ABSTRACT

This report presents a comprehensive analysis of a dataset containing diverse characteristics related to Alzheimer's disease. The main objective is to investigate the relationship between these characteristics and the diagnosis of Alzheimer's (Demented) or non-Alzheimer's (Nondemented) status using various statistical methods and machine learning algorithms in R. The analysis begins with descriptive statistics, including numerical summaries and visual representations such as summary tables, boxplots, histograms, and pair plots to illustrate variable distributions and relationships. Clustering algorithms, namely K-means clustering and Hierarchical clustering, are then applied to identify patterns and clusters in the data, providing insights through validation measures and graphical comparisons. A logistic regression model is developed to predict the Group variable (diagnosis), using remaining predictor variables, and its implications and interpretations are discussed. Additionally, a feature selection method, specifically the wrapper feature selection technique (Boruta), is implemented to identify the most significant features associated with Alzheimer's disease. The findings contribute valuable insights into the relationship between these characteristics and Alzheimer's diagnosis.

CONTENTS

Word count without cover page and appendix: 1798

INTRODUCTION

Alzheimer's disease is a progressive neurological disorder that affects millions of individuals worldwide. It is a complex condition influenced by a combination of genetic, environmental, and lifestyle factors. Understanding the relationship between various characteristics and the diagnosis of Alzheimer's is crucial for early detection and intervention. In this report, we analyze a dataset containing multiple variables related to Alzheimer's disease and investigate the association between these characteristics and the diagnosis Group (Demented or Nondemented) using statistical analysis techniques.

Research in [1], suggests that age, with advancing age being the primary risk factor, and lower levels of education and socioeconomic status are associated with an increased risk of developing Alzheimer's disease. Additionally, cognitive test scores and clinical dementia ratings provide objective measures of cognitive impairment and disease progression.

To ensure the accuracy and reliability of the analysis, the R programming language was employed, leveraging its capabilities for statistical analysis and machine learning. The statistical methods used in this study provide a comprehensive understanding of the dataset

PRELIMINARY ANALYSIS

This Informational report is based on a dataset that includes various characteristics related to Alzheimer's disease. The Dataset consist of 10 variables and 373 rows before data cleaning. The "Group" variable is the main target variable, representing the diagnosis of individuals as either "Demented" or "Nondemented". This variable will be used to investigate the relationship between the other characteristics and the diagnosis. The other 9 variables provide insights into the factors associated with Alzheimer's disease diagnosis.

In the preliminary analysis, we conducted an initial exploration of the dataset to gain insights into its structure and characteristics. We performed data cleaning tasks such as converting the "M.F" variable to numeric values, removing rows with "Group" labeled as "Converted," and missing values. We then conducted descriptive statistics of numerical variables to summarize the dataset, including measures such as mean, median, minimum, maximum,1st quartile, and 3rd quartile.

	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
	Min. :60.00	Min. : 6.00	Min. :1.000	Min. : 4.00	Min. :0.0000	Min. :1106	Min. :0.6440	Min. :0.876
	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
	Median :76.00	Median :15.00	Median :2.000	Median :29.00	Median :0.0000	Median :1476	Median :0.7320	Median :1.189
	Mean :76.72	Mean :14.62	Mean :2.546	Mean :27.26	Mean :0.2729	Mean :1494	Mean :0.7306	Mean :1.192
	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
	Max. :98.00	Max. :23.00	Max. :5.000	Max. :30.00	Max. :2.0000	Max. :2004	Max. :0.8370	Max. :1.587

Table 1: Summary of Numerical Variables

The participants' age ranges from 60 to 98 years. Most participants fall between the ages of 71 and 82, as indicated by the first and third quartiles. The years of education (EDUC) completed by participants range from 6 to 23, with an average of approximately 14.62 years. Socioeconomic status (SES) is rated on a scale from 1 to 5, with higher values indicating higher socioeconomic status. The median SES value is 2, indicating a predominantly lower socioeconomic status among the participants. The Mini Mental State Examination (MMSE) score, which assesses cognitive function, ranges from 4 to 30. The mean MMSE score is approximately 27.26, suggesting a moderate level of cognitive function overall. The Clinical Dementia Rating (CDR) measures the severity of dementia symptoms. The maximum CDR value is 2, indicating a higher level of severity. The estimated total intracranial volume (eTIV) ranges from 1106 to 2004. The mean eTIV is approximately 1494, indicating the average estimated volume inside the skull. The normalized whole brain volume (nWBV) ranges from 0.644 to 0.837, with a mean value of approximately 0.7306. The Atlas Scaling Factor (ASF) ranges from 0.876 to 1.587. ASF is used in brain imaging analysis and represents a scaling factor for brain atlases.

Plotting Histogram for continues variables to see their distribution

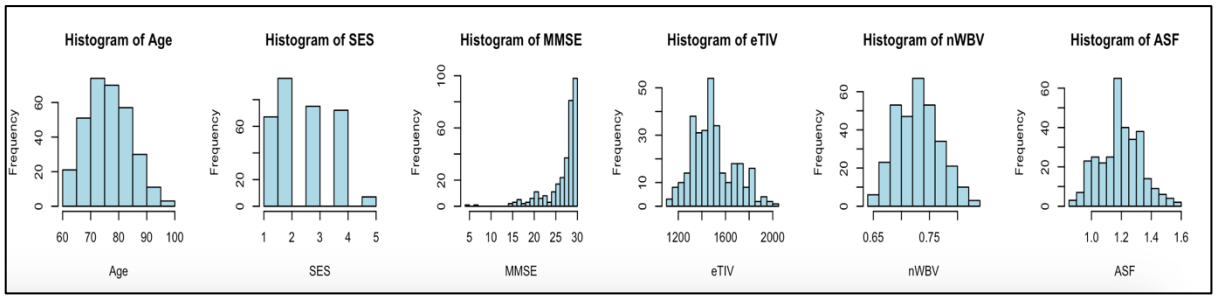

Figure 1: Histogram for Continues Variables

The five continuous-numerical variables seem to follow normal distribution except the MMSE. MMSE is left-skewed which suggests that a significant portion of the individuals in the dataset have higher MMSE scores, indicating better cognitive functioning.

We then looked at the relation of Group with Gender (M.F). The table below show the relation of Group with Gender (M.F)

Group		Gender			Demented	Nondemented	Demented	Nondemented
Demented	Nondemented	Female	Male	Female	51	129	40.15748	67.89474
127	190	180	137	Male	76	61	59.84252	32.10526

Table 2: Relation of Group with Gender

There were 317 people of which 127 were Demented and 190 were Nondemented. There were 180 females of which 51 were Demented and there were 137 males of which of which 76 were demented. Further analysis reveals that out of the 127 individuals diagnosed as

Demented, approximately 59.84% were males, while approximately 40.15% were females. This suggests that a higher proportion of males were diagnosed with Dementia compared to females.

We plotted box plots to see the relation between predictor variables and target variable.

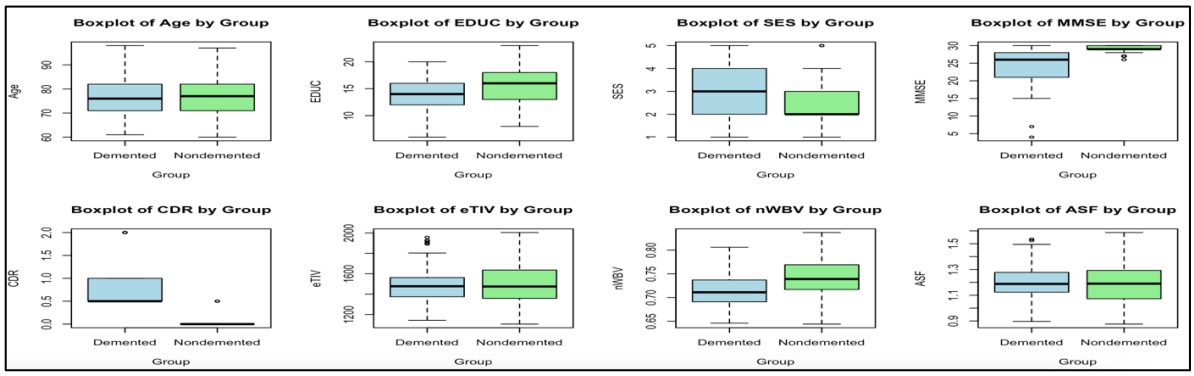

Figure 2: Boxplot between predictor and target variables

We discovered that individuals who scored above 27.5 on the Mini Mental State Examination (MMSE) and those that had a Clinical Dementia Rating (CDR) below 0.5 were classified as Nondemented. Therefore, individuals meeting the criteria of MMSE > 27.5 and CDR < 0.5 are more likely to be classified as Nondemented, indicating a healthier cognitive status.

The last thing we did is we plotted the pair plot to see the correlation of variables (figure is included in the Appendix). In the pair plot we found out that there is a highly negative correlation between ASF and eTIV hence we dropped the column ASF before clustering

ANALYSIS

To prepare for clustering, we standardized the numerical variables, ensuring they were on a common scale. Additionally, we encoded the target variable, assigning a value of 1 for Demented and 0 for Nondemented. Subsequently, we employed two clustering algorithms. The first one was K-means clustering, and using the silhouette method, we determined that the optimal number of clusters was 2.

Figure 3: K-means Clustering

DISCUSSION

The K-means clustering analysis resulted in the formation of two distinct clusters. The first cluster, labelled Cluster 1, contains 93 data points, while the second cluster, labeled Cluster 2, contains 224 data points. Upon examining the cluster means, we observe differences in the average values of the variables between the two clusters. In Cluster 1, participants have a higher value for the Group variable, indicating a higher likelihood of being classified as Demented. Cluster 1 has a higher proportion of males (M.F = 0.559) compared to Cluster 2 (M.F = 0.379). The average age in Cluster 1 is slightly higher (Age = 0.119) than in Cluster 2 (Age = -0.049), suggesting a potential association between age and the presence of dementia. In terms of education, participants in Cluster 1 have lower levels of education (EDUC = -0.500) compared to those in Cluster 2 (EDUC = 0.207). Socioeconomic status (SES) is also higher in Cluster 1 (SES = 0.424) compared to Cluster 2 (SES = -0.176), indicating a potential correlation between lower socioeconomic status and dementia. Furthermore, the average Mini Mental State Examination (MMSE) score is notably lower in Cluster 1 (MMSE = -1.109), suggesting a greater degree of cognitive impairment compared to Cluster 2 (MMSE = 0.461). The Clinical Dementia Rating (CDR) is significantly higher in Cluster 1 (CDR = 1.199), indicating a higher severity of dementia symptoms, while Cluster 2 has a lower CDR value (CDR = -0.498). we also found out that cluster 1 consists of 98.9% demented individuals. These findings suggest that Cluster 1 represents individuals with a higher likelihood of being diagnosed with dementia, characterized by older age, lower education levels, lower socioeconomic status, poorer cognitive function, and reduced brain volumes. On the other hand, Cluster 2 consists of individuals with relatively better cognitive function and brain measurements. The analysis provides valuable insights into the characteristics of these two distinct groups and their potential association with dementia. We also performed Hierarchal clustering and found similar results the figure is included in the appendix.

LOGISTIC REGRESSION

To perform logistic regression, we first checked the significance of each variable. We did not select CDR as it is clinical assessment scale that measures the severity of dementia and cognitive. The variables named M.F, Age, MMSE, and nWBV were found significant
The Logistic regression model demonstrated that gender, age, MMSE, and brain volume (nWBV) have significant impacts on predicting the "Group" category. Specifically, older age, higher MMSE scores, and larger brain volumes are associated with a higher likelihood of

belonging to the predicted group. The model's performance metrics, including 92.6% accuracy, 90% precision, 80% recall, and 88.5% F1 score, indicate its effectiveness in correctly classifying individuals into the "Group" category based on these predictors.

FEATURE SELECTION

After applying wrapper feature selection by Boruta. We got the result below

The result shows that CDR is the most important feature for predicting the Alzheimer. The second most important feature is MMSE. CDR and MMSE are used as important features because CDR capture the severity of dementia and cognitive decline, while MMSE assesses various cognitive functions. Including these features helps capture overall cognitive impairment, disease progression, specific cognitive performance, enhancing the accuracy of predicting the outcome but if we use these features only then we will not be able to see how other features predict the outcome that is why we have removed CDR from Logistic Regression model previously.

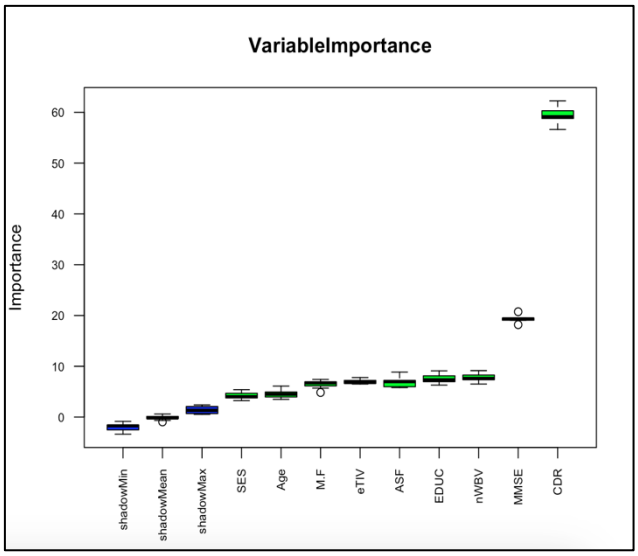

Figure 4: Importance of Features

CONCLUSION

The preliminary analysis provides valuable insights into the dataset related to Alzheimer's disease. The findings indicate that several variables, including age, gender, education, socioeconomic status, cognitive function (MMSE), and brain volume (nWBV), are associated with the diagnosis of dementia. The clustering analysis revealed two distinct groups, with Cluster 1 representing individuals with a higher likelihood of being diagnosed with dementia, characterized by older age, lower education levels, lower socioeconomic status, poorer cognitive function, and reduced brain volumes. In contrast, Cluster 2 consisted of individuals with relatively better cognitive function and brain measurements. The logistic regression model demonstrated the significant impact of gender, age, MMSE scores, and nWBV on predicting the diagnosis of Alzheimer's disease. The feature selection analysis identified CDR and MMSE as the most important features for predicting the disease. Overall, these findings contribute to our understanding of the potential risk factors and characteristics associated with Alzheimer's disease.

Important figures Extra

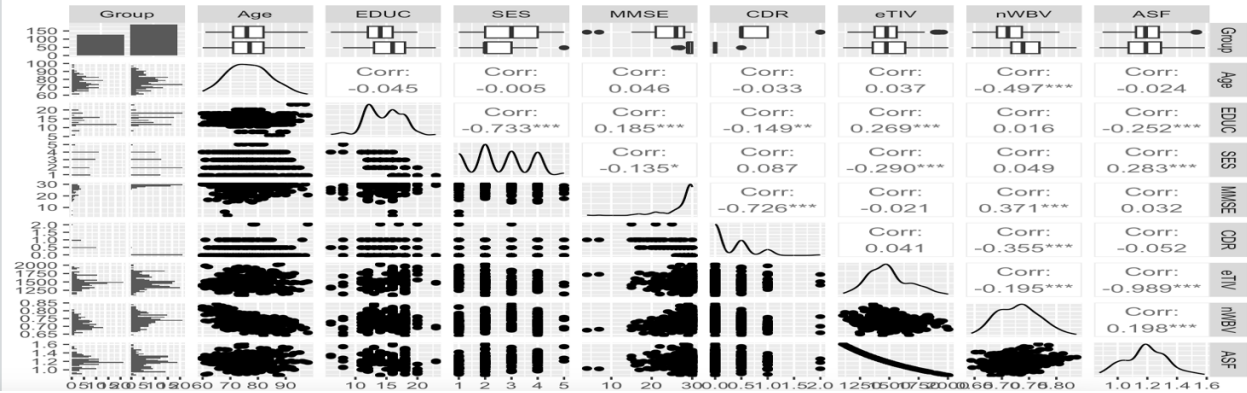

Figure 5: Pair plot to check correlation between features

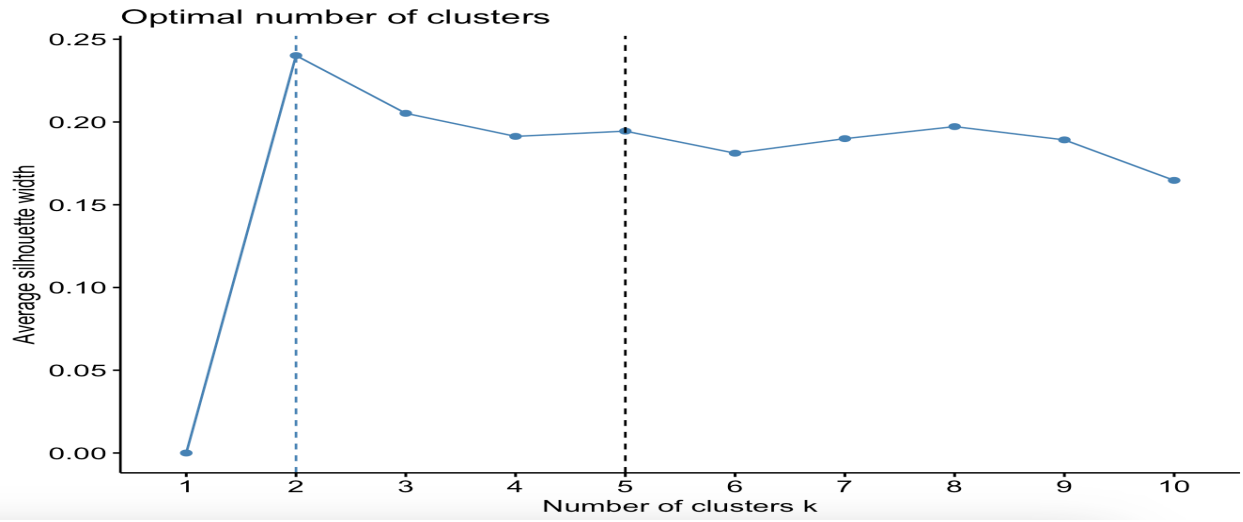

Figure 6: Optimal Number of Clusters Silhouette Method for K-Means Clustering

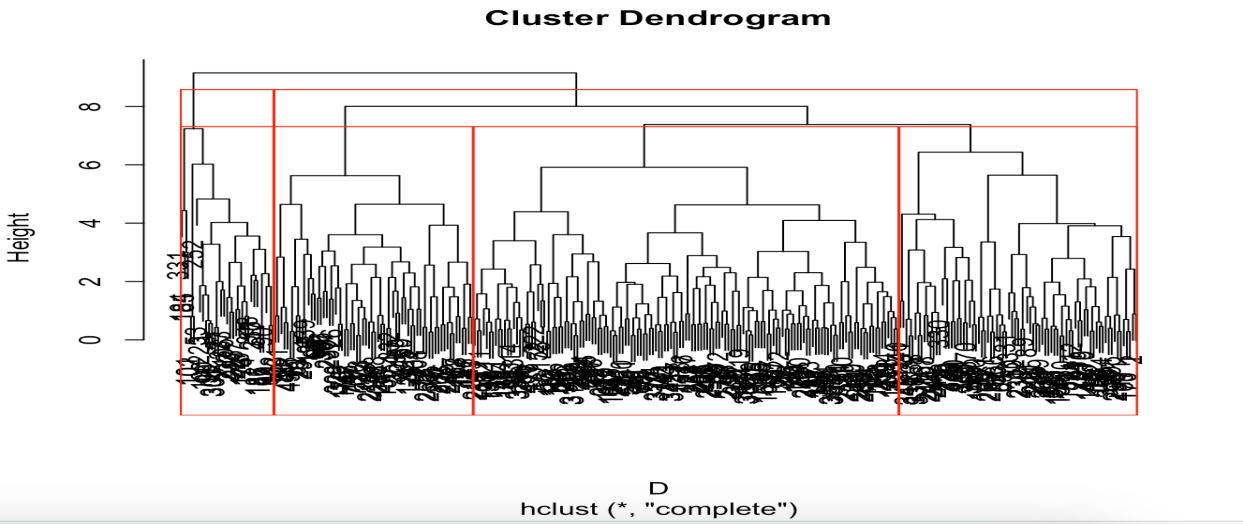

Figure 7: Dendrogram for Hierarchical Clustering k=4