MA335

Final Project

Siddhant Shrenik Patil (2201538)

MSc Applied Data Science | Date: 20 June 2023

# Abstract:

This project aimed to understand how different characteristics relate to the diagnosis of Alzheimer's disease. This dataset used included variables such as age, education, socioeconomic status, cognitive assessments (MMSE), and clinical ratings (CDR). By analyzing the data using descriptive statistics, clustering algorithms, logistic regression, and feature selection, we discovered some important insights. Our analysis revealed that age, education, and socioeconomic status are linked to a higher risk of developing Alzheimer's. As people get older and have more years of education, their chances of being diagnosed with Alzheimer's increase. Additionally, individuals with higher socioeconomic status are more susceptible to the disease.

We also found that cognitive assessments, like the MMSE, and clinical ratings, such as the CDR, play a significant role in diagnosing Alzheimer's. Higher MMSE scores and lower CDR scores indicate a lower risk of the disease. These assessments help determine the presence and severity of Alzheimer's. These findings are crucial for understanding the risk factors associated with Alzheimer's disease. Variables like age, education, socioeconomic status, MMSE, and CDR can serve as potential indicators for early detection and prevention strategies. Healthcare professionals can use this information to develop targeted interventions and delay the progression of the disease.

Content:

|  |  |  |
| --- | --- | --- |
| Sr. No. | Title | Page No. |
| 1 | Introduction | 2 |
| 2 | Data Cleaning | 2 |
| 3 | Preliminary Data Analysis | 2 |
| 4 | Demonstration of Clustering Algorithms | 4 |
| 5 | Demonstration of Logistic Regression Model | 6 |
| 6 | Demonstration of Feature Selection Method | 7 |
| 7 | Conclusion | 7 |
| 8 | References | 8 |

Word Count: 1797

*1.Introduction*:

The goal of this project is to investigate how various characteristics of Alzheimer's disease relate to the diagnosis by analyzing a dataset. Utilizing statistical analysis methods and the R programming language to investigate the dataset and gain insight into the connections between these characteristics and the diagnosis of Alzheimer's disease is the primary objective of this project. We hope to discover significant Alzheimer's disease-related patterns and features by employing descriptive statistics, clustering algorithms, fitting a logistic regression model, and feature selection. The dataset's distribution, average values, and variability are all depicted in descriptive statistics. Algorithms called clustering group cases that are like one another, making it easier to find patterns or connections between people with Alzheimer's disease. Based on other variables, we will use logistic regression to predict the "Group" variable. The effects of these variables on the likelihood of an Alzheimer's diagnosis will be demonstrated by this model. The most important variables for predicting Alzheimer's disease are identified through feature selection. We can comprehend the key factors associated with the disease by evaluating the significance of each variable. Alzheimer's disease understanding, early detection, and prevention strategies could all benefit from this project's findings.

*2.Data Cleaning:*

1. The variable "M/F," representing gender, was originally stored as categorical values ("M" for male and "F" for female). To facilitate analysis, the values were converted into numeric format. In this case, "M" was represented as 0 and "F" as 1.

2. Initially, the dataset had 373 rows, but after removing the rows with Group = "Converted," it reduced to 336 rows.

3. 19 rows were found to have missing values, specifically in the SES and MMSE columns. After this step, the dataset was left with 317 rows.

*3.Preliminary Data Analysis:*

\*T-test analysis:

|  |
| --- |
| > t\_test\_result  Welch Two Sample t-test  data: demented\_gender and nondemented\_gender  t = -5.0138, df = 260.87, p-value = 9.869e-07  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -0.3863078 -0.1684373  sample estimates: mean of x mean of y  0.4015748 0.6789474 |

In order to compare the "M/F" variable between the dataset's "Demented" and "Nondemented" groups, the t-test was used.The t-test results show a significant difference between the two groups. With a p-value of 9.869e-07, it was determined that the mean value for the "M/F" variable in the "Demented" group (mean = 0.4015748) was significantly lower than mean value in the "Nondemented" group (mean = 0.6789474). This suggests that the gender distribution in the two groups differs noticeably.The 95% confidence interval for the mean difference also covered a range of -0.3863078 to -0.1684373. This indicates that there is a 95% confidence interval around the true proportional difference between male and females in the "Demented" and "Nondemented" groups.

\*Comparison of Two Groups based on Education parameter by using Boxplot:

Both groups share the same median education level, depicted as a horizontal line within the box, indicating a similar central tendency in education levels. However, the range of education levels varies in some ways. The longer whiskers of the nondemented group indicate that members of this group come from a broader range of educational backgrounds.the box plot reveals that while the Demented and Nondemented groups share similar central tendency in terms of education levels, the Nondemented group has a slightly wider range.

A picture containing text, screenshot, diagram, line

Description automatically generated

\*Density graph comparison:

The density curve for the nondemented group is taller and wider around the age of 80.This indicates that there are more individuals in the nondemented group with an average age of around 80. The density curve for the demented group is shorter and narrower around the age of 70-72.This suggests that there are fewer individuals in the demented group with an average age of around 70-72. From this analysis, we can conclude that individuals who are diagnosed with dementia tend to have an average age of around 70-72, whereas individuals in the nondemented group have a higher average age of around 80. This implies that the likelihood of developing dementia increases as individuals approach the age of 70, with the highest probability occurring around that age range.

A picture containing diagram, plot, text, screenshot

Description automatically generated

\*Correlation Plot:

Some key points are:

* The strong negative correlation between eTIV and ASF (-0.988) suggests that ASF tends to decrease as eTIV rises. This suggests that these two variables have a strong inverse relationship.
* CDR and MMSE have a strong negative correlation (-0.726), indicating that CDR scores tend to rise in tandem with MMSE scores. This suggests a strong connection between severity of dementia (higher CDR scores) and cognitive impairment (lower MMSE scores).
* EDUC and MMSE have a weak positive correlation (0.185), suggesting that higher education levels may be associated with higher MMSE scores.

A picture containing text, screenshot, font, number

Description automatically generated

*4. Demonstration of Clustering Algorithms:*

A] K-Means Clustering Algorithm:

In this K-means clustering analysis, firstly we need to decide the number of clusters in which we are analysis over data. The elbow method recommends choosing the number of clusters at the elbow point because it offers a fair compromise between capturing the data's underlying structure and preventing overfitting. By analyzing the within-cluster sum of squares (WSS) for various values of k, the elbow method aids in determining the number of clusters. By visualizing the results with the fviz\_nbclust function, we discovered that when the number of clusters is set to 3, the clusters formed exhibit distinct separation and do not overlap. Based on these two dimensions, dim1 and dim2, the data can be divided into three distinct clusters.

A picture containing text, line, plot, diagram

Description automatically generated A picture containing text, diagram, map

Description automatically generated

The three clusters' centroids have approximate coordinates of (29.6, 26.9). Based on how close it is to the centroid, each data point is assigned to one of these clusters. The fact that the clusters are distinct and do not overlap suggests that the clustering algorithm was successful in locating significant patterns and structure in the data. We have 3 clusters identified by the analysis. The number of data points in each cluster is 138, 77, and 102, respectively. I tried this analysis with 4 cluster also because it is little bit confusing when I saw the optimum number of cluster graph. After comparing the results of 3 and 4 clusters I got following findings:

For the three-cluster approach:

WCSS: 471.6878 variance: 54.6%

For the four-cluster approach:

WCSS: 472.5668 variance: 45.4%

According to these metrics, the 3-cluster solution explains more variance than the 4-cluster solution and has a slightly lower WCSS. A lower WCSS typically indicates better cluster separation and compactness. As a result, the 3-cluster solution may be regarded as superior to the 4-cluster solution in this instance. The data were divided into three similarity-based clusters by the k-means clustering analysis. The analysis quantifies the compactness of the clusters as well as their separation from one another. Each cluster has its own profile of average values.

B] Hierarchical clustering:

Based on how different the objects were from one another, the hierarchical clustering algorithm chose four different groups. Compared to objects in various clusters, those in the same cluster are thought to be more like one another. You can group objects that are similar using the cluster assignments, which can be useful for further data analysis or interpretation.

A picture containing text, screenshot, vehicle

Description automatically generated

This hierarchical clusters appear to be more visually distinct and well-separated in dendrograms produced by the complete and average linkage methods. This indicates that compared to observations in other clusters, observations within each cluster are more like one another. As a result, it is simpler to see and understand the different clusters.

On the other hand, the dendrograms produced by the single and centroid linkage methods might have less distinct boundaries between the clusters. This suggests that it may be more difficult to visually distinguish and interpret the clusters because the observations within each cluster may be more like those in other clusters.

*5. Demonstration of Logistic Regression Model:*

|  |
| --- |
| > logistic\_model0 <- glm(Group ~M.F+eTIV+nWBV+ASF+Age+EDUC+SES+MMSE+CDR, data = halfscale\_data, family = binomial)  > summary(logistic\_model0)  Call:  glm(formula = Group ~ M.F + eTIV + nWBV + ASF + Age + EDUC +  SES + MMSE + CDR, family = binomial, data = halfscale\_data)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.196e-04 -2.100e-08 2.100e-08 2.100e-08 1.381e-04  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -462.485 750861.902 -0.001 1.000  M.F 41.762 64818.698 0.001 0.999  eTIV 76.782 318829.917 0.000 1.000  nWBV 36.182 87406.218 0.000 1.000  ASF 33.556 369413.952 0.000 1.000  Age 6.473 7726.541 0.001 0.999  EDUC -3.827 12049.400 0.000 1.000  SES 16.899 45304.188 0.000 1.000  MMSE 24.698 76429.434 0.000 1.000  CDR -126.256 61706.153 -0.002 0.998  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 4.2685e+02 on 316 degrees of freedom  Residual deviance: 5.4985e-08 on 307 degrees of freedom  AIC: 20  Number of Fisher Scoring iterations: 25 |

The logistic regression model you fitted, known as the full model, aims to predict whether a person is demented or non-demented based on several independent variables: M.F, eTIV, nWBV, ASF, Age, EDUC, SES, MMSE, and CDR. However, when we examine the results of the model, we can see that it is not fitting the data properly. None of the independent variables show a significant relationship with the dependent variable (Group). This means that none of the variables are strongly associated with the likelihood of being demented or non-demented. To address this issue, we need to perform feature selection. Feature selection helps us identify the most important variables that have a meaningful impact on the prediction of the dependent variable. By selecting a subset of variables that are more relevant, we aim to improve the fit of the model and increase its predictive accuracy.

*6. Demonstration of Feature Selection Method:*

By performing feature selection, we aim to build more efficient and effective models by focusing on the most relevant information, reducing noise and redundancy, and improving interpretability and generalization. Here in this case, I choose the REF model over wrapper methods as we known RFE tends to be computationally more efficient and less prone to overfitting compared to wrapper methods. It is also suitable when the number of features is relatively large. The RFE algorithm selects the optimal subset of variables based on the highest performance metric. In this case, the asterisk (\*) indicates that the model achieved the best performance with a subset of 9 variables, including CDR, M.F, ASF, nWBV, and eTIV

The top 5 variables (out of 9): CDR, M.F, ASF, nWBV, eTIV.

After performing feature selection using the top 5 variables (CDR, M.F, ASF, nWBV, eTIV) in a logistic regression model, it was observed that the model did not provide a good fit. To address this issue, collinearity between the variables was examined. It was found that the variables eTIV and CDR were highly correlated, indicating redundancy or multicollinearity. To improve the model's fit, these two variables were removed. Subsequently, a revised logistic regression model was fitted, which showed improved performance and a better fit to the data.

This logistic regression model suggests that the variables M.F, nWBV, and ASF have a significant impact on predicting the diagnosis of Alzheimer's disease.

After finding the good fit model, checked the accuracy of this model by splitting the given data into train and test data in 70&30 respectively. Finally, I got the AUC: 0.8388388. AUC addresses the discriminative force of the model. An AUC of 0.5 demonstrates an irregular model, while an AUC more like 1 shows a superior prescient presentation. The logistic regression model's moderate discriminative power in predicting the "Group" variable is indicated by the AUC of 0.8388388in this instance.

|  |
| --- |
| > logistic\_model3 <- glm(Group ~ M.F+nWBV+ASF, data = halfscale\_data, family = binomial)  > summary(logistic\_model3)  Call:  glm(formula = Group ~ M.F + nWBV + ASF, family = binomial, data = halfscale\_data)  Deviance Residuals:  Min 1Q Median 3Q Max  -2.1019 -0.9967 0.5206 0.8980 1.7396  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -0.4767 0.2154 -2.213 0.0269 \*  M.F 1.7135 0.3280 5.223 1.76e-07 \*\*\*  nWBV 0.7640 0.1420 5.382 7.37e-08 \*\*\*  ASF -0.6365 0.1591 -4.002 6.29e-05 \*\*\*  ---  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: 358.14 on 313 degrees of freedom  AIC: 366.14  Number of Fisher Scoring iterations: 4 |

*7.Conclusion:*

The objective was to investigate the connection between these characteristics and the diagnosis of either non-Alzheimer's disease (Non-demented) or Alzheimer's disease (Demented). I used k-means clustering in my analysis to find distinct patterns and groups in the data. I was successful in dividing the data into three clusters by selecting the optimal number of clusters based on variance and within-cluster sum of squares. In addition, I used logistic regression to make a diagnosis prediction based on several independent variables. However, there were no significant correlations between the variables and the diagnosis because the initial model did not fit the data well. I used feature selection to fix this, removing highly correlated variables to make the model fit better. In conclusion, this project shed light on how the characteristics of Alzheimer's disease and the diagnosis relate to one another.

*8.Refrances:*

* Alzheimer's Association. (2021). 2021 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 17(3), 327-406.
* Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., ... & Liu, E. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14(4), 535-562.
* Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., ... & Morris, J. C. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimer's & Dementia, 13(4), e1-e85.
* Ritchie, C. W., Muniz-Terrera, G., Kivipelto, M., Solomon, A., Tom, B., Molinuevo, J. L., & EPAD Consortium. (2017). The European Prevention of Alzheimer's Dementia (EPAD) longitudinal cohort study: Baseline data release–V500.0. Journal of Prevention of Alzheimer's Disease, 4(4), 255-263.

\*Appendix\*

# clear out the past

if(!is.null(dev.list())) dev.off()

rm(list = ls())

cat("\014")

library(caret)

library(dplyr) #data manipulation that provides easy-to-use functions for filtering, grouping, and summarizing data, among other tasks.

library(tidyr) #data tidying that provides functions for reshaping data from "wide" to "long" format, and vice versa.

library(moments)

library(ggplot2)

library(ILS)

library(gridExtra)

setwd("/Users/siddhantpatil/Desktop/MA335 /Final project-20230603")

Proj\_data <- read.csv("project data.csv")

head(Proj\_data)

#\*\*Data Cleaning\*\*#

###Convert M/F into numeric values###

Proj\_data$M.F <- ifelse(Proj\_data$M.F == "M", 0, 1)

head(Proj\_data)

###Remove rows with Group = "Converted"

dim(Proj\_data)

# Count occurrences of "Converted" in the Group column

converter\_count <- sum(Proj\_data$Group == "Converted")

converter\_count

# Remove rows with Group = "Converted"

Proj\_data1 <- Proj\_data[Proj\_data$Group != "Converted", ]

dim(Proj\_data1)

####

dim(Proj\_data1)

# Count missing values in each column

missing\_counts <- colSums(is.na(Proj\_data1))

missing\_counts

# Remove rows with missing values

Proj\_data1 <- na.omit(Proj\_data1)

dim(Proj\_data1)

head(Proj\_data1)

Proj\_data2<-Proj\_data1

Proj\_data2$Group <- ifelse(Proj\_data2$Group == "Nondemented", 1, 0)

final\_data<-Proj\_data2

final\_data1<-final\_data

fullscale\_data<-scale(final\_data)

final\_data1[,5:10]<- scale(final\_data1[,5:10])

halfscale\_data<-final\_data1

head(Proj\_data1)

head(Proj\_data2)

head(final\_data)

head(final\_data1)

head(fullscale\_data)

head(halfscale\_data)

#\*\*Preliminary Data Analysis\*\*#

# Subset the dataset for Demented and Nondemented groups

demented\_gender <- Proj\_data1$M.F[Proj\_data1$Group == "Demented"]

nondemented\_gender <- Proj\_data1$M.F[Proj\_data1$Group == "Nondemented"]

# Independent t-test

t\_test\_result <- t.test(demented\_gender, nondemented\_gender)

# Print the result

t\_test\_result

#boxplot

boxplot(EDUC ~ Group, data = Proj\_data2, main = "Box Plot of Education (year) by Group", xlab = "Group", ylab = "Education (Year)", col = c("#1f77b4", "#ff7f0e", "#2ca02c"))

# Create density graph comparison

ggplot(Proj\_data1, aes(x = Age, color = Group, fill = Group)) +

geom\_density(alpha = 0.5) +

labs(x = "Age", y = "Density") +

scale\_fill\_manual(values = c("blue", "red", "green")) +

scale\_color\_manual(values = c("blue", "red", "green")) +

theme\_minimal()

# Create correlation plot with color and numerical values

# Calculate correlation matrix

cor\_matrix <- cor(Proj\_data2[, c("Age", "EDUC", "MMSE", "CDR", "eTIV", "nWBV", "ASF")])

corrplot(cor\_matrix, method = "shade", type = "full", tl.cex = 0.7, addCoef.col = "black")

#\*\*Demonstration of Clustering Algorithms\*\*#

data1 <- scale(final\_data)

set.seed(123)

fviz\_nbclust(data1, kmeans, method = "wss")+

geom\_vline(xintercept = 3, linetype = 2)

kmeans2 <- kmeans(data1, centers = 2, nstart = 100)

kmeans3 <- kmeans(data1, centers = 3, nstart = 100)

kmeans4 <- kmeans(data1, centers = 4, nstart = 100)

f1 <- fviz\_cluster(kmeans2, geom = "point", data = data1) + ggtitle("k = 2")

f2 <- fviz\_cluster(kmeans3, geom = "point", data = data1) + ggtitle("k = 3")

f3 <- fviz\_cluster(kmeans4, geom = "point", data = data1) + ggtitle("k = 4")

grid.arrange(f1, f2, f3, nrow = 2)

#Hierarchical clustering

idx <- sample(1:dim(Proj\_data2)[1], 50);idx

df1<-Proj\_data2[idx,]

dfX<-df1[,-(c(1,2))]

dfX<- na.omit(dfX)

dfX<-scale(dfX)

#Start my calculating the distance matrix

d <- dist(dfX, method = "euclidean")

#Apply hierarchical clustering for differnt linkage methods

fit.single <- hclust(d, method="single")

fit.complete <- hclust(d, method="complete")

fit.average <- hclust(d, method="average")

fit.centroid <-hclust(d, method="centroid")

#par(mfrow=c(2,2))

plot(fit.single) # print the dendrogram

groups.fit.single <- cutree(fit.single, k=4) # cut tree into k=4 clusters # draw dendrogram with red borders around the 4 clusters rect.hclust(fit.single, k=4, border="red")

#Checking how many observations are in each cluster

table(groups.fit.single)

plot(fit.complete)

groups.fit.complete <- cutree(fit.complete, k=4)

rect.hclust(fit.complete, k=4, border="red")

table(groups.fit.complete)

plot(fit.average)

groups.fit.average <- cutree(fit.average, k=4)

rect.hclust(fit.average, k=4, border="red")

table(groups.fit.average)

plot(fit.centroid)

groups.fit.centroid <- cutree(fit.centroid, k=4)

rect.hclust(fit.centroid, k=4, border="red")

table(groups.fit.centroid)

#\*\*Perform logistic regression\*\*#

logistic\_model0 <- glm(Group ~M.F+eTIV+nWBV+ASF+Age+EDUC+SES+MMSE+CDR, data = halfscale\_data, family = binomial)

summary(logistic\_model0)

#\*\*Feature selection\*\*#

X <- halfscale\_data[, c("M.F", "Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV", "ASF")] # Replace with the column names you want as X variables

# Create the Y variable vector

Y <- halfscale\_data[["Group"]] # Replace with the column name of your target variable

set.seed(10)

control <- rfeControl(functions = lmFuncs, method = "repeatedcv",#cv

repeats = 10,

number = 10)

lmProfile <- rfe(X, Y,

sizes = c(1:6),

rfeControl = control)

lmProfile

plot(lmProfile, type = c("g", "o"))

plot(lmProfile, metric = "Rsquared", type = c("g", "o"))

predictors(lmProfile)

lmProfile$fit

#\*Again logistic regression\*#

# Perform logistic regression again using the top features

logistic\_model2 <- glm(Group ~ M.F+nWBV+ASF+CDR+eTIV, data = halfscale\_data, family = binomial)

summary(logistic\_model2)

logistic\_model3 <- glm(Group ~ M.F+nWBV+ASF, data = halfscale\_data, family = binomial)

summary(logistic\_model3)

#checking accuracy of this model

data00 <- as.data.frame(halfscale\_data)

train\_index1 <- sample(1:nrow(data00), nrow(data00) \* 0.8)

train\_data1 <- data00[train\_index1, ]

test\_data1 <- data00[-train\_index1, ]

logistic\_model3 <- glm(Group ~ M.F+nWBV+ASF, data = train\_data1, family = binomial)

# Make predictions on the test data

test\_predictions1 <- predict(logistic\_model3, newdata = test\_data1, type = "response")

# Evaluate the model performance on the test data using pROC

library(pROC)

roc\_obj2 <- roc(test\_data1$Group, test\_predictions1)

auc2 <- auc(roc\_obj2)

# Print the AUC value

cat("AUC:", auc2, "\n")