DATA-DRIVEN ANALYSIS OF ALZHEIMERS DISEASE AND PREDICTIVE MODELLING TO AID ITS CLASSIFICATION AND EARLY DETECTION.

By

ADEDOYIN ABIGAIL AGUNBIADE

ABSTRACT

Alzheimer's disease (AD) is a disorder that may affect daily lives due to progressive destruction of the brain. This report investigates the relationship between features of AD and the development of a model for predicting dementia in individuals. The study examines a dataset comprising both demented and non-demented individuals, focusing on factors such as gender, age, years of education (EDUC), socioeconomic status (SES), clinical dementia rating (CDR), estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and atlas scaling factor (ASF). Descriptive statistics, presented in tables and graphics, summarise the specific patterns and correlations among the variables. By employing K-means clustering analysis, which identifies distinct groupings within unlabelled data, the study facilitates a better understanding of varied dementia profiles. Feature selection is implemented to reduce overfitting and improve performance by selecting relevant predictors for AD. The developed prediction model, based on a generalised linear model, achieves remarkably high classification accuracy. While acknowledging the limitations of the limited dataset, this report emphasises the necessity for further research and validation using larger and more diverse datasets. Overall, this research provides valuable insights for AD prediction, contributing to early treatment provision and minimising brain damage.

1.0: INTRODUCTION

Alzheimer's disease (AD) is a form of dementia characterised by cognitive decline and functional impairment. The disease is associated with abnormal protein accumulation in the brain, leading to brain shrinkage and cognitive impairment. It is a major cause of death worldwide, with no definitive diagnosis until post-mortem (1).

Effective treatment and management, therefore, rely on early detection to help minimise brain damage (1). This report aims to use data analysis to explore the relationship between AD-related characteristics and develop a predictive model to assist clinicians in identifying dementia status. The goal is to enable early diagnosis and intervention, ultimately improving outcomes for individuals affected by AD.

2.0: PRE-ANALYSIS

Data pre-processing involves cleaning, organising, and transforming data to prepare it for analysis. To focus on demented and non-demented individuals, the "Converted" group was removed. Additionally, 19 rows with missing SES values and 2 rows with missing MMSE values were excluded. As a result, the dataset was reduced from 373 to 317 observations. The data types were checked, and the variables "Group" and "M/F" (renamed to "Gender") were converted to numeric values (0 for males, 1 for females) and (0 for demented, 1 for non-demented) to facilitate more effective analysis and processing.

3.0: ANALYSIS AND DISCUSSION

3.1: Descriptive Statistics

Table 1 and Figure 1 shows that the data analysis revealed several interesting findings. Firstly, non-demented individuals (190) were just slightly more than demented individuals (127), indicating the class can be considered balanced for effective modelling. Gender distribution showed that the demented group had a higher proportion of males, while the non-demented group had a higher proportion of females, with distinct gender patterns observed. Age variability was significant, but there was no noticeable difference in ages between the two groups, as age is common around 73 years. Education levels differed between the groups, with non-demented individuals having higher levels on average, suggesting a potential association between education and dementia risk. The distribution of educational attainment was negatively skewed for the demented group and positively skewed for the non-demented group.

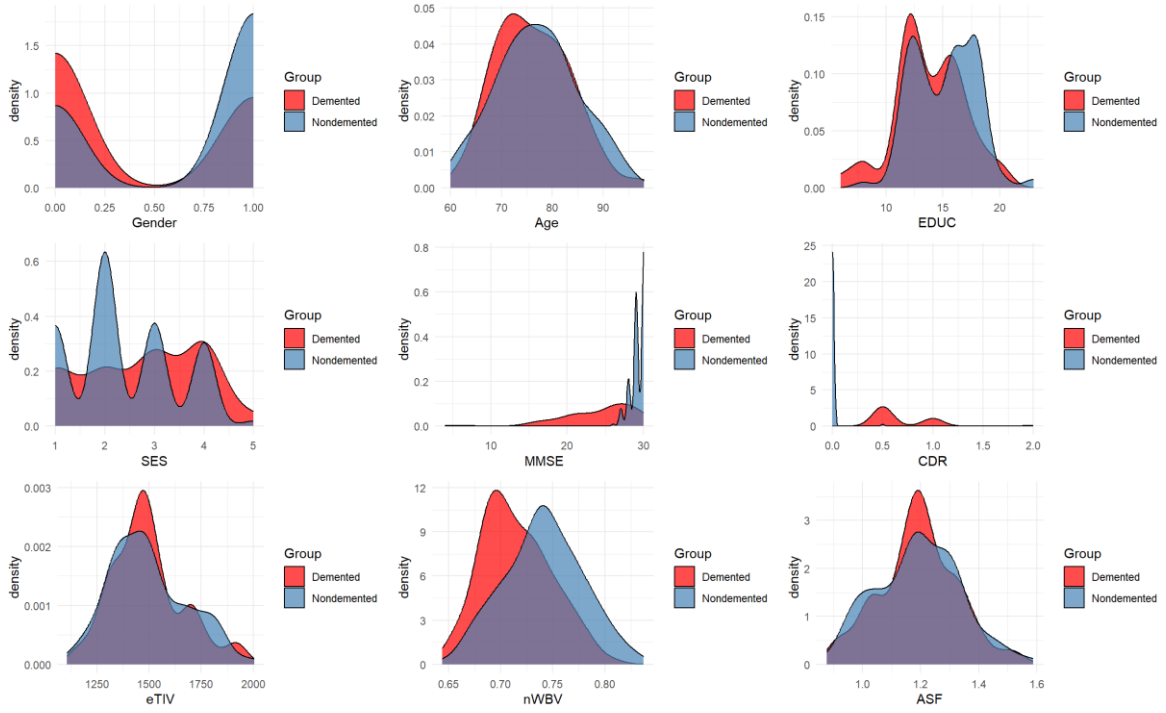
Regarding other variables, SES, nWBV, and ASF did not exhibit significant differences between the groups, except for a slightly higher mode and mean of SES in the demented group (mode = 4, mean = 2.77), indicating that there are more people with relatively higher socioeconomic status among demented individuals.

CDR scores, which measure dementia severity, showed distinct distributions between the groups. The demented group had higher average scores, a positive skewness, and a mode of 0.5 (indicating mild impairment), while the non-demented group had a mode of 0 (no impairment). MMSE scores, measuring cognitive functioning, differed significantly between the two groups. The demented group had lower average scores (24.32), ranging as low as 4, indicating more severe impairment, whereas the non-demented group had a minimum score of 26, mean of 29.23, indicating better cognitive functioning. As both CDR and MMSE differ significantly in both groups, this may be a key association with predicting dementia.

Table 1: Summary statistics of Demented and Non-demented individuals.

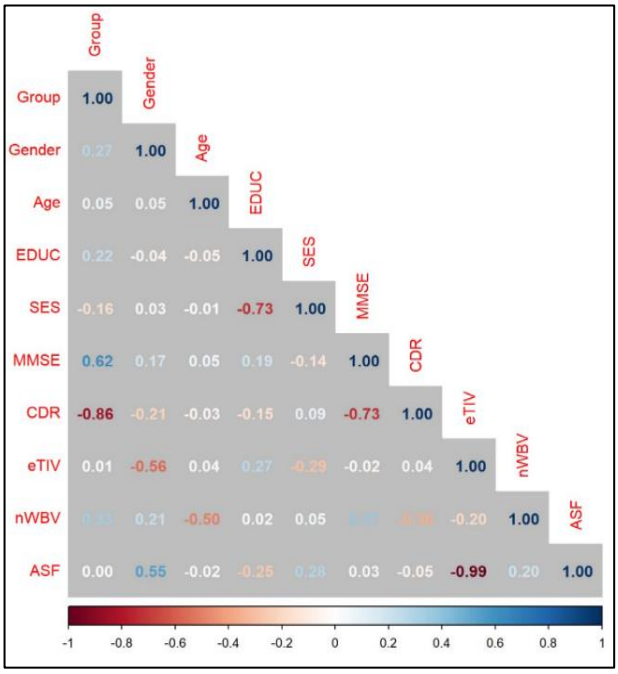
	DEMENTED					**NON-DEMENTED**				
	Min	**Max**	**Median**	**Mode**	**Mean**	**Min**	**Max**	**Median**	**Mode**	**Mean**
Gender	0	1	0	0	0.4	0	1	1	1	0.68
Age	61	98	76	73	76.2	60	97	77	73	77.06
EDUC	6	20	14	12	13.82	8	23	16	18	15.14
SES	1	5	3	4	2.77	1	5	2	2	2.40
MMSE	4	30	26	26	24.32	26	30	29	30	29.23
CDR	0.5	2	0.5	0.5	0.67	0	0.5	0	0	0.01
eTIV	1143	1957	1477	1483	1490.7	1106	2004	1474	1506	1496
nWBV	0.65	0.81	0.711	0.695	0.72	0.64	0.84	0.74	0.739	0.74
ASF	0.9	1.53	1.188	1.134	1.19	0.88	1.59	1.19	1.162	1.19

Figure 1: Density plot showing the distribution of Demented and Non-demented.



The correlation analysis (Figure 2) revealed meaningful associations in the data: education and socioeconomic status were strongly negatively correlated (-0.73), indicating higher education was linked to lower socioeconomic status. Cognitive impairment (CDR) showed a strong negative correlation with cognitive function (MMSE) (-0.73), suggesting lower function with higher impairment. Moderate negative correlations were found between eTIV and gender (-0.56), and between age and nWBV (-0.50), implying gender-related differences and volume decline with increasing age. ASF showed a moderate positive correlation with gender (0.55) and a strong negative correlation with eTIV (-0.99), suggesting gender-related scaling differences and decreased scaling factor with larger intracranial volumes. The response variable, "Group," displayed a strong negative correlation with CDR (-0.73) and a moderate positive correlation with MMSE (0.62), indicating their importance in predicting dementia.

Figure 2: Correlation plot.

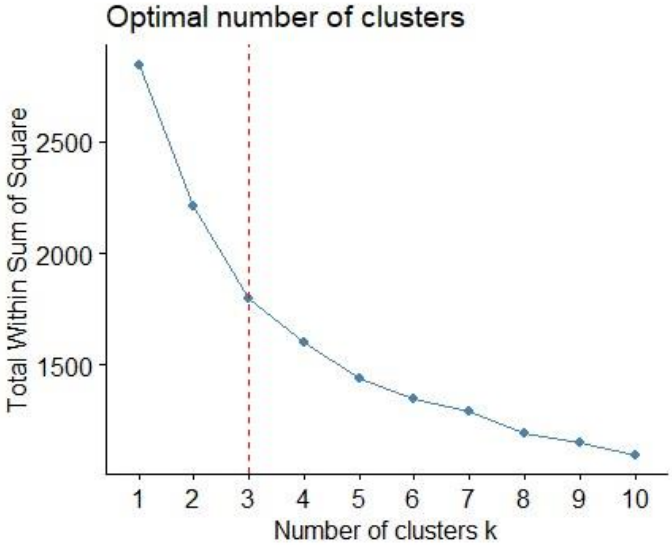


3.2: Clustering

To identify distinct groupings in the dataset based on different attributes, K-means clustering was employed to partition the data into clusters based on their similarities. Initially, the response variable was removed to create an unlabelled dataset. Subsequently, the remaining variables were scaled to normalise the data and bring all features to a similar range.

The optimal number of clusters (k) for the analysis was determined using the Elbow method. This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and looking for the "elbow" point, which indicates the optimal choice. Based on the elbow plot, it was determined that the optimal value was 3 (Figure 3). To increase the chances of finding the best clusters, the clustering function was run 20 times with different random starting points, aiming to minimise WCSS in each run. The cluster results with the lowest WCSS, achieved when k = 3, is shown in Figure 4 (Left).

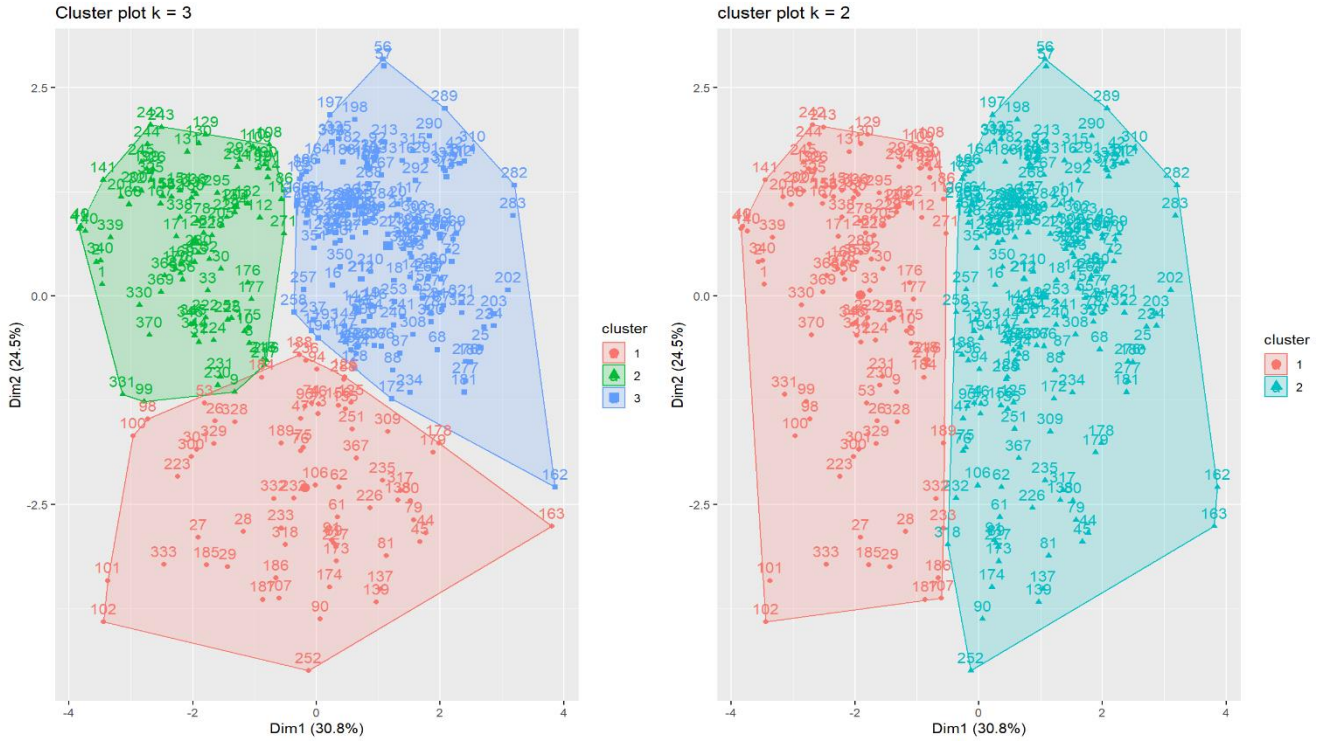
Figure 3: Elbow Method plot for optimal k value



Upon examining the optimal value of k (3), it was observed that there is a slight overlap between the clusters (Figure 4). To explore alternative solutions, the analysis was extended to include k = 2, which revealed

two distinct clusters with no overlap (Figure 4 (right)), giving a clearer boundary grouping. Therefore, this report primarily focuses on the results obtained when k = 2 due to a few reasons: domain knowledge (the aim is to identify demented and non-demented groups), no overlap, less computation cost and parsimony.

Figure 4: K-means clustering using k=3 and k=2



The k-means clustering with k = 2 resulted in two clusters: Cluster 1 (size: 112) and Cluster 2 (size: 205), suggests that there are more data in Cluster 2. As shown in Table 2, Cluster 1 is characterised by males, older adults, higher education, higher cognitive impairment (CDR), larger brain volume (ASF), lower nWBV, socioeconomic status, and lower MMSE scores, suggesting people in this cluster have cognitive decline. On the other hand, Cluster 2 primarily consists of females, younger adults, lower education levels, lower cognitive impairment, brain volume, and higher MMSE, socioeconomic status, and nWBV, indicating better cognitive health and brain structure in this Cluster.

The evaluation of clustering quality considered the within-cluster sum of squares (WCSS) and between-cluster sum of squares (BCSS). Cluster 1 and Cluster 2 had a WCSS of 859.06 and 1349.14 respectively, indicating higher similarity and compactness within Cluster 1. The proportion of BCSS to the total sum of squares was 22.4%, suggesting considerable variations between the two clusters and the possibility of overlap. The silhouette coefficient score, which measures clustering quality, was 0.244, indicating a moderate level of separation between the clusters and room for improvement thus suggesting the exploration of other robust clustering algorithms may enhance the clustering performance for this dataset.

Table 2: K-means Clustering means of two clusters

CLUSTER	Gender	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
1	-0.857	0.216	0.537	-0.478	-0.167	0.209	1.033	-0.485	-1.000
2	0.468	-0.118	-0.294	0.261	0.091	-0.114	-0.564	0.265	0.546

3.3: Feature selection

Feature selection is crucial in modelling to improve performance, reduce overfitting, efficiency, and interpretability. In this report, backward and forward wrapper techniques were used with logistic regression models to identify relevant predictors of AD based on AIC. The backward wrapper, which starts with all features and removes one variable at a time, selected Age, SES, CDR, eTIV, and nWBV as relevant predictors of AD (AIC 12). Conversely, the forward wrapper, starting with an empty set and gradually adding variables, identified CDR, ASF, EDUC, and MMSE as the optimal predictors (AIC 10). Comparing AIC values, the forward wrapper model had a lower value, indicating better fit and classification performance but overall, both models had low AIC, suggesting general good fit.

Null deviance was 426.85 on 316 degrees of freedom (df) for both wrappers, while residual deviance was smaller for backward (1.56e-07 on 311 df) than forward (2.83e-06). These close-to-zero values indicate accurate capturing of relationships between predictors and the response variable. The P-values were however not significant (>0.5) in both techniques, suggesting variables might not be influential predictors individually.

Using the ANOVA function, a likelihood ratio test was performed, and it revealed no significant difference between the models (Pr(>Chi) = 0.999). Consequently, the forward wrapper model, which aligned with the significant variables (CDR, MMSE, EDUC) identified during data exploration, was selected for logistic regression fitting. Additionally, this model had fewer variables compared to the backward wrapper model. Although backward wrapper model consists of age - a variable considered a major contributor to dementia according to many literature reviews, however, showed no significant difference in this dataset. Plus, the backward model used more variables, which increases the chances of multicollinearity, supporting its exclusion.

3.4: Logistic regression

The forward wrapper method selected predictor variables for fitting a logistic regression model, a type of generalised linear model (GLM). This logistic regression model aims to classify individuals as demented or non-demented by estimating their probability of belonging to each class.

The GLM with a binomial family was used to train the model. Using a 10-fold cross-validation approach, the dataset was divided into 10 parts, with 9 parts used for training and 1 part for testing. This process was repeated 10 times to ensure a robust estimate and generalisation to unseen data.

The model's performance was assessed using the "Accuracy" metric, measuring the proportion of correctly classified instances. Additionally, a confusion matrix was generated to evaluate the model's predictions. The results, presented in Table 3, revealed a perfect accuracy of 1 (i.e 100%), indicating no misclassifications. Specifically, the model accurately classified 127 individuals as demented and 190 individuals as non-demented, which matches the result earlier identified in the descriptive statistics.

Table 3: Confusion Matrix

		ACTUAL GROUP	
		Demented	Non-demented
PREDICTED	Demented	127	0
	Non-demented	0	190

With a 95% confidence level, the accuracy was estimated to be between 0.9883 and 1, demonstrating a high level of confidence in the model's performance. Sensitivity, specificity, and other metrics also achieved perfect scores of 1, demonstrating the model's exceptional ability to accurately identify dementia cases. Furthermore, the importance of each predictor variable in the model's classification problem was analysed. The variable "CDR" had the highest impact, with a weight of 100, indicating its crucial contribution to classifying individuals. Followed by Education with a weight of 75.61, while ASF had a lower impact (36.90), and MMSE had no influence (0.00).

It is important to note that these findings are based on a small dataset of 317 observations, thus it may have limited applicability to larger or diverse datasets, and overfitting is a potential concern.

4.0 CONCLUSION

This study explored AD-related features and developed a predictive model for dementia status. Dementia severity measured by CDR was found to be important in predicting dementia status. The clustering algorithm revealed distinct categories but with potential misclassification. Although the GLM model achieved 100% accuracy, none of the predictors was significant according to their P-values, suggesting the presence of other variables not included in this data (e.g genetics or biomarker) may have contributed to the significance of predictors or stronger associations with AD. The small dataset and the need for further study and validation are limitations of this data. Nonetheless, this research provides valuable insights for predicting and treating AD, benefiting individuals with dementia.

5.0 REFERENCES

1. Harper L, Fumagalli GG, Barkhof F, Scheltens P, O’Brien JT, Bouwman F, et al. MRI visual rating scales in the diagnosis of dementia: Evaluation in 184 post-mortem confirmed cases [Internet]. Oxford University Press; 2016 [cited 2023 Jun 19]. Available from: https://academic.oup.com/brain/article/139/4/1211/2464252