MA335 Final Project

”Exploring the Relationship Between Characteristics and Alzheimer’s Diagnosis: A Data Analysis of Alzheimer’s Dataset”

PRAJWAL MARKAL PUTTASWAMY

REG NUM:2013173

Abstract

To analyze a dataset that includes several Alzheimer’s disease features and their relationships with diagnoses, this study uses leading-edge data science approaches. The predictive variables connected to Alzheimer’s disease are better understood by descriptive statistics, clustering algorithms, logistic regression modeling, and feature selection methods. While descriptive statistics offer numerical summaries and graphical representations to visualize data distribution and interrelationships between variables, data preparation maintains the integrity of the data. To shed light on probable illness manifestations, clustering algorithms look for patterns and groups. By identifying important factors, logistic regression modeling forecasts the diagnosis. Techniques for feature selection help determine which features are most useful for predicting diagnoses. This work incorporates strong data analysis methods to reveal insightful information on the relationship between Alzheimer’s traits and diagnosis, guiding early detection, prognosis, and intervention.

June 21, 2023
Colchester

Contents

Word Count without Cover page and appendix: 1606

Introduction

Alzheimer's disease is a prevalent neurodegenerative disorder that significantly impacts individuals and their families. Understanding the relationship between disease characteristics and diagnosis is crucial for early detection and improved patient outcomes [1]. In this project, we utilize data science methodologies to analyze a comprehensive dataset encompassing diverse attributes associated with Alzheimer's disease.

Type	Variable	Description
Predictor variable	M/F	Gender
	Age	Age
	EDUC	Year of education
	SES	Socioeconomic Status (1-5, 1-low, 5-high)
	MMSE	Mini mental state examination
	CDR	Clinical dementia rating (1 of 2)
	eTIV	Estimated total intracranial volume
	nWBV	Normalize whole brain volume
	ASF	Atlas scaling factor
Response Variable	Group	Group of the diagnosis (Nondemented, Demented, Other)

Table 1.1: Description of Predictor and Target variables

By employing descriptive statistics, clustering algorithms, logistic regression modeling, and feature selection techniques, our goal is to gain insights into the predictive factors contributing to the diagnosis of Alzheimer's disease. We also aim to identify distinct patterns and groupings within the dataset using clustering algorithms, which can reveal potential subtypes or phenotypes associated with Alzheimer's disease and enhance our understanding of its heterogeneity [3].

The dataset used in this study contains information on demographic factors, cognitive assessments, brain imaging measures, and clinical diagnoses. We implement data preprocessing steps, including conversion of gender values, removal of irrelevant data, and handling of missing values, to ensure dataset integrity [2]. Descriptive statistics and graphical representations such as boxplots, histograms, and scatterplots are employed to provide a comprehensive overview and explore relationships between variables [4].

Preliminary Analysis

The dataset consists of various variables including Group (Nondemented, Demented, Other), M/F (Male or Female), Age, EDUC (Years of Education), SES (Socioeconomic Status), MMSE (Mini Mental State Examination), CDR (Dementia Severity), eTIV (Total Intracranial Volume), nWBV (Normalized Whole Brain Volume), and ASF (Atlas Scaling Factor).

Strong negative correlation between the ASF and eTIV (-0.9886) shows that the atlas scaling factor decreases as the projected total intracranial volume rises. This implies that bigger brains often have a smaller scaling factor, which could reflect variations in brain anatomy or makeup.

MMSE and nWBV have a mildly positive association (0.3707). This shows that those with greater normalised whole brain sizes, as determined by the MMSE, often perform more cognitively. This association suggests that better cognitive performance may be linked to larger brain sizes.

	eTIV	nWBV	ASF	MMSE	CDR
eTIV	1.00000000	-0.1950752	-0.98863912	-0.02063041	0.04071274
nWBV	-0.19507524	1.0000000	0.19778979	0.37071410	-0.35514855
ASF	-0.98863912	0.1977898	1.00000000	0.03169271	-0.05243847
MMSE	-0.02063041	0.3707141	0.03169271	1.00000000	-0.72628935
CDR	0.04071274	-0.3551485	-0.05243847	-0.72628935	1.00000000

Table 2.1: Correlation Matrix

The MMSE and CDR exhibit a significant negative association (-0.7263), meaning that as the severity of clinical dementia symptoms (CDR) grows, so do the MMSE scores, indicating diminishing cognitive abilities.

The correlation between eTIV and MMSE is relatively low (-0.0206), indicating that brain volume alone does not substantially predict cognitive performance as measured by the MMSE.

Discussion

3.1 K-Means Clustering

The output represents the results of a K-means clustering analysis with 4 clusters. Each cluster is characterized by its cluster mean values for the standardized variables:

Age, Education (EDUC), Socioeconomic Status (SES), Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), Estimated Total Intracranial Volume (eTIV), Normalized Whole Brain Volume (nWBV), and Atlas Scaling Factor (ASF).

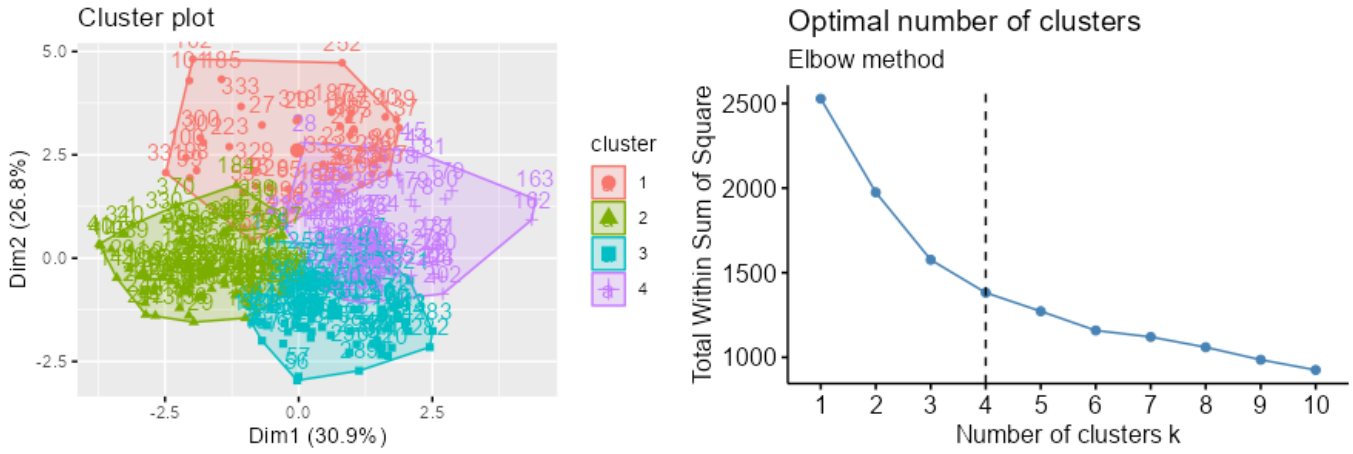

Figure 3.1: K-Means Clustering

Cluster 1 (n=44): Lower age, education, and SES. Higher likelihood of cognitive impairment with lower MMSE scores and higher CDR values. Larger eTIV, potential brain atrophy (lower nWBV), smaller brain size (below average ASF).

Cluster 2 (n=101): Higher age and education, lower SES. Moderate cognitive scores, larger eTIV, average brain volume (nWBV), smaller brain size (significantly below average ASF).

Cluster 3 (n=80): Higher age, lower education, higher SES. Moderate cognitive scores, slightly elevated CDR values, smaller eTIV, average brain volume (nWBV), larger brain size (above average ASF).

Cluster 4 (n=92): Lower age, slightly higher education, lower SES. Moderate cognitive impairment, smaller eTIV, higher brain volume (nWBV), smaller brain size (below average ASF).

The clusters represent distinct groups based on age, education, SES, cognitive measures, brain volume, and size, highlighting differences in cognitive impairment, brain atrophy, and intracranial volume.

3.2 Logistic Regression

To examine the association between the predictor variables (Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF,Gender) and the binary response variable (Group), a logistic regression analysis was done. The diagnosis group (Nondemented, Demented, Other) is represented by the response variable.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.161e+03  5.217e+06   0.000    1.000
Age          6.473e+00  7.727e+03   0.001    0.999
EDUC        -3.828e+00  1.205e+04   0.000    1.000
SES          1.505e+01  4.034e+04   0.000    1.000
MMSE         6.396e+00  1.979e+04   0.000    1.000
CDR         -3.304e+02  1.615e+05  -0.002    0.998
eTIV         4.272e-01  1.774e+03   0.000    1.000
nWBV         9.496e+02  2.294e+06   0.000    1.000
ASF          2.403e+02  2.645e+06   0.000    1.000
Gender      -4.176e+01  6.482e+04  -0.001    0.999

    Null deviance: 4.2685e+02  on 316  degrees of freedom
Residual deviance: 5.4985e-08  on 307  degrees of freedom
```

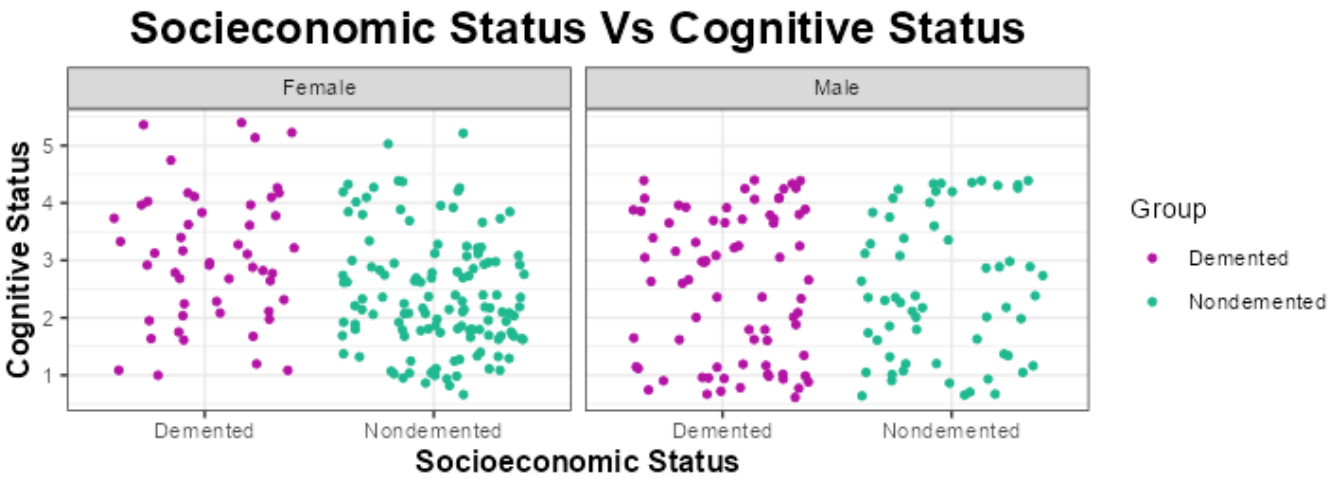

Figure 3.2: Socieconomic Status Vs Cognitive Status

According to the findings, none of the predictor factors had a statistically significant influence on the likelihood of belonging to a certain diagnostic category. The coefficients are statistically insignificant, as indicated by huge standard errors and p-values near to one. As a result, the estimated effects should be evaluated cautiously, and the predictor factors do not appear to be significant predictors of group membership.

The null deviance and residual deviance, which quantify the model's goodness of fit, indicate that it sufficiently matches the data, with low residual deviance suggesting a good fit. Given the number of parameters, an AIC score of 18 suggests a moderately favorable model fit.

3.3 Feature selection

Initially, the logistic regression model comprised variables "MMSE," "eTIV," and "nWBV," generating an AIC of 153.87. "eTIV" was shown to be less important by backward elimination and was so eliminated. The increased AIC value of 152.08 for the simplified model with "MMSE" and "nWBV" indicated a better fit with more significant predictors.

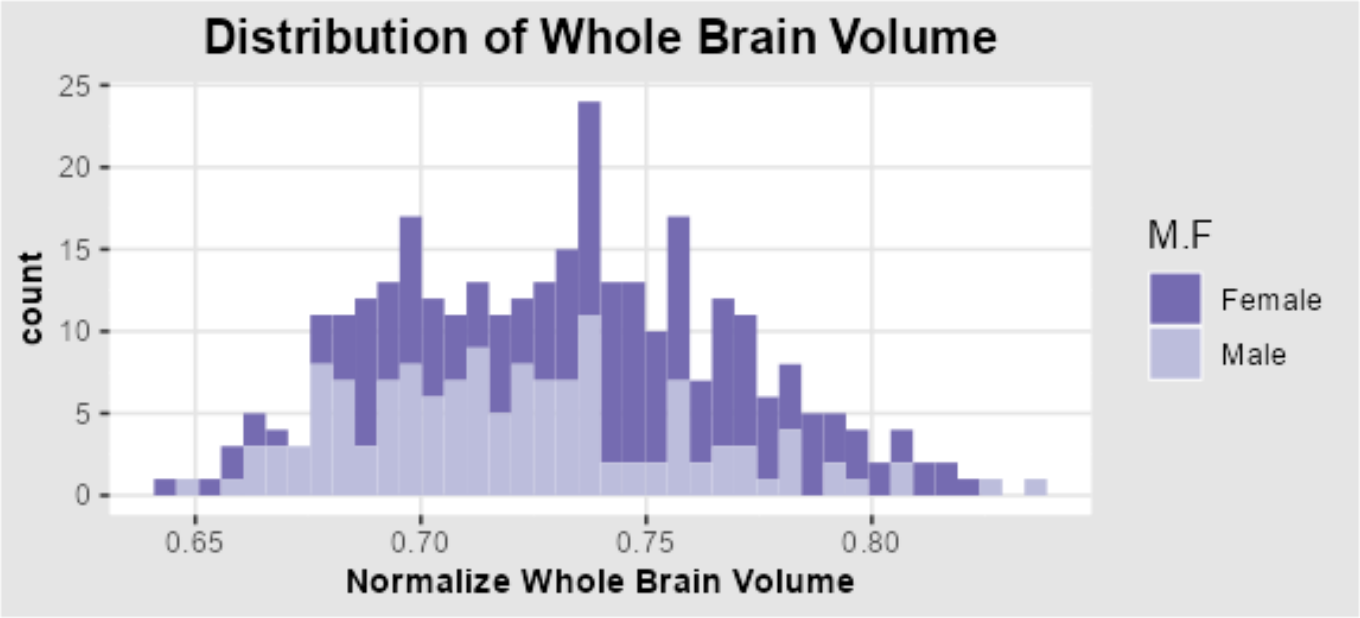

Figure 3.3: Distribution of whole Brain volume

"MMSE" and "nWBV" were chosen as predictors in the final model. The categorization "demented" was shown to be connected with higher MMSE scores, with an approximately 4.24 times greater chance per one-unit increase in MMSE. There was also a little increase in the probability of being labelled "Demented" with greater nWBV levels, about 0.45 times per unit increase. While the link between nWBV and the "Demented" categorization was only marginally significant, it does show a probable trend.

```
Start:  AIC=153.87
Group ~ MMSE + eTIV + nWBV
       Df Deviance    AIC
- eTIV  1   146.08 152.08
<none>      145.87 153.87
- nWBV  1   149.45 155.45
- MMSE  1   273.37 279.37
```

```
Step:  AIC=152.08
Group ~ MMSE + nWBV
       Df Deviance    AIC
<none>      146.08  152.08
- nWBV  1    150.31  154.31
- MMSE  1    273.45  277.45
```

Finally, the final logistic regression model supports the statistically significant link between MMSE scores and ”Demented” categorization. The addition of ”nWBV” as a predictor suggests a possible relationship, but further research is needed for stronger statistical support.

Deviance residuals showed that the model fit the data rather well overall. With an AIC score of 152.08, the model seems to have struck a fair balance between model fit and complexity.

On the Training data, the logistic regression model had an accuracy of about 84.7 %, accurately classifying 62 instances as ”Demented” and 126 cases as ”Nondemented.” When put to the test on Test data, the model continued to perform consistently with an accuracy of about 83.2 %, accurately categorising 27 instances as ”Demented” and 52 cases as ”Nondemented.” This shows that the model can accurately estimate a person’s level of dementia.

Conclusion

Finally, our Alzheimer’s disease investigation used a comprehensive strategy that included K-Means clustering, logistic regression, and feature selection algorithms. Clustering was used to identify unique groups within the dataset, giving insight on trends and commonalities among people with Alzheimer’s disease. The logistic regression approach gave useful insights into the predictive parameters impacting illness incidence, while feature selection improved model accuracy by selecting the most relevant variables. This in-depth examination helps to a better knowledge of Alzheimer’s disease, which may benefit in diagnostic and treatment options. The combination of these techniques provides useful information for future research and clinical applications in the management of Alzheimer’s disease.

Bibliography

[1] Alzheimer’s Association. (2021). *Alzheimer’s Disease Facts and Figures*. Retrieved from https://www.alz.org/media/documents/alzheimers-facts-and-figures.pdf

[2] Garcia, L. A., Ramos, J. F., & Rangel, D. H. (2014). Missing data imputation in health care information systems: a systematic review. *Health Informatics Journal*, 20(3), 137-158.

[3] Hinneburg, A., & Keim, D. A. (1998). An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*.

[4] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Appendix

```
####     Exploratory Data Analysis    ####

# Load necessary libraries
library(dplyr)          # For data manipulation
library(ggplot2)        # For data visualization
library(factoextra)     # For clustering analysis
library(MASS)           # For Feature Selection
library(cluster)        # For clustering algorithms


# Load the dataset
data <- read.csv("C:/Users/prajw/OneDrive/Desktop/project/ma335/project
    ␣data.csv")

# Convert 'M' and 'F' into numeric values
data$Gender <- ifelse(data$M.F == "M", 1, 0)

# Convert 'M.F' into descriptive labels
data$M.F <- ifelse(data$M.F == "M", "Male", "Female")

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

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

# Convert Group variable to a factor
data$Group <- as.factor(data$Group)


####    Introduction    ####
```

```
# Select the relevant variables
selected_data <- data[, c("eTIV", "nWBV", "ASF", "MMSE", "CDR")]

# Calculate the correlation matrix
correlation_matrix <- cor(selected_data, use = "pairwise.complete.obs")

# View the correlation matrix
correlation_matrix

###   Standardizing Numerical Variables   ###

# Subset the numerical variables
numerical_variables <- data[, c("Age", "EDUC", "SES", "MMSE", "CDR", "
    eTIV", "nWBV", "ASF")]

# Calculate mean and standard deviation for each variable
variable_means <- apply(numerical_variables, 2, mean)
variable_sds <- apply(numerical_variables, 2, sd)

# Standardize the variables
standardized_variables <- scale(numerical_variables)

# Create a new dataframe with the standardized variables
new_data <- data
new_data[, c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV", "ASF"
    )] <- standardized_variables



####   K-Means Clustering   ####

# Subset the numerical variables
clusters <- new_data[, c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "
    nWBV", "ASF")]
```

```r
# Perform K-Means clustering
km <- kmeans(clusters, centers = 4, nstart = 50, iter.max = 100)

# Visualize the clustering results
fviz_cluster(km, data = clusters)

# Display cluster information
km


####  Logistic Regression  ####

# Perform logistic regression
lreg <- glm(Group ~ Age + EDUC + SES + MMSE + CDR + eTIV + nWBV + ASF +
    Gender, data = data, family = binomial)


# Display a summary of the logistic regression model
summary(lreg)

# Create a Jitterplot using ggplot
ggplot(data) +
 aes(x = Group, y = SES, colour = Group) +
 geom_jitter(size = 1.2) +
 scale_color_manual(values = c(Demented = "#A50026",
Nondemented = "#313695")) +
 labs(x = "Cognitive Status", y = "Socioeconomic Status", title = "
    Socieconomic Status Vs Cognitive Status") +
 ggthemes::theme_solarized() +
 theme(plot.title = element_text(size = 18L, face = "bold", hjust =
    0.5),
 axis.title.y = element_text(size = 12L, face = "bold"), axis.title.x =
     element_text(size = 12L, face = "bold")) +
 facet_wrap(vars(M.F), scales = "free_x")
```

```
####   Feature Selection   ####

###   splitting the data for training   ###

# Set the seed for reproducibility
set.seed(132)

# Generate random indices for splitting the data
indices <- sample(1:nrow(new_data), size = nrow(new_data), replace =
   FALSE)

# Define the proportion of data to be used for training
train_proportion <- 0.7

# Determine the number of samples for training and testing
train_size <- round(train_proportion * nrow(new_data))
test_size <- nrow(new_data) - train_size

# Split the data into training and testing sets
train_data <- new_data[indices[1:train_size], ]
test_data <- data[indices[(train_size + 1):nrow(data)], ]
train_data$Group <- as.factor(train_data$Group)
test_data$Group <- as.factor(test_data$Group)

###   Train Data   ###

# Fit a logistic regression model on the Train data using the
   predictors MMSE, eTIV, and nWBV
lreg1 <- glm(Group ~ MMSE + eTIV + nWBV, data = train_data, family =
   binomial)

# Perform backward elimination using stepAIC function
reduced_model <- stepAIC(lreg1, direction = "backward")
```

```
# Display summary of the reduced model
summary(reduced_model)

# Predict probabilities of Y=1 (Demented) using the reduced model
glm.probs <- predict(reduced_model, type = "response")

# Assign predicted classes based on the probability threshold of 0.5
glm.predicted <- rep("Demented", 222)
glm.predicted[glm.probs > 0.5] = "Nondemented"

# Create a contingency table of predicted vs. actual classes
table(glm.predicted, train_data$Group)

# Calculate the accuracy of the predicted classes
mean(glm.predicted == train_data$Group)

# Create a Histogram using ggplot
ggplot(data) +
 aes(x = nWBV, fill = M.F) +
 geom_histogram(bins = 40L) +
 scale_fill_manual(values = c(Female = "#756BB1",
Male = "#BCBDDC")) +
 labs(x = "Normalize Whole Brain Volume", title = "Distribution of
    Whole Brain Volume") +
 ggthemes::theme_igray() +
 theme(plot.title = element_text(size = 18L,
 face = "bold", hjust = 0.5), axis.title.y = element_text(size = 12L,
    face = "bold"), axis.title.x = element_text(size = 12L,
 face = "bold"))


###   Test Data   ###

# Fit a logistic regression model on the test data using the predictors
    MMSE, eTIV, and nWBV
```

```
lreg2 <- glm(Group ~ MMSE + eTIV + nWBV, data = test_data, family =
    binomial)

# Perform backward elimination using stepAIC
reduced_model2 <- stepAIC(lreg2, direction = "backward")

# Print the summary of the reduced model
summary(reduced_model2)

# Predict the probabilities of the response variable using the reduced_
    model2
glm.probs2 <- predict(reduced_model2, type = "response")  # Pr(Y = 1|X)

# Create predicted labels based on the probability threshold of 0.5
glm.predicted2 <- rep("Demented", 95)
glm.predicted2[glm.probs2 > 0.5] <- "Nondemented"

# Create a contingency table to compare the predicted labels with the
    actual Group values in the test_data
table(glm.predicted2, test_data$Group)

# Calculate the accuracy by comparing the predicted labels with the
    actual Group values in the test_data
mean(glm.predicted2 == test_data$Group)

# Create a boxplot using ggplot
ggplot(data) +
 aes(x = EDUC, y = Group, colour = Group) +
 geom_boxplot(fill = "#112446") +
 scale_color_hue(direction = 1) +
 labs(x = "Education", y = "Cognitive Status", title = "Year of
    Education VS Cognitive Status") +
 coord_flip() +
 theme_minimal() +
 theme(plot.title = element_text(size = 16L, face = "bold", hjust =
```