HDSC Summer’22 Premiere Project

Presentation: Classification & Prediction of Dementia

A project by team Source-code

Introduction

When thinking, memory, and reasoning skills are lost to the point where they interfere with day-to-day tasks, this condition is known as dementia. Some dementia patients have emotional instability and personality changes. The intensity of dementia varies from the mildest stage, when it is just starting to interfere with a person's ability to function, to the most severe level, when the individual must fully rely on others for fundamental daily activities.

Various disorders and factors contribute to the development of dementia. Neurodegenerative disorders result in a progressive and irreversible loss of neurons and brain functioning. Currently, there are no cures for these diseases.

The five most common forms of dementia are:

- Alzheimer’s disease, It is caused by changes in the brain, including abnormal buildups of proteins, known as amyloid plaques and tau tangles.
- Frontotemporal dementia, It is associated with abnormal amounts or forms of the proteins tau and TDP-43.

- Lewy body dementia, a form of dementia caused by abnormal deposits of the protein alpha-synuclein, called Lewy bodies.
- Vascular dementia, a form of dementia caused by conditions that damage blood vessels in the brain or interrupt the flow of blood and oxygen to the brain.
- Mixed dementia, a combination of two or more types of dementia.

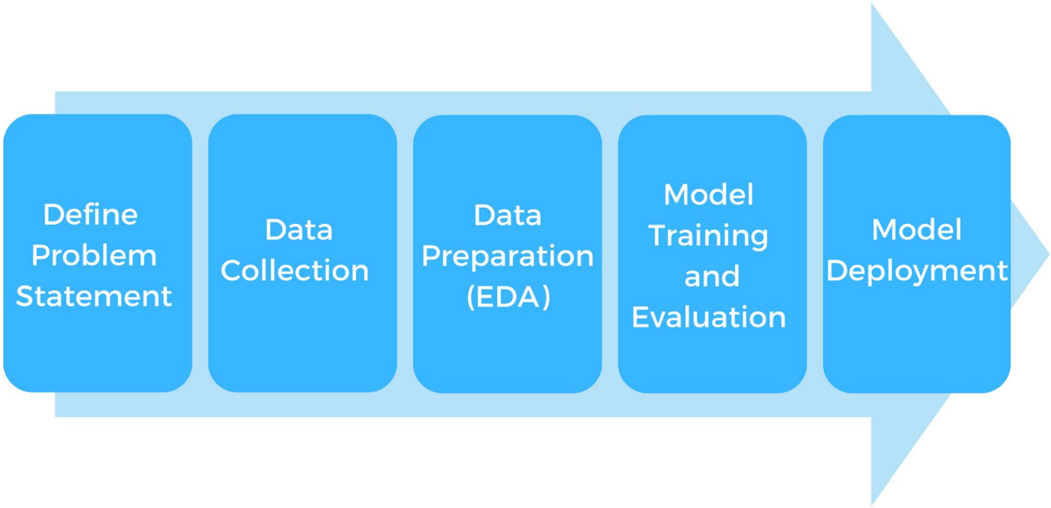
Aim and Objectives

The aim of this project is to build a machine laerning model to predict chances that a person will have dementia and the probable type they can have in order to bridge the barriers to diagnosis and care.

The specific objectives includes:

- Data gathering from a public repository (kaggle)
- Data preparation and transformation
- Training a machine learning model on the prepared and transformed data
- Evaluation of the model performance
- Deployment of the model for use by the public

Methodology (Flow Process)



Problem Statement

As the seventh leading cause of mortality and one of the main causes of disability and dependency among older people worldwide, dementia is frequently unrecognized and misunderstood, which leads to stigmatization and barriers to diagnosis and care.

Data Collection and Description

The data was collected from a public repository Kaggle and the dataset description is as follows:

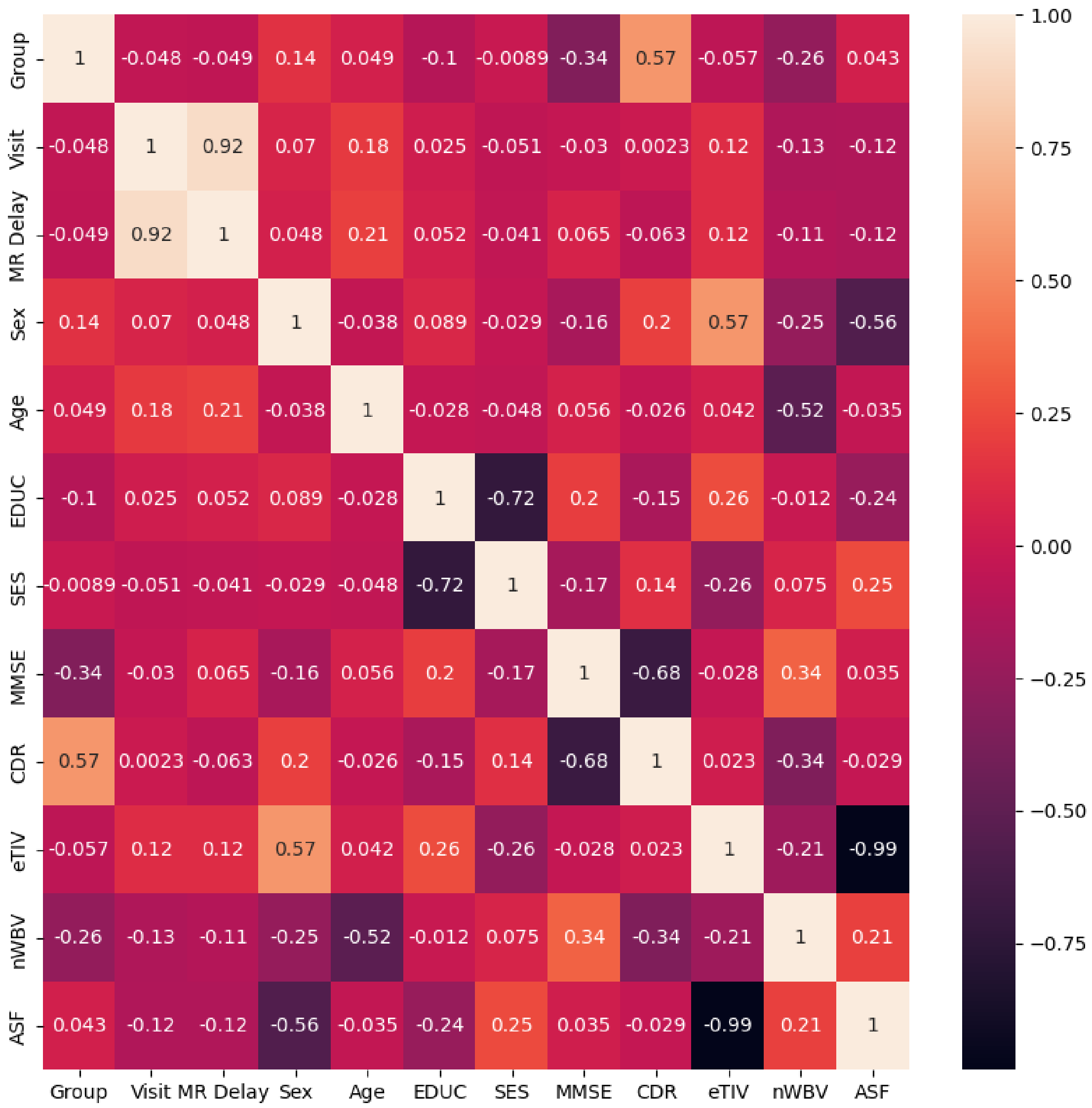
- This dataset consists of a longitudinal collection of 150 subjects aged 60 to 96.
- Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included.
- The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study.
- 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer’s disease.
- Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit

Data Preparation

The collected data had 14 columns and 373 rows of data entries. The following task were carried out during data preparation:

- We dropped the default columns, which are the `Subject ID, MRI ID, Hand`, converted the 'M/F' column name to 'Sex' and changed the categorical columns to numeric columns such that; ('Nondemented'= 0,'Demented' = 1, 'Converted' = 2) - with respect to the 'Group' column and ('M' = 1, 'F' = 0) - with respect to the 'Sex' column.

- We imputed the missing values in 'MMSE' with the median of the demented group (since all missing values in MMSE were from 'Demented' category of the 'Group' column) and imputed the missing values in 'SES' with the mode of the 'Demented' group.



- From the heat map above it was interpreted that,

  - The dementia classification group has the largest positive association with 'CDR,' as well as a moderately significant positive correlation with the subject's age.

  - The dementia classification group exhibits the highest negative correlation with 'MMSE' score of the subject and also exhibits relatively high negative correlation with the 'Sex', 'education level', and the 'nWBV' of the subject.

  - The feature ASF indicates to exhibit high correlation with the eTIV .Thus, 'ASF' is eliminated such that the model is not affected by multi-collinearity.

Hence, the following features were considered for the prediction model:

```
- Sex.

- Age.

- Education level.

- Mini-mental state examination score.

- Clinical dementia rating.

- Normalized whole brain volume.

- Estimated total intercranial volume.
```

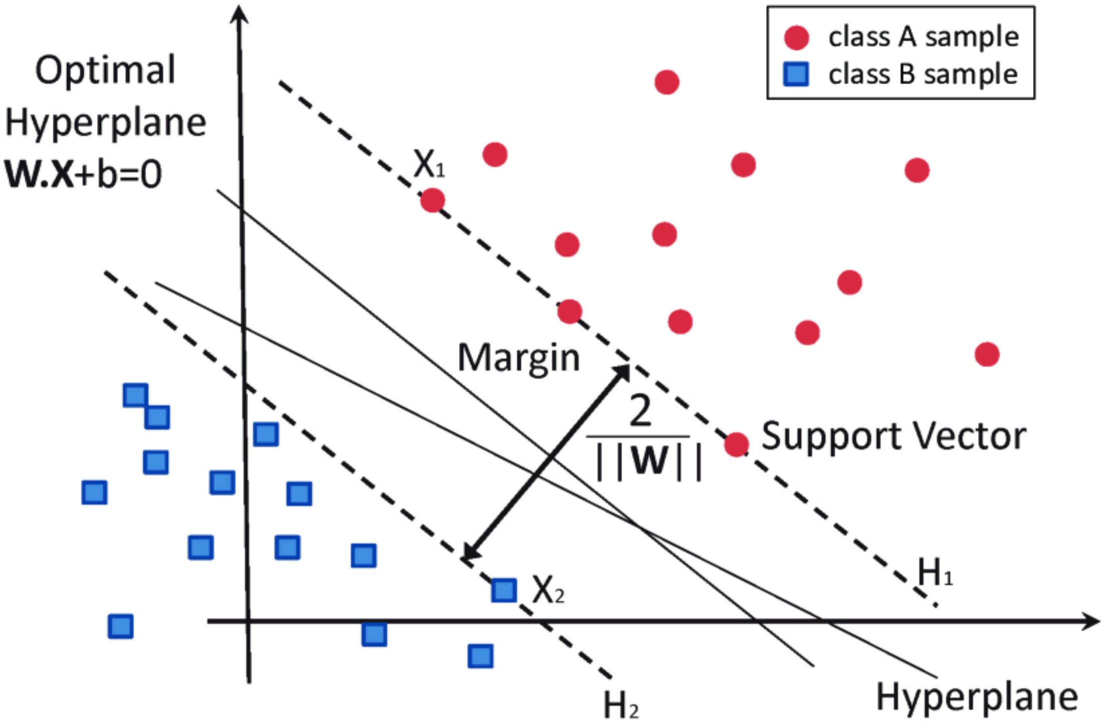
- We removed outliers present in the following columns: 'EDUC', 'MMSE' and 'eTIV'
- We normalized the dataset using 'StandardScaler' package from 'Scikit-learn' and performed a train-test split of 80% and 20% respectively.

Model Training and Evaluation

The following models was trained and validated using scikit-learn accuracy report to select the best performing model:

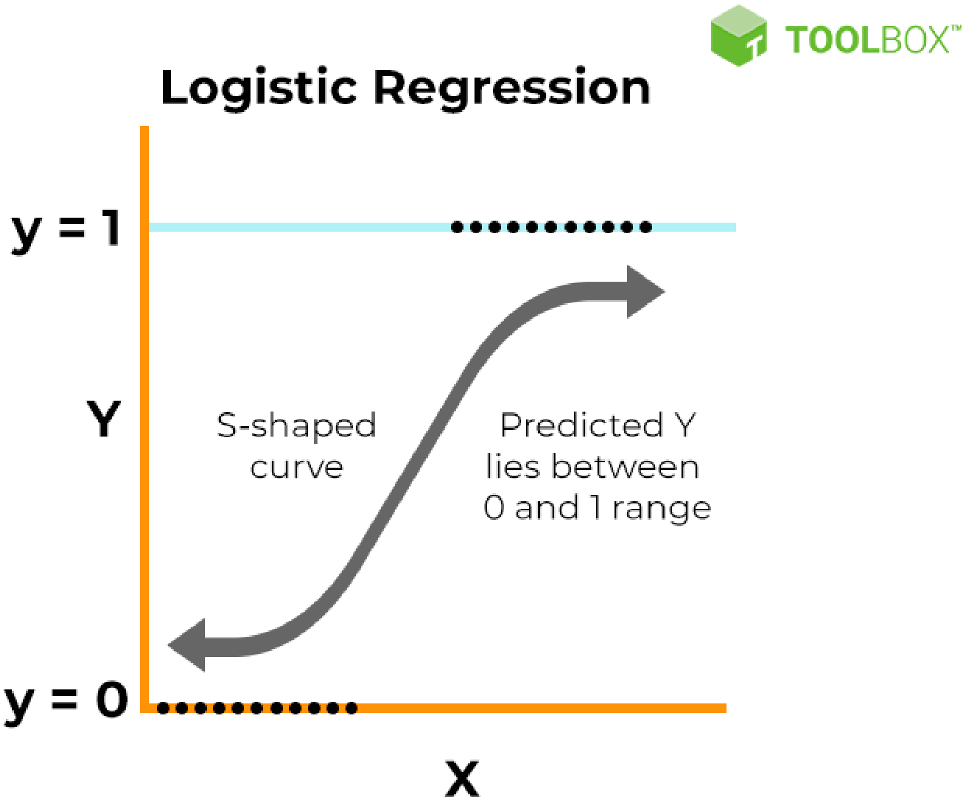
- **Support Vector Machine**

  Support Vector Machine (SVM) is a supervised machine learning algorithm, that classifies the data using largemargin classification technique. It is a vector space-based machine learning method, where, the decision boundary between two classes having the maximum distance from any point in thetraining data, is used to classify the testing data. The SVM classifies the training data to generate the Hyperplane (decision boundaries that classifies the data points), by maximizing the distance between the data and the hyperplane.



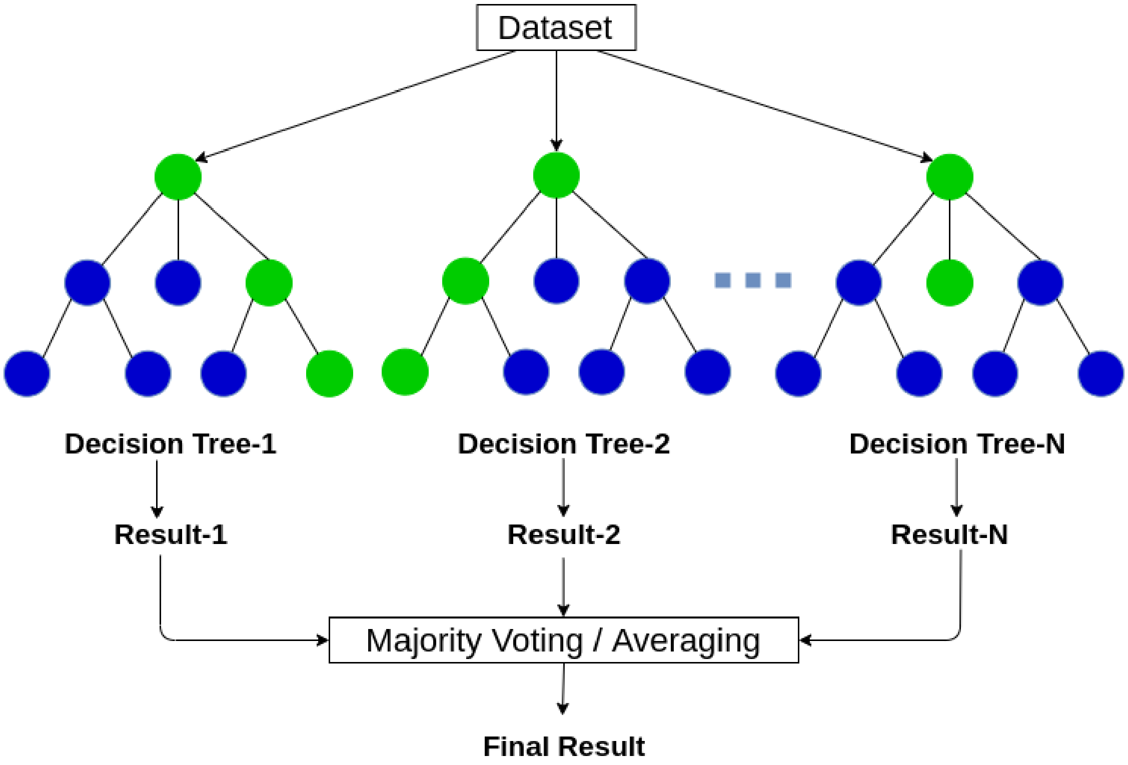
- **Logistic Regression**

  Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. However, unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model.



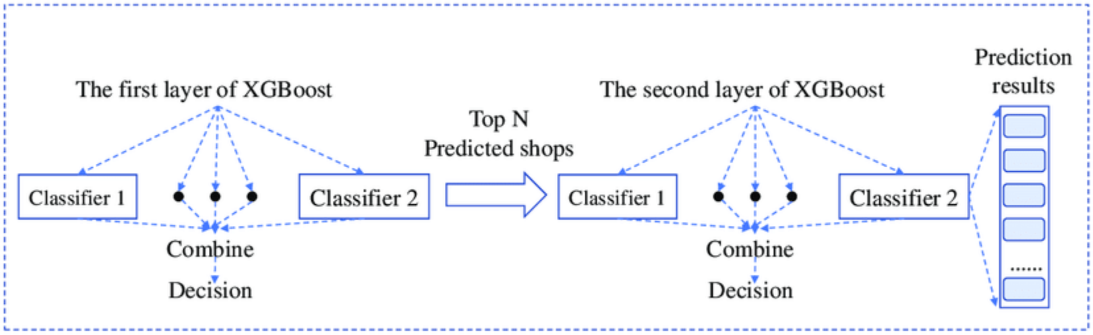
- **Random Forest**

  The random forest classifier is a supervised machine learning algorithm, that is trained using the ensemble predictions from a series of decision trees. The decision trees reach the final prediction from the features, that provide the maximum information gainat each node.



- **Extreme Gradient Boosting**

  The eXtreme Gradient Boosting (XGB) is a decision tree ensemble technique, that parallelizes and performs greedily in the tree pruning process. Using more complex models such as,least absolute shrinkage and selection operator (LASSO) and Ridge regularization, the XBG algorithm prevents overfitting. The XGB model, utilizes the distributed weighted Quantile Sketch algorithm, to effectively find the optimal split points amongst weighted datasets. The XGB algorithm is a more robust version of the Random Forest algorithm.

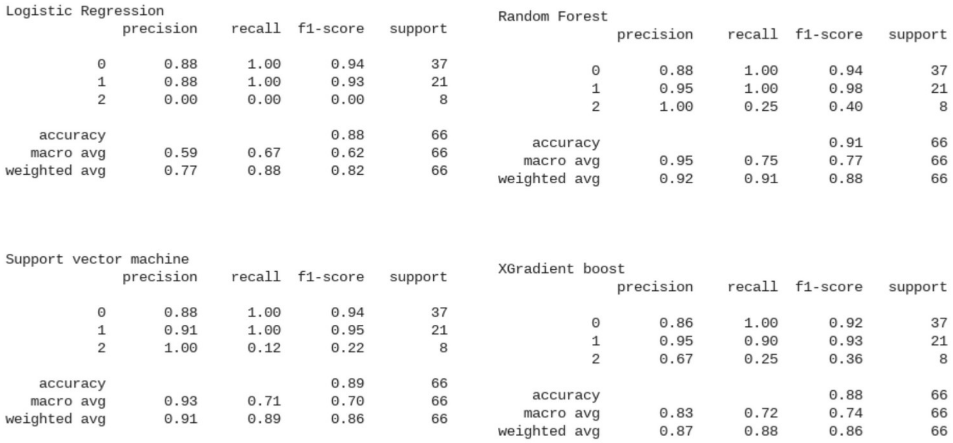


Model Deployment

The final model was deployed on streamlit - Streamlit is an open source app framework in Python language. It helps data scientist create web apps for data science and machine learning in a short time. Here is the link

Results

The accuracy report of the models that is tested with the test dataset is shown below:



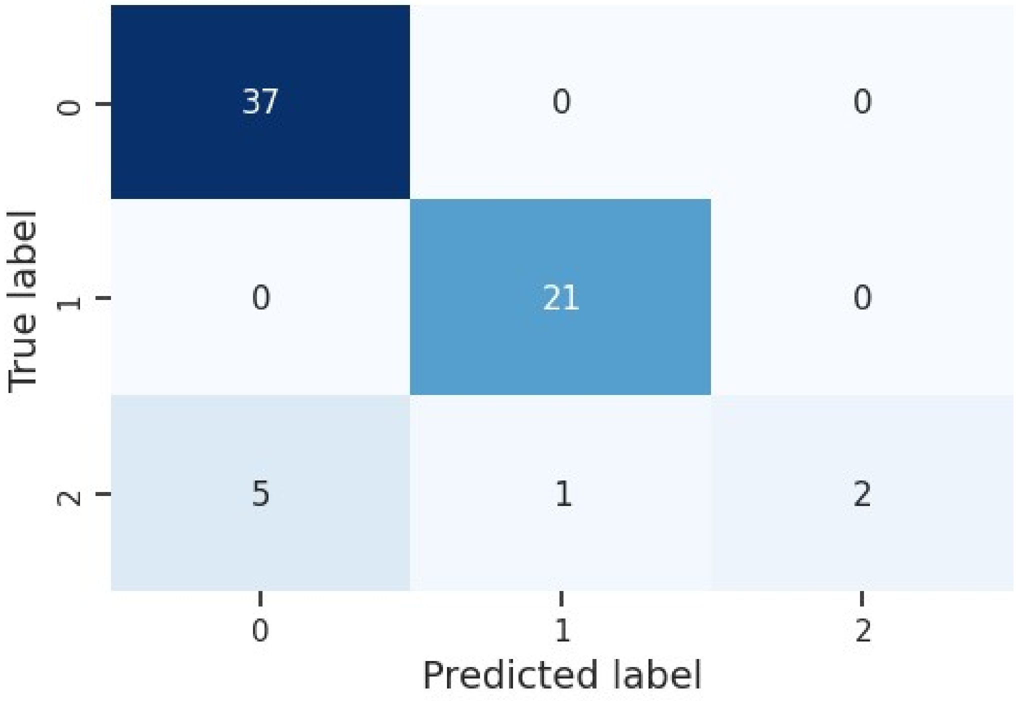
```
Logistic Regression
              precision    recall  f1-score   support

           0       0.88      1.00      0.94        37
           1       0.88      1.00      0.93        21
           2       0.00      0.00      0.00         8

    accuracy                           0.88        66
   macro avg       0.59      0.67      0.62        66
weighted avg       0.77      0.88      0.82        66
```

```
Random Forest
              precision    recall  f1-score   support

           0       0.88      1.00      0.94        37
           1       0.95      1.00      0.98        21
           2       1.00      0.25      0.40         8

    accuracy                           0.91        66
   macro avg       0.95      0.75      0.77        66
weighted avg       0.92      0.91      0.88        66
```

```
Support vector machine
              precision    recall  f1-score   support

           0       0.88      1.00      0.94        37
           1       0.91      1.00      0.95        21
           2       1.00      0.12      0.22         8

    accuracy                           0.89        66
   macro avg       0.93      0.71      0.70        66
weighted avg       0.91      0.89      0.86        66
```

```
XGradient boost
              precision    recall  f1-score   support

           0       0.86      1.00      0.92        37
           1       0.95      0.90      0.93        21
           2       0.67      0.25      0.36         8

    accuracy                           0.88        66
   macro avg       0.83      0.72      0.74        66
weighted avg       0.87      0.88      0.86        66
```

The Random Forest model, performs with the maximum accuracy score of 91%. The Confusion matrix for the Random Forst model is shown below:



Conclusion

The Random Forest classification model is used to estimate the dementia classification group for the dataset. When compared to other machine learning models, the model delivers the highest testing accuracy of 91% in categorizing the dataset. Furthermore, we can determine the probable type of dementia for each classification by looking at the CDR value, where;

- 0 means the patient is cognitively normal.
- 0.5 means the patient has very mild dementia.
- 1 means the patient has mild dementia.
- 2 means the patient has moderate dementia.

Recommendations

- The converted category of the group column should be changed to either demented or nondemented based on the CDR result in order to deploy the machine learning model in realtime. This is due to the fact that a real-time program cannot use the converted category because it was obtained after the fact, after patients had received at least two diagnoses..

- Additionally, it was found that the models performed better without the converted category..

Team Members

1. Abdulmalik Adeyemo
2. Glory Eke Kelechi
3. Odelola Solomon Oluwatobi
4. Elizabeth Okon
5. Onaolapo Sunday Akintunde
6. Chidiebere Nnadiegbulam
7. Osukoya Oluwatosin
8. Oswald Ohiole Ojo
9. Joseph Michael
10. Kalyani Pusadkar
11. Adewunmi Solomon
12. Akinyemi A.A
13. Moses Ojunba
14. Udeh Sandra Nkem
15. Godson E.
16. Jesujana Kayode
17. Justus Ilemobayo
18. Amao Jacobs
19. Emuejevoke E.
20. Teminijesu Jesufemi