Alzheimer's Disease Prediction

BY
Sarang Sanjay Tirmanwar

Introduction

The dataset I used is from OASIS i.e., Open Access Series of Imaging Studies whose action is to make datasets to openly make available to the scientific community. They want to help discoveries in scientific and clinical neuroscience by assembling and openly publishing MRI data sets. Then, the data is being trained by multiple machine learning algorithms for training and testing for detecting low to moderate madness or Alzheimer's in people of different age and coitus. The data is originally being gutted to remove the null values and converted to needed type. After that, visualizations are being done between different attributes to know the relation between the parameters. Lastly, different types of models are being fitted in order to compare the perfection and delicacy.

Alzheimer's is an issue which is common and is a form of brain memory issue. The complaint is gradational, starting with mild memory loss and potentially progressing to the loss of communication and environmental mindfulness. The brain regions that are responsible for language, memory and study are affected by Alzheimer's complaint. It can significantly vitiate a person's capacity to carry out daily tasks. Alzheimer's complaint doesn't do naturally as people age. One of the original symptoms of Alzheimer's and other issues is constantly memory loss. An individual with Alzheimer's symptoms in addition to the memory issues might also have any of the following issues

- Loss of memory which interferes with diurnal life, similar as asking the same questions constantly or getting lost in a analogous or familiar places.
- Difficulty managing finances and paying bills.
- A challenge finishing routine duties at work, at home, or in rest.
- reduced or bad judgment
- losing effects and not being suitable to go back and find them.
- mood, station, or behavioural changes.

Hence, Alzheimer's complaint can be considered as a really serious issue and conducting new exploration and work is really important. Hence the data about the MRI of the brain is being taken for the analysis of the complaint. The discovery of the Alzheimer's is being done. However, also the inflexibility of the complaint is being calculated, If the Alzheimer's is present.

Methodology

There are 371 observations and 15 columns, as can be clearly seen after generating various descriptions of the data. The two ID columns can be taken out of the list, leaving the two columns that indicate if a patient has Alzheimer's disease and how severe it is. As was already indicated, our main objective is to establish if a patient has Alzheimer disease or not. CDR can

be transformed into a binary variable that shows whether a person has Alzheimer's disease or not. We can also see that handedness has a value of 1, making it irrelevant for all observations. The dataset we've chosen is oasis_longitudinal.csv. The Structure of the dataset is shown below.

VARIABLE	TYPE	DESCRIPTION
Subject.ID	char	Identification number
MRI.ID	char	Identification number
Group	int	Demented/Non-Demented
Visit	int	Number of visits
MR.Delay	int	Delay
M.F	chr	Male/Female
Hand	chr	Right hand/Left Hand
Age	int	Age in Years
EDUC	int	Years of Education
SES	int	Socioeconomic Status
MMSE	int	Mini Mental State Examination
CDR	num	Clinical Dementia Rating
eTIV	int	Estimated Total Intracranial Volume
nWBV	Num	Normalize Whole brain volume
ASF	Num	Atlas Scaling Factor

Data Cleaning

The process of locating and eliminating inaccurate or flawed data from a database is known as data cleansing. The method is typically applied to databases to identify any incorrect, missing, unreliable, or irrelevant data pieces and then modify, replace, or remove them. When it comes to maintaining the accuracy of client addresses or ensuring that bills are sent to recipients in a timely manner via email or postal delivery, business operations heavily rely on data. Business enterprises must prioritize data quality if they want to make sure that

customer data is used in the most beneficial and effective ways possible to raise the intrinsic value of the brand.

Managing missing values

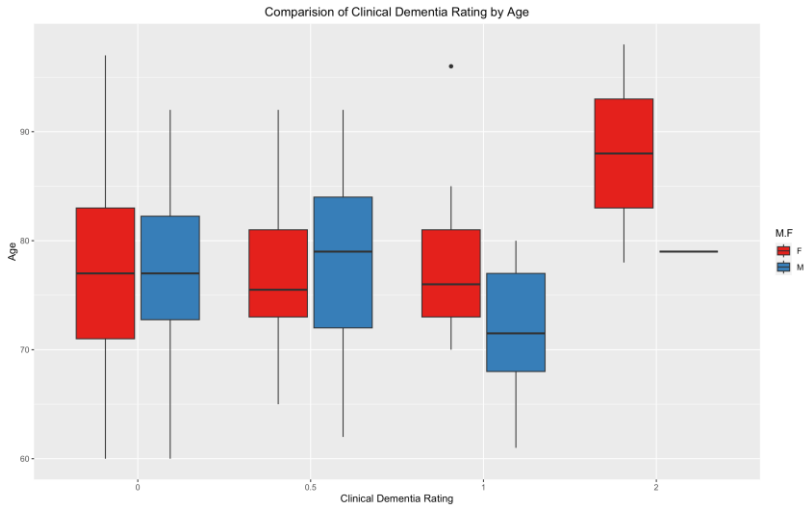
Managing missing values is the first basic step that we need to follow in the data cleaning process. In this data set there are certain missing values in the columns named "SES" which contains 19 null values and "MMSE" contains 2 null values. As "SES" and "MMSE" are of integer data type, we can replace the null values with the mean of the entire column.

Removing unwanted columns

The columns which doesn't show any relation to any of the other columns in the entire dataset should be removed from the dataset. In this data set, the columns which doesn't show any relation for the statistical visualization are "Subject ID", MRI.ID, MR.Delay, Hand. So these columns are removed from the dataset.

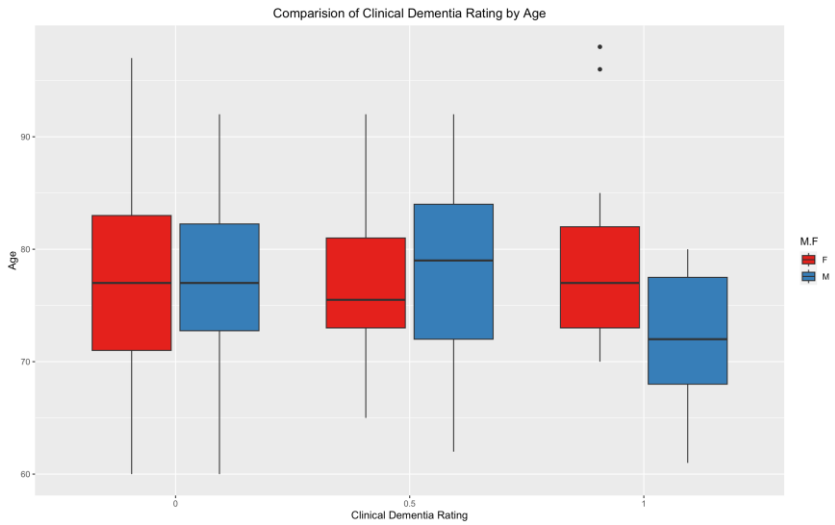
Exploratory Data Analysis and Visualization.

1. Relation between Male/Female and CDR.



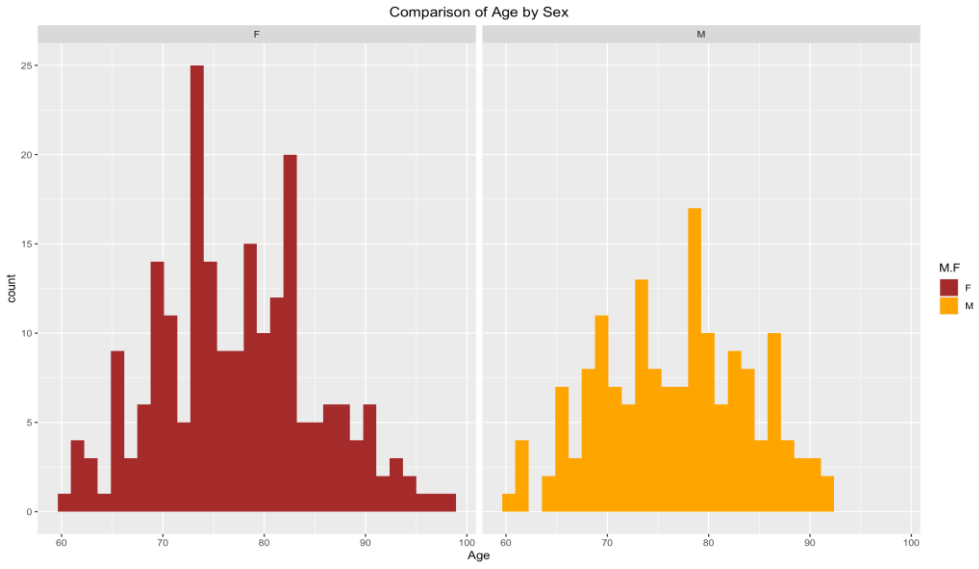
From the above figure, we can clearly understand that, The red box is for Females and the Blue plot is for Males, the line in the middle represents the median value, the first box plot represents the male and females are having same median values for '0' Clinical Dementia Rating. And for '0.5' Clinical Dementia Rating, the median value for Females is less when compared with the males. For '1' Clinical Dementia Rating, the median value for Females is more when compared with the males, and for the '2' Clinical Dementia Rating, the median value for females are more when compared to other clinical Dementia Rating, but as there is no sufficient data for the '2' Clinical Dementia Rating, we need to group by Clinical Dementia Rating '1' and Clinical Dementia Rating '2'.

2. Relation between Male/Female and CDR after modifications done.



- After grouping Clinical Dementia Rating of '1' and '2'. We can observe that, the median values for Male and Females are same for Clinical Dementia Rating '0'. The Clinical Dementia Rating for '0.5' has a median values which is higher for Males when compared to Females and for Clinical Dementia Rating of '1' has a median values which is higher for Females and lower for the Males.
- The above box plot is done in R programming language, with the help of ggplot library.

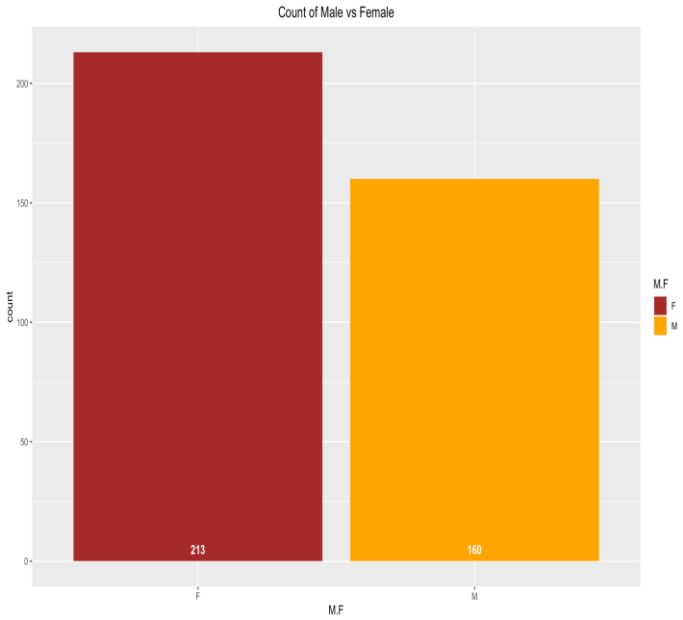
3. Comparison of Age by Sex



- From the above histogram, we can clearly understand the distribution of the Age by Male and Females affected. The Age increases gradually after the age 73 for the Females, and increases after the age of 78 to 80 and decreases gradually after the age 80 for the males.

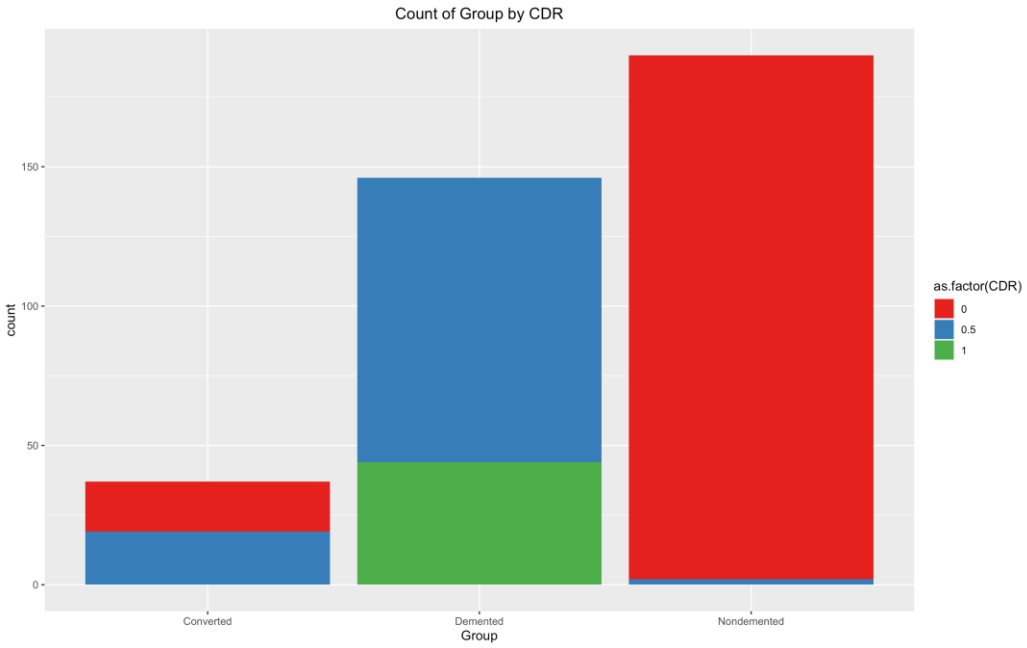
- We can also observe that, the distributions are almost same. The above histograms are done In R programming language with the help of ggplot library.

4. Count of Males and Females:



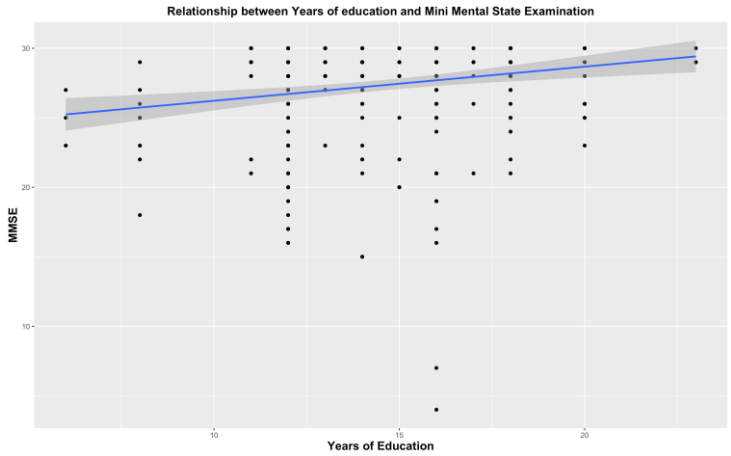
- From the above figure i.e., count of Males and females, the red coloured bar represents Females and yellow colour represents Male, by this we can understand that there are more number of females than males who are affected with the disease.

5. Count of Group by Clinical Dementia Rating:



The above figure represents the count of Group by clinical Dementia Rating, Here Red colour represents Clinical Dementia Rating of '0' and Blue colour represents Clinical Dementia Rating of '0.5' and Green represents Clinical Dementia Rating of '1'. From the above figure, we can also conclude that Converted group has equal number of people who has Clinical Dementia Rating of '0' and Clinical Dementia Rating of '0.5'. Demented group has more number of people who has Clinical Dementia Rating of 0.5 when compared to 1 and from the Non Demented group we can say that there are more number of people who has CDR of 0 and very less number of people who has CDR of 0.5.

6. Relation between Years of Education and Mini Mental State Examination

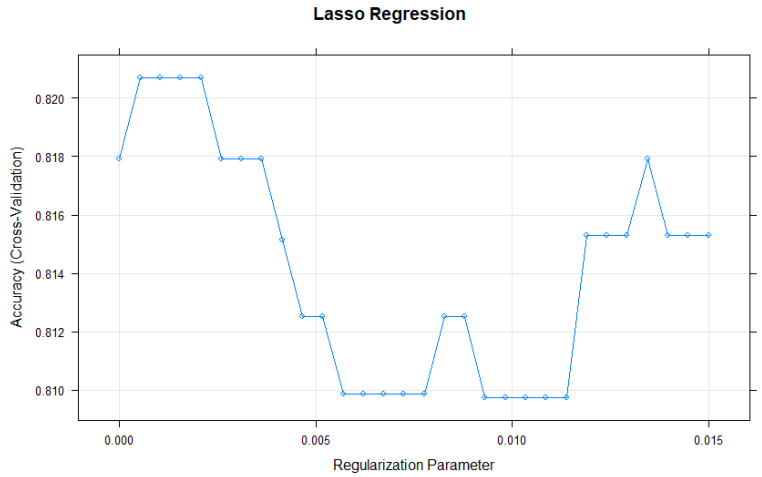


From the above figure, we can observe that, there is linear distribution between Mini Mental State Examination and Years of Education. EDUC and MMSE have a positive correlation, as shown by the scatter plot with linear regression lines for Male and Female.
We can say that, One's MMSE score will often increase with the number of years of education they have.

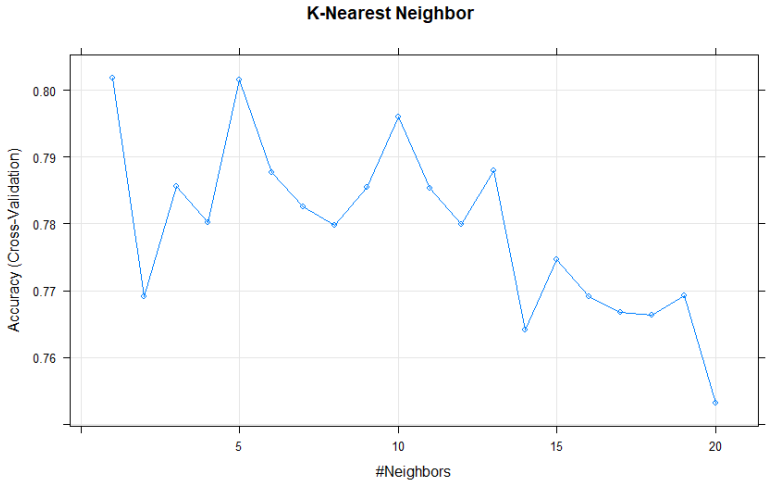
Modelling

The problem that we are trying to solve is – predicting if a person is positive or negative for Alzheimer's disease. The algorithms for modelling which we used are Lasso Regression, Decision Tree, KNN, Random Forest and SVM. For all the above models, we are using different metrics to calculate the test results. And, we are calculating the accuracy of each model which acts as the primary metric for comparison of all the models.

1) **Lasso Regression:** Lasso Regression is being used here and the plot is shown in the figure. We can find the accuracy levels ranging from 0 to 81.67 with respect to the regularization parameter. We can conclude that the accuracy of the model is 81.67. 'Alpha' which is the tuning parameter is held constant for the value of 1. Accuracy is being used to select the optimal model which has the highest accuracy i.e., 81.67 percent. After selecting the model, we could see that the lambda value is 0.0020.

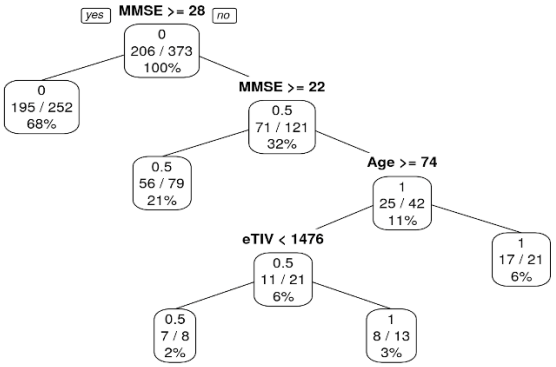


2) K-Nearest Neighbor : K-Nearest Neighbor is a lazy learning algorithm. Firstly, the test data is being loaded and the accuracy is being calculated for different models. Th value of K is being changed constantly and the accuracy is being calculated. By using the value of k as 1, the model has the highest accuracy of 81.9%.

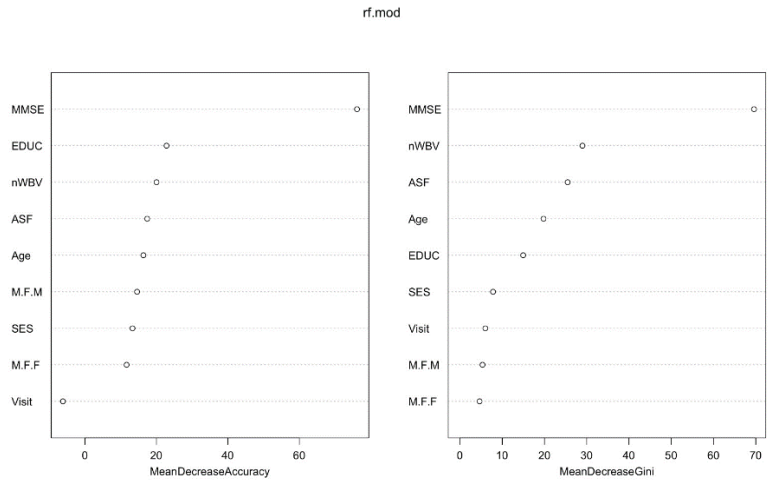


3) **SVM** : Firstly the model is divided into training and testing data and then the accuracy or precision is calculated. Based on the training data, the model is built. After that the accuracy is being calculated on the model built. Here we are using SVM model to predict is a person is testing positive or negative for the disease. Sigma which is the tuning parameter is held constant for the value of 0.097. Accuracy is being used to select the optimal model which has the highest accuracy i.e., 85.41 percent. After selecting the model, we could see that the sigma value that was used is 0.097 and 'C' value is 8.

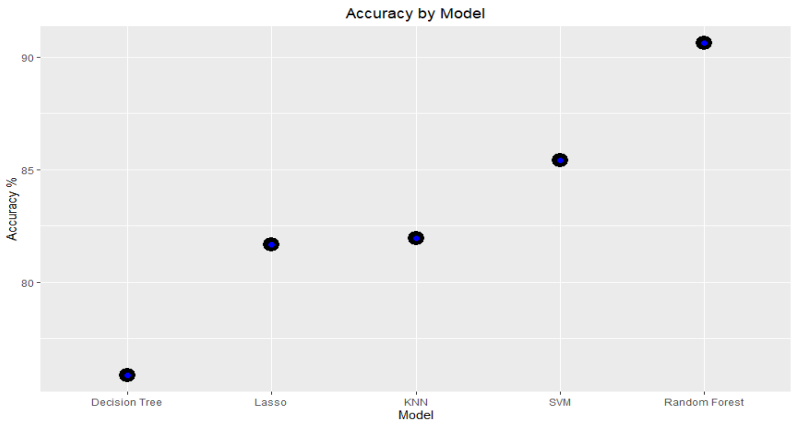
4) **Decision Tree :** Decision tree model is being built based on the training data. We are building a test dataset and train datset from the initial dataset. We are generating a 5-folds cross validation plan. The features used for the decision tree are MMSE, Age, eTIV. The final decision tree is being built, the accuracy being 75.87.



5) **Random Forest :** When compared with other models, the accuracy provided by the bagging technique used by Random Forest is higher, and its interpretability is also strong. Both the variance and the overfitting issue are avoided. Feature selection is being done after the model fitting. We can conclude that the features MMSE, EDUC and nWBV are the most import features. Also, the error is being calculated for different number of trees. We can understand that the error is constantly being decreased everytime when the number of trees is increased.



6) Accuracy Comparison:



The accuracy is being compared for all the models done. We can see that the accuracy is being highest for the Random Forest model i.e., 90.62%. The least accurate model is Decision Tree which has 75.87% accuracy. Hence, we can conclude that the model Random Forest performed really well, and can further used to test a patient for Alzheimer's detection.

Conclusion : In conclusion, this study sought to shed some light on the models that are used to detect the Alzheimer's. It concludes that Random forest is the highly accuracy model. According to the report, detection is being done which is about 90% accurate and measures must be taken once the Alzheimer's is detected to be positive.

References:

[1] "Decision trees : ," *GeeksforGeeks*, 06-Dec-2022. [Online]. Available: https://www.geeksforgeeks.org/decision-tree-in-r-programming/.

[2] Sejaldua, "Glioma-detection: MATLAB detection of Alzheimer's ," *GitHub*, 05-Jul-2019. [Online]. Available: https://github.com/sejaldua/glioma-detection/blob/master/oasis%20dataset/oasis_longitudinal.csv.

[3] Zach, "Lasso regression : Definition, Methods, Conclusion.," *Statology*, 13-Nov-2020. [Online]. Available: https://www.statology.org/lasso-regression-in-r/.