International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

Diagnosis Of Alzheimer's Disease Using Machine Learning

Goulikar Laxmi Narasimha Deva [1], Ramesh Ponnala[2]
[*] *Department of MCA, Chaitanya Bharati Institute of Technology, India.*
[*]*Ascend Author, Assistant Professor, Department of MCA, Chaitanya Bharati Institute of Technology, India.*

Abstract
This project applies the paramount machine learning techniques for the early detection and effective diagnosis of severe AD. It is vital to diagnose the disease in initial stages for more effective and beneficial treatment.
Machine Learning is becoming an area of great interest because it is showing the remarkable achievement in the present, advance and crucial decision making. Medical diagnosis is one of the crucial areas that have paramount importance where various learning algorithms can be contributed for the improvement in disease diagnosis. Due to the evolution of computation technology, the generation of data is increased exponentially, especially in medical field. To cope up this problem, this project tests various approaches like Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, Ada boosting , to identify the best parameters for the Alzheimer's Disease prediction using OASIS_Longitudinal MRI Data.
Experimental result is analyzed in terms of accuracy, recall, and AUC (Area Under Curve). The analysis shows that the Random Forest and AdaBoost have same accuracy but Random Forest performs better than AdaBoost in terms of recall and AUC as well.

1. INTRODUCTION

AD is a grievous neurodegenerative disease which is one of the chronic diseases that needs early detection, so that the treatment can be effective. Usually it starts slowly but with time worsens. This disease primarily affects the older people and can become the cause for dementia. The task of detecting this disease at the early stage is very difficult but equally important, so there is a necessity of intelligent system for supporting the clinicians in the early diagnosis of this disease. To address the mentioned problem, this paper elaborates the machine learning concept. Due to the handiness of improved technology, data exponentially increases and the world becomes a data rich society. This enormous amount of data takes the learning algorithms of machine at paramount height. Analyzing the immense data to get the fruitful result is the area of budding research. The target of all learning technologies is to retrieve the unseen patterns that can be further helpful in taking decision. Learning techniques are abundantly found in different sectors like media, health, agriculture, etc. intelligent learning models. Intelligent methods of learning.

1.1 Motivation

The System uses Jupyter notebook GUI as a platform for coding, where the dataset can be loaded and thereafter some exploratory data analysis will be performed on the dataset like analyzing various features in the dataset. Then, data pre-processing is performed on the dataset like filing the missing values etc., This dataset is further divided into training and testing data. Then various machine learning models are trained and tested on this dataset. The experimental results are then compared in terms of accuarcy, recall and AUC. The best model based on the above metrics is used to develop a web application using Tkinter GUI.

1.2 TERMINOLOGIES

EDUC: Number of years of education.

SES: Socio-Economic Status.

DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

MMSE: Mini Mental State Examination.

CDR: Clinical Dementia Rating.

eTIV: Estimated Total Intracranial Volume

nWBV: Normalize Whole Brain Volume.

ASF: Atlas Scaling Factor.

2. LITERATURE SURVEY

Machine Learning-Based Method for Personalized and Cost Effective Detection of Alzheimer's Disease

Thus, it is desirable to develop new approaches to support clinicians in the early, more effective (in terms of number of tests and/or cost), and personalized detection of disease. Alzheimer's disease (AD) is the most common neurodegenerative disease in older people. There is a considerable delay between the start of AD pathology and the clinical diagnosis of AD dementia, which can only be confirmed by autopsy. Thus, it is very difficult to detect AD early and accurately, and there is a need for intelligent means to support clinicians in the personalized diagnosis of this disease. To address such challenges, we test a proof-of-concept personalized classifier for AD dementia and mild cognitive impairment (MCI) patients based on biomarkers. We extend previous analyses to AD, including new feature selection approaches, classifier, and measures of similarity between subjects suitable for continuous variables. Our aim is to support the clinician in the diagnosis process by providing him or her with information about the patient's probability of disease and which biomarkers may be more informative.

An Optimal Decisional Space for the Classification of Alzheimer's Disease and Mild Cognitive Impairment

This paper proposes to combine MRI data with a neuropsychological test, mini- mental state examination (MMSE), as input to a multi-dimensional space for the classification of Alzheimer's disease (AD) and it's prodromal stages—mild cognitive impairment (MCI) including amnestic 13 MCI (aMCI) and nonamnestic MCI (naMCI). The decisional space is constructed using those features deemed statistically significant through an elaborate feature selection and ranking mechanism. FreeSurfer was used to calculate 55 volumetric variables, which were then adjusted for intracranial volume, age and education. The classification results obtained using support vector machines are based on twofold cross validation of 50 independent and randomized runs. The study included 59 AD, 67 aMCI, 56 naMCI, and 127 cognitively normal (CN) subjects. The study shows that MMSE scores contain the most discriminative power of AD, aMCI, and naMCI. For AD versus CN, the two most discriminative volumetric variables (right hippocampus and left inferior lateral ventricle), when combined with MMSE scores, provided an average accuracy of 92.4% (sensitivity: 83.0%; specificity: 95.1%). MMSE scores are found to improve all classifications with accuracy increments of 8.2% and 12% for aMCI versus CN and naMCI versus CN, respectively. Results also show that brain atrophy is almost evenly seen on both sides of the brain for AD subjects, which is different from right- side dominance for aMCI and left-side dominance for naMCI. Furthermore, hippocampal atrophy is seen to be the most significant for aMCI, while Accumbens area and ventricle are most significant for naMCI**2012:** In this author proposes an enhanced LSB algorithm for image steganography. In this proposed work they only embed secret information in the blue component of the RGB color space. In their technique first M×N size cover image is selected. After the selection of the cover image, only the blue component

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

A Survey on Machine-Learning Techniques in Cognitive Radios

In this survey paper, we characterize the learning problem in cognitive radios (CRs) and state the importance of artificial intelligence in achieving real cognitive communications systems. We review various learning problems that have been studied in the context of CRs classifying them under two main categories: Decision- making and feature classification. Decision-making is responsible for determining policies and decision rules for CRs while feature classification permits identifying and classifying different observation models. The learning algorithms encountered are categorized as either supervised or unsupervised algorithms. We describe in detail several challenging learning issues that arise in cognitive radio networks (CRNs), in particular in non-Markovian environments and decentralized networks, and present possible solution methods to address them. We discuss similarities and differences among the presented algorithms and identify the conditions under which each of the techniques may be applied

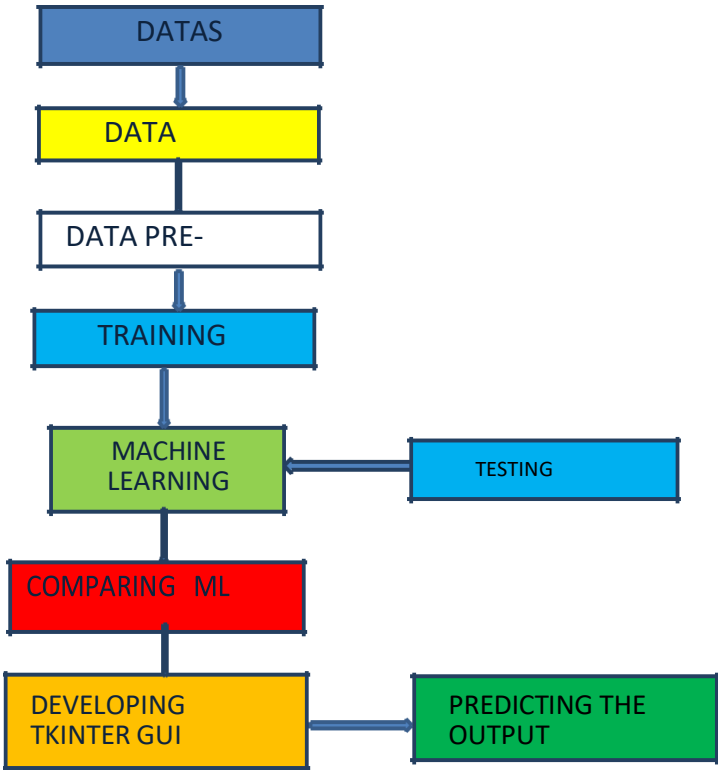


Figure.1 **System Architecture of Diagnosis of Alzheimer's Disease.**

2.1 Dataset Analysis:

This project uses data generated by the Open Access Series of Imaging Studies (OASIS) project that is available both, on their website and kaggle that can be utilized for the purpose of training various machine learning models to identify patients with mild to moderate dementia. So the project uses OASIS_Longitudinal MRI data.

- The dataset consists of 373 records.
- The dataset contains 15 attributes in total.
- The dataset for cross-sectional data was obtained from OASIS. Cross-sectional MRI is the imaging based on axial slices.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

- The cross-sectional data has multiple attributes categorized as follows:

2.2 Removing Unnecessary Attributes

- The number of visits of a patient is dropped because it has no role to play in Alzheimer's disease and the project focuses on whether the person is suffering from Alzheimer's or not.
- The handedness attribute is also removed because the dataset contains only righthanded people.
- The MRI ID, Subject_ID attributes are also removed.

2.3 Splitting of Dataset :

Each dataset taken for the study is divided into train and test datasets. The training dataset contains 80% of the data and the test dataset contains 20% of the data in the way the dataset is divided. The dataset is divided in such a way that the result attribute has equally divided the classes Demented and NonDemented. This process is done by the train_test_split module in sklearn Library.

Ex:

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=20)

2.3 Data Pre-Processing :

A) Finding the missing values in the dataset : At first, any missing values are present in the dataset are checked. The dataset taken from the OASIS or Kaggle has the missing values in SES column. The total number of values that are missing in each column is noted. To check the missing values at each column there is a function isnull() which checks each value in the dataset is null or not. The sum of missing values at each column is given by data.isnull().sum() this gives the sum

Ex:

```
pd.isnull(df).sum()
```

```
Subject ID    0
Group         0
MR Delay      0
M/F           0
Age           0
EDUC          0
SES           8
MMSE          0
CDR           0
eTIV          0
nWBV          0
ASF           0
dtype: int64
```

.

B) Filling Missing Values using Imputation : The project identified 8 rows with missing values in SES column. To deal with this issue there are 2 approaches. One is just to drop the rows with missing values. The other is to replace the missing values with the corresponing values, also known as 'Imputation'. Since the project uses the dataset that has only 373 data tuples, I assume imputation would help the performance of our model. Since SES(Socio Economic Status) attribute has discrete values, the missing values are replaced by the median of the SES column

3. Machine Learning Algorithms:

In the study done here, five machine learning algorithms are used. These algorithms are trained and tested over the dataset and compared.

3.1.1LOGISTIC REGRESSION

Logistic regression is a classifier that separates the data linearly. With logistic regression, the data can be linearly separated. In this study, the logistic Regression used is imported from sklearn. To train the model fit() function is used. To predict target values predict() function is used. The logistic Regression is a sigmoid function.

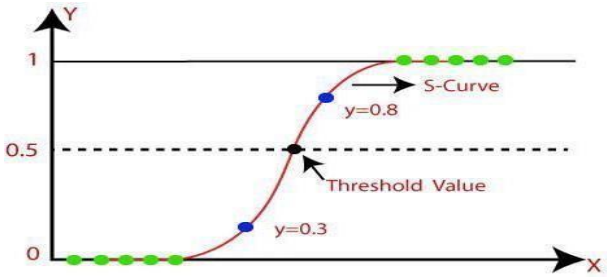
International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

The formulae of the logistic function are represented below h(x) = sigmoid (Z)
Sigmoid = 1/(1+e^-x)
X=w0+w1x1+w2x2…+wnxn+b W
here wi = weights, xi = input and b is bias.

By changing the w and b values the curve can be fit and make the classification required. By changing the value of b the curve can be moved left or right on the x -axis. By changing the value of the w the slope of the curve can be changed as required to make the classification of the data

STEPS :

- In this proposed system, the Logistic Regression model iterates through some random values for inverse of regularization factor 'c'.



- At each iteration, the accuracy is noted for that particular value of 'c'.
- All the accuarcies are noted and finally the best parameter 'c' is selected.
- Then, using the best parameter, accuaracy, recall and AUC are calculated and displayed.

3.2 SUPPORT VECTOR MACHINE

Support vector machine is a classification algorithm that separates the data into two groups. The lung cancer data is divided into groups and classification is done as patient as cancer or not. It creates two hyperplanes one is close to positive points and the other to the negative points. The middle line is the margin line which helps in classifying the data as shown in
figure 3.5.
Margin line (y)= (W^T)*X + b where W is the slope of the line, X is input and b is the intercept.

Graphical Representation of SVM

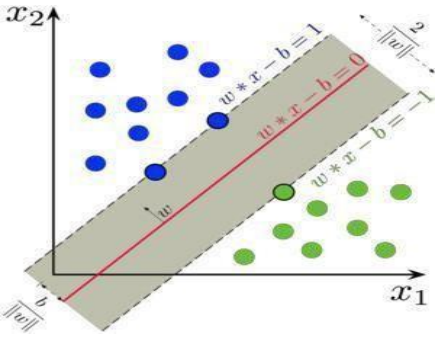


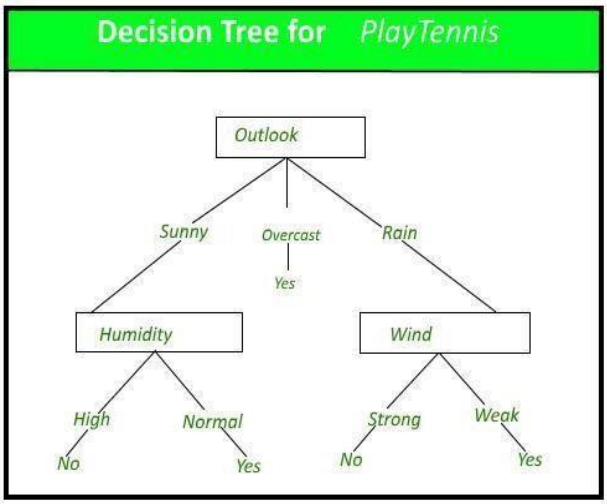
Figure 4.7.3 represents the Support Vector Machine Algorithm using graphical represenation. STEPS :

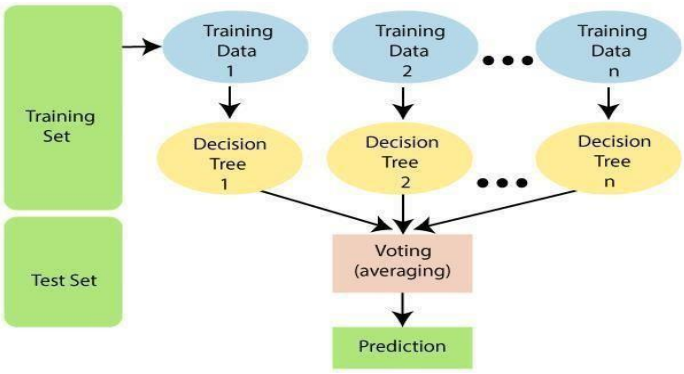
- In this proposed system, the SVM model iterates through the dataset over some random values for regularization factor **'c'** and also different types of kernel functions like Linear, polynomial, sigmoid, RBF.
- It also iterates over some random values for kernel coefficient.
- At each iteration , accuracies are noted and compared and finally, the best parameters for Regularization factor, Kernel coefficient and Kernel function are determined.
- These values are used to train and test the SVM model and accuracy, recall and AUC

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

DECISION TREE CLASSIFIER:

The decision tree is a tree-like structure and which consists of a root node, internal node, and leaf node. The root node indicates the resultant class. The internal node indicates the conditions which lead to the root nodes that are to the resultant class. The decision is a set of if - else statements where





several if-else statements are put together as shown in figure 3.7 in the form of a tree. This set of statements checks the values and form conditions.

STEP:

The number of features used in this project are 8. So, the maximum depth of thedecision tree can be upto 8.

The Decision Tree model iterates over different depths upto 8 and finds the bestdepth by calculating feature importance and classifies the dataset.

Using the best depth, accuarcy, recall and AUC are calculated along with the feature imporatance

RANDOM FOREST ALGORITHM:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learningwhich is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Working of Random Forest Algorithm

represents work flow of Random Forest Algorithm, where it generates a large number of decision trees and compares their accuracies and gives the best decision tree as output.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

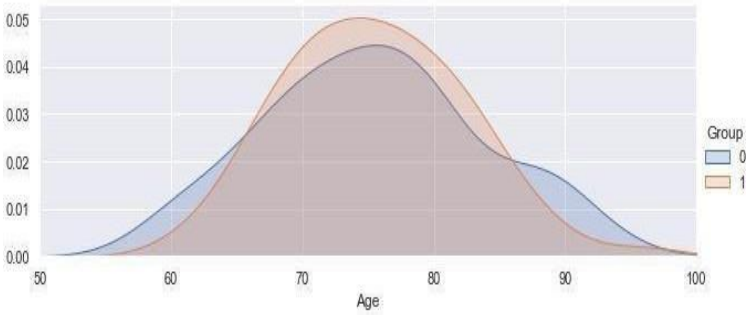
- Select random K data points from the training set.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

- Build the decision trees associated with the selected data points.
- Choose the number N for decision trees that you want to build.
- Repeat 1 & 2 steps.
- For new data points, find the predictions of each decision tree, and assign the newdata points to the category that wins the majority votes.

- In this project, the Random Forest model iterates through dataset over some randomvalues to identify :

M - Number of trees in the forest
d - Number of features to be considered when looking for the best split
m - the maximum depth of the tree



4. FEATURES OF THE PROPOSED SYSTEM

In this project, various Machine Learning models are used to detect the Alzheimer's disease using OASIS_Longitudinal data set. The models are trained by selecting some features of the data set based on exploratory data analysis on the data set. The experimental results of each model are compared in terms of their accuaracy, recall, Area Under Curve(AUC). The best Machine Learning Model based on the above 3 metrics is used to develop a web application using Tkinter, where a user can enter his/her medical details and check whether he/she is suffering from Alzheimer's disease or not. This project also uses MMSE(Mini Mental State Examination) rating as a part of its selected features to detect Alzheimer's, which was ignored in earlier studies

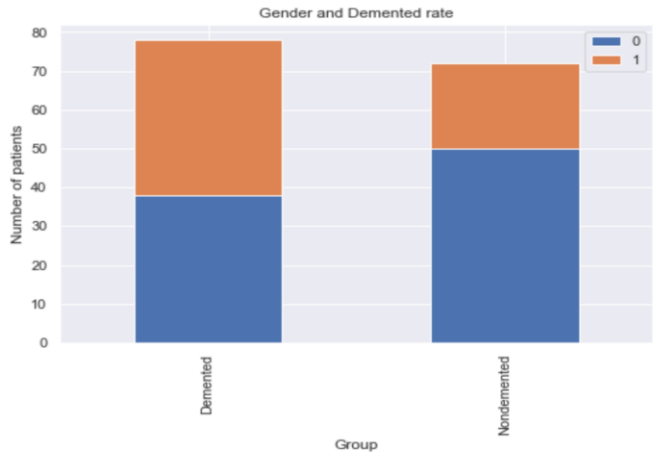
The proposed system consists of the following goals and advantages:
Goals:
• To analyze different features of the OASIS_Longitudinal Data set
• To compare various Machine Learning Algorithms in diagnosing Alzheimer's Disease.
• To build Effective model to diagnose early stage Alzheimer's Disease.
• To get high accuracy, recall, AUC.

4.1 SCREENSHOTS

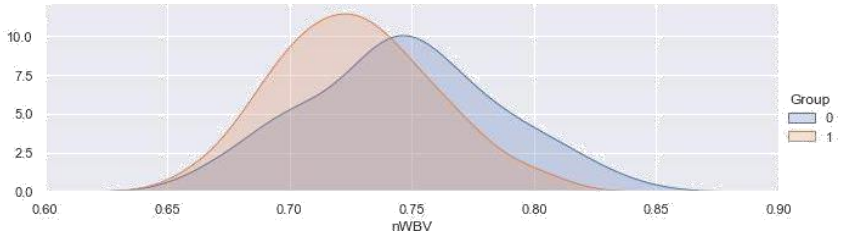
	Model	Accuracy	Recall	AUC
0	Logistic Regression (w/ imputation)	0.763158	0.70	0.766667
1	SVM	0.815789	0.70	0.822222
2	Decision Tree	0.815789	0.65	0.825000
3	Random Forest	0.868421	0.80	0.872222
4	AdaBoost	0.868421	0.65	0.825000

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

```
Text(0.5, 1.0, 'Gender and Demented rate')
```



The output indicating Gender versus Dementia



The output screen indicating nWBV versus Dementia .

The Output Screen showing relationship between Age and Alzheimer's Disease

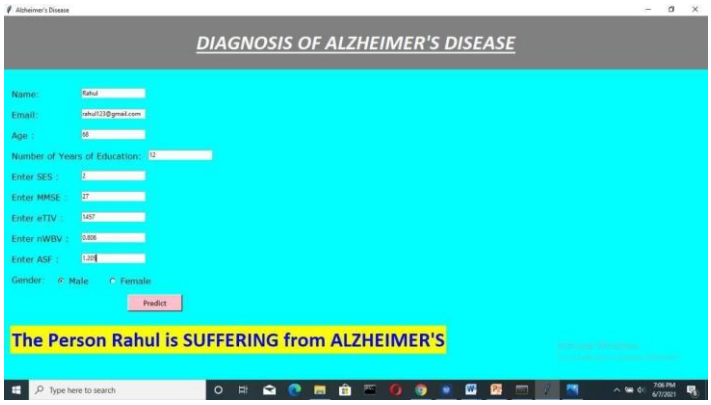
```
pd.isnull(df).sum()
```

```
Subject ID    0
Group         0
MR Delay      0
M/F           0
Age           0
EDUC          0
SES           8
MMSE          0
CDR           0
eTIV          0
nWBV          0
ASF           0
dtype: int64
```

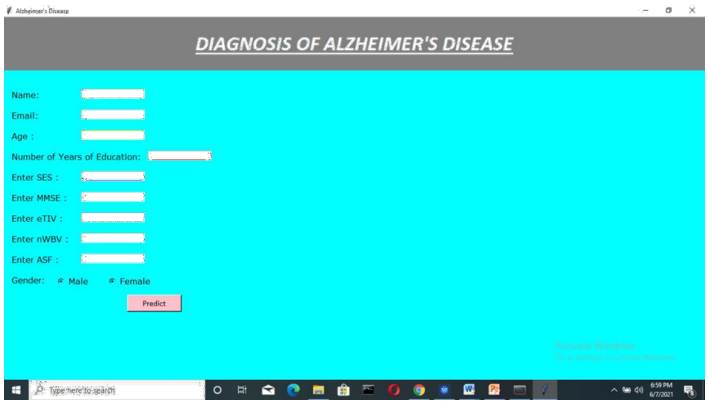
The Output Screen indicating missing values in the dataset

Comparison of Machine Learning Models

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621



Comparison of Machine Learning Models based on Recall



The Output Screen Indicating GUI of Diagnosis of Alzheimer's Disease

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

The GUI of Diagnosis of Alzheimer's Disease predicting Alzheimer's Disease

The Figure 5.17 indicating the prediction of GUI of the system that the person is suffering from Alzheimer's Disease according to the deatils he entered in the application.

The GUI of Diagnosis of Alzheimer's Disease predicting Non-Alzheimer's Disease The Figure 5.18 is indicating the output predicted by the GUI of Diagnosis ofAlzheimer'sDisease that the person is not suffering from Alzheimer's Diseas

Conclusion

 The paper presents the comprehensive overview of AD and how various learning approaches can analyze them. With the advancement in computational technology, data is enhancing day by day. It becomes difficult to handle this bulk of data. Machines as well as deep learning models explained in this study are the tools to tackle and analyze this bulky data. These learning models are able to analyze the data and can further classify or predict the result. Different learning model's analysis through experimental work are also shown in this article. The potential in terms of accuracy, recall, AUC, and time requirement to execute is analyzed efficiently in tabular as well as graphical form. With the comparative study of five ML models on oasis_longitudinal datasets, this paper concludes:

- Random Forest and AdaBoost achieves high accuracy as compared with others. This is because AdaBoost has the capability to turn weak classifier to strong one. Random Forest strength to overcome the overfitting problem, makes it better than the others.
- Random Forest gets high recall or true positive rate due to reduction of overfitting problem.
- AUC is a performance measure parameter, which is high with Random Forest.
- Along with good performance, Random Forest classifier takes more time to execute. This classifier training speed is low as compared to othersEncryption Within the tool is not possible as there are issues with JAVA base64 decoding while performing AES decryption's last block permutation. So, dependency is being used.

Future Scope

There are limitations in implementing a complex model because of the quantity of the dataset. Even though the nature of each feature is evident, the ranges of each group's test value are not classified well. In other words, we should have identified more clearly the differences in the variables which might have played a role in the result.The predicted value using the random forest model is higher than the other models. It implies there is a potential for higher p rediction rate if we pay more attention to develop the data cleaning and analysis process. Moreover, the perfect recall score 1.0 of SVM 1.0. Indicates that the quality and accuracy of the classification might decrease dramatically when we use different dataset.The main takeaway for us is that there are several key factors which are caused by Dementia and we should continue to check it and clear the process in different ways.For the further study, it is necessary for us to improve our understanding through more sophisticated EDA process with a larger sample size. For instance, we would try not only the age itself but also group it into generation, or grade volume of brain tissue or exam scores. If the results from this process are reflected in the data cleaning process and positively affect the decision making of the model, the accuracy of the prediction model can be further improved

REFERENCE

[1] J. Escudero, E. Ifeachor, J.P. Zajicek, C. Green, J. Shearer, S. Pearson, "Machine learning-based method for personalized and cost -effective detection of Alzheimer's disease", IEEE transactions on biomedical engineering, Vol. 60, no. 1, (2013), pp. 164 - 168.
[2] Q. Zhou, M. Goryawala, Cabrerizo, J. Wang, W. Barker, D.A. Loewenstein, R. Duara, M. Adjouadi, "An optimal decisional space for the classification of Alzheimer's disease and mild

DOI:10.46647/ijetms.2022.v06i06.110 ISSN: 2581-4621

cognitive impairment", IEEE Transactions on Biomedical Engineering, Vol. 61, no. 8, (2014), pp. 2245-2253.

[3] S. Athmaja, M. Hanumanthappa, V. Kavitha, "A survey of machine learning algorithms for big data analytics", Proceedings of the IEEE International Conference on Innovations in Information, Embedded and Communication Systems, Coimbatore, India, (2017) March 17-18.

[4] M. Bkassiny, Y. Li, S.K. Jayaweera, "A survey on machine-learning techniques in cognitive radios", IEEE Communications Surveys & Tutorials, Vol. 15, no. 3, (2012), pp. 1136-1159.

[5] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, V. C. M. Leung, "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View," IEEE Access, vol. 6, (2018), pp. 12103- 12117.

[6] Y. Ji, Q. Wang, X. Li, J. Liu, "A Survey on Tensor Techniques and Applications in Machine Learning" IEEE Access, Vol. 7, (2019), pp. 162950-162990.

[7] S. Das, M.J. Nene, "A survey on types of machine learning techniques in intrusion prevention systems", Proceedings of IEEE International Conference on Wireles