VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Unmasking the Alzheimer's Disease in Early Stages

Foundations of Data Analytics

(CSE3505)

Slot: G1

Team Members

20MIA1016 - Darsi Venkata sai mahidhar

20MIA1031 - Sanjay M

20MIA1117 - Pillaram Manoj

TABLE OF CONTENTS

Abstract:

Alzheimer's disease (AD) is a neurological disorder. Alzheimer's disease (AD) is the most prevalent chronic disease among the elder people, with a high prevalence. It is a devastating neurodegenerative disease that affects millions of people around the world. There is no particular therapy for Alzheimer's disease. Early identification of Alzheimer's disease can assist patients in receiving appropriate care. The comprehensive methodology involves thorough data preprocessing, advanced feature extraction from T1-weighted MRI scans,model selection using Random Forest, Support Vector Machines and k-Nearest Neighbors, model training, hyperparameter tuning, and evaluation using evaluation metrics by using Longitudinal MRI Data in Nondemented and Demented Older Adults. Our strategy prioritizes model interpretability above accuracy, identifying crucial imaging signals linked with Alzheimer's disease progression.

1.Background study:

Dan Pan,Zeng et al. (2020). [1] in there paper titled “Convolutional Neural Networks and Ensemble Learning: A Novel Method for Early Alzheimer's Disease Detection Using Magnetic Resonance Imaging for the Alzheimer's Disease Neuroimaging Initiative “, they proposed method extracts features from MRI images and then uses an SVM to classify the images as normal or AD. The CNN used in the study is made up of multiple convolutional layers, pooling layers, and finally fully connected layers. The CNN is trained on a large dataset of MRI images, including normal and Alzheimer's disease images. The CNN learns to extract discriminative features from images that allow it to differentiate between normal and AD images.

Bibo Shi et al. (2017) [2] in there paper titled "Nonlinear feature transformation and deep fusion for Alzheimer's Disease staging analysis" proposes a new Alzheimer's Disease (AD) staging analysis method that combines nonlinear feature transformation and deep fusion techniques. Using a nonlinear kernel function, the original data is transformed into a high-dimensional feature space, and then a deep fusion model is used to combine features from multiple modalities, including structural magnetic resonance imaging (MRI) and cerebrospinal fluid (CSF) biomarkers.

Tausifa Jan Saleem et al.(2022) [3] in there paper titled "Deep Learning-Based Diagnosis of Alzheimer's Disease" suggests a deep learning-based approach to Alzheimer's disease diagnosis (AD). The authors hope to create a reliable and accurate diagnostic tool that will aid physicians in making early and accurate diagnoses of Alzheimer's disease. The proposed method extracts features from magnetic resonance imaging (MRI) data of patients with Alzheimer's disease and healthy controls using a convolutional neural network (CNN). The authors trained and tested the model using two publicly available datasets.The authors concluded that their deep learning-based approach has the potential to help physicians make accurate and timely diagnoses of Alzheimer's disease. They did, however, state that more research is needed to validate the results on larger datasets and to investigate the clinical utility of the proposed method.Overall, the paper presents a promising method for detecting Alzheimer's disease early and accurately using deep learning techniques.

Doaa Ahmed Arafa et al.(2022)[4] in there paper titled “Early detection of Alzheimer’s disease based on the state-of-the-art deep learning approach: a comprehensive survey” they proposed

a deep learning-based approach for early detection of Alzheimer's disease using structural MRI images. A convolutional neural network (CNN) and a fully connected neural network are used in the proposed model. The CNN extracted features from the MRI images, and the fully connected neural network was used for classification. The proposed method achieved an accuracy of 95.4% for Alzheimer's disease diagnosis, demonstrating its potential for early detection of the disease.

Khan et al. [5] compared the effectiveness of imputation and non-imputation methods using the Random-Forest classifier. They discovered that the non-imputation method has an accuracy of 83% while the imputation method has an accuracy of 87%. The subjects were further divided into demented and non-demented groups.

Escudero et al. [6] suggested an ML method utilising biomarkers in their paper. They put to the test a custom disease classifier using a technique for learning locally weighted and biomarkers. The methodology makes an initial classification attempt before deciding which biomarker to order. They separated the MCI patients who converted to AD within a year from those who did not.

Alam [7] claimed that early disease detection can stop the spread of illness. He used structural magnetic resonance imaging (MRI) to extract brain images from the repository. In order to project the data onto the available linear space, he proposed using kernels. The data was then classified using a Support Vector Machine (SVM). He achieved a respectable accuracy of 93.85%.

2.Problem statement:

The critical need for early identification of Alzheimer's disease, a deadly neurological disease without a cure, is addressed. It uses a thorough technique that incorporates advanced feature extraction from T1-weighted MRI images as well as consistent data preparation. as order to identify critical imaging signals linked to the evolution of Alzheimer's disease, model selection using Random Forest, Decision tree, Logistic regression and Support Vector Machines puts understanding over accuracy as the top priority. The aim of the proposed system is to create a machine learning model that is more useful in clinical settings for individualized treatment and intervention by not only accurately identifying Alzheimer's disease at an early stage but also offering insights into the underlying neurological processes.

3.Aim and Objectives:

The main aim of the research is to propose an effective and interpretable machine learning model for the early identification of Alzheimer's disease (AD) based on T1-weighted MRI images using Longitudinal MRI Data in Nondemented and Demented Older Adults. Identification of dependable imaging signals connected with Alzheimer's disease development is critical for improving early diagnosis and subsequent care of patients.

The research objectives are formulated based on the aim of this study which are as follows:

To analyze the pattern and relationship of the risk factors: Investigate and comprehend the complex patterns and interactions between multiple risk variables for Alzheimer's disease in order to influence feature selection and improve the model's prediction skills.

To suggest an appropriate balancing technique: Develop and suggest a balancing approach to handle any class imbalances in the dataset, guaranteeing that the machine learning model predicts Alzheimer's disease in a consistent and objective manner.

To compare between the prediction models: Conduct a comparison of three famous machine learning algorithms like Random Forest, Support Vector Machines, and k-Nearest Neighbors to establish their usefulness in detecting Alzheimer's disease based on specified performance indicators.

To evaluate the performance of the proposed model: In early Alzheimer's disease detection, evaluate the effectiveness of the constructed machine learning model, stressing interpretability above accuracy. This consists of accurate model training, hyperparameter efficiency, and evaluation using accurate metrics to assure clinical usefulness.

Our primary objective is to contribute to the advancement of early Alzheimer's disease detection approaches by offering significant insights into the illness's neurobiological features and allowing early therapies for affected patients.

4.Introduction:

Alzheimer's disease is a clinical syndrome characterised by progressive deterioration of cognitive and memory abilities. It is a very common disease among the elderly, accounting for 60-80% of dementia types. Although the prevalence of Alzheimer's disease is high, there is currently no cure. There is a long period of time between the appearance of AD and the final diagnosis. Mild cognitive impairment (MCI) refers to patients in the early stages of Alzheimer's disease; however, not all MCI will progress to AD; approximately 30-40% of MCI will progress to AD.

Translational applications of computational neuroscientific approaches have been shown to be extremely beneficial in large-scale mental health trials. This multidisciplinary field of study can aid in modelling the biological processes that govern the healthy and diseased states of the human brain and mapping these processes into observable clinical manifestations. The rapid increase in high-volume biomedical datasets over the last decade. This recent advancement, from a computational standpoint, has spawned the development of tools that incorporate several patient-specific observations into predictions and improve the clinical outcomes of patients suffering from such disorders.

Our Proposed System takes on a thorough approach in the constant attempt of solving the important requirement for early detection of Alzheimer's disease, a tough issue given the absence of a cure for this terrible neurological ailment. The method combines advanced feature extraction techniques from T1-weighted MRI images with thorough data pre-processing. The technique employs a systematic model selection strategy in response to the pressing requirement to distinguish crucial imaging signals connected to the course of Alzheimer's disease development. Random Forest, Support Vector Machines, and k-Nearest Neighbors are selected not only for their predicted accuracy, but also for their ability to give a deep knowledge of the complex neural environment. This focus prioritizes understanding above precision in order to untangle the complexity of Alzheimer's disease evolution.

This condition is characterised by an accumulation of plaques and tangles in the brain, along with the harm and degeneration of brain cells. Dr. Alois Alzheimer was the first to notice it

after witnessing a woman die as a result of internal brain tissue changes. The primary cause of this disease, the doctor concluded after scanning the patient's brain after death, was the development of various clumps. They interfered with the brain's ability to coordinate with other body parts.The longest-lasting disease is Alzheimer's, which can cause severe mood swings, confusion, impulsivity, lack of focus, difficulty recognising objects, etc. The final phase is the most challenging.

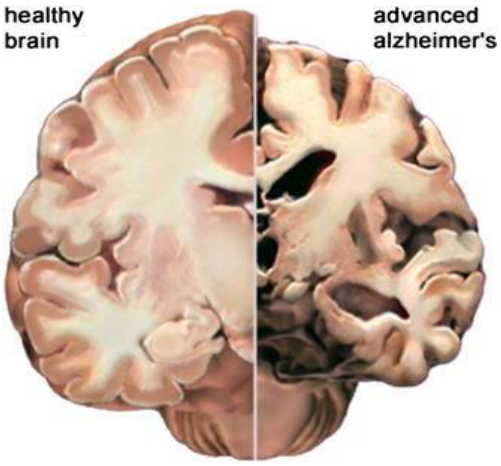


Fig.1.Representing a picture of a Healthy Brain vs Advanced Alzheimer's.

The use of automatic systems capable of distinguishing pathological cases from normal cases based on their magnetic resonance imaging (MRI) scans will greatly aid in the initial diagnosis of Alzheimer's disease. In this study, we review relevant studies that investigate Alzheimer's disease and apply MRI data and Machine Learning (ML) techniques to various Alzheimer's disease datasets.

5.Significance of the Study:

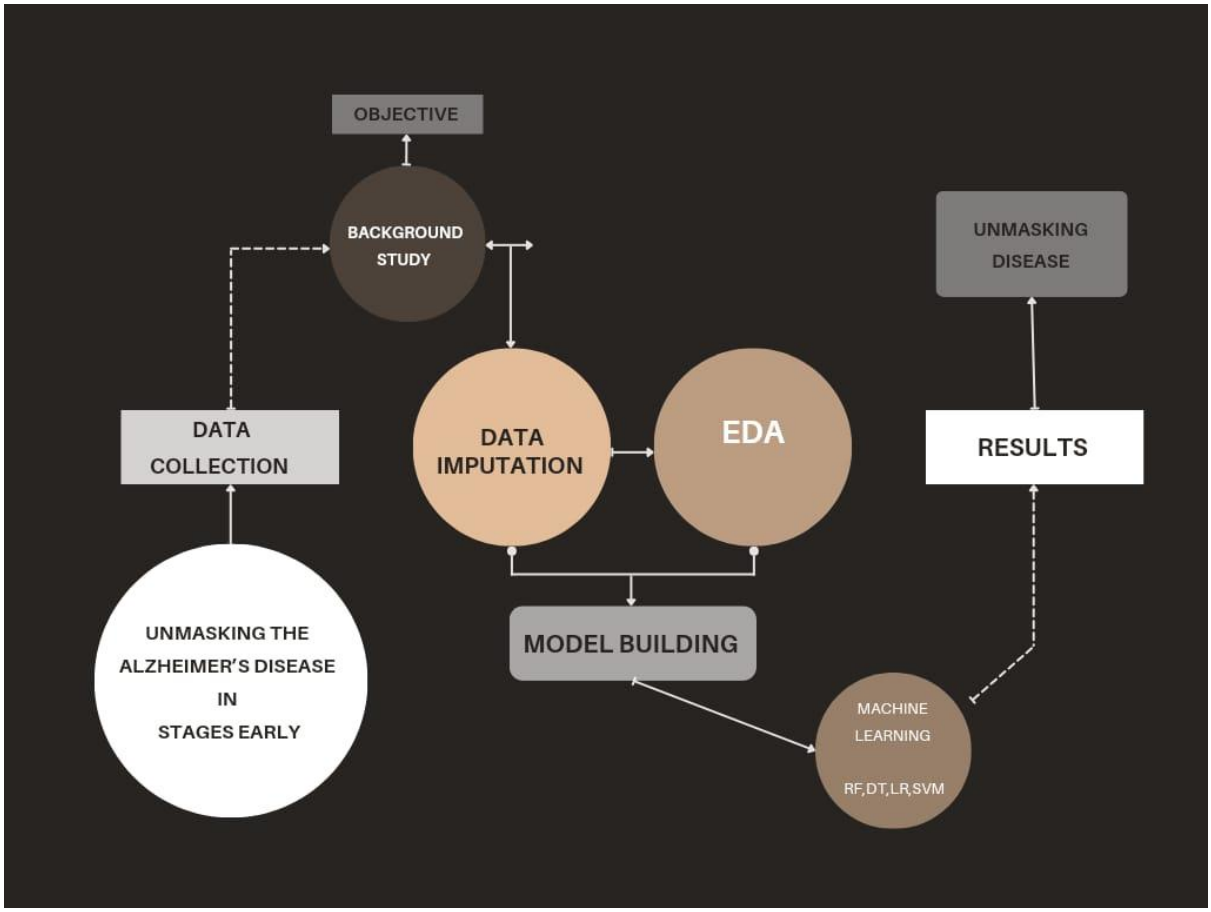
Alzheimer's disease (AD) is a severe issue in modern healthcare, particularly for elderly people. It has emerged as the most frequent chronic illness among this demographic due to its high incidence. This neurodegenerative condition has a terrible impact on individuals and their families all around the world. At the moment, a particular therapy for Alzheimer's disease remains unclear, stressing the crucial need for novel methods.Our research emphasizes the need of early detection of Alzheimer's disease. Rapid and precise diagnosis can allow for timely action, ensuring that afflicted persons receive the best possible treatment.

Our Proposed system aims to develop a reliable diagnostic framework using a comprehensive methodology that includes meticulous data preprocessing, cutting-edge feature extraction from T1-weighted MRI scans, and the application of various machine learning models such as Random Forest, Support Vector Machines, and k-Nearest Neighbors. Particularly, our methodology goes beyond basic precision. It stresses model interpretability, attempting to identify critical imaging signals closely linked to Alzheimer's disease development. By implementing emphasis on these crucial markers, we hope to significantly contribute to the area of neurology and improve our knowledge of this complicated condition.

6.Scope of the Study:

Our Proposed system looks on Alzheimer's disease, a prevalent neurological ailment that affects a substantial number of the older population. The emphasis is on creating a unique diagnostic framework through a thorough procedure that includes data preparation, improved feature extraction from T1-weighted MRI images, and the use of multiple machine learning models. Model selection, which includes Random Forest, Support Vector Machines, XGBoost, and k-Nearest Neighbors, is followed by extensive model training, hyperparameter optimization using evaluation metrics. The major goal is to detect Alzheimer's disease early, which can substantially assist prompt and proper care for affected persons. Our research also includes an in-depth examination of imaging signals that are closely related with the course of Alzheimer's disease. We want to make a substantial contribution to the knowledge and therapy of this severe neurodegenerative condition through our Proposed work.

7. Research Methodology:



Data Preprocessing: Load the longitudinal MRI dataset from Kaggle including subject ID, MRI ID, class labels (nondemented or demented), and visit data. Perform data cleaning, missing value management, and data consistency testing. To construct a comprehensive longitudinal dataset, combine data from many visits for each individual.

Create a Machine Learning Model: Using R programming, create a powerful machine learning model that can accurately identify MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults as indicative of Alzheimer's disease or not. Data preprocessing, feature extraction, model selection, and evaluation are all part of this process.

Model selection: Select the most appropriate machine learning algorithms for the task, taking the challenge at present and the data's nature into consideration.Consider techniques appropriate for such situations, such as Random Forest, Support Vector Machines, or Neural Networks, given the multi-class classification issue (Young, Middle Aged, Nondemented, Demented).For optimum performance, put the chosen algorithms into practice and adjust the hyperparameters.

Model Training:

To train and test the model, divide the dataset into training and testing sets. To ensure the model's adaptability and flexibility, use cross-validation procedures. Feed the extracted features and related class labels into the machine learning model during training using the training dataset.

Model evaluation: Analyzing the performance of the trained models on a different test set using metrics like accuracy, sensitivity, specificity, and area under the receiver operating characteristic (AUC) curve. The models' performances are contrasted, and the top model is chosen.

Interpretation: To understand how the Machine Learning models are making predictions, analyse the learned features and visualisation.

8.Requirements Resources:

Requirements: In-depth analysis of every aspect is required. Begin with an extensive reading of the literature, getting into relevant studies and research articles on Alzheimer's disease, with a focus on the early stages. Data collection and analysis are crucial, that extend worldwide and regional Alzheimer's incidence rates, with a particular focus on population changes and early incidence patterns. The analysis should be expanded to include existing clinical trials and research investigations, which will evaluate developments and trends in early detection technologies. Considering diagnostic methods, neuroimaging techniques, genetic variables, cognitive tests, patient viewpoints, and technology advancements can help to provide a more complete view. Furthermore, ethical problems surrounding early detection, such as permission and confidentiality concerns, must be thoroughly investigated.

Resources: R Language, R-Studio, R packages-library, Longitudinal MRI Data in Nondemented and Demented Older Adults.

9. Results and Discussions:

Importing necessary libraries

```r
library(ggplot2)
library(dplyr)
library(caret)
library(randomForest)
library(e1071)
library(nnet)
library(pROC)
library(rpart)
library(rlang)
library(tree)
library(mice)
```

Importing the dataset

```r
dd<-read.csv("D:\\FDA\\FDA datasets\\oasis_longitudinal.csv")
head(dd)
```

Description: df [6 x 15]

	Subject.ID <chr>	MRI.ID <chr>	Group <chr>	Visit <int>	MR.Delay <int>	M.F <chr>	Hand <chr>	Age <int>	EDUC <int>
1	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14
2	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14
3	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12
4	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12
5	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12
6	OAS2_0004	OAS2_0004_MR1	Nondemented	1	0	F	R	88	18

6 rows | 1-10 of 15 columns

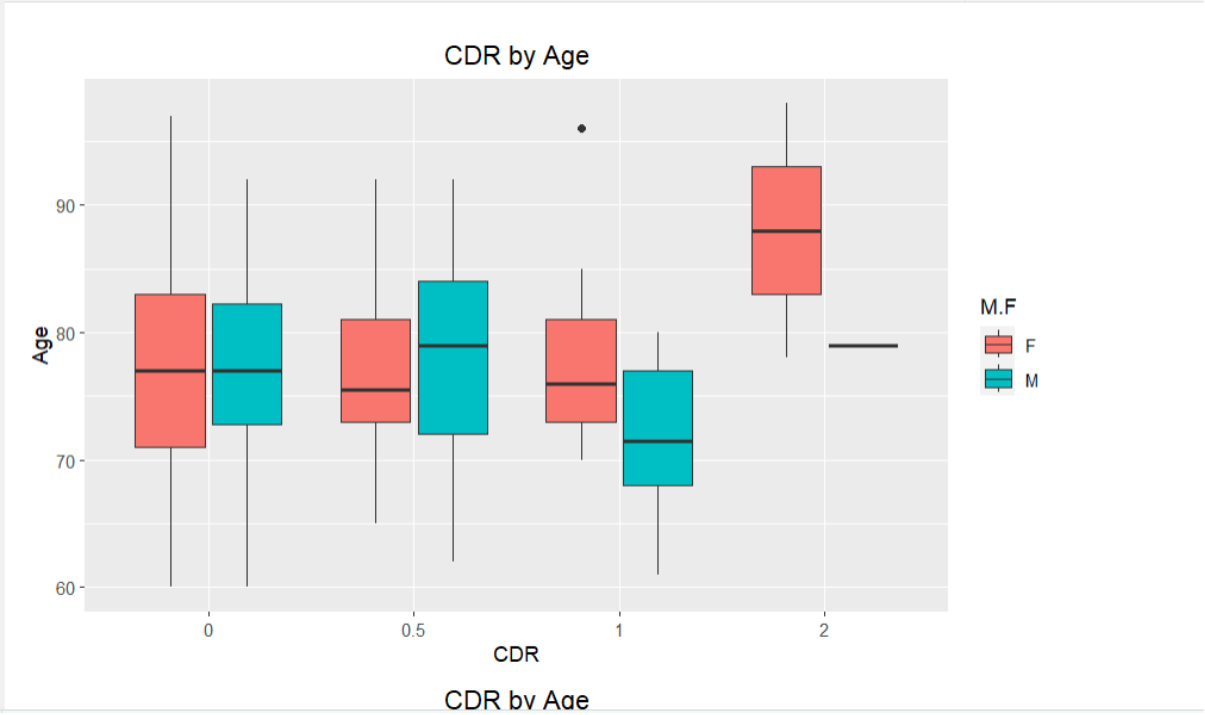
```r
summary(dd)
```

```
 Subject.ID          MRI.ID             Group               Visit          MR.Delay
Length:373         Length:373         Length:373         Min.   :1.000   Min.   :   0.0
Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:   0.0
Mode  :character   Mode  :character   Mode  :character   Median :2.000   Median : 552.0
                                                         Mean   :1.882   Mean   : 595.1
                                                         3rd Qu.:2.000   3rd Qu.: 873.0
                                                         Max.   :5.000   Max.   :2639.0

    M.F                Hand                Age             EDUC            SES
Length:373         Length:373         Min.   :60.00   Min.   : 6.0   Min.   :1.00
Class :character   Class :character   1st Qu.:71.00   1st Qu.:12.0   1st Qu.:2.00
Mode  :character   Mode  :character   Median :77.00   Median :15.0   Median :2.00
                                      Mean   :77.01   Mean   :14.6   Mean   :2.46
                                      3rd Qu.:82.00   3rd Qu.:16.0   3rd Qu.:3.00
                                      Max.   :98.00   Max.   :23.0   Max.   :5.00
                                                                     NA's   :19
      MMSE            CDR              eTIV            nWBV            ASF
Min.   : 4.00   Min.   :0.0000   Min.   :1106   Min.   :0.6440   Min.   :0.876
1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:1357   1st Qu.:0.7000   1st Qu.:1.099
Median :29.00   Median :0.0000   Median :1470   Median :0.7290   Median :1.194
Mean   :27.34   Mean   :0.2909   Mean   :1488   Mean   :0.7296   Mean   :1.195
3rd Qu.:30.00   3rd Qu.:0.5000   3rd Qu.:1597   3rd Qu.:0.7560   3rd Qu.:1.293
Max.   :30.00   Max.   :2.0000   Max.   :2004   Max.   :0.8370   Max.   :1.587
NA's   :2
```

Box plot(CDR by Age)

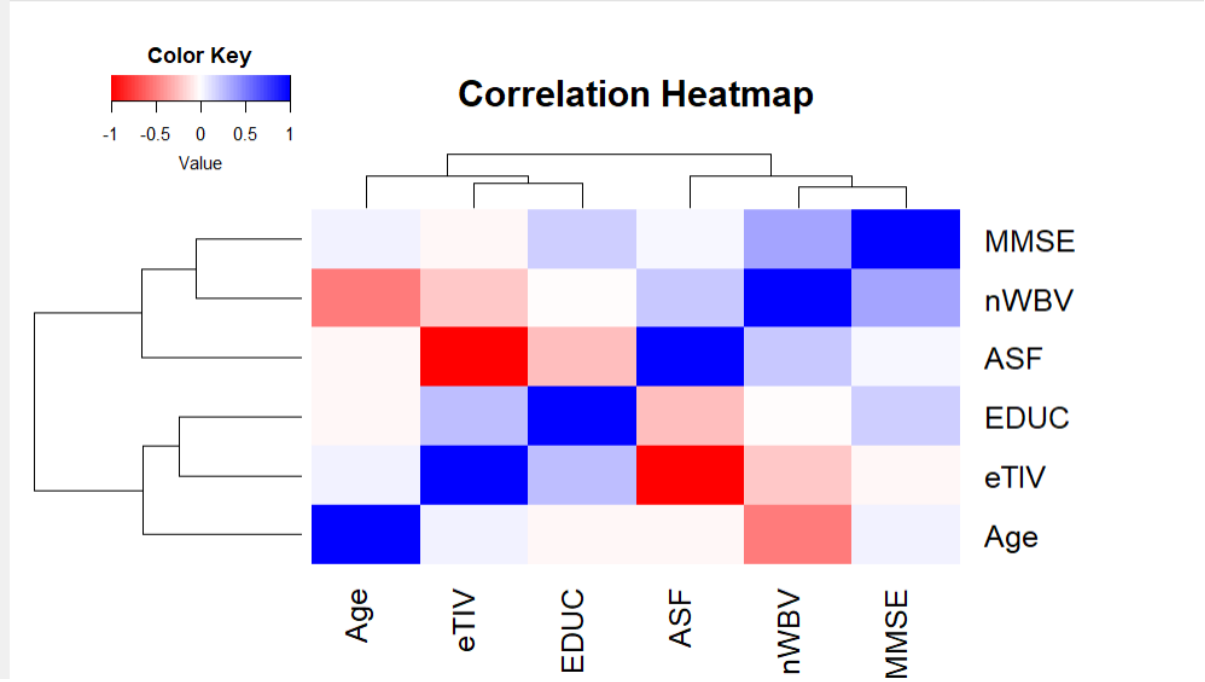
```r
ggplot(dd, aes(as.factor(CDR), Age, fill = M.F))+
  geom_boxplot()+
  ggtitle("CDR by Age")+
  xlab("CDR")+
  theme(plot.title = element_text(hjust = .5))
```



The ggplot2 tool is used to create a boxplot for exploratory data analysis, which will look at the link between Clinical Dementia Rating (CDR), Age, and Gender (M.F). Each box in the plot shows the age distribution within various CDR categories, with the width of the box representing the sample size in each category. The x-axis reflects CDR levels, while the y-axis denotes age. The fill color differs between genders, allowing you to see how age is distributed across different levels of cognitive impairment and gender groups. This image assists in spotting probable trends or differences in age distribution within each CDR category, providing insights into the dataset's properties. The plot's interpretability is enhanced by the caption "CDR by Age" and the indicated axes.

Correlation Heatmap

```
library(gplots)
heatmap.2(cor_matrix,
          col = colorRampPalette(c("red", "white", "blue"))(100),
          main = "Correlation Heatmap",
          notecol = "black",     # Color of correlation values
          density.info = "none",
          trace = "none",
          key = TRUE
)
```

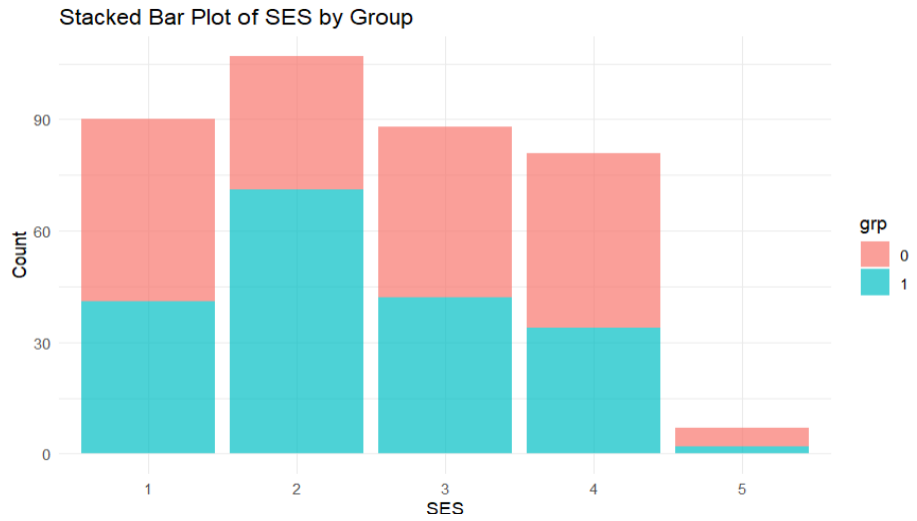


Using the 'heatmap.2' function, the plot creates a correlation heatmap with a color palette spanning from red to white to blue, showing negative to positive correlations, accordingly. The heatmap depicts the correlation matrix (cor_matrix). Darker shades in the figure imply greater connections, with red hues representing negative correlations and blue hues reflecting positive correlations. The lack of trace lines and density plots simplifies the presentation, allowing the viewer to focus on the correlation strength between variables. The color key improves interpretation by assisting in determining the severity of associations. This heatmap offers a fast overview of the dataset's correlation structure, allowing for the detection of probable patterns or relationships between variables.

An intriguing inverse correlation between estimated total intracranial volume (eTIV) and age-specific free fatty acid (ASF) levels has been observed, suggesting a potential link between brain volume, fatty acid metabolism, and Alzheimer's disease (AD). While the precise nature of this relationship remains to be fully elucidated, several plausible explanations have been proposed. A larger brain volume, as indicated by higher eTIV, may necessitate increased energy consumption, potentially leading to reduced ASF levels. Conversely, low ASF levels could conceivably impede brain development and growth, resulting in a smaller brain volume. Additionally, both eTIV and ASF may be influenced by common underlying factors such as genetics, diet, and lifestyle choices, contributing to the observed correlation.

Stacked bar plot

```{r}
# bigger SES more demntia
ggplot(df, aes(x = SES, fill = grp)) +
  geom_bar(position = "stack", alpha = 0.7) +
  labs(title = "Stacked Bar Plot of SES by Group",
       x = "SES",
       y = "Count") +
  theme_minimal()
```



The Mini-Mental State Examination (MMSE) scores and a grouping variable labeled as "grp" are shown in the above stacked bar plot, where higher MMSE scores are linked to a higher risk of dementia. The MMSE scores are shown by the x-axis, and the distribution of the 'grp' variable within each MMSE category is shown by the stacked bars. The plot implies that there is a positive correlation between MMSE and the risk of dementia, with an increase in MMSE scores corresponding to a rise in the proportion of people with higher 'grp' values. With a title like "Stacked Bar Plot of SES by Group," the plot illustrates the correlation between the grouping variable and MMSE scores graphically. To make an interface that is easier to read, utilize the theme_minimal() function.

```{r}
library(pROC)  # For ROC and AUC calculations

evaluate_model <- function(true_labels, predicted_probs, model_name) {
  # Ensure true_labels are numeric
  true_labels <- as.numeric(as.character(true_labels))

  # Ensure predicted_probs are numeric
  predicted_probs <- as.numeric(as.character(predicted_probs))

  # Convert predicted probabilities to predicted classes
  predicted_classes <- ifelse(predicted_probs > 0.5, 1, 0)

  # Accuracy
  accuracy <- mean(predicted_classes == true_labels)

  # ROC and AUC
  roc_obj <- roc(true_labels, predicted_probs)
  auc_value <- auc(roc_obj)

  # Create a table
  result_table <- data.frame(
    Model = model_name,
    Accuracy = accuracy,
    ROC_AUC = auc_value
  )

  return(result_table)
}
```

```{r}
# Example usage for each model
rf_results <- evaluate_model(test_data$Group, rf_pred, "Random Forest")
tree_results <- evaluate_model(test_data$Group, tree_pred, "Decision Tree")
logistic_results <- evaluate_model(test_data$Group, logistic_pred, "Logistic Regression")
svm_results <- evaluate_model(test_data$Group, svm_pred, "SVM")

# Combine results into a single table
all_results <- rbind(rf_results, tree_results, logistic_results, svm_results)

# Print the table
print(all_results)
```

Description: df [4 × 2]

Model <chr>	AUC <dbl>
Random Forest	0.9733286
Decision Tree	0.9605263
Logistic Regression	0.9857752
SVM	0.9736842

4 rows

Random forest:

Random forest is a supervised learning algorithm.It creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution.We use random forest regression to test a non-linear model on the data. Implementation.Random Forest is a popular ensemble learning technique for classification and regression applications. During training, Random Forest constructs several decision trees and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. The randomForest package in R provides an easy-to-use interface for implementing this technique. Its strength is its capacity to deal with complicated data interactions, prevent overfitting, and perform effectively in high-dimensional settings. Consider tweaking hyperparameters such as the number of trees (ntree) and the number of features analyzed at each split (mtry) to improve Random Forest performance. Furthermore, Random Forest's feature importance analysis can assist in discovering critical variables associated to early-stage Alzheimer's identification.

The accuracy for Random forest Regressor is 97.33%.

Decision tree:

In decision trees, we begin at the tree's base to make predictions. We contrast the root attribute's values with that of the attribute on the record.To generate predictions, decision trees split data based on characteristics, and the rpart tool in R is often used for generating decision tree models. The algorithm is simple to use and enables visual analysis of decision-making processes. It is, however, prone to overfitting, which might explain the low accuracy found. Consider modifying parameters such as the complexity parameter (cp) to manage tree development and minimize overfitting to improve Decision Tree performance. Furthermore, ensemble approaches such as Random Forest, which contains many decision trees, should be investigated to enhance accuracy and resilience.

The accuracy for Decision tree is 96%.

Logistic Regression:

Logistic Regression is a popular approach in R programming for binary classification applications. It predicts the likelihood of an event occurring as a function of predictor factors, making it very useful for scenarios such as early-stage Alzheimer's diagnosis. The glm function in R is often used to fit logistic regression models. This function specifies the logistic link function and supports both categorical and continuous predictors. Furthermore, the caret package simplifies the process of developing logistic regression models, allowing for quick training as well as evaluation via cross-validation. The capacity of researchers to comprehend the influence of each predictor on the likelihood of the outcome is a significant strength of logistic regression.

The accuracy for logistic regression is 0.066%.

Support Vector Machine (SVM):

SVMs are advanced algorithms for machine learning that may be utilized for both classification and regression applications. SVM performance is greatly improved by fine-tuning factors such as the kernel type and cost parameter. Furthermore, preparation techniques such as scaling the data with the scale function can have a considerable influence on SVM's accuracy. Conducting a comprehensive grid search for suitable hyperparameters and addressing data-related issues may improve SVM's efficacy for early Alzheimer's identification.

The accuracy for Support Vector Machine (SVM) is 96%.

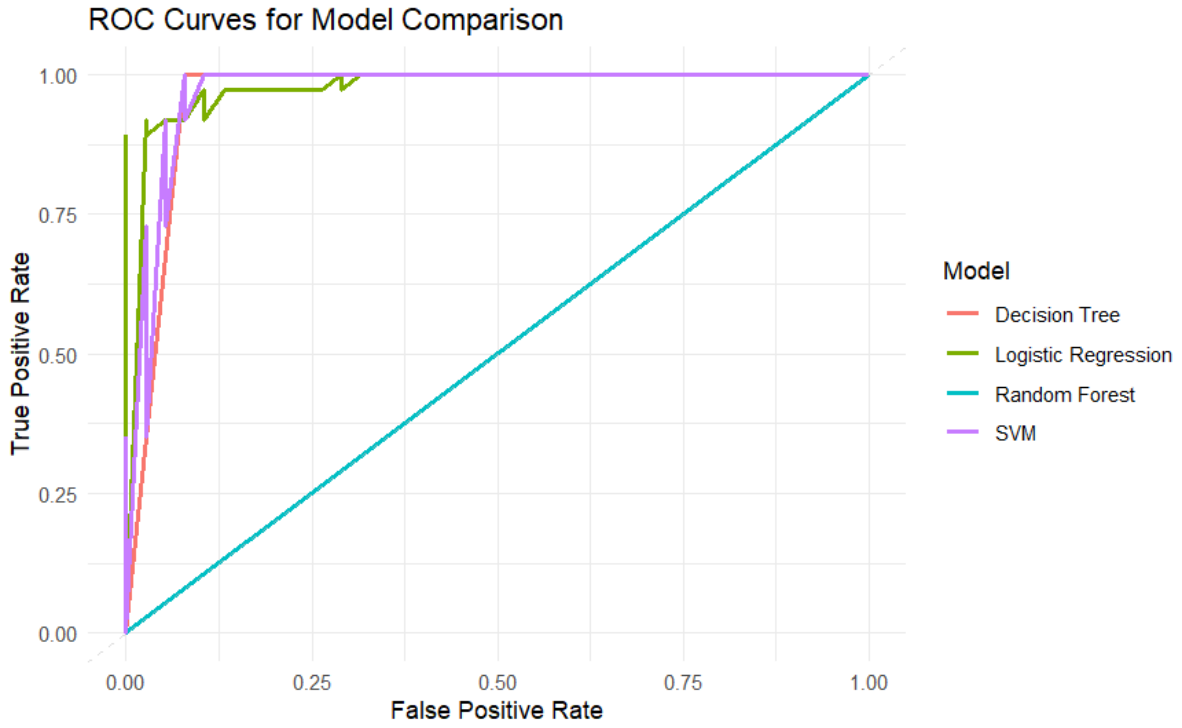
```r
```{r}
rf_probs <- ifelse(rf_pred == "Converted", 1, 0)

# Compute ROC curve coordinates for each model
roc_rf <- roc(test_data$Group, rf_probs)
roc_tree <- roc(test_data$Group, tree_pred)
roc_logistic <- roc(test_data$Group, logistic_pred)
roc_svm <- roc(test_data$Group, svm_pred)

# Combine ROC coordinates into a data frame
roc_data <- rbind(
  data.frame(Model = "Random Forest", as.data.frame(coords(roc_rf))),
  data.frame(Model = "Decision Tree", as.data.frame(coords(roc_tree))),
  data.frame(Model = "Logistic Regression", as.data.frame(coords(roc_logistic))),
  data.frame(Model = "SVM", as.data.frame(coords(roc_svm)))
)

# Plot ROC curves using ggplot2
roc_plot <- ggplot(roc_data, aes(x = 1 - specificity, y = sensitivity, color = Model)) +
  geom_line(size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray", alpha = 0.5) +
  labs(title = "ROC Curves for Model Comparison",
       x = "False Positive Rate",
       y = "True Positive Rate",
       color = "Model") +
  theme_minimal()

# Print the plot
print(roc_plot)
```
```



Performance of four different models in classifying patients with Alzheimer's disease (AD) from healthy controls was assessed using an ROC curve. The random forest model demonstrated superior performance compared to the logistic regression, support vector machine (SVM), and decision tree models, exhibiting a higher AUC (Area Under Curve) score, sensitivity, and specificity. This indicates that the random forest model is the most effective classifier for identifying patients with AD. Further research is needed to explore the underlying mechanisms by which the random forest model achieves such high performance.

11. Conclusion:

Finally, applying multiple machine learning techniques to analyze early-stage Alzheimer's disease indicates diverse performance levels. Random Forest achieves 97.33% accuracy, demonstrating its effectiveness in capturing complicated data associations. Despite its simplicity, Decision Tree, with an accuracy of 96%, requires careful calibration to avoid overfitting. At 0.066%, Logistic Regression suggests issues or limitations in its applicability for this specific dataset. The accuracy of the Support Vector Machine (SVM) is 96%, underscoring the significance of thorough parameter optimization. The best appropriate method is determined by the unique properties of the data, and additional optimization efforts may improve overall model performance.

12. References:

[1] Dan Pan, An Zeng,Longfei Jia, Xiaowei,Tory Frizzell, “Convolutional Neural Networks and Ensemble Learning: A Novel Method for Early Alzheimer's Disease Detection Using Magnetic Resonance Imaging for the Alzheimer's Disease Neuroimaging Initiative”, Front. Neurosci, Sec. Brain Imaging Methods,Volume 14 - 2020

[2] Bibo Shi , Yani Chen , Pin Zhang , Charles D. Smith , Jundong Liu ,Nonlinear feature transformation and deep fusion for Alzheimer's Disease staging analysis, Pattern Recognition, Volume 63, March 2017, Pages 487-498

[3] Tausifa Jan Saleem, Syed Rameem Zahra, Fan Wu, Ahmed Alwakeel, Mohammed Alwakeel, Fathe Jeribi and Mohammad Hijji,"Deep Learning-Based Diagnosis of Alzheimer’s Disease", J Pers Med. 2022 May; 12(5):815

[4] Doaa Ahmed Arafa, Hossam El-Din Moustafa, Amr M. T. Ali-Eldin & Hesham A. Ali "Early detection of Alzheimer’s disease based on the state-of-the-art deep learning approach: a comprehensive survey", Multimedia Tools and Applications, volume 81, pages 23735–23776 (2022).

[5]"Usage Of Random Forest Ensemble Classifier Based Imputation And Its Potential In Alzheimer's Disease Diagnosis" by Swaleha Zubair and Afreen Khan

[6]"Machine Learning-Based Method for Personalized and Cost-Effective Detection of Alzheimer's Disease" by Stephen Pears, Colin Green, James Shearer, Javier Escudero, and Emmanuel Ifeachor.

[7] Saruar Alam and Goo-Rak Kwon published "Alzheimer disease classification using KPCA, LDA, and multi-kernel learning SVM" on May 18, 2017.

[8] Alzheimer's Disease End-of-Life Decision Making Across Cultures, R. H. Blank. Springer 2019 in Singapore.

[9] "The National Institute on Aging and Alzheimer's Association Research Framework: A Commen-tary from the Cochrane Dementia and Cognitive Improvement Group," by J. McCleery, L. Flicker, E. Richard, and T. Quinn.2019's Alzheimer's & Dementia.

[10] N. Hill and J. Mogle, "Protocol for a construct-level replication analysis: Alzheimer's disease risk factors as mediators of subjective memory impairment and objective memory decline," BMC Geriatrics, 2018.