Disease Prediction Using Machine Learning Classification Algorithms

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

By

Harshit Saxena (191369)

Under the supervision of

Dr. Hari Singh

to

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology
Waknaghat,
Solan-173234, Himachal Pradesh

Candidate’s Declaration

I hereby declare that the work presented in this report entitled **“Disease Prediction Using Machine Learning Classification Algorithms”** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from July 2022 to May 2023 under the supervision of **Dr. Hari Singh, Assistant Professor (SG), Department of Computer Science and Engineering**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Harshit Saxena, 191369

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Hari Singh
Assistant Professor (SG)
Computer Science and Engineering
Dated:

Acknowledgement

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for us to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisor **Dr. Hari Singh, Assistant Professor**, Department of CSE Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of **Machine Learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Group No. : 52
Harshit Saxena
191369

Table of Content

List of Figures

List of Tables

Abstract

Alzheimer's disease is the main cause of dementia in elderly people. The current generation, a great deal of interest in applying machine learning to explore metabolic diseases that affect large numbers of people worldwide, such as Alzheimer's disease and diabetes. Its incidence is increasing at an alarming rate each year. Neurodegenerative changes affect the brain in Alzheimer's disease. As the population ages, more and more people, their families, and health care workers will suffer from diseases that affect memory and function. The impact is severe socially, financially and economically. Alzheimer's disease is difficult to predict in its early stages. Treatment in the early stages of Alzheimer's disease is more effective and causes less damage than treatment in the later stages. Several techniques such as decision trees, random forests, support vector machines, XGBoost, extra-tree classifiers, gradient boosting, AdaBoost, and voting classifiers were used to identify the optimal parameters for predicting Alzheimer's disease. After this, the dataset was trained with Neural Networks to depict that neural nets are not recommended with small datasets. These predictions of the disease are based on Open Access Series of Imaging Studies (OASIS) data, and performance is measured using parameters such as ML model accuracy, recall, accuracy, and F1 score. The proposed classification scheme can be used by clinicians to diagnose these diseases. Reducing annual mortality from Alzheimer's disease through early detection using these ML algorithms is highly beneficial. The proposed work shows better results with the highest average validation accuracy of almost 90% on the AD test data. This test accuracy is significantly higher compared to existing work.

Chapter-1

INTRODUCTION

1.1 Introduction

Millions of people throughout the world are afflicted by the progressive neurological ailment known as Alzheimer's disease. Early detection and diagnosis of Alzheimer's disease can help in better management and treatment of the disease. This project aims to compare the performance of various classification algorithms in predicting Alzheimer's disease using the OASIS-2: Longitudinal MRI Data in Nondemented and Demented Older Adults dataset.

About Dementia :

Dementia isn't a specific complaint. It’s an overall term that describes a group of symptoms associated with a decline in memory or other thinking chops severe enough to reduce a person’s capability to perform everyday conditioning. Alzheimer’s complaints account for 60 to 80 percent of cases. Vascular dementia, which occurs after a stroke, is the alternate most common dementia type. But there are numerous other conditions that can beget symptoms of dementia, including some that are reversible, similar as thyroid problems and vitamin scarcities

Dementia is presently the seventh leading cause of death among all conditions and one of the major causes of disability and reliance among aged people worldwide. dementia has physical, cerebral, social and profitable impacts, not only for people living with dementia, but also for their careers, families and society at large. There's frequently a lack of mindfulness and understanding of dementia, performing in stigmatization and walls to opinion and care.

Key Facts :

- Dementia is a pattern in which there's deterioration in cognitive function beyond what might be anticipated from the usual consequences of natural aging.
- Although dementia substantially affects aged people, it isn't an ineluctable consequence of aging.
- Presently more than 55 million people live with dementia worldwide, and there are nearly 10 million new cases every time.
- Dementia results from a variety of conditions and injuries that primarily or digressively affect the brain. Alzheimer's complaint is the most common form of dementia and may contribute to 60 - 70 percent of cases.
- Dementia is presently the seventh leading cause of death among all conditions and one of the major causes of disability and reliance among aged people encyclopedically.
- Dementia has physical, cerebral, social and profitable impacts, not only for people living with dementia, but also for their careers, families and society at large.

Alzheimer :

Alzheimer's is a type of madness that causes problems with memory, thinking and behavior. Symptoms generally develop sluggishly and get worse over time, getting severe enough to intrude with diurnal tasks.

Alzheimer's isn't a normal part of aging. The topmost known threat factor is adding age, and the maturity of people with Alzheimer's are 65 and aged. But Alzheimer's isn't just a complaint of old age. Roughly 200,000 Americans under the age of 65 have younger - onset Alzheimer’s complaints (also known as early - onset Alzheimer’s).

Alzheimer's is the sixth leading cause of death in the United States. Those with Alzheimer's live a normal of eight times after their symptoms come conspicuous to others, but survival can range from 4 to 20 times, depending on age and other health conditions.

1.2 Problem Statement

- The most frequent cause of dementia and a neurodegenerative ailment, Alzheimer's disease (AD), with no known cause or pathogenesis. It mostly affects older adults.
- Patients who are suspected of having Alzheimer's disease (AD) are evaluated with brain imaging by means of MRI.
- The brain tissue has shrunk both locally and globally, according to the MRI results.
- According to some studies, the features of an MRI may be able to predict the rate of AD's decline and serve as a guide for future treatment.
- However, in order for clinicians and researchers to reach that point, they will need to employ ML methods that accurately predict a patient's progression from dementia to moderate cognitive impairment.
- We propose creating a reliable model that can assist doctors in doing this and predicting early Alzheimer's disease.

1.3 Objectives

The specific objectives of this project are as follows:

- To explore the literature on machine learning algorithms for Alzheimer's disease prediction.
- To preprocess the OASIS-2 dataset and extract relevant features for classification.
- To compare the performance of eight classification algorithms, including Random Forest, SVM, Decision Tree, XGBoost, Voting, ExtraTrees, Gradient Boosting, and Ada Boosting.
- To analyze the results and identify the most accurate algorithm for Alzheimer's disease prediction.
- Under the current conditions, mortal instinct and standard measures don't frequently coincide. In order to break this problem, we need to

work on innovative approaches similar to machine literacy, which are computationally ferocious and non-traditional.

- Machine literacy ways are increasingly being used in complaint validation and visualization to offer visionary and tailored conventions.
- In addition to perfecting cases' quality of life, this drift aids croakers in making treatment opinions and health economists in making their analyses.
- Viewing medical reports may lead radiologists to miss other complaint conditions. As a result, it only considers many causes and conditions.
- The thing then is to identify the knowledge gaps and implicit openings associated with ML frameworks and EHR deduced data.

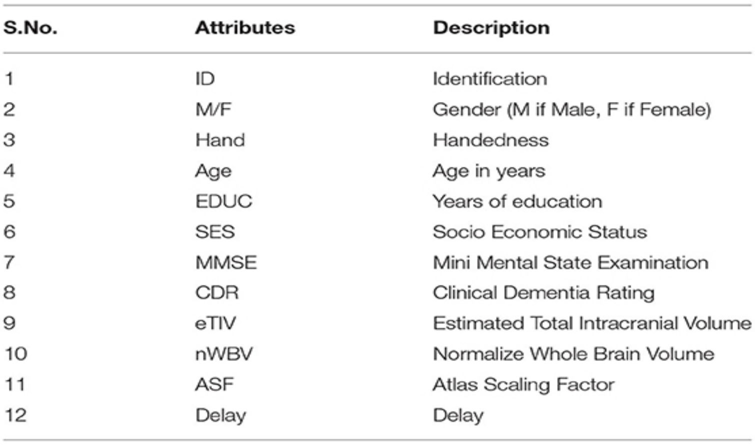
1.4 Methodology

The project followed a standard methodology for machine learning projects, which includes the following steps:

- Literature survey: A comprehensive review of relevant literature on machine learning algorithms.
- data : gathered data on Alzheimer's patients and nondemented older adults from the OASIS-2 longitudinal MRI dataset. This dataset includes imaging and clinical data from over 150 participants, and selected the relevant features for the analysis.
- Data Preprocessing: The collected data contained missing values, outliers, or noisy data, which had to preprocess before feeding it to the classification models.
- Model Selection: Eight classification models were selected for the analysis, including Random Forest, SVM, Decision Tree, XGBoost, Voting, ExtraTrees, Gradient Boosting, and Ada Boosting. These models were selected based on their ability to handle high-dimensional data, ability to handle imbalanced data, and their performance in previous studies.
- Model Development: Developed each of the selected models using the preprocessed data. The model development process involved selecting

the appropriate hyperparameters, training the models on the training data, and validating the models on the test data. This process was repeated for each of the eight models.

- Performance Evaluation: Once the models were developed, performance was evaluated using accuracy as the primary performance metric. accuracy was then compared of each model and analyzed the results to identify which models performed the best.
- Result Analysis: Analyzed the results to determine the strengths and weaknesses of each model.



S.No.	Attributes	Description
1	ID	Identification
2	M/F	Gender (M if Male, F if Female)
3	Hand	Handedness
4	Age	Age in years
5	EDUC	Years of education
6	SES	Socio Economic Status
7	MMSE	Mini Mental State Examination
8	CDR	Clinical Dementia Rating
9	eTIV	Estimated Total Intracranial Volume
10	nWBV	Normalize Whole Brain Volume
11	ASF	Atlas Scaling Factor
12	Delay	Delay

Table 1. Dataset Description

The Machine Learning ways were applied to Alzheimer's complaint datasets to bring a new dimension to prognosticate disease at an early stage. The raw Alzheimer's complaint datasets are inconsistent and spare, which affects the accuracy of algorithms. Before assessing machine - learning algorithms, data must be effectively prepared for analysis by removing unwanted attributes, missing values, and spare records. Building a machine- literacy model requires unyoking the data into training and testing sets. In the following data medication step, the training data were used to produce a model, which was also applied to test data to prognosticate Alzheimer's Disease. The model was trained from training set data, and test set data were used to test unseen data. Cross-validation was carried out by dividing the dataset into three subsets. Model prognostications are made using one subset of the data(test data) and model performance is estimated using the other subsets (training and confirmation) of the data. The data had been preprocessed, and we aimlessly divide it into an 80:20 rate, with 80 going to training and 20 going to testing. Fig 1, describes the workflow of the system

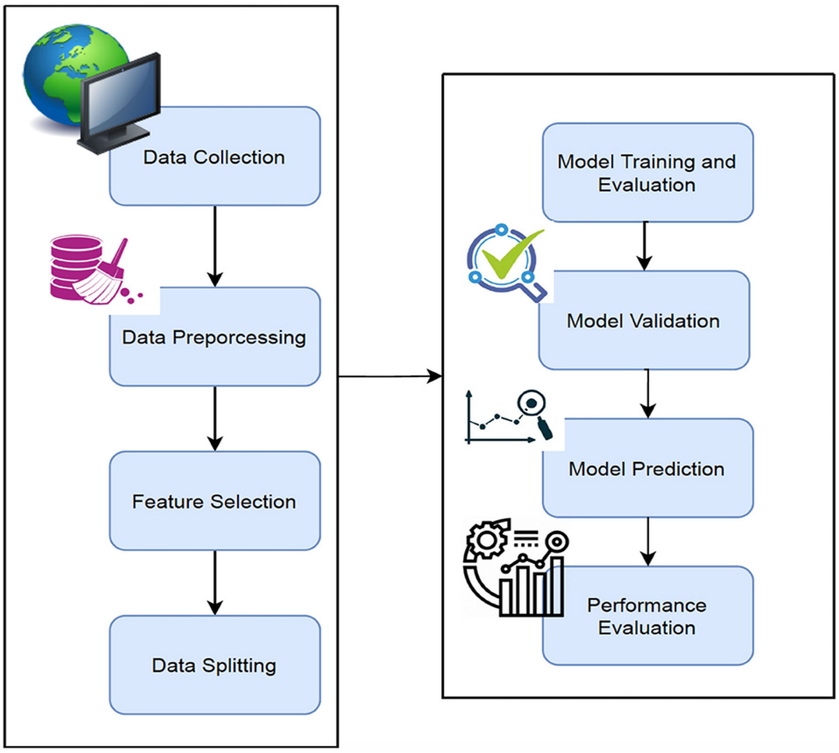


Figure 1. Workflow

1.4 Organization

The report is organized into five chapters. Chapter 1 provides an introduction to the project, including the problem statement, objectives, methodology, and organization. Chapter 2 presents a literature survey on machine learning algorithms for Alzheimer's disease prediction. Chapter 3 describes the system development, including data preprocessing and model development. Chapter 4 presents the experiments and result analysis, including a comparison of the performance of the eight classification algorithms. Chapter 5 provides conclusions, future scope, and applications contributions of the project.

Chapter-2

LITERATURE SURVEY

The literature survey focused on recent developments in machine learning algorithms for Alzheimer's disease prediction. The survey found that machine learning algorithms have shown promising results. Various algorithms, including Random Forest, SVM, Decision Tree, XGBoost, and ensemble algorithms, have been used for Alzheimer's disease prediction. The survey also highlighted the importance of feature selection and preprocessing in improving the accuracy of classification algorithms.

The following are examples of works that were discovered in the literature:

1. Comparison of classification algorithms for Early Alzheimer's disease detection through structural MRI and cognitive assessments.

This paper compares the performance of several classification algorithms for early detection of Alzheimer's disease using structural MRI and cognitive assessments. The authors used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and evaluated the performance of seven classification algorithms: logistic regression, decision tree, random forest, support vector machine, k-nearest neighbors, naive Bayes, and gradient boosting. The results showed that random forest and gradient boosting had the highest accuracy, while naive Bayes had the lowest accuracy. The study suggests that machine learning algorithms can be useful tools for Early Alzheimer's disease detection.

3. A comparison of machine learning classification algorithms for predicting Alzheimer’s disease” by A. Al-Duwairi et al.

The paper "a machine learning comparison classification algorithms for predicting Alzheimer's disease" by A. Al-Duwairi et al. published in the Journal of Healthcare Engineering in 2021 compares the performance of six machine learning algorithms in predicting Alzheimer's disease using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The algorithms tested include Naive Bayes, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Logistic Regression (LR), and Artificial Neural Network (ANN). The study found that SVM and kNN had the highest accuracy rates in predicting Alzheimer's disease, with SVM being the most sensitive and kNN being the most specific. The paper concludes that SVM and kNN can be used as effective tools for early prediction of Alzheimer's disease, and that further studies are needed to optimize the accuracy of these models.

3. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects.

In order to distinguish between age-related cognitive decline and Alzheimer's disease, the authors of this paper sought to identify mild cognitive impairment (MCI). Using machine learning methods, the group proposes a novel MRI-based biomarker. They used information from the ADNI Database, which is part of the Alzheimer's Disease Neuroimaging Initiative.According to the paper, their aggregate biomarker distinguished between stable MCI (sMCI) and progressive MCI (pMCI) with a cross-validation area under the curve (AUC) score of 0.9020.

Noteworthy Methods:

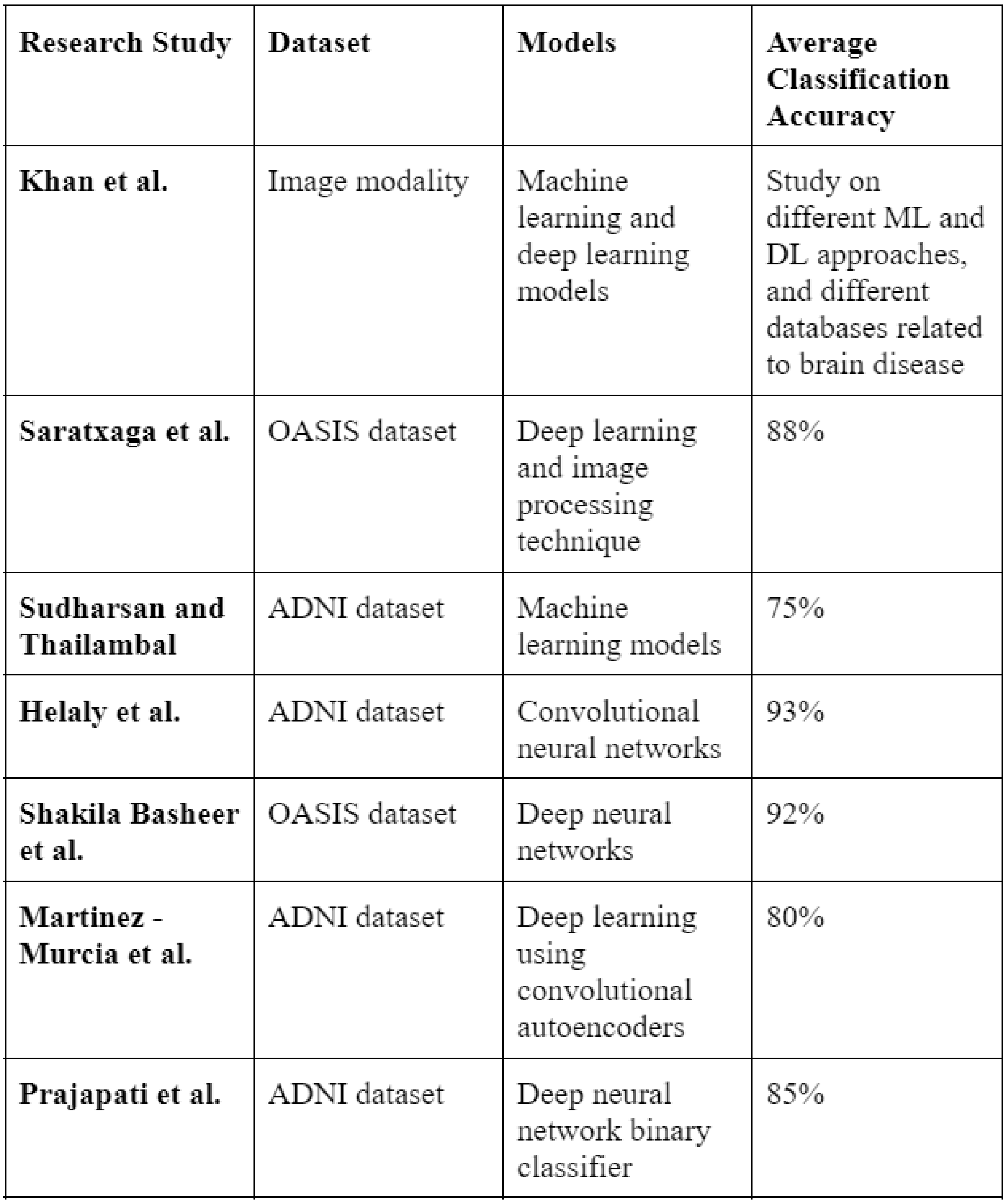
- To aid in the sMCI/pMCI classification, semi-supervised learning was used with data from AD patients and healthy controls without MCI patients.
- Used regularized logistic regression to select features.
- Before training the classifier, they removed aging effects from the MRI data to avoid confusion between changes caused by AD and normal aging.

- Finally, an aggregate biomarker was created by employing a random forest classifier to combine MCI subjects' age and cognitive measures with a separate MRI biomarker.

No matter what the effect of these elements at any phase of life, dispensing with them is useful. Early intervention and treatment of modifiable Alzheimer's risk factors, according to studies [6], can either delay or prevent 30 percent of Alzheimer's cases.

The Lifestyle for Brain Health (LIBRA) index is one method for calculating Alzheimer's risk based on risk factors, as stated by the Innovative Midlife Intervention for Alzheimer's Deterrence (In-MINDD) project [7]. The three main types of dementia intervention, according to the National Academy of Medicine, were cognitive training, controlling hypertension, and increasing physical activity. The most prevalent form of Alzheimer's disease is known as AD.

In their study, Tatiq and Barber[8] suggested that modifiable vascular risk factors could be used to prevent Alzheimer's. Williams and others used four different model to make predictions about functioning based on demographic and neuropsychological data: NN, Naive-Bayes, SVM, and Decision Tree.



Research Study	Dataset	Models	Average Classification Accuracy
Khan et al.	Image modality	Machine learning and deep learning models	Study on different ML and DL approaches, and different databases related to brain disease
Saratxaga et al.	OASIS dataset	Deep learning and image processing technique	88%
Sudharsan and Thailambal	ADNI dataset	Machine learning models	75%
Helaly et al.	ADNI dataset	Convolutional neural networks	93%
Shakila Basheer et al.	OASIS dataset	Deep neural networks	92%
Martinez - Murcia et al.	ADNI dataset	Deep learning using convolutional autoencoders	80%
Prajapati et al.	ADNI dataset	Deep neural network binary classifier	85%

Table 2. Recent work related to AD.

Chapter-3

SYSTEM DEVELOPMENT

This chapter describes the system development, including data preprocessing and model development.

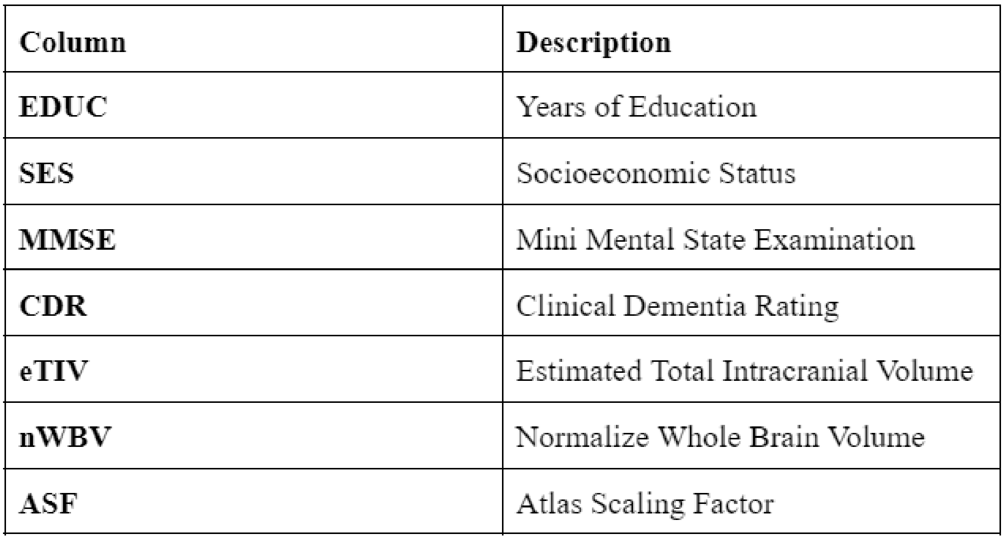
Data [9]

Summary :

A longitudinal collection of 150 subjects between the ages of 60 and 96 make up this set. For a total of 373 imaging sessions, each subject was scanned at least twice over the course of a year. Three or four individual T1-weighted MRI scans from one scan session are included for each subject. All of the subjects are male and female, and they are all right-handed. Throughout the course of the study, 72 of the participants were identified as not demented. 64 of the participants, including 51 with mild to moderate Alzheimer's disease, were identified as demented at their initial visits and remained so at subsequent scans. At the time of their initial visit, 14 subjects were categorized as nondemented, but at a subsequent visit, they were categorized as demented.

Dataset Description

- The longitudinal MRI data is used.
- 150 people between the ages of 60 and 96 provided longitudinal MRI data for the dataset.
- At least once was scanned for each subject.
- Everyone uses their right hand.
- During the entire research, a total of 72 participants were classified as "Nondemented."
- At their first visit, 14 subjects were classified as "Nondemented", but at subsequent visits, they were classified as "Demented". These are classified as "Converted".



Column	Description
EDUC	Years of Education
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

Table 3. Columns Description

Mini–Mental State Examination (MMSE) :

The Mini-Mental State Examination (MMSE) is a widely used tool for cognitive assessment that evaluates a person's cognitive functioning, including orientation, memory, attention, language, and visual-spatial skills. The test consists of a series of questions and tasks that are designed to assess cognitive function and is commonly used to screen for dementia and other cognitive impairments.

The MMSE typically takes around 10-15 minutes to complete and consists of a series of questions and tasks, such as asking the patient to repeat a series of words, to identify the current date and location, and to perform simple calculations.

The MMSE is not a definitive diagnostic tool for dementia or other cognitive impairments, but it is useful in screening for cognitive impairment and can help to identify individuals who may require further evaluation or treatment. It is important to note that the MMSE is just one tool used in the assessment of cognitive function and should be used in conjunction with other assessment measures and clinical judgment.

A 30-point questionnaire is frequently utilized in clinical and research settings to measure cognitive impairment. In medicine and other related fields, it is frequently used as a dementia screening tool. The MMSE has not been designed to diagnose any particular nosological entity on its own.

A normal cognitive state is generally indicated by a score of 24 points or higher on the MMSE. However, it is important to note that this is just a general guideline and that other factors, such as age and educational attainment, may need to be taken into account when interpreting scores. For example, a person who is older or has less formal education may score lower on the MMSE than someone who is younger or has a higher level of education, even if both individuals have similar levels of cognitive function.

Scores below 24 points on the MMSE may indicate cognitive impairment, with scores between 19 and 23 points indicating mild cognitive impairment, scores between 10 and 18 points indicating moderate cognitive impairment, and scores less than 9 points indicating severe cognitive impairment.

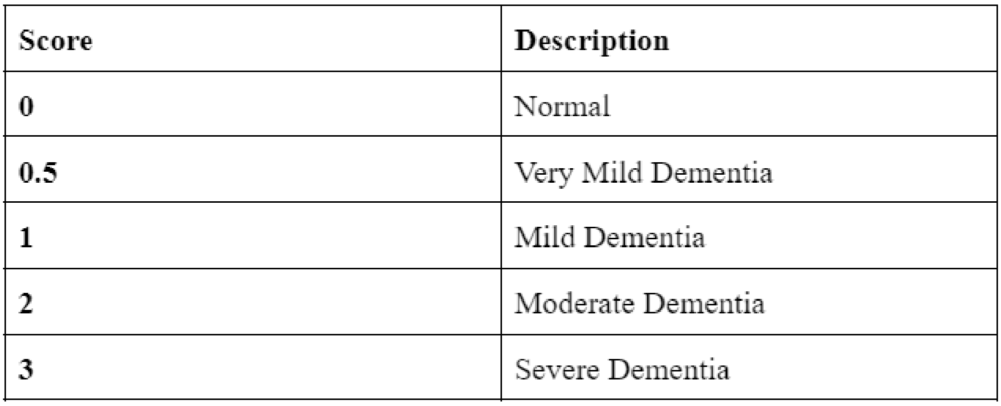
Clinical Dementia Rating (CDR) :

The Clinical Dementia Rating (CDR) is a 5-point scale used to evaluate the cognitive and functional performance of individuals with Alzheimer's disease and related dementias.

The memory domain assesses the patient's ability to recall past events and information. The orientation domain assesses the patient's awareness of their surroundings and ability to understand and respond to questions. The judgment and problem-solving domain assesses the patient's ability to make decisions and solve problems. The community affairs domain assesses the patient's ability to participate in social and community activities. The hobbies and home life domain assesses the patient's ability to engage in leisure activities and perform household tasks. The personal care domain assesses the patient's ability to take care of themselves, such as bathing and dressing.

Scores for each of the six domains are assigned on a scale ranging from 0 to 3, higher score meaning more significant cognitive and functional decline. A rating of 0 indicates no impairment, while a rating of 3 indicates severe impairment. The scores for each domain are combined to produce an overall score on the CDR, which ranges from 0 (no impairment) to 5 (severe impairment).

The descriptive anchors in the CDR table help the clinician use clinical judgment and interview data to make appropriate ratings. Notwithstanding appraisals for every space, an in general CDR score might be determined using a calculation.



Score	Description
0	Normal
0.5	Very Mild Dementia
1	Mild Dementia
2	Moderate Dementia
3	Severe Dementia

Table 4. CDR Table

Estimated Total Intracranial Volume (eTIV) :

eTIV is a measure of the total volume of the brain and cerebrospinal fluid (CSF) within the skull. It is an important measurement in neuroscience research and clinical practice, as it can help researchers and healthcare professionals to better understand brain structure and function, and to diagnose and treat neurological conditions.

eTIV is typically measured using magnetic resonance imaging (MRI) of the brain. The process involves using an MRI scanner to capture high-resolution images of the brain, which can then be analyzed using specialized software to calculate the total volume of the brain and CSF.

There are several different methods used to estimate eTIV from MRI scans, but they all involve some variation of the following steps:

- Segmentation: The MRI images are segmented into different tissue types, such as gray matter, white matter, and CSF, using image processing algorithms. This allows the software to differentiate between the different components of the brain.

- Registration: The segmented images are then registered or aligned with a standardized brain template. This helps to ensure that the different brain structures are accurately identified and measured across different individuals.

- Calculation: The software then calculates the total volume of the brain and CSF based on the segmented and registered images. This provides an estimate of eTIV for the individual being scanned.

eTIV is an important measurement in neuroscience research and clinical practice for several reasons. It can help to normalize other brain measurements, such as cortical thickness or gray matter volume, to account for differences in brain size across individuals. It can also be used to study the relationships between brain structure, function, and behavior, and to identify biomarkers for neurological conditions.

Overall, eTIV is an important measurement in neuroscience research and clinical practice, providing valuable information about brain structure and function that can help to improve our understanding and treatment of neurological conditions.

Atlas Scaling Factor (ASF) :

ASF is a normalization technique commonly used in neuroimaging studies to correct for differences in brain size between individuals. It is particularly useful in studies that compare brain structures across groups or populations, where differences in brain size can make it difficult to interpret the results.

The ASF is calculated by dividing the size of the individual's brain by the size of a reference brain, which is typically a standardized brain template. The size of the brain is usually estimated using a measure of brain volume, such as Estimated Total Intracranial Volume (eTIV), which is calculated from magnetic resonance imaging (MRI) scans.

The formula for calculating ASF is as follows:

ASF = (eTIV)^(1/3) / (reference brain size)^(1/3)

In this formula, the exponent of 1/3 is used to account for the fact that brain size scales with the cube of body size.

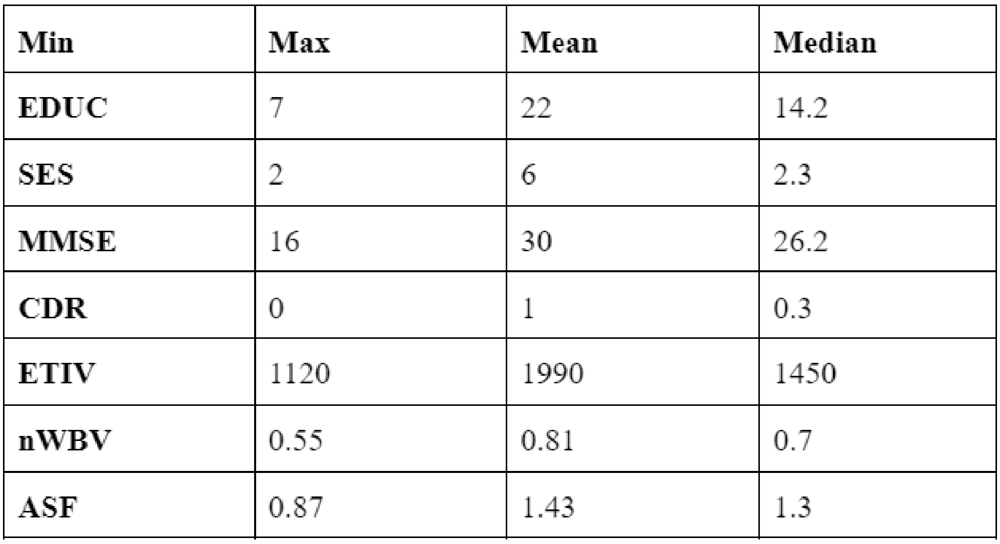
The ASF is then used to scale the size of brain structures in each individual's MRI scan to a common space, which is typically the reference brain template. This involves multiplying the size of each brain structure by the ASF, which adjusts for differences in brain size between individuals.

Preparation of Data

In this phase, the data were cleaned and preprocessed using a variety of data mining methods. Missing values are taken care of, features are taken out, features are changed, and so on. We found nine rows in the SES column with missing values. There are two approaches to this problem. Dropping the rows that don't have any values is the simplest solution. Given that we only have 140 measurements, we should be able to impute better results for the model. In the SES attribute, the nine rows with missing values are removed.

Analysis of data

Prior to data extraction and analysis, the Exploratory Data Analysis method was utilized to determine the relationships between data points through a graph. This information was subsequently employed to determine how best to analyze and interpret the data.



Min	Max	Mean	Median
EDUC	7	22	14.2
SES	2	6	2.3
MMSE	16	30	26.2
CDR	0	1	0.3
ETIV	1120	1990	1450
nWBV	0.55	0.81	0.7
ASF	0.87	1.43	1.3

Table 5. Dataset values

Feature Selection

Feature selection is a process in machine learning and statistical modeling that involves selecting a subset of the most relevant features (also known as variables, inputs, or predictors) from a larger set of available features. The objective of feature selection is to improve the accuracy and efficiency of the model by reducing the dimensionality of the dataset, and avoiding overfitting.

Feature selection can be done in two ways:

- Univariate Feature Selection: This approach selects features based on their individual relationship with the target variable. Univariate selection methods include statistical tests like Chi-Squared Test, ANOVA, t-tests, and F-tests, among others.

- Wrapper Feature Selection: This approach evaluates the performance of the model using different subsets of features. It involves training and evaluating the model on different subsets of features and selecting the subset that produces the best performance.

During the pre-processing stage, one commonly used technique is the filter method. Another method that cores the feature subset is wrapped methods. Finally, the wrapper and filter methods are combined in the embedded method.

Information gain, correlation coefficient and chi-square feature selection methods were selected because they are commonly used and widely recognized in the field.

Correlation Coefficient

To rephrase, the covariance of two variables X and Y is being referred to:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Correlation coefficient is a statistical measure that quantifies the degree of linear association between two variables. It is a number between -1 and 1, where -1 indicates a perfectly negative linear correlation, 0 indicates no linear correlation, and 1 indicates a perfectly positive linear correlation.

where x and y are the two variables, n - number of observations, Σxy is the sum, Σx and Σy are the sums of the values of x and y, Σx^2 and Σy^2 are the sums of the squares of the values of x and y.

The correlation coefficient measures the direction and strength of the linear relationship between two variables. If the correlation coefficient is positive, it indicates that the two variables have a positive linear relationship. If the correlation coefficient is negative, it indicates that the two variables have a negative linear relationship.

The correlation coefficient[10] has some important properties that make it useful in statistical analysis. Firstly, it is a standardized measure, which means that it is independent of the scale of measurement of the variables. Secondly, it measures only the linear relationship between two variables, and may not capture nonlinear relationships. Thirdly, the correlation coefficient is affected by outliers, which can distort the relationship between the variables.

Information gain

Information gain is based on the concept of entropy, which is a measure of the amount of uncertainty or randomness in a dataset. Entropy is calculated as follows:

$$H(X) = -\Sigma p(x)\log2p(x)$$

where X - random variable, p(x) - probability of occurrence of value x of the variable X, and log2 is the logarithm to the base 2.

Mathematically, it can be expressed as:

$$IG(D, F) = H(D) - \Sigma(|Dj|/|D|)H(Dj)$$

where IG is the information gain, D is the original dataset, F is the feature being considered for splitting, Dj is the subset of D resulting from splitting D using feature F, |Dj| is the number of instances in Dj, and |D| is the total number of instances in D.

The information gain of a feature measures the reduction in entropy achieved by splitting the dataset using that feature. A higher information gain indicates that splitting the dataset using that feature results in more distinct and homogeneous subsets, and therefore, the feature is more informative for the decision tree algorithm.

Chi-Square

Chi-square (χ^2) is a statistical test used to determine the independence of two categorical variables. The test is based on the difference between the observed frequency and the expected frequency of each category.

The chi-square test statistic is calculated as follows:

$$\chi^2 = \Sigma (O - E)^2 / E$$

where O is the observed frequency, E is the expected frequency, and the sum is taken over all categories.

The expected frequency is calculated by multiplying the total number of observations by the probability of each category. If the two variables being analyzed are independent, then the expected frequency of each category can be calculated using the formula:

E = (row total x column total) / grand total

where the row total is the total number of observations in a particular row, the column total is the total number of observations in a particular column, and the grand total is the total number of observations in the entire dataset.

The chi-square test[11] statistic follows a chi-square distribution with (r-1) x (c-1) degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table. The p-value associated with the test statistic is calculated using the chi-square distribution table or statistical software.

If the p-value is less than the level of significance (usually 0.05), then the null hypothesis of independence is rejected, indicating that the two variables are dependent.

The chi-square test can be used for various applications, such as testing the relationship between gender and political affiliation, or analyzing the association between smoking and lung cancer. It can also be used in feature selection, where features with a low chi-square value are eliminated from the dataset as they are not significantly related to the target variable.

We can use this strategy to investigate categorical variables like the connection between food and obesity.

$$Chi - Square \; = \; \frac{(Observed - Expected)^2}{expected}$$

Preparation and Splitting of Data

Preparing and splitting data[12] is a crucial step in machine learning (ML) that involves dividing the dataset into training, validation, and test sets to enable the ML model to learn and generalize better. The process of preparing and splitting data involves the following steps:

- Data Collection: The first step is to collect the data from various sources, such as databases, spreadsheets, web scraping, or data feeds.
- Data Cleaning: Once the data is collected, it needs to be cleaned to remove any missing, duplicate, or irrelevant data. This step is important to ensure the accuracy and integrity of the data.
- Data Transformation: The next step is to transform the data into a format that can be used for training the ML model. This step involves converting the data into numerical or categorical values, standardizing or normalizing the data, and encoding categorical variables using techniques such as one-hot encoding.
- Feature Selection: Feature selection is the process of selecting the most relevant features that can improve the performance of the ML model. This step involves selecting features based on their relevance to the

target variable, removing redundant features, and transforming the features to improve their quality.

- Data Splitting: Once the data is prepared, it needs to be split into training, validation, and test sets. The training set is used to train the ML model, the validation set is used to optimize the hyperparameters of the model, and the test set is used to evaluate the performance of the model.
- Cross-Validation: Cross-validation is a technique used to assess the performance of the ML model by partitioning the data into subsets and using each subset for testing and the remaining subsets for training. Cross-validation helps to prevent overfitting and ensures that the model can generalize well to new data.
- Data Augmentation: Data augmentation is a technique used to increase the size and diversity of the training dataset by applying various transformations to the existing data, such as cropping, flipping, rotating, or adding noise. Data augmentation helps to improve the performance and robustness of the ML model.

In summary, preparing and splitting data is a critical step in ML that involves cleaning, transforming, and selecting relevant features from the data, as well as dividing the data into training, validation, and test sets

Figure 2. describes the data splitting and preparation.

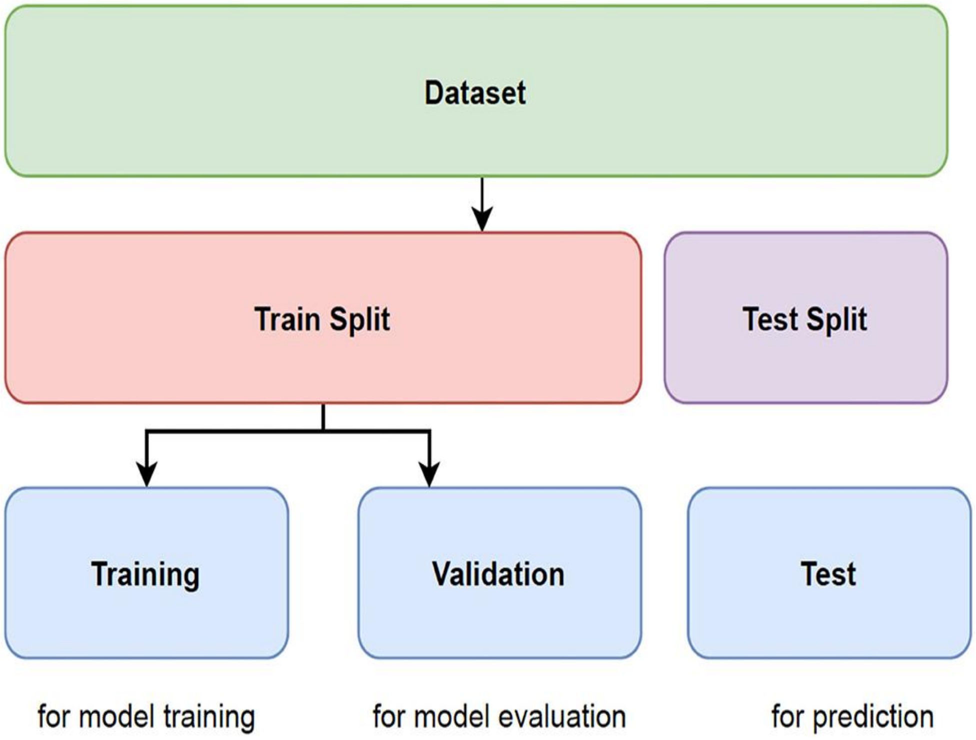


Figure 2. Data Splitting and Preparation

Classifier Models:

Decision Tree (DT)

The decision tree is a machine learning algorithm that is used for supervised learning tasks, such as classification and regression. It is a graphical representation of all the possible decisions and outcomes that can be made based on a series of input features. Decision trees are commonly used in business, finance, medicine, and many other fields where decision-making is crucial.

The decision tree algorithm functions through the iterative division of data based on input features and their respective values. In the decision tree algorithm, the feature that provides the highest information gain or the best split based on a selected criterion (such as Gini impurity or entropy) is chosen at each node of the tree.

Decision trees have several advantages, including their interpretability, ability to handle both categorical and numerical data, and their robustness to outliers and missing data. However, decision trees are prone to overfitting, especially when the tree is too deep or too complex. To prevent overfitting, pruning techniques such as reduced error pruning or cost-complexity pruning can be used.

The decision tree algorithm uses different formulas to measure the impurity of a node and to calculate the information gain for each feature. The two most commonly used formulas are:

- Gini Impurity:

  Gini Impurity is a measure of the probability of misclassification if a random sample is assigned to the wrong category based on the distribution of labels in a given node.

$$\text{Gini impurity} = 1 - \sum(pi)\hat{}2$$

where pi is the proportion of instances in the ith class label in a node.

- Entropy:

  Entropy quantifies the degree of ambiguity or randomness present in a dataset. It is used to calculate the information gain of a feature, which is the reduction in entropy when a dataset is split on a particular feature.

$$\text{Entropy} = - \sum(pi)\log2(pi)$$

where pi is the proportion of instances in the ith class label in a node.

Advantages :

- Rules that are easy to understand can be generated by decision trees.
- Classification is performed by decision trees without requiring a lot of computation.
- Decision trees can deal with categorical as well as continuous variables.
- Using decision trees, it is easy to see which fields are most crucial for classification or prediction.

Disadvantages :

- When it comes to prediction tasks involving the estimation of a continuous attribute, decision trees are not as effective.
- When faced with classification problems that feature a significant number of classes and a comparatively limited number of training examples, decision trees are susceptible to making mistakes.
- Training a decision tree can be time-consuming and costly.Growing a decision tree takes a lot of computational power.Before the best split can be determined, each candidate splitting field at each node must be sorted.Field combinations are used in some algorithms, and the best combining weights must be sought.Due to the need to create and compare a large number of candidate sub-trees, pruning algorithms can also be costly.

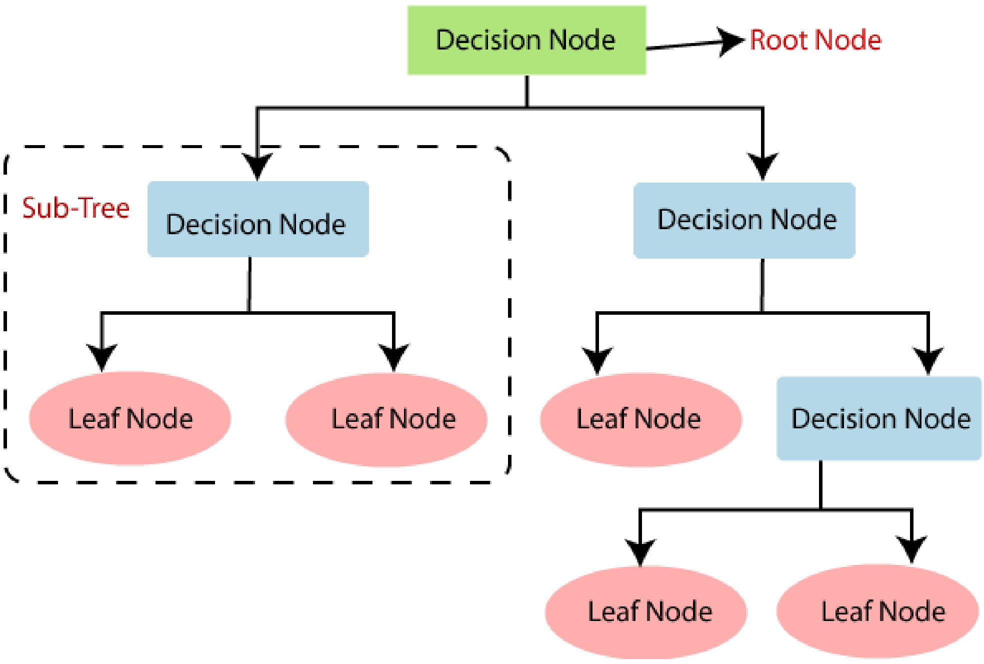


Figure 3. General Structure of DT

Random Forest

Random forests, which are a form of ensemble learning, can be utilized for tasks such as classification, regression, and others.. It works by building a lot of decision trees during training. For arrangement undertakings, the result of the irregular timberland is the class chosen by most trees. The individual trees' mean or average prediction is returned for regression tasks. Though their accuracy is lower than that of gradient-boosted trees, random forests generally perform better than decision trees. However, their performance can be affected by data characteristics.

Decision trees of numerous types, each slightly different from the others comprise the models based on random forests. The ensemble technique combines the predictions of individual decision tree models by employing the majority voting method to generate forecasts. As a result, less overfitting occurs while still keeping each tree's capacity for prediction.

The basic steps of the Random Forest algorithm are:

- Data Preparation: The dataset is first prepared by cleaning, preprocessing, and transforming the data to a suitable format for model training.
- Random Sampling: A random sample of the dataset is taken for each tree in the forest. The size of the sample can be controlled by hyperparameters.
- Decision Tree Creation: For each sample, a decision tree is created using a subset of features. The features are selected at random at each node of the tree.

The Random Forest algorithm offers several benefits in comparison to a single decision tree:

- Random Forest can handle missing values, outliers, and noisy data because it uses multiple decision trees.
- Random Forest can deal with high-dimensional datasets because it randomly selects a subset of features for each decision tree.
- Random Forest is easy to interpret and visualize because it is based on decision trees.

The general approach known as bootstrap aggregating, or bagging, is utilized in the random forest training algorithm for tree learners. Given a training set -> X = xl,x2,..., xn with responses Y = y1, y2,..., yn, Bagging selects a random sample that replaces the training set and fits trees to these samples:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Additionally, the standard deviation of the predictions from each individual regression tree on x' can be used to estimate the prediction's uncertainty:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f}\,)^2}{B-1}}.$$

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a frequently used supervised learning technique that can be utilized for both regression and classification problems, though its primary usage is for classification.

Its goal is to identify the most effective decision boundary or line for classifying n-dimensional space, enabling quick categorization of new data points in the future. The best decision boundary is represented by a hyperplane, which is created using extreme points and vectors selected by the SVM.

These extreme points are known as support vectors, giving the algorithm its name. The diagram below illustrates the separation of two distinct categories by a decision boundary or hyperplane:

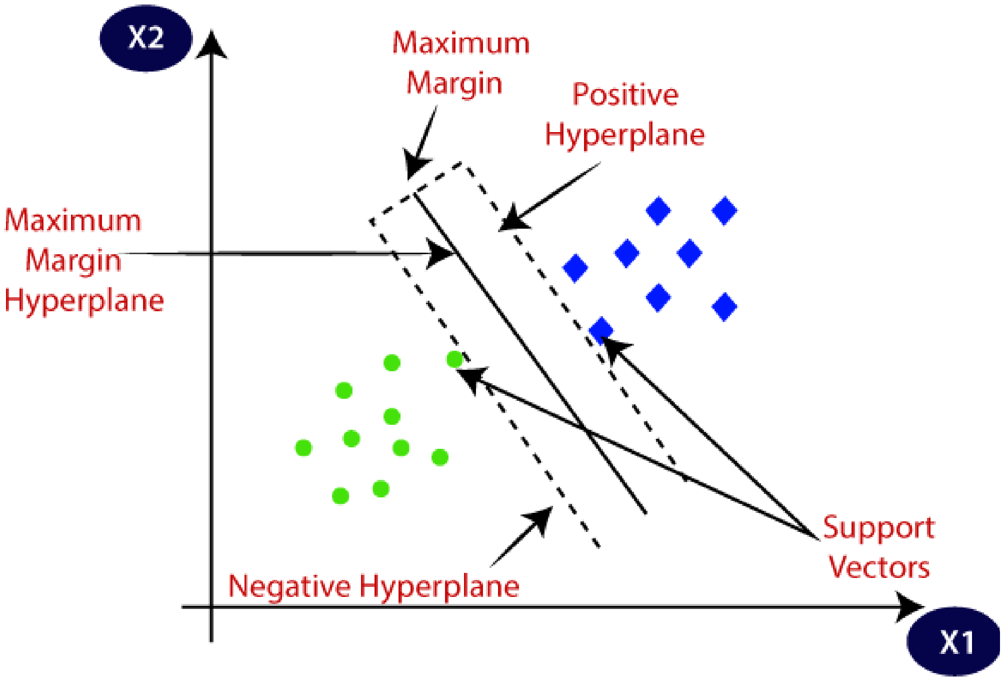


Figure 4. Hyperplane

The basic steps of the SVM algorithm are:

- Data Preparation: The dataset is first prepared by cleaning, preprocessing, and transforming the data to a suitable format for model training.
- Hyperplane Selection: A hyperplane is chosen in such a way that the distance between the classes is maximized. In the case of linearly separable data, the hyperplane is a line that separates the two classes. For nonlinear data, SVM uses a kernel trick to transform the data into a higher-dimensional space where it becomes linearly separable.
- Margin Calculation: The goal is to maximize the margin, which makes the decision boundary more robust.
- Optimization: The SVM algorithm finds the optimal hyperplane by minimizing the classification error and maximizing the margin. This is done using a convex optimization function.

The SVM algorithm has several advantages over other classification algorithms:

- SVM is less prone to overfitting because it maximizes the margin between the classes, which reduces the variance of the model.
- SVM can handle datasets with high-dimensional features because it only needs to consider the support vectors to make predictions.

Types :

- **Linear SVM** : Linear SVM is used for linearly separable data, which means that a classifier is called a Linear SVM classifier if a dataset can be divided into two classes by drawing a single straight line between them.
- **Non-linear SVM** : In cases where a dataset cannot be separated by a linear boundary, it is classified as non-linear data and a non-linear SVM classifier is employed. This method is applied for data that is not linearly separable.

XGBoost

XGBoost represents Outrageous Angle Helping, which was proposed by the analysts at the College of Washington. It is a library written in C++ that helps Gradient Boosting training run more smoothly.

Decision trees are constructed sequentially using this algorithm. In XGBoost, weights play a significant role. All independent variables receive weights, which are then fed into a decision tree that predicts outcomes. The second decision tree is fed the variables that the first tree incorrectly predicted by increasing the weight of those variables. The ensemble of these individual classifiers and predictors produces a robust and more precise model. Regression, classification, ranking, and user-defined prediction problems are all possible applications.

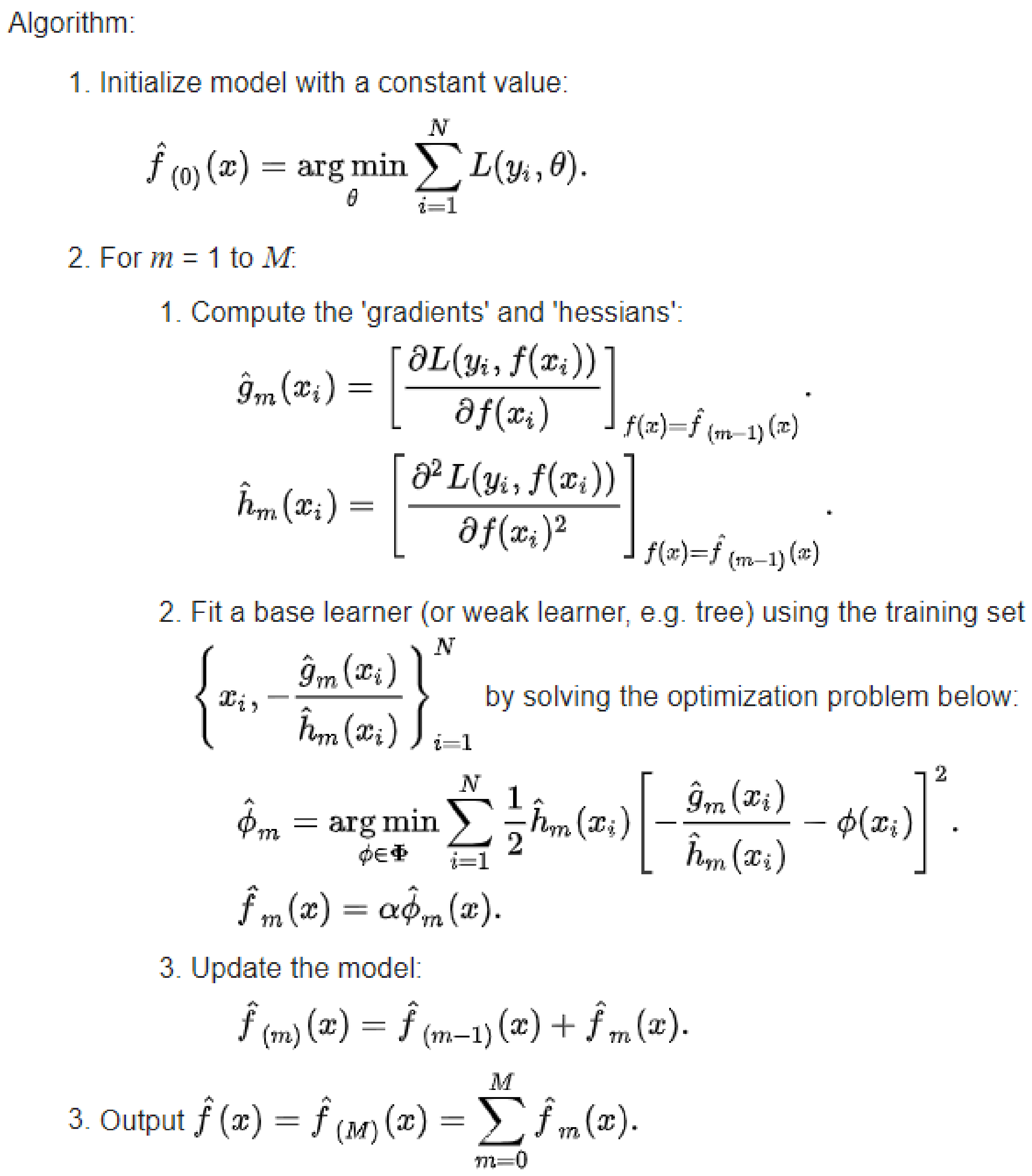
In contrast to gradient boosting, which functions as gradient descent in function space, XGBoost uses a second-order Taylor approximation in the loss function to connect to the Newton-Raphson method.

The basic steps of the XGBoost algorithm are:

- Data Preparation: The dataset is first prepared by cleaning, preprocessing, and transforming the data to a suitable format for model training.
- Initial Model: A simple model is first created, such as a decision tree, that can make predictions on the data.
- Tree Building: A fresh decision tree is constructed to estimate residuals by utilizing the gradient and Hessian.
- Model Update: The predictions of all the decision trees are combined to create the final prediction. The weights of the decision trees are adjusted to minimize the loss function using a process called gradient descent.

XGBoost has several advantages over other gradient boosting algorithms:

- XGBoost is highly scalable and efficient because it uses a parallelized tree boosting algorithm that can handle large datasets and high-dimensional features.
- XGBoost can handle missing values and imbalanced classes by using a technique called gradient-based sampling.
- XGBoost can automatically handle regularization by using L1 and L2 regularization to prevent overfitting.



Algorithm:

1. Initialize model with a constant value:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta).$$

2. For m = 1 to M:

    1. Compute the 'gradients' and 'hessians':

    $$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

    $$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

    2. Fit a base learner (or weak learner, e.g. tree) using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ by solving the optimization problem below:

    $$\hat{\phi}_m = \arg\min_{\phi\in\Phi} \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2.$$

    $$\hat{f}_m(x) = \alpha\hat{\phi}_m(x).$$

    3. Update the model:

    $$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x).$$

3. Output $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x).$

Voting

A voting classifier is a type of ensemble learning algorithm in machine learning that combines multiple individual models to make a final prediction. The idea behind a voting classifier is to combine the predictions of several different models to make a more accurate prediction than any single model would be able to make on its own.

There are two main types of voting classifiers: hard voting and soft voting.

- Hard voting:
  Hard voting involves aggregating the predictions of each individual model and selecting the class that receives the highest number of votes as the final prediction. For example, if we have three individual models that predict the classes A, B, and A, then the hard voting classifier would predict Class A as it received the most votes.

Hard voting is useful when the individual models are similar in their accuracy and have the same level of importance in the final prediction. It can also help to reduce the impact of individual model errors by taking into account the opinions of multiple models.

- Soft voting:
  In soft voting, the final prediction is made by taking the average probability of the individual models. This means that the class with the highest average probability among the individual models is chosen as the final prediction. For example, if we have three individual models that predict the probabilities of the classes A and B as follows:
  Model 1: A (0.8), B (0.2)
  Model 2: A (0.7), B (0.3)
  Model 3: A (0.6), B (0.4)

  Then the soft voting classifier would predict Class A as the final prediction, as the average probability of A is (0.8 + 0.7 + 0.6) / 3 = 0.7, which is higher than the average probability of B.

Soft voting is useful when the individual models produce probability estimates, and the averaging process can help to improve the accuracy of the final prediction. It can also be more effective when the individual models are diverse in their accuracy, as it takes into account the strength of each model's prediction.

Some popular types of models that can be combined in a voting classifier include decision trees, random forests, support vector machines (SVMs), and neural networks. It is important to choose models that are diverse and have different strengths and weaknesses, so that they can complement each other and produce a more accurate final prediction.

There are also different strategies for selecting the individual models to be used in the voting classifier, such as selecting models that have the highest individual accuracy or selecting models that are the most diverse.

The objective is to construct a single model that is trained on these models and generates outputs based on their collective majority of votes for each output class, instead of developing separate models and evaluating their performance individually.

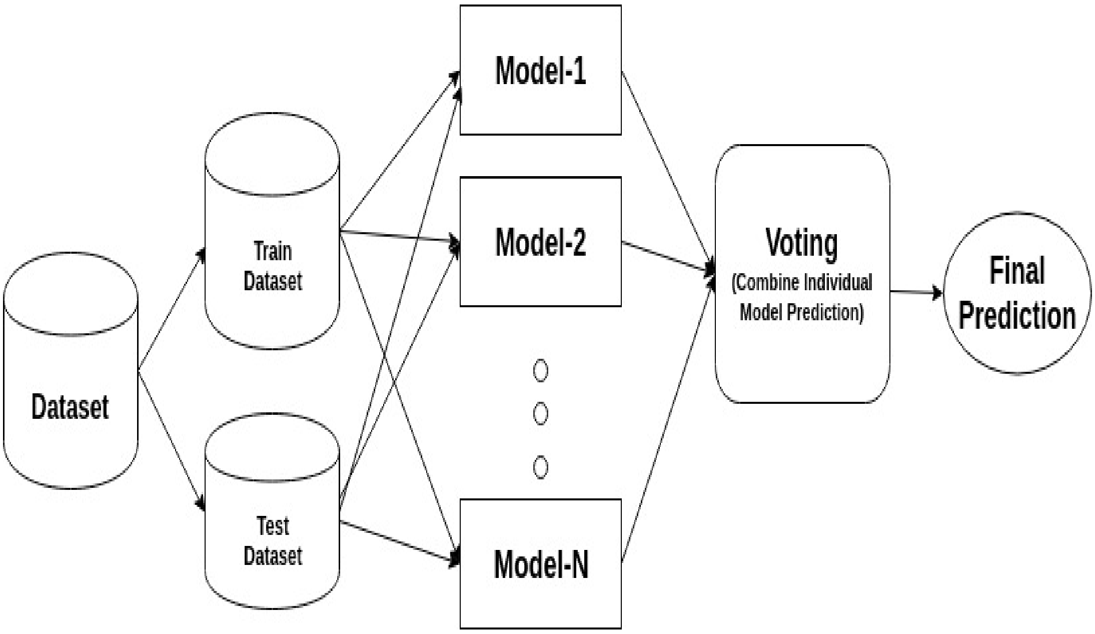


Figure 5. Voting Classifier

ExtraTrees

Extra Trees, also known as Extremely Randomized Trees, is an ensemble learning approach that shares similarities with Random Forest. Extra Trees, like Random Forest, merge the outputs of multiple decision trees to make a final prediction. Nonetheless, there are notable distinctions between the two methods.

The main difference between Extra Trees and Random Forest is the way that the trees are built. In Random Forest, the decision trees are built using a bootstrap sample of the training data, and the best split is selected for each node based on a random subset of the features. In Extra Trees, the decision trees are also built using a bootstrap sample of the training data, but the splits are chosen at random from a predefined set of values, rather than finding the best split. This means that Extra Trees is even more random and less biased than Random Forest, as it does not rely on any specific algorithm to choose the best split.

Extra Trees has several advantages over other ensemble learning algorithms, including:

- Reduced Variance: Because Extra Trees combines the predictions of multiple decision trees, it has lower variance.
- Low Bias: Because Extra Trees selects splits at random from a predefined set of values, rather than finding the best split, it has lower bias than other decision tree algorithms.
- Fast Training: Because the trees in Extra Trees are built independently, the algorithm can be parallelized and can be trained faster than other ensemble learning algorithms. This makes it well-suited for large datasets or problems with many input features.
- Feature Importance: Like other decision tree algorithms, Extra Trees can also be used to estimate the importance of each input feature.

In summary :

- By default, it samples without replacing and builds multiple trees with bootstrap = False.
- A random subset of the features selected at each node is used to split the nodes.

Randomness in Extra Trees is generated by the random splits of all observations rather than by bootstrapping the data.

The Extra Trees algorithm offers a faster execution time and lower computation costs compared to other algorithms. This algorithm saves time because the entire process is the same, but it does not calculate the best split point because it chooses it at random.

Gradient Boosting

Gradient Boosting is a popular machine learning algorithm that is widely used for both regression and classification problems. The Extra Trees algorithm employs an ensemble approach, which involves combining several weaker models to form a more powerful model that can make more precise predictions. The algorithm works by iteratively improving the predictions of a model by minimizing the loss function.

Here is how Gradient Boosting works:

- Initialize the model: The algorithm starts by initializing the model with a single decision tree, which is trained on the training data.
- Predict: The model predicts the outcomes for the training data, and then measures the discrepancy between these predicted values and the true values. This difference is called the residual.
- Train a weak model on the residuals: A new decision tree is trained on the residuals, rather than the original target variable. This decision tree

is known as a weak model because it does not capture all the underlying patterns in the data.

- Repeat steps 2-4: Steps 2-4 are repeated until the desired number of weak models has been trained, or until the loss function is minimized.

Some advantages of Gradient Boosting include its ability to handle a wide range of data types, and its ability to handle missing data. However, Gradient Boosting can be computationally expensive and may require tuning of many hyperparameters to achieve optimal performance.

This algorithm's main idea is to build models in order, with each model trying to reduce errors from the previous one. But how do we achieve that? How can we cut down on errors? On the basis of the errors or residuals of the previous model, a new model is constructed for this purpose.

Gradient Boosting Classifier is used when the target column is a classification problem, while Gradient Boosting Regressor is used when the target column is continuous. The "Loss function" is the only thing that separates the two. By utilizing gradient descent, weak learners will be added with the intention of minimizing this loss function. Since it is based on a loss function, we will use different loss functions for regression problems like Mean squared error (MSE) and classification problems like log-likelihood.

Loss function :

$$L = -\sum_{i=1}^{n} y_i \log(p) + (1-p)\log(1-p)$$

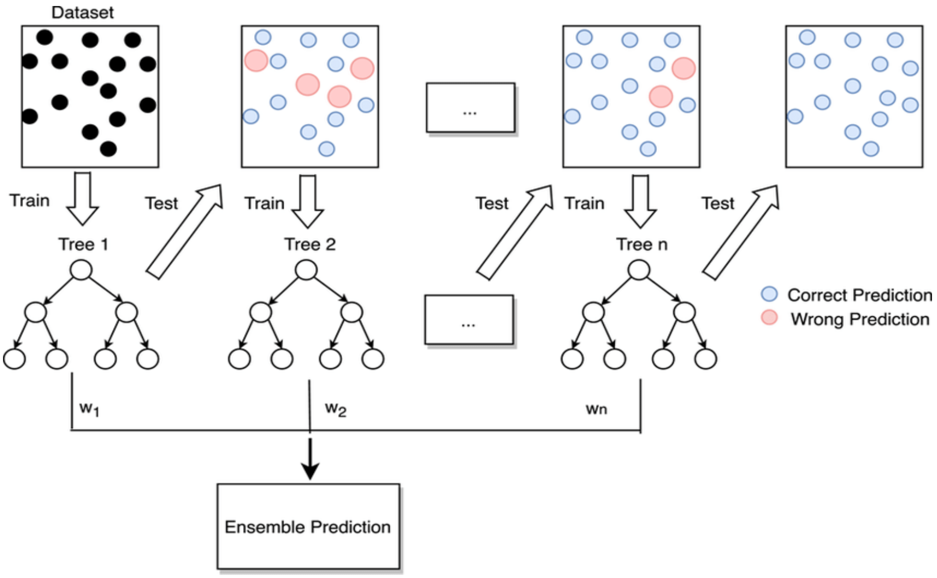


Figure 6. Gradient Boosting Classifier

AdaBoost

AdaBoost is a commonly used technique in ensemble learning, which merges multiple weak models to create a robust model that can provide accurate predictions. The algorithm works by iteratively training a sequence of weak models on weighted versions of the data, and then combining their predictions to make a final prediction.

Here is how AdaBoost works:

- Initialize the weights: Each data point in the training set is given an equal weight.
- Train a weak model: A weak model is trained on the weighted training data, where the weights are used to emphasize the importance of the misclassified samples from previous iterations.
- Repeat steps 2-4: Steps 2-4 are repeated until the desired number of weak models has been trained, or until the error rate has converged to a satisfactory level.

- Combine the weak models: The forecasts made by the weak models are integrated by using a weighted average method, where the weights are determined based on the accuracy of each model.

Some advantages of AdaBoost include its ability to handle a wide range of data types and its ability to handle noisy data. However, AdaBoost can be sensitive to outliers and may require careful tuning of the hyperparameters to achieve optimal performance.

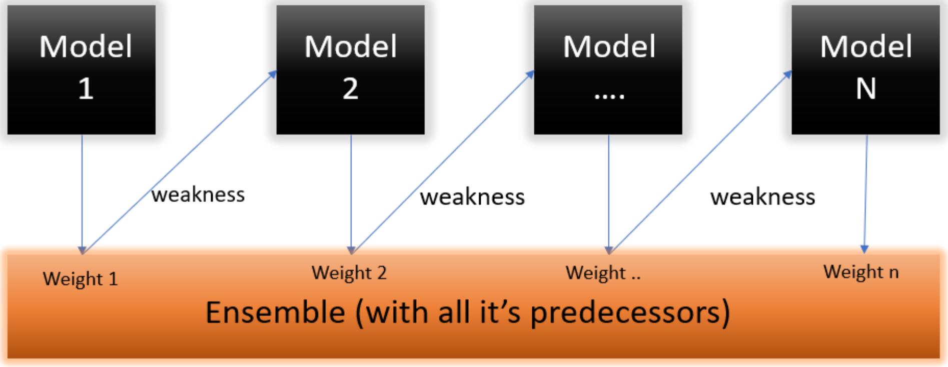


Figure 7. AdaBoost

AdaBoost is referred to as adaptive due to its ability to assign a higher priority to instances that were incorrectly classified by prior weak learners in subsequent iterations. It may be less susceptible to the overfitting issue in some situations than other learning algorithms. The final model can be shown to converge to a strong learner if the performance of each individual learner is slightly better than random guessing. The individual learners may be weak.

Neural Networks

Neural networks are a type of machine learning model that are designed to emulate the functionality and structure of the human brain. These models are composed of numerous layers of artificial neurons, which can identify patterns in data and make predictions or decisions based on those patterns.

Each neuron in a neural network is typically represented by a mathematical function that takes in one or more inputs and produces an output.

The weights in the neural network determine the strength of the connections between neurons, and are typically learned from data using an optimization algorithm such as gradient descent.

There are many different types of neural networks. Some common types include feedforward neural networks, CNN’s, RNN’s and autoencoders. The selection of the appropriate architecture is a crucial aspect of constructing an effective neural network as each type of network is appropriate for different data types and tasks.

Neural networks can be mathematically represented using linear algebra and calculus. Here is a simplified mathematical representation of a feedforward neural network:

Let X be the input to the network, and Y be the output. The network has L layers, with layer l having N_l neurons. Each neuron j in layer l is represented by a vector of weights w_lj and a bias b_lj.

The output of a neuron j in layer l is given by:

a_lj = f_l(Σ_i(w_lj,i * a_l-1,i) + b_lj)

where Σ_i is the sum over all inputs i to neuron j in layer l1, and f_l is the activation function for layer l.

The output of layer l is a vector A_l, where A_lj is the output of neuron j in layer l.

The output of the network is given by:

Y = f_L(Σ_i(w_L,i * A_L-1,i) + b_L)

where Σ_i is the sum over all inputs i to neuron j in the final layer L, and f_L is the activation function for the final layer.

Chapter 4

EXPERIMENTS AND RESULT ANALYSIS

We look at accuracy, precision, recall, and the F1 score as performance metrics. We carry out five fold cross-validation in order to identify the ideal parameters for each model: SVM, Random Forests, Decision Trees, XGBoost, Voting, ExtraTrees, AdaBoost, gradient boosting and neural network. Finally, we compare each model's accuracy. After the models were made, a number of metrics and methods were used to find issues with overfitting and parameter tuning. The confusion matrix is used to describe both binary and multiclass performance evaluations. A novel Machine Learning classifier was developed and validated to predict and distinguish true Alzheimer's disease patients from a given population, and a learning model was developed to distinguish true Alzheimer's disease patients from a given population. Using these components, the following evaluation metrics were calculated: accuracy, ROC, and confusion matrix is the proportion of people accurately identified as having Alzheimer's disease, as determined by this study. The proportion of people correctly classified as not having Alzheimer's is the accuracy of the diagnosis. Alternately, accuracy is the proportion of people correctly classified, while F1 is the weighted average of recall and precision. As per the outcomes, the patient gets a report that lets that person know the phase of Alzheimer's Illness the individual presently in. It is vital to distinguish the stages on the grounds that the stages depend on the reactions of the patients. Additionally, doctors benefit from having a better understanding of the stage of the disease.

For the purposes of experiments and analysis, this study made use of the following environments, tools, and libraries:

- Python 3
- Jupyter Notebook
- Scikit-learn libraries

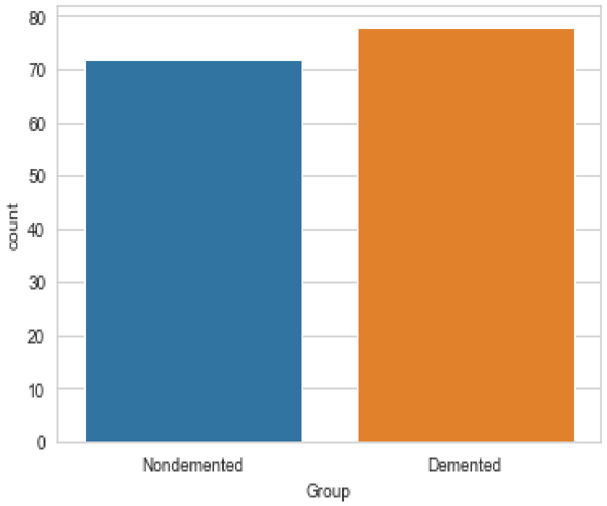


Figure 8. Non demented and demented groups

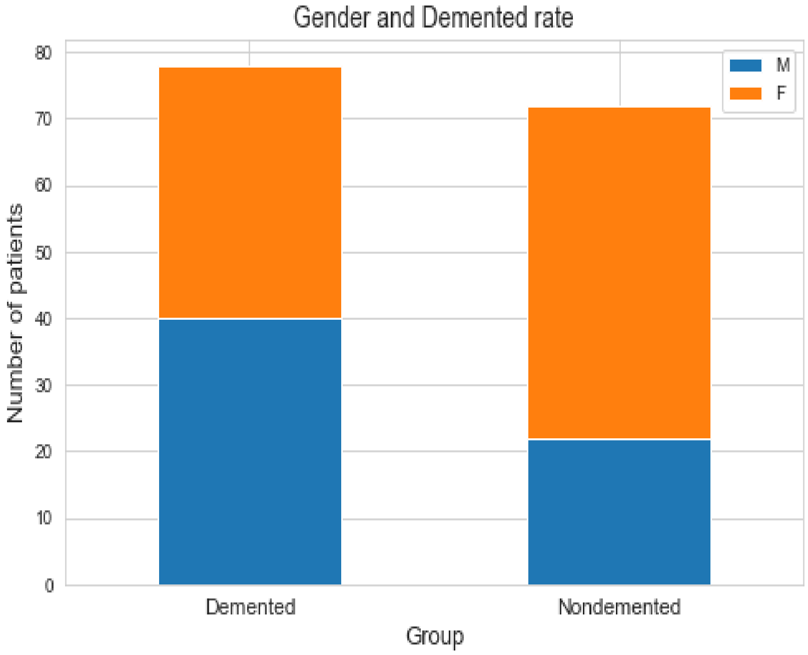


Figure 9. An examination of the prevalence of dementia among individuals based on their gender.

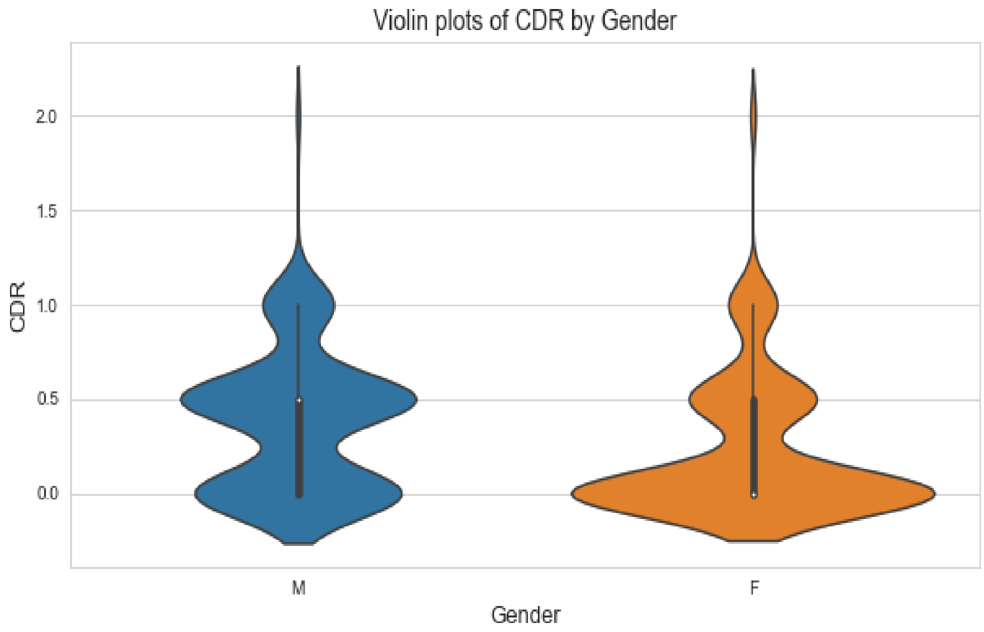


Figure 10. CDR By Gender

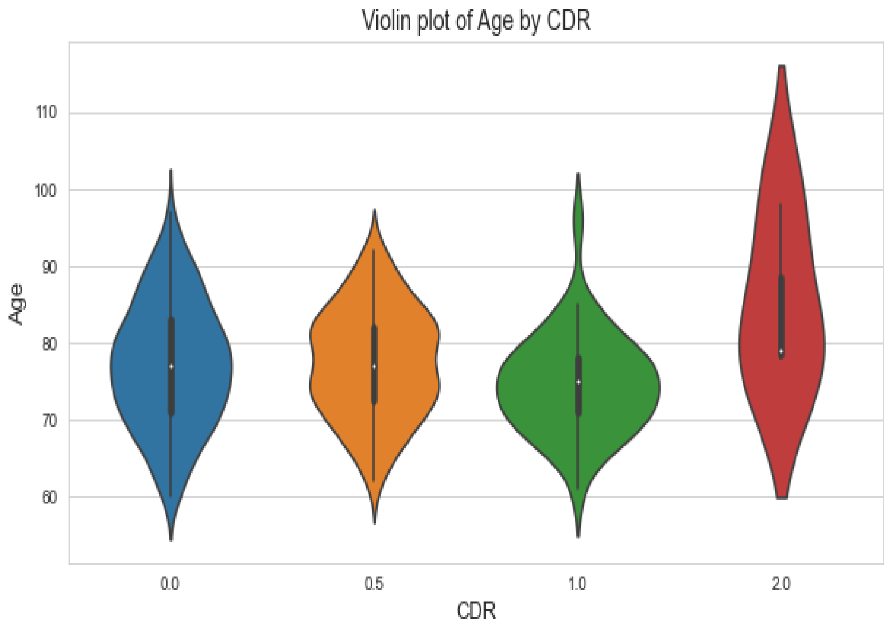


Figure 11. CDR By Age

Figure 11 illustrates the analyzed data of ASF, eTIV, and nWBV for both the Demented and Non-Demented groups. The Non-Demented group has a greater ratio of brain volume compared to the Demented group, which is a result of the diseases causing the shrinkage of brain tissues. Figure 12 displays the impact of EDUC on Demented and Non-Demented individuals.

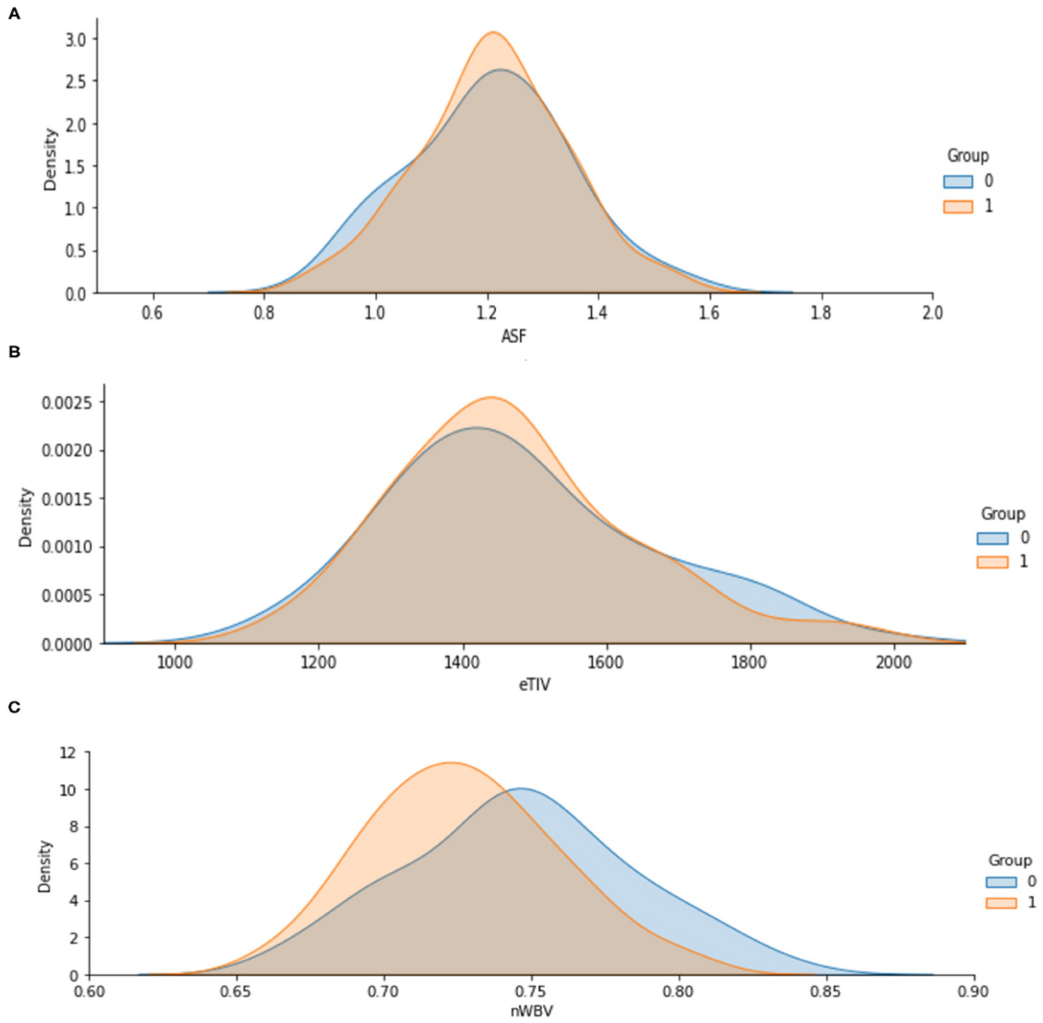


Figure 12. ASF, eTIV, nWBV

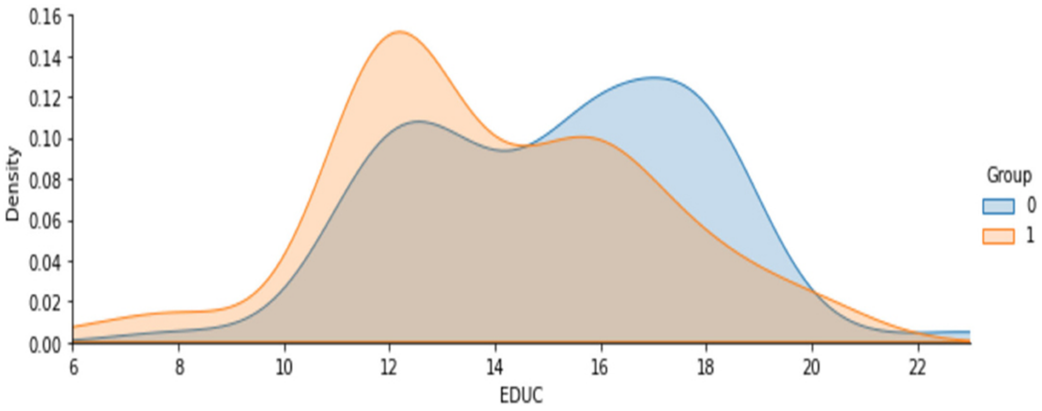
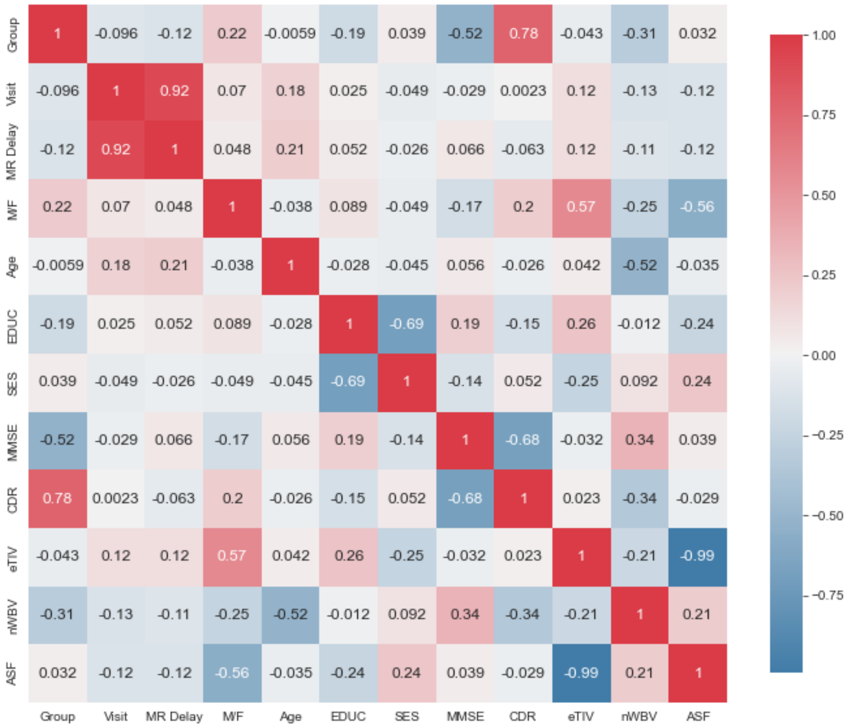


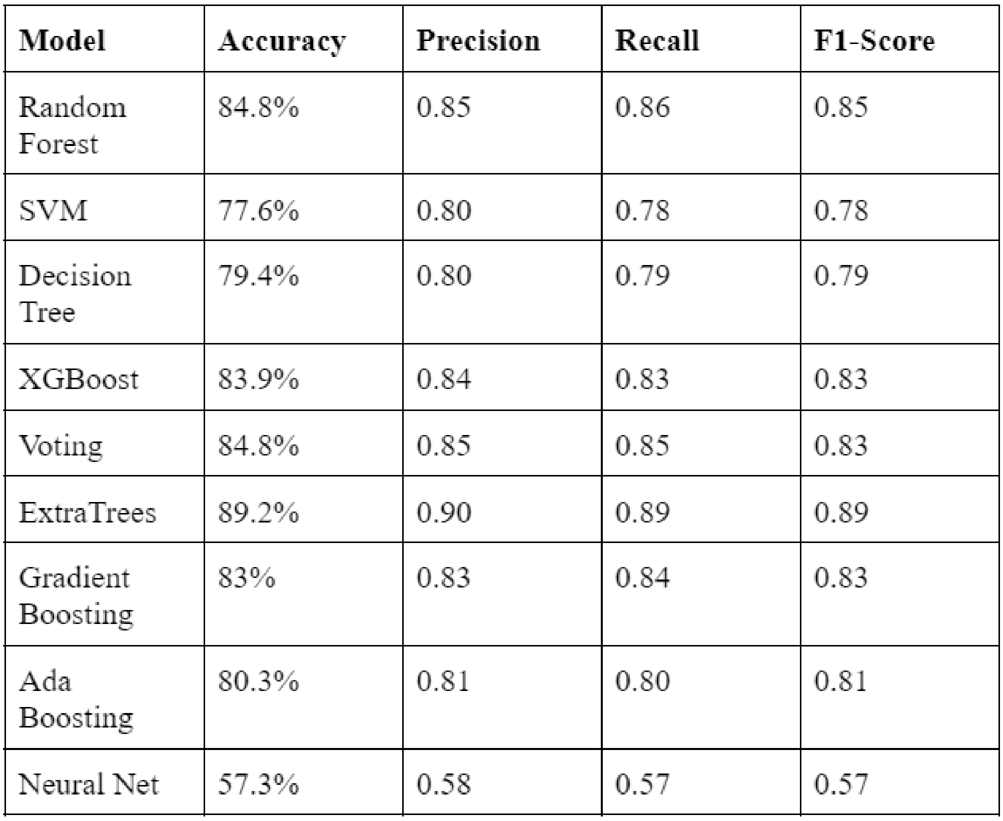
Figure 13. EDUC analysis



	Group	Visit	MR Delay	M/F	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
Group	1	-0.096	-0.12	0.22	-0.0059	-0.19	0.039	-0.52	0.78	-0.043	-0.31	0.032
Visit	-0.096	1	0.92	0.07	0.18	0.025	-0.049	-0.029	0.0023	0.12	-0.13	-0.12
MR Delay	-0.12	0.92	1	0.048	0.21	0.052	-0.026	0.066	-0.063	0.12	-0.11	-0.12
M/F	0.22	0.07	0.048	1	-0.038	0.089	-0.049	-0.17	0.2	0.57	-0.25	-0.56
Age	-0.0059	0.18	0.21	-0.038	1	-0.028	-0.045	0.056	-0.026	0.042	-0.52	-0.035
EDUC	-0.19	0.025	0.052	0.089	-0.028	1	-0.69	0.19	-0.15	0.26	-0.012	-0.24
SES	0.039	-0.049	-0.026	-0.049	-0.045	-0.69	1	-0.14	0.052	-0.25	0.092	0.24
MMSE	-0.52	-0.029	0.066	-0.17	0.056	0.19	-0.14	1	-0.68	-0.032	0.34	0.039
CDR	0.78	0.0023	-0.063	0.2	-0.026	-0.15	0.052	-0.68	1	0.023	-0.34	-0.029
eTIV	-0.043	0.12	0.12	0.57	0.042	0.26	-0.25	-0.032	0.023	1	-0.21	-0.99
nWBV	-0.31	-0.13	-0.11	-0.25	-0.52	-0.012	0.092	0.34	-0.34	-0.21	1	0.21
ASF	0.032	-0.12	-0.12	-0.56	-0.035	-0.24	0.24	0.039	-0.029	-0.99	0.21	1

Figure 14. Correlation Map

Table 4 compares the performance of various ML models in terms of accuracy, precision, recall, and F1 score. The following is a definition of the performance measures:



Model	Accuracy	Precision	Recall	F1-Score
Random Forest	84.8%	0.85	0.86	0.85
SVM	77.6%	0.80	0.78	0.78
Decision Tree	79.4%	0.80	0.79	0.79
XGBoost	83.9%	0.84	0.83	0.83
Voting	84.8%	0.85	0.85	0.83
ExtraTrees	89.2%	0.90	0.89	0.89
Gradient Boosting	83%	0.83	0.84	0.83
Ada Boosting	80.3%	0.81	0.80	0.81
Neural Net	57.3%	0.58	0.57	0.57

Table 6. Performance Analysis

Why did the neural networks give the worst accuracy among others?

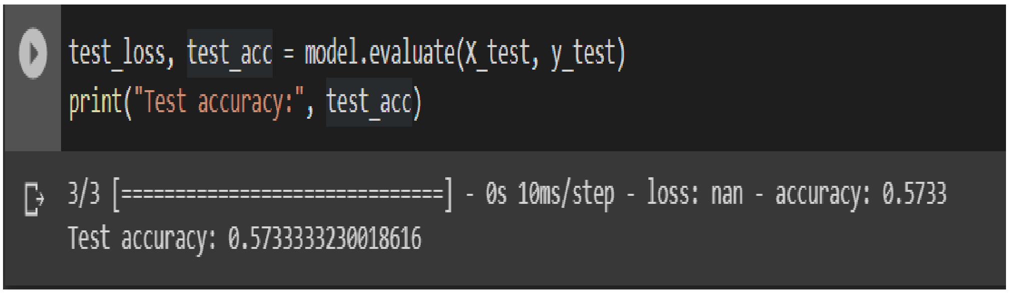
Neural networks can be very powerful models for machine learning, but they often require large amounts of data to train effectively. There are several reasons why neural networks may not work well on small datasets:

- Overfitting: Neural networks are very adaptable and have a tendency to overfit the training data if there is not enough data to enable them to generalize effectively. Overfitting occurs when the network learns to memorize the training examples rather than learning to generalize to new, unseen data. This can result in poor performance on the validation and test sets, even if the network achieves high accuracy on the training set.

- Lack of diversity: Small datasets may not provide enough diversity to capture the full range of variation in the data. Insufficient data can pose a challenge for the network to acquire knowledge.
- Limited model capacity: Small datasets may not provide enough information to fully parameterize the neural network. If the network is too large relative to the size of the dataset, it may be prone to overfitting or may not converge to a good solution.
- Noisy or biased data: A small dataset can be vulnerable to noise or bias, which can pose a challenge for the neural network to learn and form accurate representations.

It's important to note that neural networks are highly flexible and can achieve excellent performance on a wide range of problems, but they require careful tuning and preprocessing of the data, and often require more data than other machine learning models

Figure 16 shows the neural network’s accuracy.



```
test_loss, test_acc = model.evaluate(X_test, y_test)
print("Test accuracy:", test_acc)
```

```
3/3 [==============================] - 0s 10ms/step - loss: nan - accuracy: 0.5733
Test accuracy: 0.5733333230018616
```

Figure 15. Neural Net Accuracy

Figure 17 represents the accuracy comparison of the models :

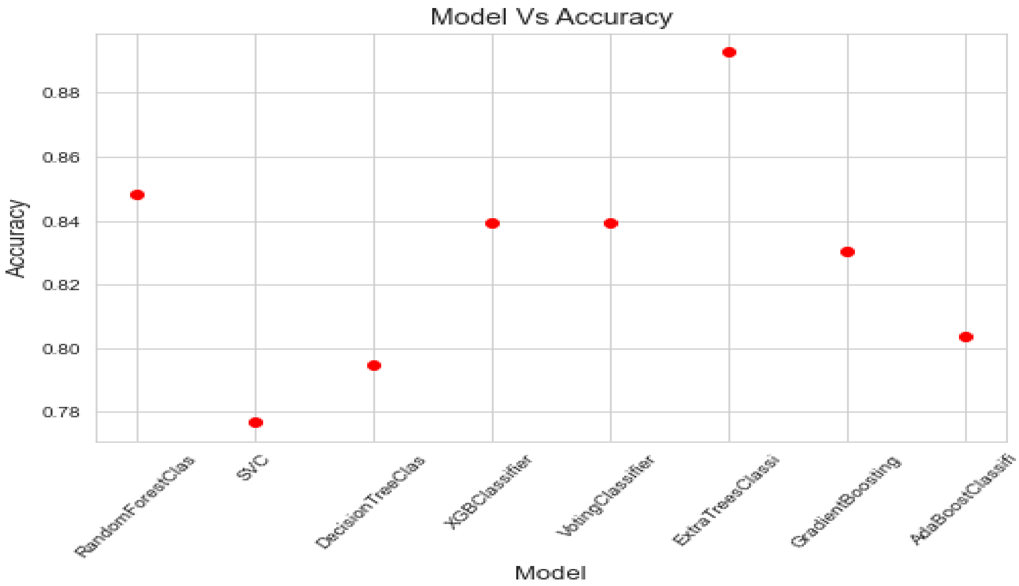


Figure 16. Comparison

Chapter 5:

CONCLUSIONS

5.1 Conclusions

Since Alzheimer's disease is a major health concern, prevention, early intervention, and accurate diagnosis of symptoms are more important than finding a cure. The literature review reveals that numerous efforts have been made to identify Alzheimer's disease using a variety of machine learning algorithms and micro-simulation techniques; However, identifying relevant characteristics that can detect Alzheimer's disease very early remains a difficult task. The extraction and analysis of new features that are more likely to aid in the detection of Alzheimer's disease will be the primary focus of subsequent research, as will the removal of redundant and irrelevant features from existing feature sets in order to boost detection accuracy. We will be able to train our model to distinguish between healthy adults and those with Alzheimer's disease by including metrics like MMSE and education.

The system's primary objective is to predict Alzheimer's disease. For anticipating Alzheimer's sickness or Dementia in grown-up patients, the "OASIS Longitudinal" dataset has been utilized, which has been given by the Open Access Series of Imaging Studies (OASIS) project. The dataset has been visualized and the missing values have been added.Data has been preprocessed by getting rid of some features that aren't needed. Standardization was done to ensure that the ML models could easily use the values.

After that, the dataset was used to train decision trees, random forest, adaboost, gradient boosting, XGBoost, Voting, extra trees, SVM and neural network. The accuracy, recall, ROC, and confusion matrix have been utilized as evaluation metrics. Using ExtraTress, the system produced the best results for this particular dataset. Overfitting was a problem with more complex models like the random forest classifier. The ExtraTrees model has been used for deployment because it produces the best results out of all the models. A larger dataset and a greater number of machine learning models, such as KNN, Majority Voting, and Bagging, could be used in the future to enhance the system models. The system's performance and dependability will both improve as a result of this. By simply entering MRI data, the ML system can assist the general public in determining the likelihood of dementia in adult patients. It is hoped that this will assist patients in receiving early treatment for dementia and enhance their lives.

5.2 Future Scope

Moving forward, the research will focus on extracting and analyzing new features that could aid in the identification of Alzheimer's disease, removing redundant and irrelevant features from the current feature set, and enhancing the accuracy of detection. To improve system models, a larger dataset and a greater number of machine learning (ML) models, such as AdaBoost, KNN, Majority Voting, and Bagging, could be utilized. Additionally, the model can be improved by incorporating additional metrics like MMSE and education to differentiate between healthy individuals and those with Alzheimer's disease.

REFERENCES

[1] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka. “Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects”, 2015 Jan 1;104:398-412. doi: 10.1016/j.neuroimage.2014.10.002

[2] Yudong Zhang, Zhengchao Dong, Preetha Phillips, Shuihua Wang, Genlin Ji, Jiquan Yang and Ti-Fei Yuan. “Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning”, Front. Comput. Neurosci., 02 June 2015, https://doi.org/10.3389/fncom.2015.00066

[3] Benoît Magnin, Lilia Mesrob, Serge Kinkingnéhun, Mélanie Pélégrini-Issac, Olivier Colliot, Marie Sarazin, Bruno Dubois, Stéphane Lehéricy, Habib Benali. “Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI”, 2009 Feb;51(2):73-83. doi: 10.1007/s00234-008-0463-x. Epub 2008 Oct 10.

[4] Prajapati R, Khatri U, Kwon GR. “An efficient deep neural network binary classifier for Alzheimer's disease classification,” In: International Conference on Artificial Intelligence in Information and Communication (ICAIIC). (2021), p. 231–234.

[5] Helaly HA, Badawy M, Haikal AY. Deep learning approach for early detection of Alzheimer's disease. Cogn Computing. (2021) 21:1–17. doi: 10.1007/s12559-021-09946-2

[6] Yaffe K. Modifiable risk factors and prevention of dementia: what is the latest evidence. JAMA Intern Med. (2018) 178:281–2. doi: 10.1001/jamainternmed.2017.7299

[7] Marcus, DS, Fotenos, AF, Csernansky, JG, Morris, JC, Buckner, RL, “Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in