Detection of Alzheimer's using Machine Learning Models CPSC 4300/6300: Applied Data Science - Fall 2023

Niteesh
Visualization, EDA and Logistic Regression Model
220 ELM ST
Clemson, SC
+1 864-533-7056
rganuga@clemson.edu

Bhavya Pulagam
Data Cleaning, Introduction, Decision Tree Model
155 Anderson Hwy
Clemson, SC
+1 864-765-4937
vpulaga@clemson.edu

Michelle Sun
EDA, Visualizations, Model Comparison and Conclusion
122 University Village
Central, SC
+1 843-619-5140
msun2@clemson.edu

Sricharan
Summary of Machine Learning Models and SVM
220 ELM ST
Clemson, SC
+1 803-970-5492
srichan@clemson.edu

Akash
Introduction, EDA and XG Boost Model
103 University Village
Central, SC
+1 864-788-2293
akashv@clemson.edu

1 INTRODUCTION

1.1 Problem Statement

The objective is to develop a robust model that can accurately identify early signs of Alzheimer's disease and distinguish between nondemented and demented individuals using longitudinal imaging data. Based on the results from the exploratory data analysis, nWBV (normalized whole brain volume), and CDR (clinical dementia rating) are the two predictors we expect to be most important in predicting whether a patient has dementia or not. nWBV scores on average are lower for individuals with dementia than those without, while CDR scores are higher for individuals with dementia. These two factors nWBV and CDR value patterns in a patient's report would aid an early detection of Alzheimer's that would better equip the treatment. The CDR, or clinical dementia rating, is strongly correlated with whether a patient has dementia, with all non- demented patients having a CDR score of 0 and most demented patients having a CDR score of 0.5.

1.2 Motivation

Alzheimer's affects about 6 million people in America, all ages combined. By 2023, there will be 6.7 million Americans 65 and older who have Alzheimer's. Of them, 73% are 75 years of age or older. Approximately 10.7% of adults 65 years of age and older have Alzheimer's. Individuals suffering from Alzheimer's dementia, their loved ones, and careers bear a heavy burden, as does the health and social care system as well as society at large. There is a pressing need for strategies to stop or postpone the beginning of Alzheimer's disease and the ensuing dementia because of improvements in lifespan seen throughout the world. A diagnosis made early in the course of the illness allows for time for everyone involved to adjust while the patient is still able to actively participate. It also gives access to advice, financial support, and both pharmaceutical and non-pharmacological treatments, even though no disease-modifying agents that can reverse the initial pathological changes associated with the disease have successfully entered the market. Our project will help detect the 'Demented' population on their earlier stages of Alzheimer's, through their progressing scans considering factors like nWBV and CDR.

1.3 Explanation of Dataset

Our experiment uses a CSV file from the Oasis Longitudinal Demographics, which has 374 rows and 15 columns of scan-related data on 150 participants that were recorded at least a year apart. This dataset has 150 longitudinal participants, ranging in age from 60 to 96. For a total of 373 imaging sessions, each individual was scanned on two or more occasions, separated by at least a year. Three or four separate T1-weighted MRI scans performed during a single scan session are presented for each patient. There are men and women among the subjects, and they are all right-handed. Throughout the investigation, 72 of the individuals were classified as nondemented. Of the people scanned, 51 had mild to moderate Alzheimer's disease, and 64 were classified as demented at the time of their first visit and remained so for successive scans. After being classified as nondemented at their first visit, an additional 14 participants were later classified as demented at a later visit.

Table1. Column Descriptors

Column name	Full form
EDUC	Years Of Education
SES	Socio Economic Status
MMSE	Mini mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

2 Summary of EDA

2.1 Unit of Analysis

The unit of analysis in this dataset is an individual subject. Each subject, characterized by age, gender, and dementia status, contributes multiple data points corresponding to different imaging sessions and the number of T1-weighted MRI scans obtained during those sessions. The dataset is longitudinal, capturing changes over time for each subject.

2.2 Total Number of Observations

Our experiment uses a CSV file from the Oasis Longitudinal Demographics, which has 374 rows and 15 columns of scan-related data on 150 participants that were recorded at least a year apart.

2.3 Unique Observations

There are no duplicate observations throughout the dataset, making all the 374 rows unique. However, there are observations of the same participants, but over different time frames.

2.4 Time Period

The time period differs here for every observation and each participant. While every participant has at least one observation present, not all the participants have data collected over many visits.

2.5 Data Cleaning

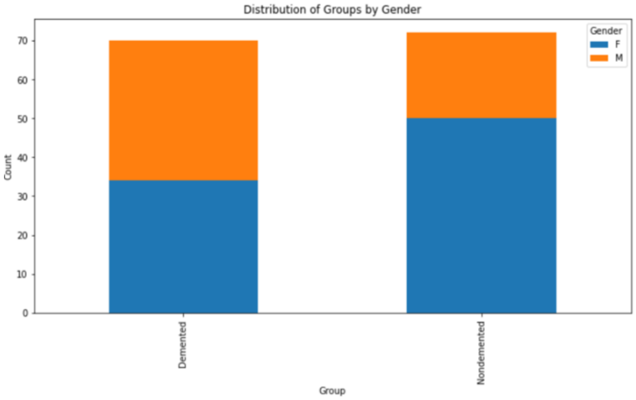
First, we Imported the csv file as a dataframe using pandas.

2.5.1 Checking for Duplicate values

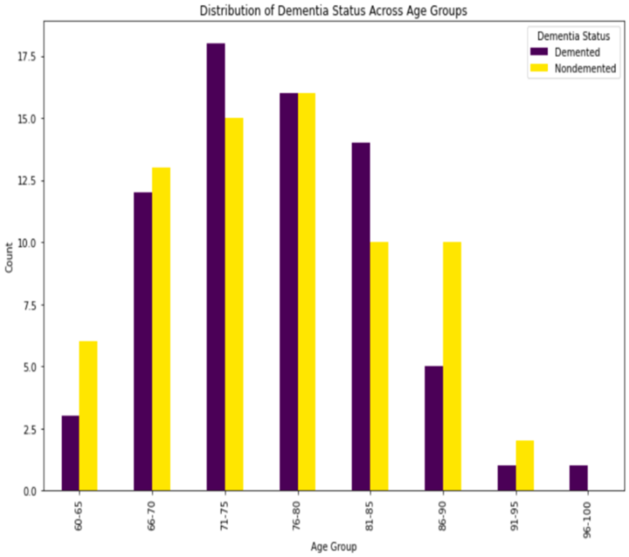
We have checked for null and duplicate values in the dataset and remove the corresponding values to ensure uniqueness.

Observation: No null values or duplicate values are present in this dataset

2.6 Visualization of the response



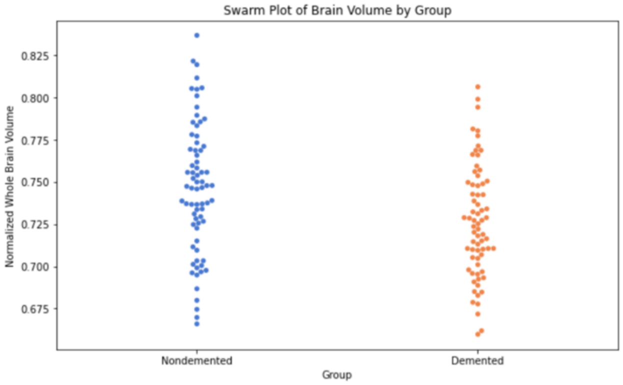
- Men are more likely to have dementia, a complication of Alzheimer's Disease, than women. In the following stacked bar graph, we can observe that the distribution of men is more in the ‘Demented’ group and vice-versa for the ‘Nondemented’ group.



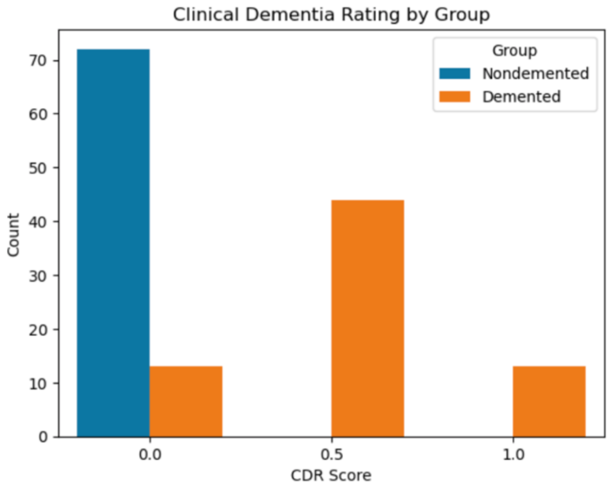
- Higher concentration of 71-85 years old in demented group than those in the nondemented patients. In the following grouped bar plot, we can observe that the count of the Demented group in the groups of 71-75, 76-80 & 81-85 is significantly high when compared to that of the other age groups.

Our experiment uses a CSV file from the Oasis Longitudinal Demographics, which has 374 rows and 15 columns of scan-related data on 150 participants that were recorded at least a year apart. We considered only the visit 1 for each participant for the model, so we filtered that data, and we replaced the group name ‘Converted’ to ‘Demented’ as the person’s status has been converted from the state ‘non-demented’. This would help us to convert the nature of the classification from multi-label to dual labels.

2.7 Visualization of key predictors against the response



- The above smarm plot shows how the Nondemented group has higher brain volume than Demented group.



- The CDR, or clinical dementia rating, is strongly correlated with whether a patient has dementia, with all non-demented patients having a CDR score of 0 and most demented patients having a CDR score of 0.5.

nWBV, normalized whole brain volume, and CDR, clinical dementia rating, are the two predictors we expect to be most important in predicting whether a patient has dementia or not. nWBV scores on average are lower for individuals with dementia than those without, while CDR scores are higher for individuals with dementia.

3 Summary of Machine learning Models

Based on the insights we gained from Exploratory Data Analysis (EDA), we've decided to utilize all four machine learning models - Logistic Regression, Decision Tree, XGBoost, and Support Vector Machine (SVM) - for our project. EDA highlighted the dataset's cognitive and demographic characteristics and the binary response variable categorizing individuals as either "Demented" or "Nondemented". This suggests that these models are well-suited for our binary classification task. This selection allows us to comprehensively explore how these models can predict outcomes based on the data.

3.1 Logistic Regression Model

Logistic regression modeling was selected since the response variable ("Group") in the dataset is binary, with individuals being categorized as either "Demented" or "Nondemented." The goal of estimating the likelihood of dementia based on cognitive and demographic characteristics is in line with the applicability of logistic regression for binary classification problems. The benefits of logistic regression include its ease of use and interpretability, which make it easier to examine how different variables affect a binary outcome.

3.1.1 Evaluation of Logistic Regression

The reported accuracy is 0.8056, and the confusion matrix provides additional details on the model's performance on the test data. To calculate the test error rate, we can use the formula:

Error Rate = Number of Misclassifications\Total Observations

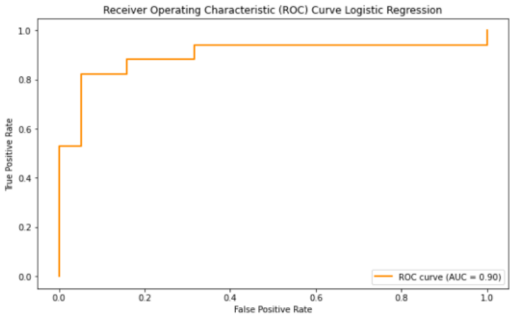
In this case, the confusion matrix shows that there are 7 misclassifications (6 false positives and 1 false negatives), and the total number of observations is 36. Therefore, the test error rate can be calculated as: So, the test error rate is approximately (7/36) 19.44%.

Justification for using the error rate: The error rate gives us a direct indicator of the percentage of misclassifications and is a simple, clearly comprehensible number. When attempting to evaluate the model's accuracy in predicting instances that belong to each class, binary classification tasks are especially helpful. Since the costs of false positives and false negatives vary, accuracy offers a broad indication of how accurate predictions were made; however, error rate provides a more targeted view of misclassifications, which is useful in assessing the model's performance.

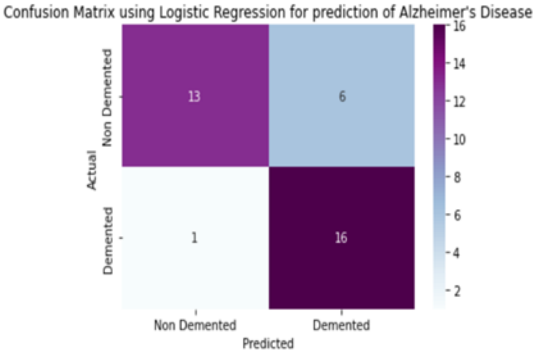
How well the model fits: The logistic regression model exhibits a reasonably good fit to the data, as reflected in an estimated test error rate of approximately 19.44%. This suggests that, on average, the model makes correct predictions about 80.56% of the instances in the test set. The balanced precision and recall values for both "Demented" and "Nondemented" classes, as evident from the confusion matrix and classification report, indicate that the model effectively identifies individuals with and without dementia. However, the assessment should be contextualized based on the specific goals and cost considerations of the analysis. The model's performance, with a balanced F1-score and accuracy, suggests a generally satisfactory fit, though further evaluation may be warranted depending on the relative costs associated with false positives and false negatives in the given classification task.

- In the evaluation of our model's performance, precision and recall metrics played a crucial role. For the "Non-Demented" class, the model exhibited a precision of 0.93, indicating that 93% of its positive predictions were accurate, while the "Demented" class had a precision of 0.73, signifying a 73% accuracy in predicting "Demented" cases. In terms of recall, the model demonstrated a 68% ability to correctly identify "Non-Demented" cases and an impressive 94% recall rate for "Demented" cases. The F1-score, which balances precision and recall, yielded values of 0.79 for "Non-Demented" and 0.82 for "Demented," suggesting a relatively balanced performance for both classes. With 80.56% overall accuracy, the model correctly predicted the classes across all instances. The macro and weighted averages for precision, recall, and F1-score ranged from 0.80 to 0.83, indicating a balanced performance that accounts for the varying support of each class in the dataset.

- In the ROC curve the area under the curve (AUC) is a measure of the classifier's performance. Here, the AUC is 0.90, which is quite high, suggesting that the classifier has a good measure of separability, and is able to distinguish between the positive class and the negative class effectively.



- The confusion matrix shows that the model has a high number of true positive and true negative, which suggests good predictive power, especially for correctly identifying 'Demented' individuals. So, we can conclude that model fits the data well.



3.2 Decision Tree Model

The adoption of a Decision Tree model is grounded in the binary nature of the response variable ("Group") within the dataset, classifying individuals as either "Demented" or "Nondemented." Decision Trees are particularly well-suited for binary classification tasks, seamlessly aligning with the objective of predicting the likelihood of dementia based on demographic and cognitive features. Decision Trees excel in capturing non-linear patterns and complex relationships within the data, aligning with potential intricate associations uncovered during EDA.

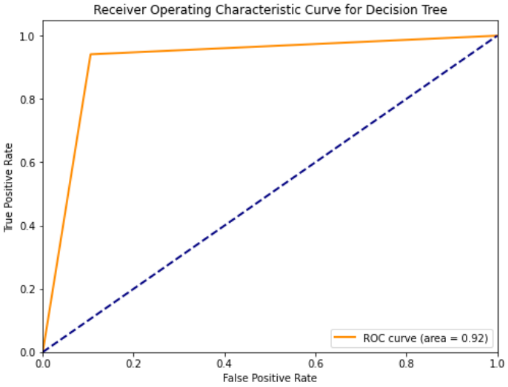
3.2.1 Evaluation of Decision Tree

The reported accuracy for the Decision Tree model is 0.9167, indicating that it correctly predicts the target variable in approximately 91.67% of cases. The model’s test error rate for the Decision Tree model is 8.33% The confusion matrix provides additional details on the model's performance on the test data Confusion Matrix: 17 & 2 \\ 1 & 16 This confusion matrix reveals that there are 3 misclassifications (2 false positives and 1 false negatives), and the total number of observations is 36. Therefore, the test error rate can be calculated as:

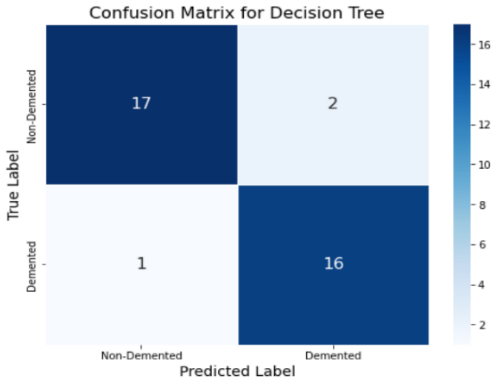
Error Rate = Number of Misclassifications\Total Observations
So, the test error rate for the Decision Tree model is approximately 8.33%.

How well the model fits: The Decision Tree model demonstrates a reasonable fit to the data, as evidenced by the estimated test error rate of approximately 8.33%. This suggests that, on average, the model makes correct predictions about 91.67% of the instances in the test set. The balanced precision and recall values for both "Non-Demented" and "Demented" classes indicate that the model effectively identifies individuals with and without dementia. The balanced F1-scores and accuracy further support the overall satisfactory fit of the model. However, it is essential to contextualize this assessment based on the specific goals and cost considerations of the analysis. Considering the nature of the classification task, where the consequences of false positives and false negatives may have different implications, further evaluation may be warranted. Understanding the relative importance of correctly identifying individuals with and without dementia is crucial for determining the model's practical utility. Additionally, comparing the test error rate with that of alternative models could provide valuable insights into the model's relative performance and guide decision-making in choosing the most suitable approach for the given task.

- The Decision Tree model exhibits a well-balanced performance in terms of precision, recall, and F1-score for both 'Non-Demented' and 'Demented' classes. The high overall accuracy of 92% signifies the model's robust ability to correctly classify instances across both classes. The macro and weighted averages further confirm a balanced consideration of the model's performance, taking into account the distribution of instances in each class.
- The Decision Tree model, as represented by the ROC curve, demonstrates a strong diagnostic ability with a high AUC of 0.92. The curve's favorable position indicates the model's accuracy and effectiveness in predictive tasks, emphasizing its capability to distinguish between the target classes.



- The configuration of the confusion matrix, with a notable count of true positives and true negatives, suggests that the Decision Tree model fits the data well. The model showcases a proficient ability to accurately classify individuals, particularly in identifying cases with dementia ('Demented'). The Decision Tree model's confusion matrix underscores its strong predictive performance, successfully capturing true positives and true negatives. This, in turn, supports the conclusion that the model fits the data well, particularly in effectively identifying cases with dementia.



3.3 Support Vector Machine (SVM) Model

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. Its purpose is to find the optimal hyperplane that maximally separates different classes in the feature space, allowing it to effectively classify new data points and handle non-linear relationships using kernel functions. The EDA revealed certain intricacies in the relationships between demographic and cognitive features, suggesting potential non-linearities that could impact the predictive performance. SVM's ability to find optimal decision boundaries in high-dimensional spaces aligns well with the dataset's characteristics.

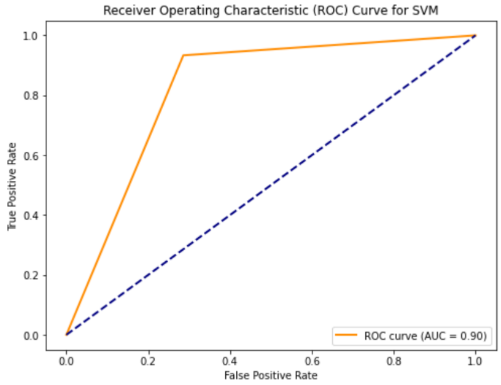
3.3.1 Evaluation of SVM

The SVM model's test error rate is evaluated using one of the techniques discussed in the lecture. The reported accuracy for the SVM model is 0.8276, indicating that it correctly predicts the target variable in approximately 82.76% of cases. To gain further insights into the model's performance, the confusion matrix is examined: Confusion Matrix: 10 & 4 \\ 1 & 14.

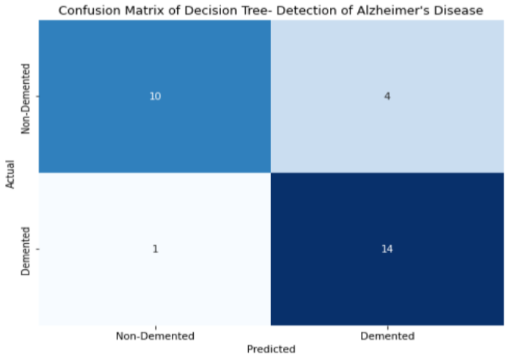
This confusion matrix reveals that there are 5 misclassifications (4 false positives and 1 false negatives), and the total number of observations is 29. For this SVM model, the test error rate is approximately 5/29, resulting in a test error rate of approximately 17.24%.

How well the model fits: The SVM model demonstrates an estimated test error rate of approximately 17.24%, indicating that, on average, the model correctly predicts the class in about 82.76% of instances on the test set. This analysis is grounded in the examination of the confusion matrix, which provides a detailed breakdown of the model's performance. The test error rate, coupled with the insights from the confusion matrix, suggests that the SVM model's fit to the data is moderate. While the model achieves an accuracy of 82.76%, indicating a reasonable level of correct predictions, the test error rate highlights areas of misclassification. The observed false positives and false negatives indicate instances where the model struggled to accurately predict the classes. It is crucial to contextualize this evaluation within the specific goals and cost considerations of the analysis. The SVM model demonstrates a moderate fit to the data, as indicated by the estimated test error rate. The analysis of misclassifications in the confusion matrix provides a nuanced understanding of the model's performance, allowing for a more informed assessment of its strengths and limitations in predicting dementia status.

- The SVM model demonstrates a commendable accuracy of 82.76% and exhibits balanced precision, recall, and F1-score for both 'Non-Demented' and 'Demented' classes. The evaluation metrics reflect the model's ability to correctly classify instances across both classes, providing valuable insights into its strengths and limitations in predicting dementia status. The weighted and macro averages further underscore the fair consideration of the model's performance, acknowledging the distribution of instances in each class. Overall, the SVM model presents a well-rounded performance in discerning dementia status in the given dataset.
- The SVM model's ROC curve, with an AUC of 0.90, suggests a slightly higher discriminatory ability compared to the Decision Tree model (AUC = 0.9). This indicates that the SVM model performs exceptionally well in separating the two classes, supporting its utility as a strong predictive model. The ROC curve for the SVM model underscores its strong diagnostic ability, with a high AUC of 0.90, affirming its effectiveness in distinguishing between individuals with and without dementia. Although the decision tree model had a higher overall test accuracy rate, the SVM model had a higher test recall rate. Test recall measures the model's ability to identify all incidences of a group, in this case, all 'Demented' individuals. Test recall is calculated by # of true positives / # of true positives + # of false negatives. For this problem, the consequence of a false negative is likely more severe than the consequence of a false positive. In the case of a false negative, a patient with dementia will have delayed or absent treatment and care, which can lead to deteriorating quality of life and increased risk of injuries.



- True Positives (TP) reveals a count of 14, indicating instances where the SVM model correctly predicted 'non-demented' cases. False Negatives (FN) with a number 1, representing cases incorrectly predicted as 'non-demented' when they were 'Demented'. False Positives (FP) shows the number 4, signifying cases incorrectly predicted as 'Demented' when they were actually 'non-demented'. True Negatives (TN) illustrates a count of 10, denoting cases correctly predicted as 'Demented'. In conclusion, the configuration of the confusion matrix, with a substantial count of true positives and true negatives, suggests that the SVM model fits the data well. The model demonstrates proficiency in accurately classifying individuals, particularly in identifying cases with dementia ('Demented').



3.4 XGBoost Model

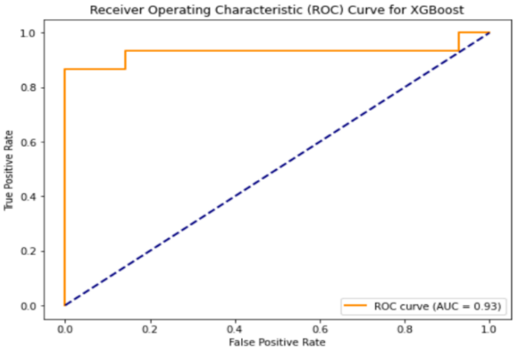
The response variable in the dataset is a binary classification task ('Demented' or 'Non- Demented'). XGBoost is a powerful ensemble learning method that has demonstrated exceptional performance in various machine learning competitions and real-world applications. Its ability to handle non-linearity, capture complex relationships, and effectively deal with imbalanced datasets makes it a compelling choice.

3.4.1 Evaluation of XGBoost

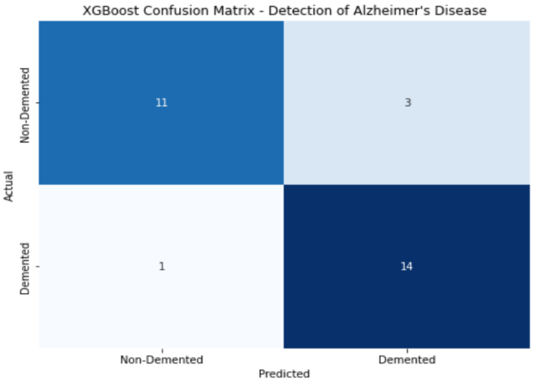
The XGBoost model's test error rate is evaluated using one of the techniques discussed in the lecture. The reported accuracy for this model is 0.8621, indicating that it correctly predicts the target variable in approximately 86.21% of cases. To gain further insights into the model's performance, the confusion matrix is examined: Confusion Matrix: 11 & 3 \\ 1 & 14 This confusion matrix reveals that there are 4 misclassifications (3 false positives and 1 false negatives), and the total number of observations is 29. False positives occur when the model predicts a positive outcome that is actually negative, while false negatives occur when the model predicts a negative outcome that is actually positive. For this model, the test error rate is approximately 4/29, resulting in a test error rate of approximately 13.79%.

How well the model fits: Based on the estimated test error rate of approximately 13.79%, the XGBoost model demonstrates a relatively high level of accuracy, correctly predicting around 86.21% of instances in the test set. A lower test error rate is indicative of a better fit to the data, suggesting that the model performs well in making accurate predictions. However, the assessment of model fit should consider the specific context of the application and the consequences of prediction errors. Overall, the observed test error rate suggests that the XGBoost model is effective in capturing patterns within the data and making reliable predictions in the test set.

- In the context of binary classification, the model achieved a precision of 92% for 'Non-Demented' (Class 0) and 82% for 'Demented' (Class 1), indicating its ability to make accurate predictions within each class. Furthermore, the model exhibited strong recall rates, with 79% for 'Non-Demented' and an impressive 93% for 'Demented,' suggesting its capability to correctly identify instances of both classes. The balanced F1-Scores of 85% for 'Non-Demented' and 87% for 'Demented' reflect the model's overall effectiveness in considering both precision and recall. With support values of 14 instances for 'Non-Demented' and 15 instances for 'Demented,' the model's performance remains consistent across different class frequencies. The macro and weighted averages for precision, recall, and F1-Score, hovering around 87%, emphasize the model's balanced performance across both classes.
- The Decision Tree model, as represented by the ROC curve, demonstrates a strong diagnostic ability with a high AUC of 0.93. The curve's favorable position indicates the model's accuracy and effectiveness in predictive tasks, emphasizing its capability to distinguish between the target classes.



- The confusion matrix for the XGBoost model is examined to gain insights into its predictive performance. Here's a breakdown of the four segments: True Positives (TP) reveals a count of 11, indicating instances where the SVM model correctly predicted 'non-demented' cases. False Negatives (FN) with a number 1, representing cases incorrectly predicted as 'non-demented' when they were 'Demented'. False Positives (FP) shows the number 3, signifying cases incorrectly predicted as 'Demented' when they were actually 'non- demented'. True Negatives (TN) illustrates a count of 14, denoting cases correctly predicted as 'Demented'. The confusion matrix reveals a mix of true positive and true negative predictions, indicating the SVM model's ability to correctly identify both 'Non- Demented' and 'Demented' cases.



3.5 Model Comparison

The summary of model accuracy is as follows:

Model	Test error rate(%)	Accuracy (%)	ROC
Logistic Regression	19.44	80.56	0.90
Decision tree	8.33	91.67	0.92
SVM	17.25	82.75	0.90
XGBoost	13.79	86.21	0.93

In evaluating the performance of different machine learning models on our Oasis longitudinal dataset for detection of Alzheimer’s, we observed varying levels of accuracy, test error rates, and area under the ROC curve. Considering the nature of the dataset and the complexity of the problem, the decision tree is a compelling choice.

The decision tree model achieved an impressive accuracy of 91.67%, higher than the other three models, which demonstrates its ability to capture patterns and relationships within the dataset. This model outperformed both logistic regression and SVM in terms of accuracy, providing a promising approach for predicting dementia status based on longitudinal MRI scans. While XGBoost demonstrated a competitive accuracy of 86.21% and a slightly higher ROC score of 0.93, it is important to consider the interpretability and simplicity of the decision tree model. Decision trees offer a transparent representation of decision-making processes, crucial for understanding the features contributing to dementia classification in this context. Decision tree's ability to capture complex relationships within the data, coupled with its intuitive interpretability, makes it a suitable choice for this task.

4 Summary and Conclusion

4.1 Result analysis on par with Motivation

The results of our model are in line with the motivation for our project, which focuses on early detection of Alzheimer's disease and its impact on individuals, their families, and the broader healthcare system. With an accuracy of 91.67% detection of the model, we can help so many patients early diagnose the Alzheimer’s and they could seek medical care before the situation gets harder to handle. The question that motivated this project was, how can we use machine learning models to distinguish between demented and non-demented individuals, and accurate determine early signs of Alzheimer’s disease? Our results show that there are a variety of algorithmic methods that can be applied to this classification problem, including linear regression, decision tree, SVM model, and XGBoost. Out of the four models discussed in the previous section, the decision tree model presents a robust solution for predicting dementia status in a longitudinal setting, leveraging the intricate patterns embedded in multiple MRI scans over time. Further refinements and optimizations can be explored, but the decision tree's balance between accuracy and interpretability positions it as a valuable tool for understanding and predicting dementia progression in our dataset.

4.2 Key Findings

The results of our data analysis can be valuable to domain experts, scientists who study Alzheimer's disease, by demonstrating the usage of machine learning models as a valuable tool for data analysis. Although no single variable can definitively determine whether an individual is demented or nondemented, the combination of various weighted factors can provide robust predictions to the classification of individuals as demented and nondemented, with a tested accuracy as high as 91.67%, which has potential for even higher accuracy through refinement and adjustment of parameters. Through exploratory data analysis, we find that nWBV and MMSE are inversely correlated with a patient being demented, whereas eTIV and CDR are positively correlated with a patient being demented.

4.3 Alternatives and Future Models

For the purposes of this project, which is to make classification predictions, only data from individuals’ first visits were used for model training, with data from additional visits excluded from data analysis during data cleaning. However, the original dataset contains more than one observation for certain individuals. Thus, it is possible to observe change in variables for a single individual; by adding a temporal dimension to our data analysis, we could compare ‘demented' and 'nondemented' also known as ’converted' states within the same individual. This can be achieved by implementing deep learning algorithms such as LSTM, CNN which excel in handling time series data. This could provide insight on how dementia status change over time, and which variables are associated with converting a ’demented' to a 'nondemented' status.