CALIFORNIA STATE UNIVERSITY
FULLERTON™

Department of Computer Science

This project has been satisfactorily demonstrated and is of suitable form.

This project report is acceptable in partial completion of the requirements for the Master of Science degree in Computer Science.

Alzheimer's Disease Prediction using Machine Learning

Project Title (type)

Sanjana Nellutla

Student Name (type)

Dr. Bin Cong

Advisor's Name (type)

Advisor's signature Date

Reviewer's name

Reviewer's signature Date

Alzheimer's Disease Prediction using Machine Learning

By:

Sanjana Nellutla

Student ID: 887453520

CPSC – 597: Graduate Project

Professor: Dr. Bin Cong

Department of Computer Science

California State University, Fullerton

Spring, 2021

ABSTRACT

Alzheimer's Disease is a progressive and degenerative disease which happens to destroy the memory and all kinds of mental functionalities of the brain. This disease is usually seen among people aged between 30 and 60. It has no cure and hence detection and prevention of this disease is essential and critical for such patients. This can be done using Machine Learning algorithms to predict the presence of Alzheimer's Disease like SVM, Random Forest, etc. by using clinical data like symptoms, MRI scans, presence of APOE4 molecule, test scores such as RAVLT, MoCA, ADAS, etc. to understand the trends and their contribution towards the chances of acquiring this disease. Critical analyses using Tableau for visualization is also helpful in determining the presence of the disease as well. This project is useful to determine the presence or absence of the Alzheimer's Disease and also to analyze the behavior of the patients and take steps accordingly to prevent this disease in the future.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 ABOUT

Alzheimer’s disease is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest tasks. In most people with the disease—those with the late-onset type—symptoms first appear in their mid-60s. Early-onset Alzheimer’s occurs between a person’s 30s and mid-60s and is very rare. Alzheimer’s disease is the most common cause of dementia among older adults.[1]

This disorder is named after Dr. Alois Alzheimer. In 1906, Dr. Alzheimer saw changes in the brain tissue of a lady who had died of a strange mental sickness. Her symptoms included memory loss, language issues, and unusual behavior. After she passed on, Dr. Alzheimer studied her brain and discovered numerous unusual lumps (presently called amyloid plaques) and tangled groups of fibers (presently called neurofibrillary, or tau, tangles).

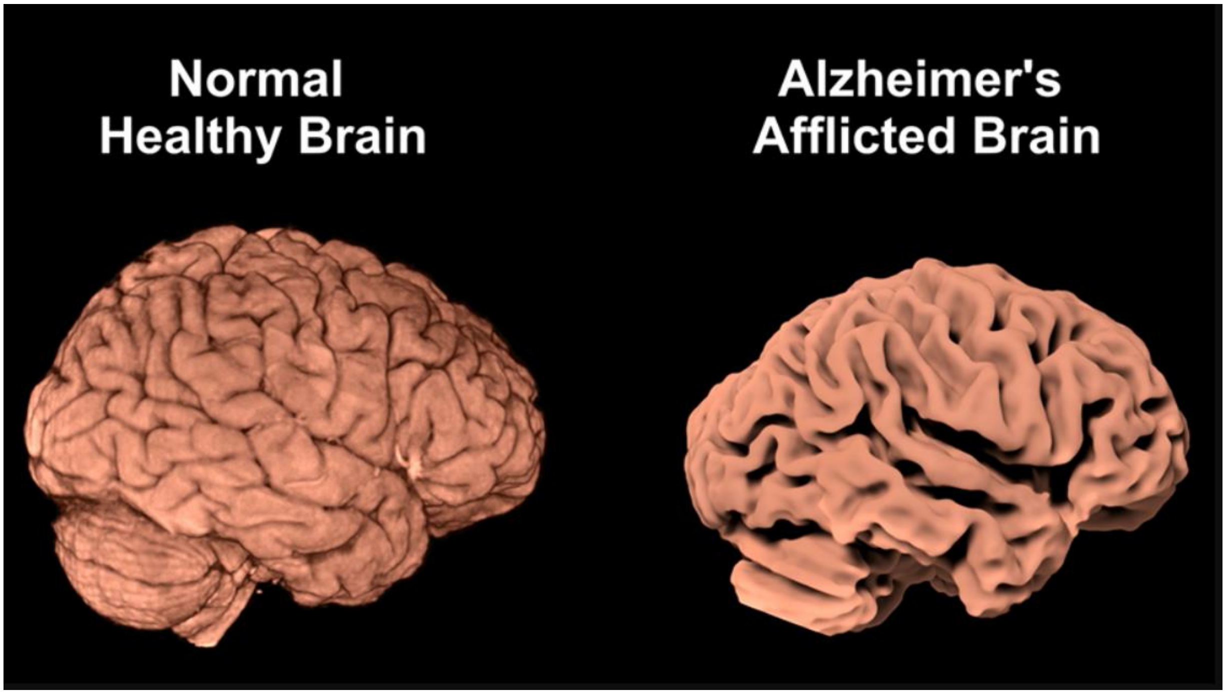


Figure 1-1 How an Alzheimer's Disease afflicted brain would look like when compared to a normal brain [2]

These plaques and tangles in the brain are still viewed as some of the primary features of Alzheimer's disease. Another component is the loss of associations between the nerve cells (neurons) in the brain. Neurons transmit messages between various pieces of the brain, and from the brain to muscles and organs in the body. Numerous other complex brain changes are thought to be involved in Alzheimer's, as well. This harm at first seems to occur in the hippocampus, the piece of the brain which plays a vital role in framing memories. As neurons die, other parts of the brain are impacted. By the last phase of Alzheimer's disease, the harm is extensive, and brain tissue has contracted significantly. [3]

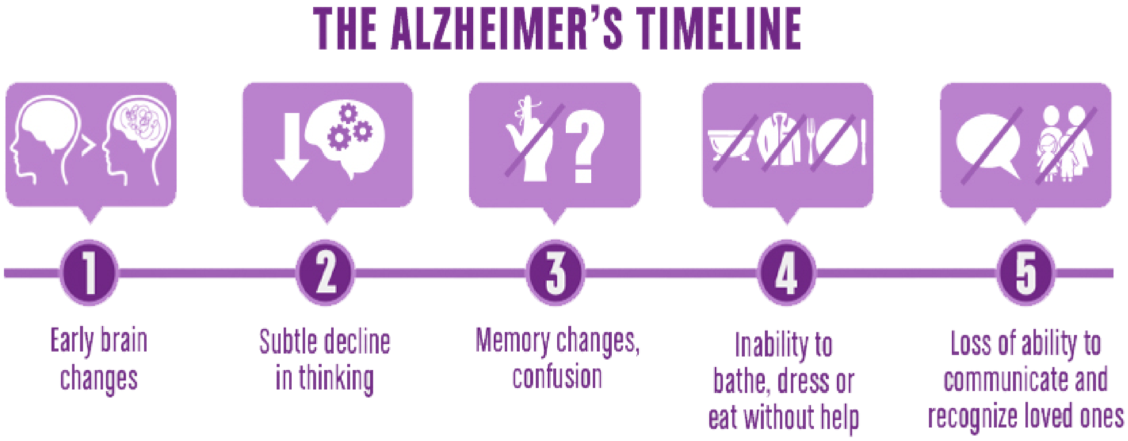

THE ALZHEIMER'S TIMELINE
1
Early brain changes
2
Subtle decline in thinking
3
Memory changes, confusion
4
Inability to bathe, dress or eat without help
5
Loss of ability to communicate and recognize loved ones

Figure 1-2 The timeline of Alzheimer's Disease [4]

What is the existing method for diagnosis and treatment of this disease?

Initially, this disease is identified by the care takers of the patients when traces of memory loss are identified in the patient. A doctor is then consulted based on the severity of the symptoms and a series of examinations and tests are done for the analysis of presence of Alzheimer’s Disease in the brain. These tests include physical and neurological examination of the patient where the muscles, senses and the neurological functions of the body are tested, lab tests like blood test which are used to detect any kind of vitamin deficiencies or thyroid disorders, mental status and neuropsychological tests and brain imaging where the images of the brain are captured and analyzed. Genetic test of the patient is usually not recommended because the results cannot be interpreted easily based on the genetic information as there can be traces of the disease in the ancestors which may show its presence in the patient even when he is not affected by it.[4]

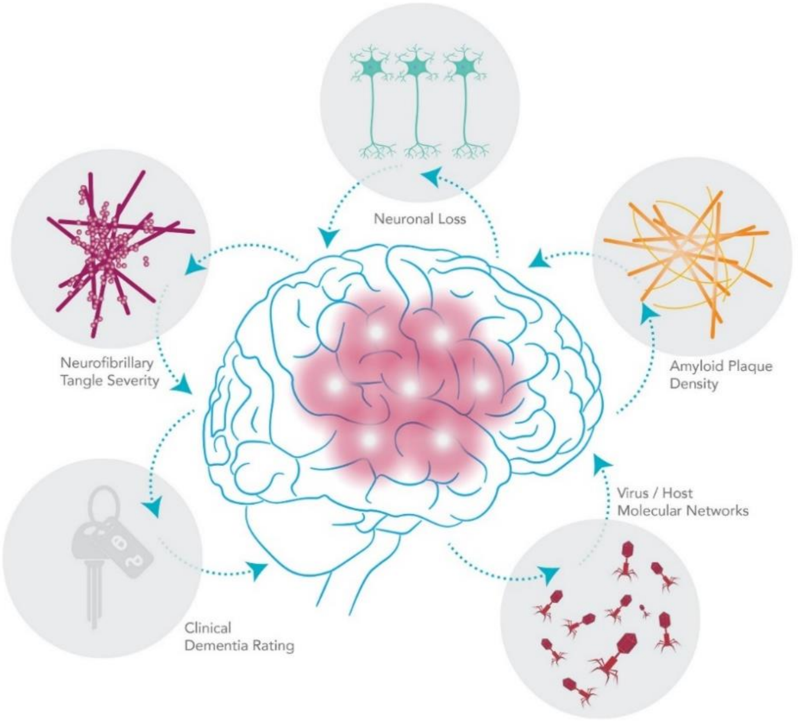

Neuronal Loss
Amyloid Plaque Density
Virus / Host Molecular Networks
Clinical Dementia Rating
Neurofibrillary Tangle Severity

Figure 1-3 Image showing factors that are seen in a brain due to Alzheimer's Disease [5]

How can this disease be cured?

Unfortunately, there is no cure for Alzheimer’s Disease, and it has become a very common and prevalent than cancer. Therefore, there is a high need to try and get better and efficient techniques for detection, treatment, and cure for this disease.

What are the considerations to implement this project?

Some of the underlying key issues of this disease are that it has been growing to become a major health issues in the current trend. This disease has no proper reason as to why it has occurred to a person which makes it even worse to be diagnosed. There is no cure for this disease which means that the potential risk of losing life is very high and the number of people dying due to this disease needs to be controlled as soon as possible. Also, another important reason is that the investment for health initiatives like this is very less which makes it even more challenging for scientists and researchers to develop cure or preventive measures for Alzheimer’s Disease.

Considering such factors, it can be concluded that early detection of the disease is very crucial as it can be prevented before the severity of the disease increases, creating a huge loss of lives. In this way, the person affected by the disease can be treated with care and medication in the early stages which could help seize the seriousness of this disease.

1.2 About Machine Learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.[6] “If computers can accurately detect debilitating conditions such as Alzheimer’s disease using readily available data such as a brain MRI scan, then such technologies have a wide-reaching potential, especially in resource-limited settings,” explained corresponding author Vijaya B. Kolachalama, PhD, assistant professor of medicine. “Not only can we accurately predict the risk of Alzheimer’s Disease but also this algorithm can generate interpretable and intuitive visualizations of individual Alzheimer’s disease risk en-route to accurate diagnosis”.[7]

1.3 Problem Statement

The primary objective of this project is to develop a Machine Learning model using algorithms like Random Forest to analyze the activities of the brain with respect to memory of the patient. Using clinical data generated as a result from various examinations and tests, prediction can be done whether the patient has been affected by Alzheimer's Disease or not.

As this method shall be used in the early stages of the disease, the patient can be taken care of right from the beginning which would consequently lead to a potential reduction in the severity of Alzheimer's Disease.

The machine learning model is going to use ADNI clinical datasets as input which will be further discussed in the later sections. This would help the model to generate accurate results without consultation with the physician.

The data collected is going to be classified into training dataset and testing dataset. The Machine Learning model is going to learn the classification method using the training data and then the model is implemented on the testing dataset to check for the accuracy of the data. This project is intended to generate an approximate accuracy of at least 95%.

The required features are going to be selected using Correlation matrix. Then, algorithms like SVM, Random Forest, etc. are going to be used to classify the attributes from the input which is further going to be balanced. Then the results, predicting the presence of Alzheimer's Disease are generated as the output in a confusion matrix per each classification.

The final objective is to visualize the given dataset into valuable and interesting insights using Tableau dashboards and stories that would also be helpful in understanding the symptoms and other factors of patient samples to give an overall view of the disease patterns.

2 Design and Architecture

This project is designed to obtain the best machine learning model that generates the maximum accuracy while predicting the presence of Alzheimer's Disease in a patient using data patterns and pre-processing methods.

Following is the architecture of the project in the form of a flowchart which is broadly classified into four stages:

1. Collecting the data
2. Pre-processing the data
3. Designing the data model
4. Generating and visualizing the results

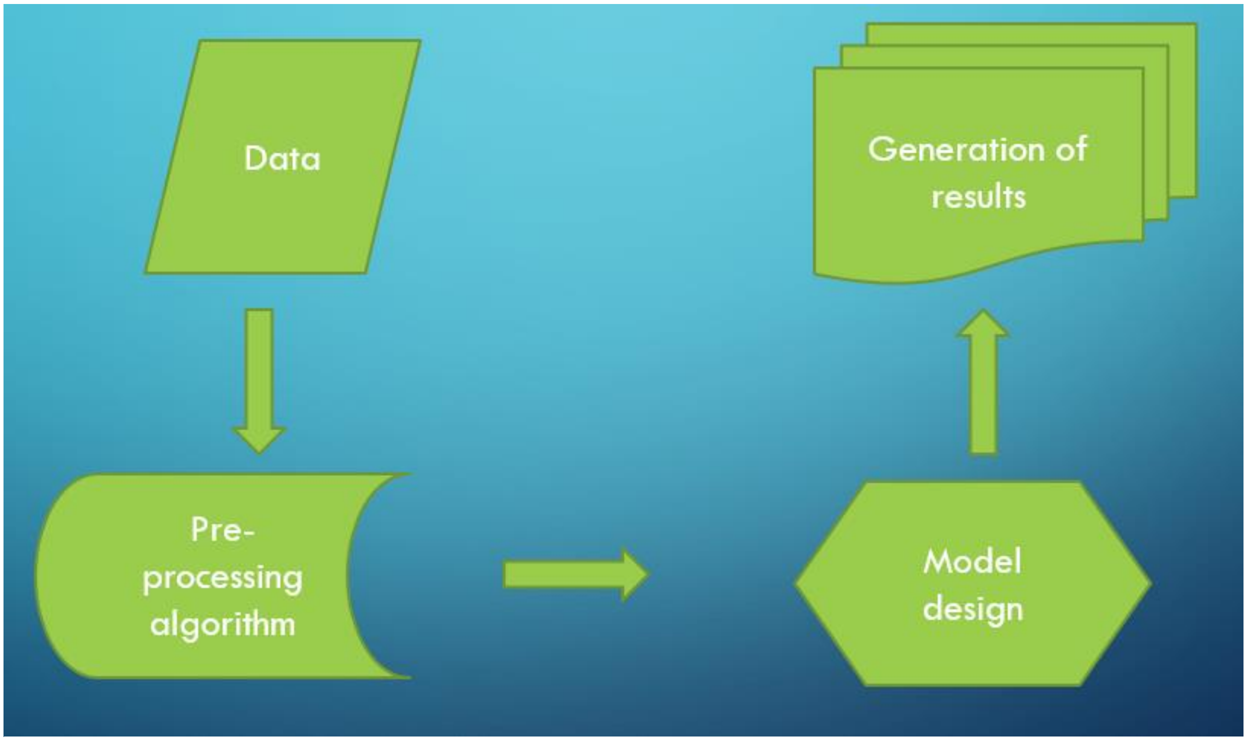


Figure 2-1 Flowchart of the activities performed in the project

The hardware and software specifications used in this project are as follows:

The minimum hardware specifications are 8 GB RAM and a processor with dual core capacity along with 64-bit OS (Windows/Linux).

The software specifications include the training and testing data from ADNI1, ADNI2 and ADNI GO. The IDE used in this project is Google Colab, which is used to write and execute Machine Learning code and is stored on the cloud. The language used to implement this project is Python.

For pre-processing the data, building and designing the model, Pandas, SciPy, NumPy and ‘model_selection’, ‘metrics’, ‘ensemble’, ‘linear_model’ and ‘neighbors’ packages from scikit-learn for Machine Learning are used.

For the visualization of data, packages such as Matplotlib, etc. and Tableau Desktop 2021.1 are used. Tableau is used to depict raw data in terms of graphical and pictorial representations.

3 IMPLEMENTATION

3.1 COLLECTING DATA

The first and foremost step while implementing this project is collecting the required data from an appropriate and reliable source to make the machine learning model as efficient and as close to the real-time application for the model as possible. Alzheimer's Disease Neuroimaging Initiative (ADNI) has provided real-time data of patients affected by the disease and their symptoms starting right from their physical examinations to attributes from the neuro-images of the brain of those patients. In this project, ADNI, ADNI2 and ADNI-GO are used as the training data and the testing data.

Field Name	Description
DX_bl	Diagnosis at Baseline
PTGENDER	Sex
PTETHCAT	Ethnicity
CDRSB	Memory Score (CDR)
RAVLT_immediate	RAVLT Immediate (sum of 5 trials)
APOE4	Presence of ApoE4
Hippocampus	UCSF Hippocampus
Ventricles	USCF Ventricles
DX	Final Diagnosis (target variable/output)

Table 3-1 Data Dictionary Examples

There are several types of attributes that are present in the ADNI dataset are described below.

1. Scores from tests like RAVLT (Rey Auditory Verbal Learning Test), MoCA (Montreal Cognitive Assessment) test, FDG – PET (Flouro-Deoxy-glucose test) analysis, CDR (Clinical Dementia Rating) test, etc.
2. Results of MRI scans that constitute the neuro imaging of the brain
3. Physical diagnosis like age, race, etc.

3.2 Pre-Processing of Raw Data

The next step is to collect all the raw data and send it for pre-processing. In this step, all the noisy data is removed using the Correlation matrix and the related features with a score of more than 0.3 are only applied for modelling the classification algorithms. For any imbalances in the data, normalization of the columns is done to balance the classes.

3.3 Modelling the Dataset

Then the design model is created and built to classify the data which divides the results into three classes – whether the patient is affected by the disease or not. This method can be done using Random Forest, Support Vector Machine (SVM), Logistic Regression and K-Nearest Neighbors algorithms.

3.4 Applying Machine Learning Models

Finally, the algorithms are fitted accordingly using the above design model on the training dataset. The model is then applied against the testing dataset to predict the presence of the disease in the patients further by visualizing the confusion matrix with the predicted labels against the true labels. In this project, the following machine learning models are applied:

1. Random Forest
2. Support Vector Machine
3. Logistic Regression
4. K-Nearest Neighbors

3.5 Determining the Best Model

Among all the applied Machine Learning models applied to the dataset, the best model is obtained by identifying the algorithm which attained the maximum accuracy.

3.6 Visualizing Data using BI Tools

The dataset is visualized using Tableau for providing data insights to analyze the Alzheimer’s Disease trends and behavioral patterns of individuals.

4 Implementation

4.1 Initial Setup

4.1.1 Downloading the ADNI Dataset

The Alzheimer's Disease Neuroimaging Initiative has provided multiple datasets to perform varied analyses that define the progression of Alzheimer's Disease in patients. This data archive requires access approval from ADNI after verifying the purpose of usage. It can be obtained using the link: https://ida.loni.usc.edu/pages/access/studyData.jsp

4.1.2 Creating a Google Colab account

Google Colab account requires access to a Google account and the files are stored in the cloud. Using https://colab.research.google.com/ link to access the IDE platform would serve the purpose.

4.1.3 Installing Tableau Desktop 2021.1

Tableau Desktop 2021.1 application can be installed with the following link: https://www.tableau.com/products/desktop/download

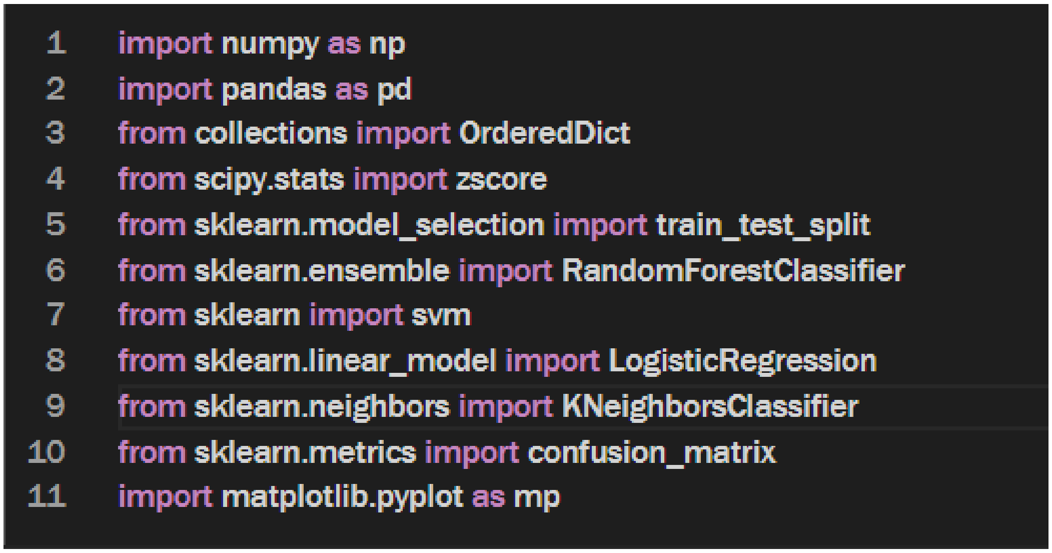
4.2 Procedure

4.2.1 Importing required packages

The following packages need to be imported in the Google Colab notebook created for this project:

1. Pandas – For Dataframe functions and libraries
2. Numpy – For numeric functions and libraries

3. Collections – For performing dictionary functions
4. Sklearn.metrics – For importing Confusion matrix
5. Scipy.stats – For preprocessing the data
6. Sklearn.model_selection – To split the dataset
7. Sklearn.ensemble – For importing Random Forest libraries
8. Sklearn (svm) – For importing SVM libraries
9. Sklearn.linear_model – For importing Logistic Regression libraries
10. Sklearn.neighbors – For importing K-Nearest Neighbors libraries
11. Matplotlib.pyplot – For data visualizations



```
import numpy as np
import pandas as pd
from collections import OrderedDict
from scipy.stats import zscore
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as mp
```

Figure 4-1 Importing required packages

4.2.2 Importing the data file

Pandas read_csv() method is used to import the csv file 'ADNIMERGE.csv' containing the dataset with 15267x101 rows and columns.

4.2.3 Mapping miscellaneous values using Dictionaries

Some features containing mixed datatypes are handled by mapping them to strict datatypes.

4.2.4 Handling missing values

Dropping missing values from the target feature and categorical values, filling numerical values by grouping them with respect to the patient ID (‘RID’) is done in this section.

4.2.5 Maintaining consistency among all numerical columns

This is done by normalizing the column values with the ‘zscore’ attribute from scipy.stats package.

4.2.6 Splitting the dataset into training and testing dataframes

To use the same dataset for training and testing purpose, it is split into 75% and 25% of the dataset for training and testing respectively using ‘train_test_split’ attribute from sklearn.model_selection library.

4.2.7 Encoding categorical features

Categorical features need to be modified for the data model to understand its importance and classification pattern. Encoding them using get_dummies() method from pandas will allocate a column for each feature and assign values respectively. This encoding must be done for training and testing datasets separately.

4.2.8 Feature Selection

This is done by correlating the encoded training data against the target variable with a score of more than 0.3.

4.2.9 Applying Machine Learning Classification algorithms

After selecting the related features, classification algorithms are applied on the training dataset and the accuracy is calculated against the testing data. Each classification algorithm is fitted on the training data with the most suitable attribute values.

4.2.10 Displaying the confusion matrix

Confusion matrix is used to represent the accuracy of the predicted values against the true or the actual values in the test dataset. This is obtained by importing ‘confusion_matrix’ from sklearn.metrics.

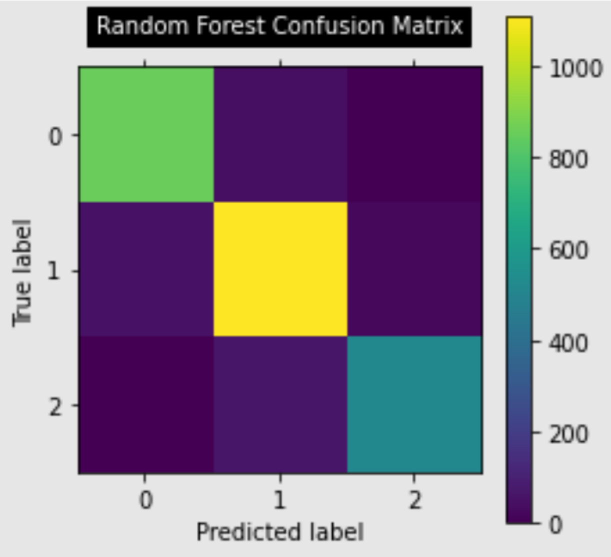


Figure 4-2 Sample confusion matrix of Random Forest

4.2.11 Showing the accuracy graph

Using a histogram, the accuracy of all the models is displayed with Matplotlib package.

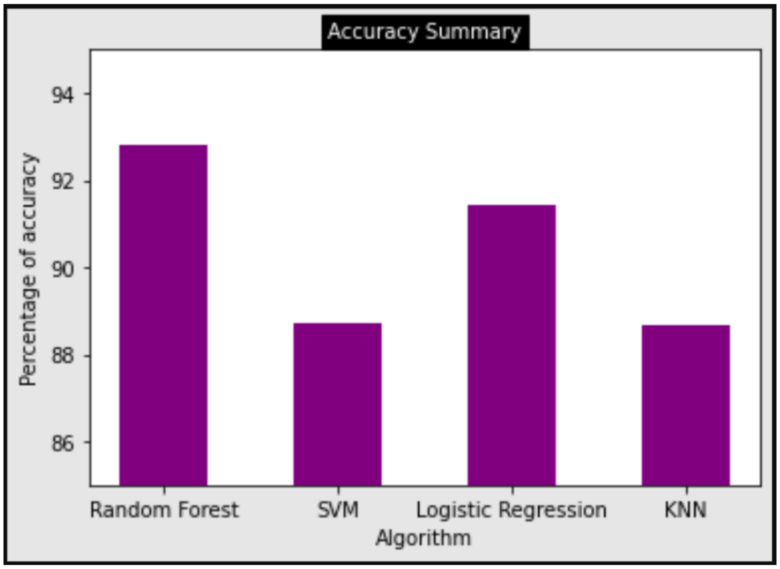


Figure 4-3 Sample output of accuracy bar graph

4.2.12 Selecting the best model for Alzheimer's Disease prediction

Depending on the accuracies of each of the algorithms, the best model with the highest accuracy is identified and selected.

4.2.13 Visualization using Tableau 2021.1

Using various charts and attributes, interactive Tableau dashboards are designed to depict the data patterns and analyze the Alzheimer's Disease data of patients.

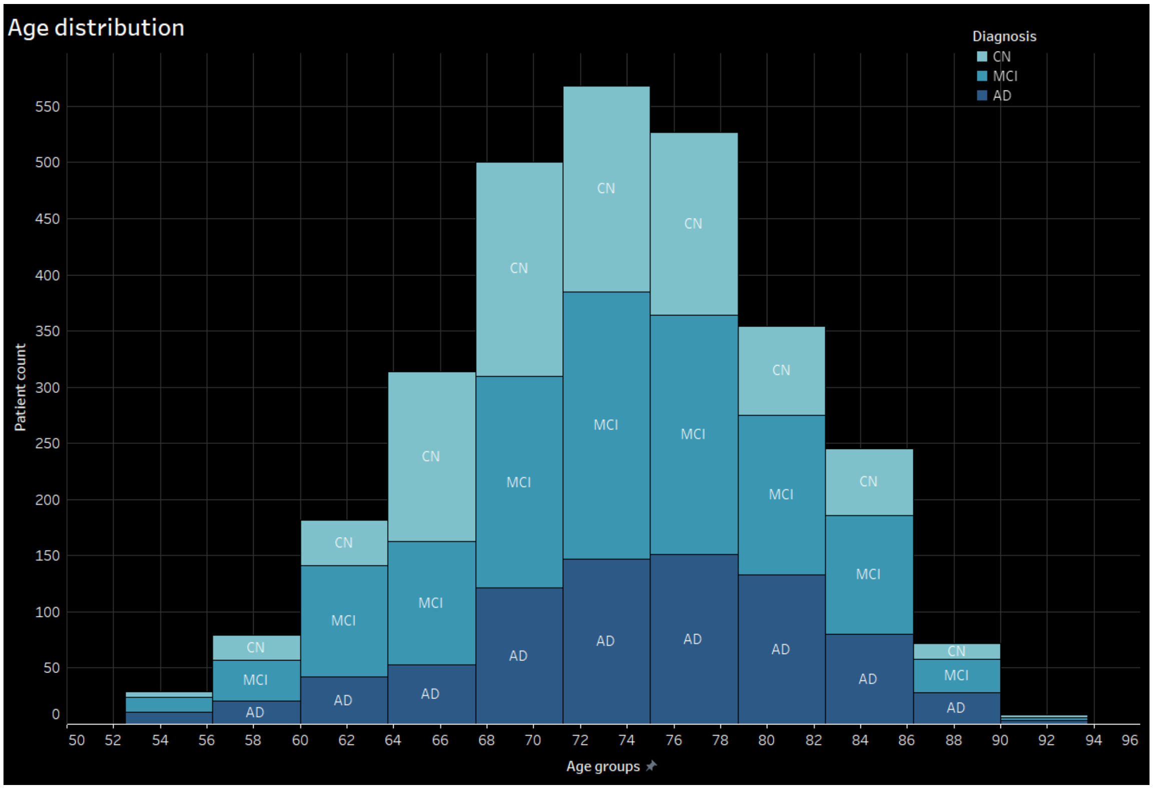
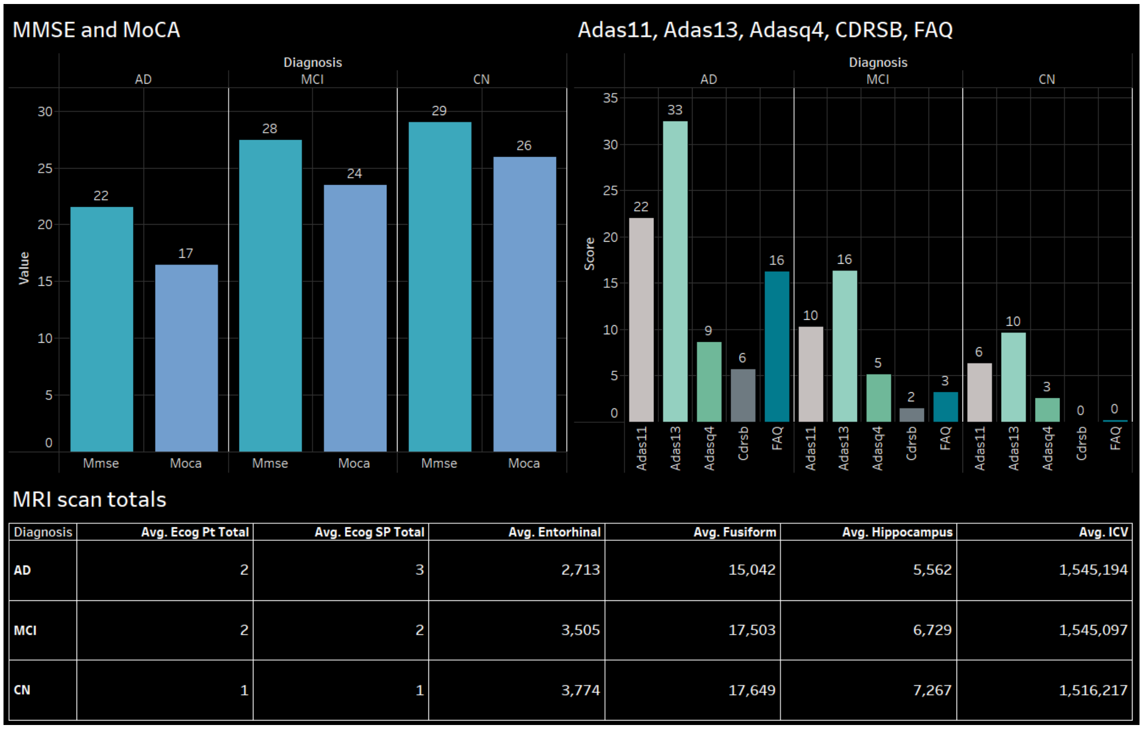


Figure 4-4 Sample Tableau Worksheet



Diagnosis	Avg. Ecog Pt Total	Avg. Ecog SP Total	Avg. Entorhinal	Avg. Fusiform	Avg. Hippocampus	Avg. ICV
AD	2	3	2,713	15,042	5,562	1,545,194
MCI	2	2	3,505	17,503	6,729	1,545,097
CN	1	1	3,774	17,649	7,267	1,516,217

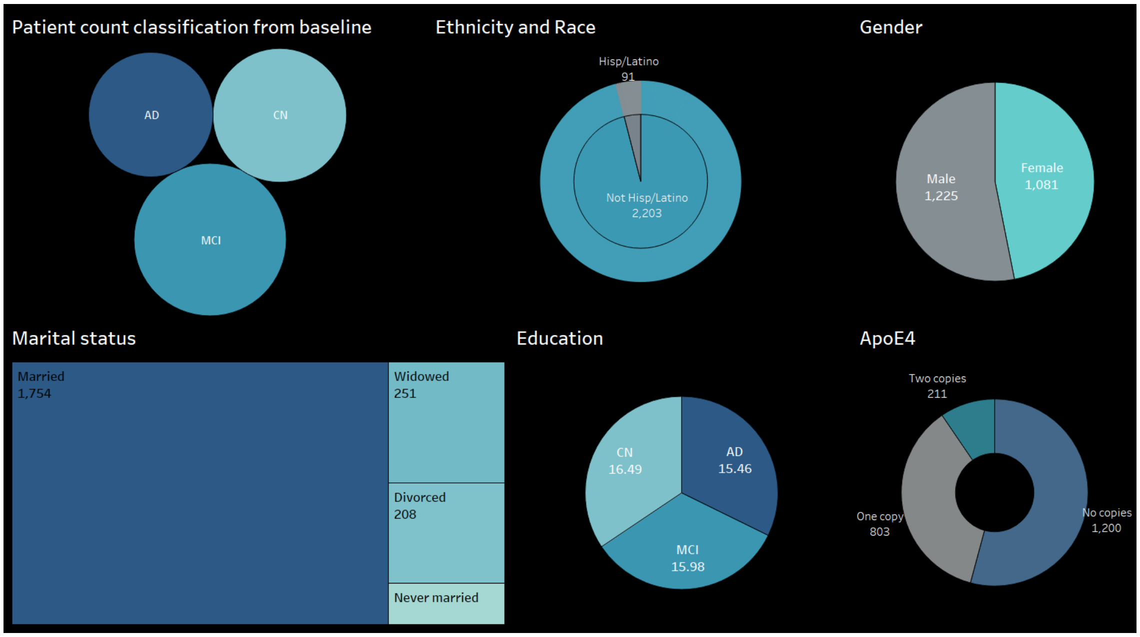


Figure 4-5 and Figure 4-6 Sample Tableau Dashboards

5 Summary and Deliverables

To detect the presence of Alzheimer's Disease in a patient, clinical data of the samples are utilized. This data is processed in terms of cleaning, handling missing values, balancing classes, modifying and maintaining consistency among all features and choosing the best features to fit into the machine learning algorithms. Random Forest Algorithm and Logistic Regression are proven to have the maximum accuracy than the other models. Using these models to the Alzheimer's Disease data provides instant and accurate results. Therefore, by detecting the presence of this disease in the early stages itself using these models will help patients in recovering from this disease by proactively treating the symptoms along with care and concern. Further, it also helps in understanding the clinical and physiological patterns of the patient and help minimize complications.

The output of this project is the data model, the training results validated against the test data and the best model along with accuracy of predicted labels against the true labels.

The deliverables include:

1. The source code
2. The dataset on which machine learning algorithms were performed
3. The visualizations using Tableau dashboards
4. The detailed documentation of the project
5. All other related reports and manuals

The entire outcome of the solution attained the maximum accuracy of 93% using Random Forest Algorithm, 90% using Support Vector Machine with RBF kernel, 92% with Logistic Regression and 88% with K-Nearest Neighbors.

6 Future Scope

Applying further complex data models using Neural Networks and other Machine Learning classification algorithms can help in improving the accuracy of prediction of Alzheimer’s Disease in patients.

Including different types of data such as MRI images of the brain, physical and extended physiological characteristics of samples, etc. and larger datasets with huge cardinality and lesser missing values can help the machine learning algorithm learn and understand complex data and provide better results.

Furthermore, various analyses can be drawn considering the varied features and the target variables to understand the complexity of the brain and the nervous system which can be useful to proactively predict other diseases and syndromes in patients.

7 R0065FERENCES

1. https://www.nia.nih.gov/health/what-alzheimers-disease
2. https://www.loni.usc.edu/SVG
3. https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed
4. https://www.wesleylife.org/memory-care/get-the-facts/
5. https://medicalxpress.com/news/2018-06-viral-alzheimer-disease.html
6. https://www.sas.com/en_us/insights/analytics/machine-learning.html
7. https://www.bumc.bu.edu/busm/2020/05/04/ai-algorithm-can-accurately-predict-risk-diagnose-alzheimers-disease/
8. Mesrob, Lilia & Magnin, Benoît & Colliot, Olivier & Sarazin, Marie & Hahn-Barma, Valérie & Dubois, Bruno & Gallinari, Patrick & Lehericy, Stéphane & Kinkingnéhun, Serge & Benali, Habib. (2008). Identification of Atrophy Patterns in Alzheimer’s Disease Based on SVM Feature Selection and Anatomical Parcellation. 5128. 124-132. 10.1007/978-3-540-79982-5_14.
9. Magnin, Benoît & Mesrob, Lilia & Kinkingnéhun, Serge & Pélégrini-Issac, Mélanie & Colliot, Olivier & Sarazin, Marie & Dubois, Bruno & Lehéricy, Stéphane & Benali, Habib. (2008). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology. 51. 73-83. 10.1007/s00234-008-0463-x.
10. Albright, Jack. (2019). Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. Alzheimer's & Dementia: Translational Research & Clinical Interventions. 5. 483-491. 10.1016/j.trci.2019.07.001.
11. Tanveer, M. & Richhariya, Bharat & Khan, Riyaj & Rashid, A.H. & Prasad, Mukesh & Khanna, Pritee & Lin, Chin-Teng. (2019). Machine learning techniques for the diagnosis of Alzheimer’s

disease: A review. ACM Transactions on Multimedia Computing, Communications and Applications.

12. https://ida.loni.usc.edu/pages/access/studyData.jsp
13. https://colab.research.google.com/
14. https://www.tableau.com/products/desktop/