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2 
259 260 261 267 268 269 270 271 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 
  2   2   2   2   2   2   2   2   2   2   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2 
293 294 295 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 
  1   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2 
322 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 350 351 
  2   1   1   1   2   2   1   1   2   1   2   2   1   1   1   1   1   2   2   2   1   1   1   2   2 
```

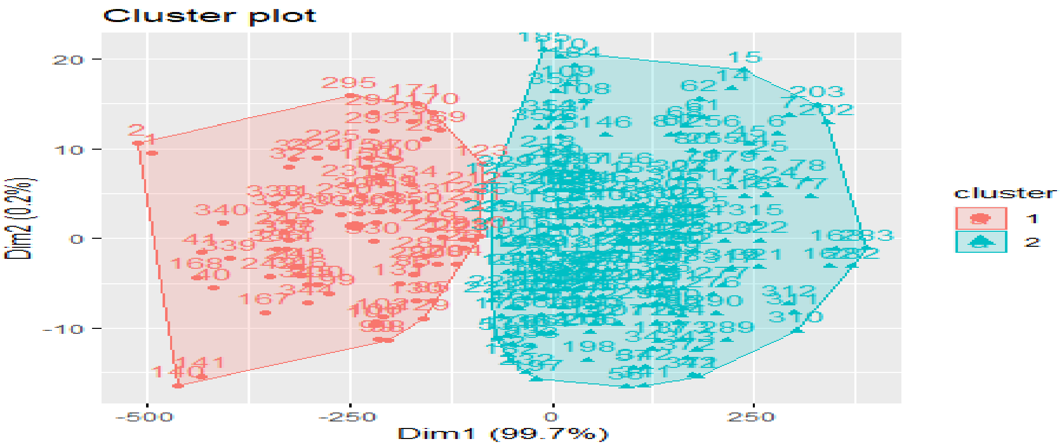


```
352 353 354 355 356 362 363 364 365 366 367 368 369 370 371 372 373 
  2   2   2   2   2   2   2   2   2   2   2   1   1   1   2   2   2 

Within cluster sum of squares by cluster:
[1]  859290.3 2474551.4
 (between_SS / total_SS =  67.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"   
[7] "size"         "iter"         "ifault"      
```



The K-means clustering algorithm has divided the dataset into two distinct clusters based on the given features. Let's refer to these clusters as Cluster 1 and Cluster 2.

Cluster means:

Cluster 1: This cluster has a higher average value for features such as M.F (gender), Age, EDUC (education level), and eTIV (estimated total intracranial volume). It indicates that the individuals in Cluster 1 tend to be older, have higher education levels, and potentially have larger brain volumes.

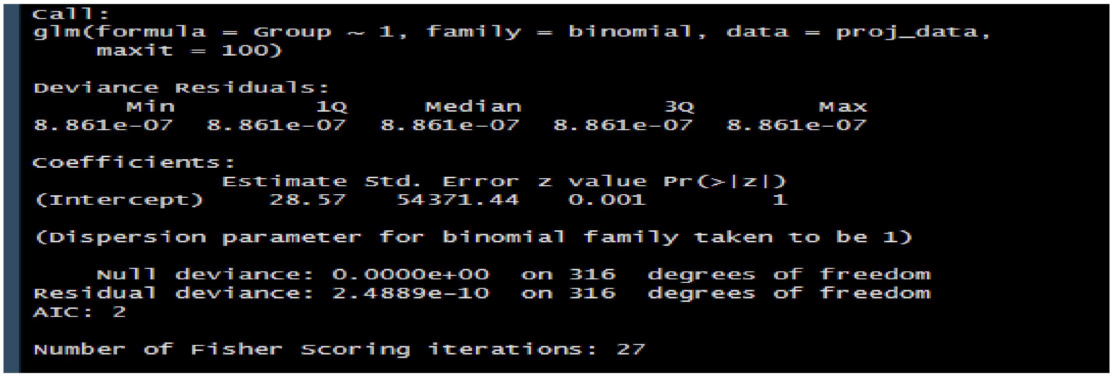
Cluster 2: This cluster has a lower average value for features such as M.F (gender), Age, EDUC (education level), and eTIV (estimated total intracranial volume). It suggests that the individuals in Cluster 2 tend to be younger, have lower education levels, and potentially have smaller brain volumes. These differences in feature means between the clusters can provide insights into the characteristics and demographics of the two groups. The within-cluster sum of squares (withinss) is a measure of the compactness or similarity of data points within each cluster. In this case, Cluster 1 has a within-cluster sum of squares of approximately 859,290.3, while Cluster 2 has a higher value of around 2,474,551.4. This suggests that the data points within Cluster 1 are more similar to each other in terms of their feature values compared to the data points in Cluster 2.

The between-cluster sum of squares (betweenss) is a measure of the separation or dissimilarity between the clusters. The proportion of between-cluster sum of squares to the total sum of squares (total_SS) is approximately 67.4%. This indicates that the clusters are somewhat distinct and reasonably separated.

In the scatter plot he larger symbols represent the cluster centers (centroids) of each cluster, showing the average feature values of the data points within that cluster.

By examining the cluster plot, you can observe the distribution and separation of the data points. The clusters are well-separated with two distinct groups of data points corresponding to each cluster.

c. **Implementing feature selection method**;



```
Call:
glm(formula = Group ~ 1, family = binomial, data = proj_data,
    maxit = 100)

Deviance Residuals:
      Min          1Q      Median          3Q         Max
8.861e-07   8.861e-07   8.861e-07   8.861e-07   8.861e-07

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    28.57   54371.44   0.001        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 316  degrees of freedom
Residual deviance: 2.4889e-10  on 316  degrees of freedom
AIC: 2

Number of Fisher Scoring iterations: 27
```

Using the backward selection method, there are differences between the observed response variable and the predicted values from the model. Here, the residuals are very small, indicating that the model fits the data well. The intercept is estimated to be 28.57. However, the standard error for this estimate is quite large (54371.44), meaning that the estimate is not very precise. The z-value and associated p-value (Pr(>|z|)) indicate that the intercept is not statistically significant, as the p-value is 1 (which is larger than the conventional threshold of 0.05).

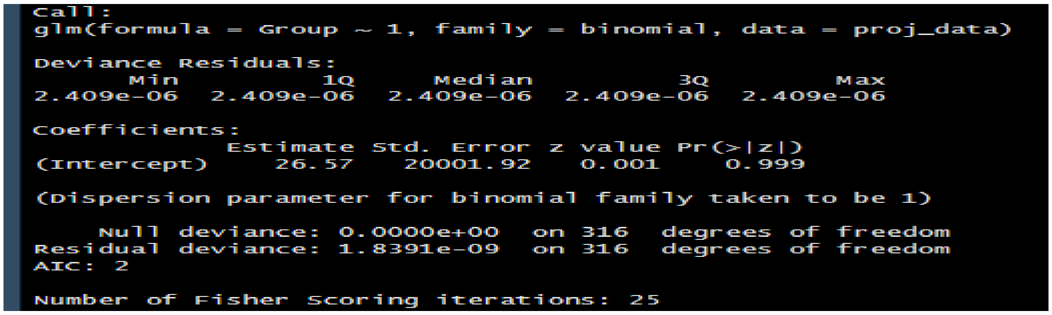
The dispersion parameter is set to 1, indicating that the model assumes a binomial distribution with a constant dispersion.

The null deviance is 0, indicating that the intercept-only model explains all the variability in the data. The residual deviance on the other hand is very small (2.4889e-10), suggesting that the model explains almost all of the variability in the data.

A lower AIC value indicates a better-fitting model. In this case, the AIC is 2, suggesting a good fit.

Number of Fisher Scoring iterations indicates the number of iterations required for the algorithm to converge and find the estimated coefficients. This took 27 iterations.

Overall, the model seems to have a good fit to the data, but the intercept coefficient is not statistically significant.



```
Call:
glm(formula = Group ~ 1, family = binomial, data = proj_data)

Deviance Residuals:
      Min          1Q      Median          3Q         Max
2.409e-06   2.409e-06   2.409e-06   2.409e-06   2.409e-06

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    26.57   20001.92   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 316  degrees of freedom
Residual deviance: 1.8391e-09  on 316  degrees of freedom
AIC: 2

Number of Fisher Scoring iterations: 25
```

The forward selection on the other hand is the intercept-only model (Group ~ 1). The coefficients section shows that the estimated intercept is 26.57, but the high standard error and insignificant p-value (0.999) indicate that the intercept is not statistically significant or informative in explaining the relationship between the predictors and the response. The null deviance is zero, indicating that the intercept-only model perfectly predicts the response variable. The residual deviance is very small (1.8391e-09), suggesting that the intercept-only model captures most of the information in the data.

The number of iterations performed is 25. The results indicate that the forward feature selection process did not add any predictors to the intercept-only model, as the AIC did not improve. This suggests that the intercept-only model provides the best fit to the data according to the AIC criterion.

d. Fitting a logistic regression model

The logistic regression model was fitted using the variables "EDUC," "MMSE," "CDR," and "ASF" to predict the "Group" variable. The model's deviance residuals, which measure the differences between the observed and predicted values, are extremely small. This indicates that the model fits the data very well also all the estimated coefficients have p-values greater than 0.05, indicating that they are not statistically significant and they may not have a meaningful relationship with the outcome variable. The AIC value of 10 suggests a reasonable balance between model fit and simplicity.

Overall, the model shows a good fit to the data based on the deviance residuals and AIC value. However, the coefficients for the predictor variables are not statistically significant, suggesting that these variables may not have a significant impact on predicting the "Group" variable.

In the context of a confusion matrix, the count of 64 in the row indicates that the model correctly classified 64 instances as "Demented" out of all the instances that are truly "Demented".

CONCLUSION

The logistic regression model with the selected variables did not provide strong evidence of significant associations between the independent variables and the group classification. The model had a good fit to the data, but the coefficients of the variables were not statistically significant.

REFERENCES

MA 335(2023). Modelling experimental and observational data: Course Module. Applied Data Science. University of Essex

APPENDIX

```
# Load required libraries
library(dplyr)
library(ggplot2)
library(reshape2)
library(summarytools)
library(caret)
library(factoextra)
# Load the dataset
setwd("C:/Users/HP/Desktop")
proj_data <- read.csv("project data.csv", header = T)
# Preliminary analysis
proj_data <- proj_data[proj_data$Group %in% c("Nondemented", "Demented"), ]
proj_data$M.F <- ifelse(proj_data$M.F == "M", 1, 0)
proj_data <- proj_data[proj_data$Group != "Converted", ]
proj_data <- na.omit(proj_data)
attach(proj_data)
# descriptive statistics table
table(M.F)
table(Group)
summary(proj_data)
# Histograms of all variables
par(mfrow = c(3, 3))
for (i in 3:10) {
  hist(proj_data[, i], main = "", col = "red", xlab = names(proj_data)[i])
}
# Scaling of all variables
Group <- ifelse(Group == "Nondemented", 0, 1)
proj_scale <- scale(proj_data[,2:10], center = TRUE, scale = FALSE)
# k means clustering
set.seed(123)
```

```
kmeans_clusters <- kmeans(proj_scale, centers=2, nstart = 1)
kmeans_clusters
fviz_cluster(kmeans_clusters, data = proj_scale,stand = FALSE)
# feature selection
Group <- factor(Group)
# backward selection
model_1 <- glm(Group ~ ., data = proj_data, family = binomial, maxit = 100)
step_1 <- step(model_1, method = "backward")
summary(step_1)
# forward selection
model_2 <- glm(Group ~ 1, data = proj_data, family = binomial)
step_2<-step(model_2,scope = ~ Group + M.F + Age + EDUC +    SES +  MMSE + CDR + eTIV    + nWBV +      ASF, method='forward')
summary(step_2)
#split data into training and test data
set.seed(123)
trainingRowIndex <- sample(1:nrow(proj_data), 0.8*nrow(proj_data))  # row indices for 80% training data
trainingData <- proj_data[trainingRowIndex, ]  # model training data
testData  <- proj_data[-trainingRowIndex, ]
y_test<- testData$Group
testData<- testData[,-1]
# Fit logistic regression model
model <- glm(Group ~ EDUC +        MMSE + CDR  +       ASF, data = trainingData, family = binomial)
summary(model)
#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
```