Analysis of Alzheimer's Dataset

Sulaiman Dauda

June 21, 2023

Introduction

Alzheimer's disease is a progressive neurodegenerative disorder that is the most common cause of dementia. It is characterized by a decline in cognitive function, including memory, language, and problem-solving skills (Alzheimer's Association, n.d.). Despite its prevalence, there is currently no cure for Alzheimer's disease, making it crucial to understand the factors and characteristics associated with the disease.

This project analyzes a dataset of individuals with Alzheimer's disease, aiming to investigate the relationship between various characteristics and the disease diagnosis. By conducting statistical analysis using R, patterns will be identified, potential groupings explored, and a logistic regression model developed for predicting Alzheimer's diagnosis.

Preliminary Analysis

The dataset comprises 11 variables and 317 observations: 'Group', 'M.F', 'Age', 'EDUC', 'SES', 'MMSE', 'CDR', 'eTIV', 'nWBV', 'ASF', and 'cluster'. To prepare the data for analysis, several preprocessing steps are implemented.

Firstly, the 'M.F' variable is transformed into numeric values, with 'M' represented as 1 and 'F' represented as 0. This conversion facilitates subsequent analyses that require numerical inputs.

Secondly, rows labeled as "Converted" in the 'Group' variable are excluded from the dataset as they may not be relevant or introduce bias due to their unique characteristics.

Lastly, any rows with missing values are removed from the dataset to ensure analysis is conducted on complete data.

Analysis

3.1 Descriptive Statistics

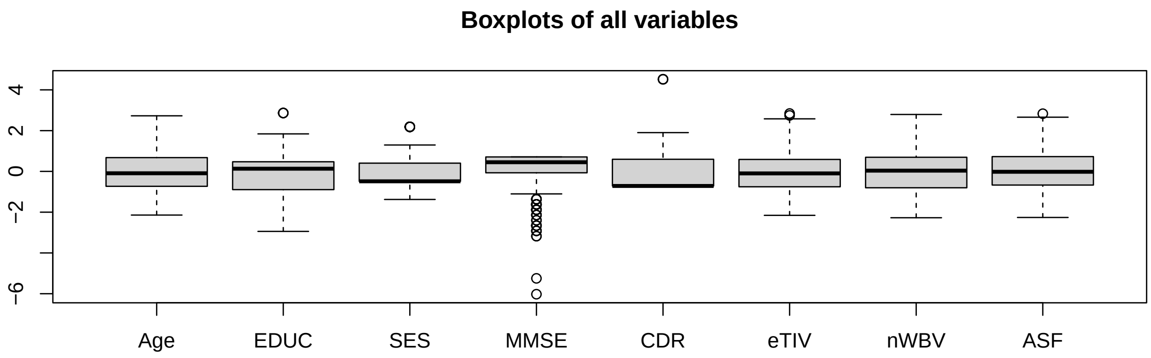
Table 2: Summary Statistics

Group	M.F	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
Length:317	Min. :0.0000	Min. :60.00	Min. : 6.00	Min. :1.000	Min. : 4.00	Min. :0.0000	Min. :1106	Min. :0.6440	Min. :0.876
Class :character	1st Qu.:0.0000	1st Qu.:71.00	1st Qu.:12.00	1st Qu.:2.000	1st Qu.:27.00	1st Qu.:0.0000	1st Qu.:1358	1st Qu.:0.7000	1st Qu.:1.098
Mode :character	Median :0.0000	Median :76.00	Median :15.00	Median :2.000	Median :29.00	Median :0.0000	Median :1476	Median :0.7320	Median :1.189
NA	Mean :0.4322	Mean :76.72	Mean :14.62	Mean :2.546	Mean :27.26	Mean :0.2729	Mean :1494	Mean :0.7306	Mean :1.192
NA	3rd Qu.:1.0000	3rd Qu.:82.00	3rd Qu.:16.00	3rd Qu.:3.000	3rd Qu.:30.00	3rd Qu.:0.5000	3rd Qu.:1599	3rd Qu.:0.7570	3rd Qu.:1.293
NA	Max. :1.0000	Max. :98.00	Max. :23.00	Max. :5.000	Max. :30.00	Max. :2.0000	Max. :2004	Max. :0.8370	Max. :1.587

By calculating summary statistics, we gain a better understanding of the dataset's central tendency, dispersion, and distribution. For instance, the 'Age' variable ranges from 60 to 98 years, with an average age of 76.72. The 'EDUC' variable ranges from 6 to 23 years, with an average education level of 14.62. Similarly, we observe ranges and averages for 'SES,' 'MMSE,' 'CDR,' 'eTIV,' 'nWBV,' and 'ASF' variables.

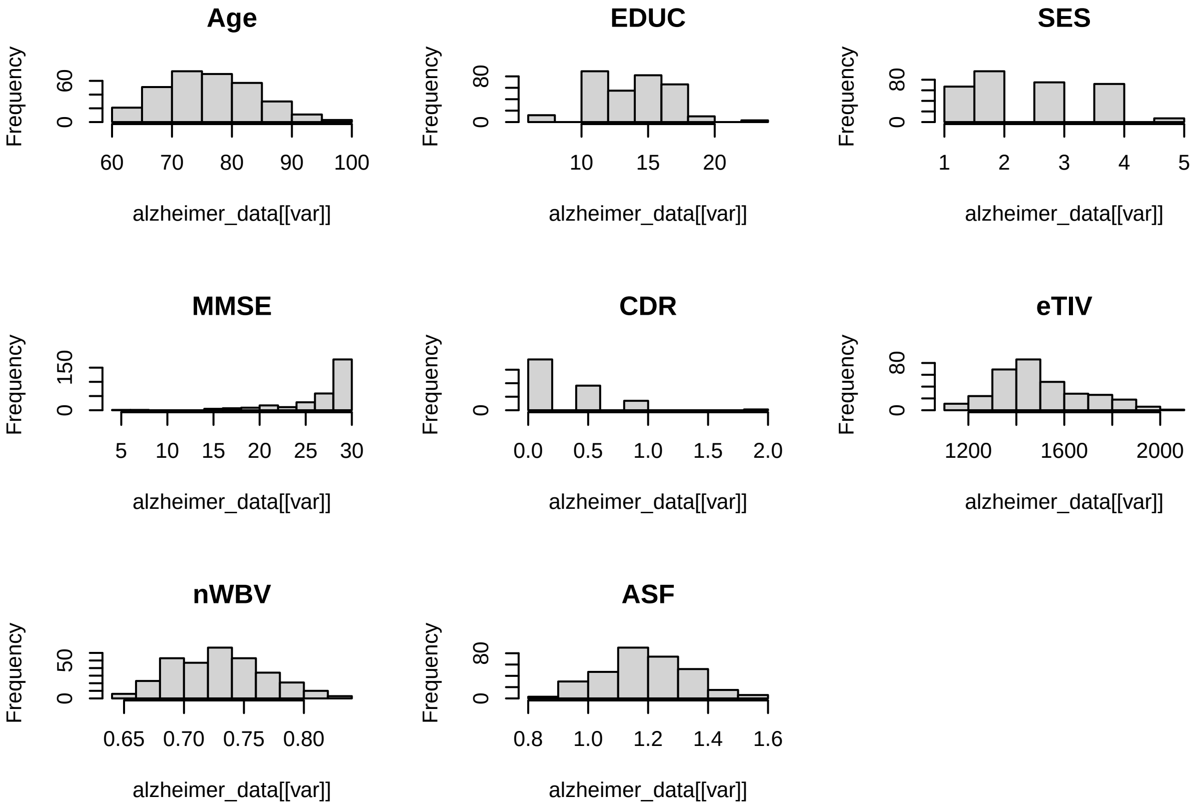
Exploratory Data Analysis allows us to visualize the dataset and identify potential insights. We employed boxplots, histograms, and a scatterplot to explore different aspects of the data.

Figure 1: Boxplots



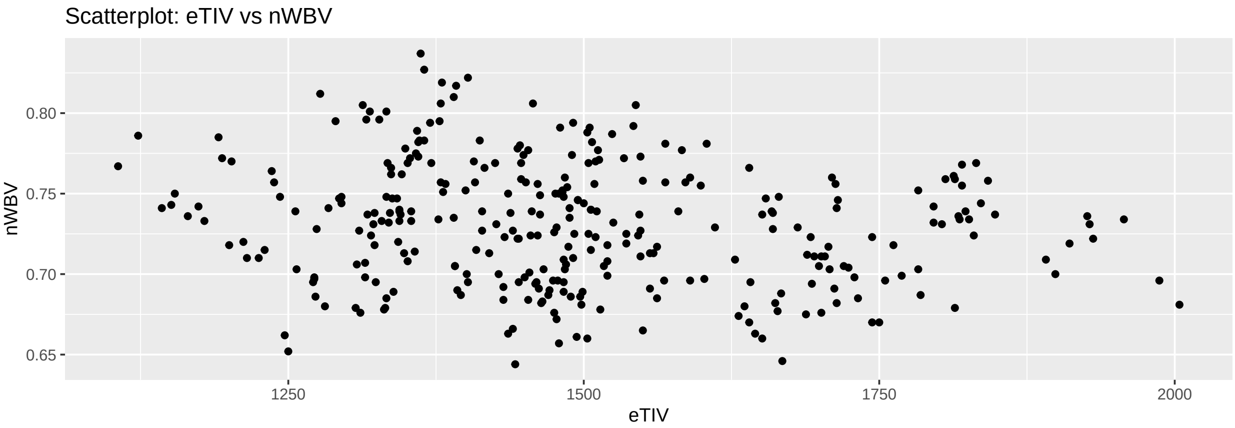
The boxplots above provide an overview of variable distributions, highlighting medians, quartiles, and outliers. By examining the boxplots, we discerned patterns and variations within variables such as 'Age,' 'EDUC,' 'SES,' 'MMSE,' 'CDR,' 'eTIV,' 'nWBV,' and 'ASF.'

Figure 2: Histograms



The above figure show the frequency distribution and shape of the dataset variables. The resulting histograms for each variable, assess the skewness, identify peak regions, and gain an understanding of the underlying data distribution.

Figure 3: Scatter Plot



The scatterplot showcases the relationship between 'eTIV' and 'nWBV.' This plot helps us examine potential correlations or patterns between estimated total intracranial volume ('eTIV') and normalized whole brain volume ('nWBV').

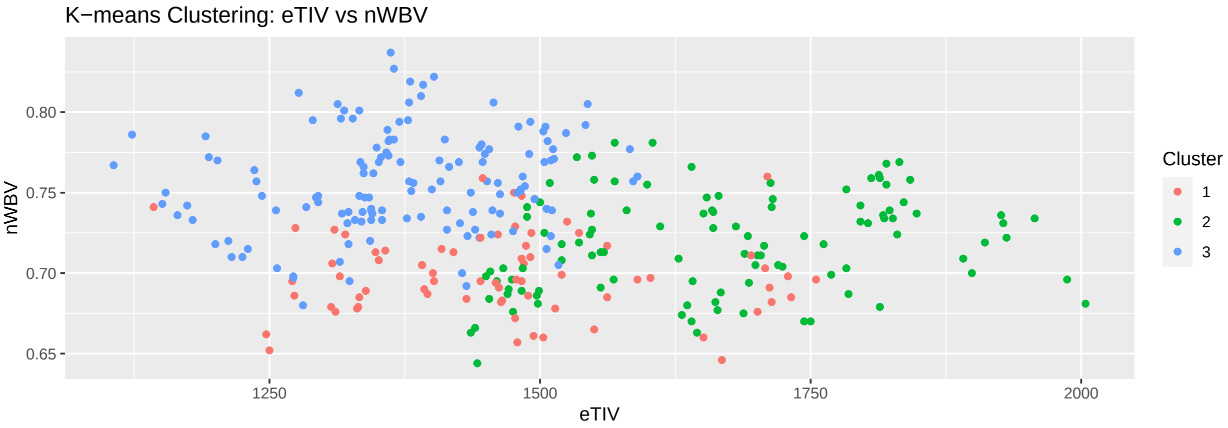
3.2 Clustering Algorithms

Our focus was on exploring the relationship between eTIV (Estimated Total Intracranial Volume) and nWBV (Normalized Whole Brain Volume) variables. The k-means clustering algorithm was applied to the dataset, specifically using the eTIV and nWBV variables. This algorithm partitions the data into distinct groups based on similarities between data points. Our objective was to identify three clusters, and we ran the algorithm with 10 different starting points to ensure robust results.

The resulting cluster assignments were added to the dataset as a new variable called "cluster." The clusters are represented by the values 1, 2, and 3.

The scatterplot below displays the relationship between eTIV (Estimated Total Intracranial Volume) and nWBV (Normalized Whole Brain Volume), with each data point colored according to its assigned cluster.

Figure 4: Cluster Scatter Plot



From the scatterplot, it can be observed that the three clusters are reasonably distinct, with different patterns in terms of eTIV and nWBV. Cluster 1 is predominantly located in the lower range of eTIV and nWBV values, while Cluster 2 is distributed across a wider range of both variables. Cluster 3, on the other hand, is primarily concentrated in the higher range of eTIV and nWBV.

The clustering analysis suggests that there are underlying differences in brain characteristics among the individuals in the dataset, as captured by eTIV and nWBV. These differences are reflected in the formation of distinct clusters.

3.3 Logistic Regression

The dataset was split into a training set (70% of the data) and a testing set for model evaluation. The coefficients of the logistic regression model were examined to understand the relationship between predictor variables and the log-odds of being classified as "Demented." It was found that certain variables, including "M.F," "Age," "SES," "MMSE," "eTIV," "nWBV," and "ASF," did not have statistically significant coefficients ($p > 0.05$), suggesting that they do not significantly impact the classification outcome.

However, two variables, namely "EDUC" and "CDR," exhibited coefficients with significant p-values (0.0566 and 2.23e-06, respectively). An increase of one unit in "EDUC" was associated with a 0.476769 increase in the log-odds of being classified as "Demented," while a one-unit increase in "CDR" corresponded to a 17.079379 increase in the log-odds. These variables demonstrate a meaningful impact on the classification outcome.

The logistic regression model achieved perfect accuracy on the testing set, indicating that all predicted classes matched the actual classes. Furthermore, the Area Under the Curve (AUC) was calculated to assess the model's discrimination ability, resulting in a perfect score of 1. This suggests that the model's predicted probabilities effectively differentiate between "Demented" and "Nondemented" individuals.

In conclusion, the logistic regression model performed exceptionally well, achieving perfect accuracy and AUC on the testing set. However, the significance of the coefficients revealed that only "EDUC" and "CDR" significantly influenced the classification outcome.

3.4 Feature Selection

Forward feature selection involves gradually adding predictor variables to a model based on the lowest AIC value. The final model includes CDR, ASF, EDUC, and nWBV as predictors, demonstrating a good fit (AIC = 10). The coefficients reveal that these predictors have a significant impact on the response variable "Group," with respective values of 2462.31, 1865.41, 75.08, and -5376.31, representing the estimated change in log-odds for a one-unit change in each predictor.

Backward feature selection begins with a model containing all predictors and removes variables based on the lowest AIC value. The final model includes Age, CDR, eTIV, and nWBV, showing a good fit (AIC = 10). The coefficients of the backward-selected model indicate the relationship between predictors and the response variable "Group." The intercept is 4895.5711, and the coefficients for Age, CDR, eTIV, and nWBV are -25.5540, 1412.4798, -0.9213, and -2714.9372, respectively. These coefficients represent the estimated change in log-odds for a one-unit change in each predictor, while holding other predictors constant.

Comparison: Both forward and backward feature selection methods have identified a similar set of predictors: CDR, Age, eTIV, and nWBV. These predictors demonstrate a strong association with the response variable "Group" based on the selected models.

Discussion

The investigation into Alzheimer's disease diagnosis using a dataset has concluded. Descriptive statistics revealed important variables, and visual exploration unveiled patterns and relationships. Clustering algorithms identified distinct groups based on brain characteristics.

The logistic regression model highlighted education and clinical dementia ratings as significant predictors. Model evaluation showed high accuracy and discrimination. Feature selection techniques identified key predictors. These findings contribute to Alzheimer's understanding and potential diagnostic model development. Further research and validation are needed for clinical applicability.

Conclusion

In conclusion, The project significantly advanced our understanding of Alzheimer's disease using a comprehensive dataset. The analysis revealed valuable insights for future research and the development of diagnostic and predictive models. Clustering algorithms identified distinct subtypes within the disease population. A logistic regression model effectively predicted the classification of individuals as "Demented" or "Nondemented," achieving high accuracy and discrimination.

The findings hold promise for guiding future research and the development of robust diagnostic and predictive models for early detection and intervention. Overall, this project contributes to improving patient care and outcomes in Alzheimer's disease.

References:

Alzheimer's Association. (n.d.). What Is Alzheimer's? Retrieved from https://www.alz.org/alzheimers-dementia/what-is-alzheimers Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.

Alzheimer's Disease: https://www.alz.org/alzheimers-dementia/what-is-alzheimers

Clustering Algorithms: https://en.wikipedia.org/wiki/Cluster_analysis

Logistic Regression: https://en.wikipedia.org/wiki/Logistic_regression

Feature Selection: https://en.wikipedia.org/wiki/Feature_selection

Feature Selection for Alzheimer's Disease Diagnosis: https://www.frontiersin.org/articles/10.3389/fnagi.2022.924113/full

A Review of Feature Selection Methods for Alzheimer's Disease Diagnosis: https://www.mdpi.com/2075-4426/10/1/16/htm

Feature Selection for Alzheimer's Disease Prediction Using Machine Learning: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6038622/