# ALZEIMER’S DETECTION

**INFO 7390 SUMMER 2019 NORTHEASTERN UNIVERSITY**

**Chinmay Arolkar**

## ABSTRACT:

Alzheimer's disease (AD) is a neurodegenerative disorder of uncertain cause and pathogenesis that primarily affects older adults and is the most common cause of dementia. The earliest clinical manifestation of AD is selective memory impairment and while treatments are available to ameliorate some symptoms, there is no cure currently available.

Brain Imaging via magnetic resonance imaging (MRI), is used for evaluation of patients with suspected AD. However, in order to reach that stage clinicians and researchers will have to make use of machine learning techniques that can accurately predict progress of a patient from mild cognitive impairment to dementia. During this reaserch project, I am trying build models using different machine learning techniques and at the end to compare which performs the best.

## INTRODUCTION

**Alzheimer’s Disease:**

Alzheimer’s disease is currently ranked as the sixth leading cause of death in the United States, but recent estimates indicate that the disorder may rank third, just behind heart disease and cancer, as a cause of death for older people.

Alzheimer’s is the most common cause of dementia among older adults. Dementia is the loss of cognitive functioning—thinking, remembering, and reasoning—and behavioral abilities to such an extent that it interferes with a person’s daily life and activities. Dementia ranges in severity from the mildest stage, when it is just beginning to affect a person’s functioning, to the most severe stage, when the person must depend completely on others for basic activities of daily living.

The causes of dementia can vary, depending on the types of brain changes that may be taking place. Other dementias include Lewy body dementia, frontotemporal disorders, and vascular dementia. It is common for people to have mixed dementia—a combination of two or more types of dementia. For example, some people have both Alzheimer’s disease and vascular dementia.

## DATASET:

The data is taken from 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.

Content in the data:

We will be using the longitudinal MRI data.

The dataset consists of a longitudinal MRI data of 150 subjects aged 60 to 96.

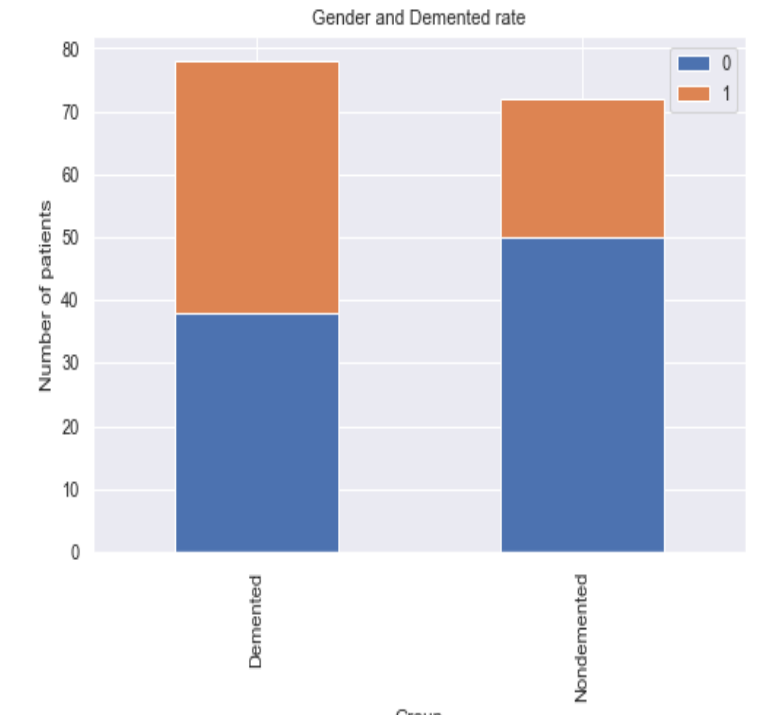
Each subject was scanned at least once.

Everyone is right-handed.

72 of the subjects were grouped as 'Nondemented' throughout the study.

64 of the subjects were grouped as 'Demented' at the time of their initial visits and remained so throughout the study.

14 subjects were grouped as 'Nondemented' at the time of their initial visit and were subsequently characterized as 'Demented' at a later visit. These fall under the 'Converted' category.



## GOAL

Goal of our project is

* To predict Alzheimer’s disease
* To use different models for the prediction
* To find out best prediction model for the data

## PROCESS:

We inspect the data to identify the data problems, null values and outliers. We have to solve identified problem to begin the analysis without any problem.

There a few null values which were taken care of. We have used numerous libraries for an entire process like data exploration and visualization.

Let us understand the process to get the best hyperparameters.

First, we have to make sure the data is clean i.e we have to remove null values in the datasets.

There are numerous methods to remove the null values in the data sets.

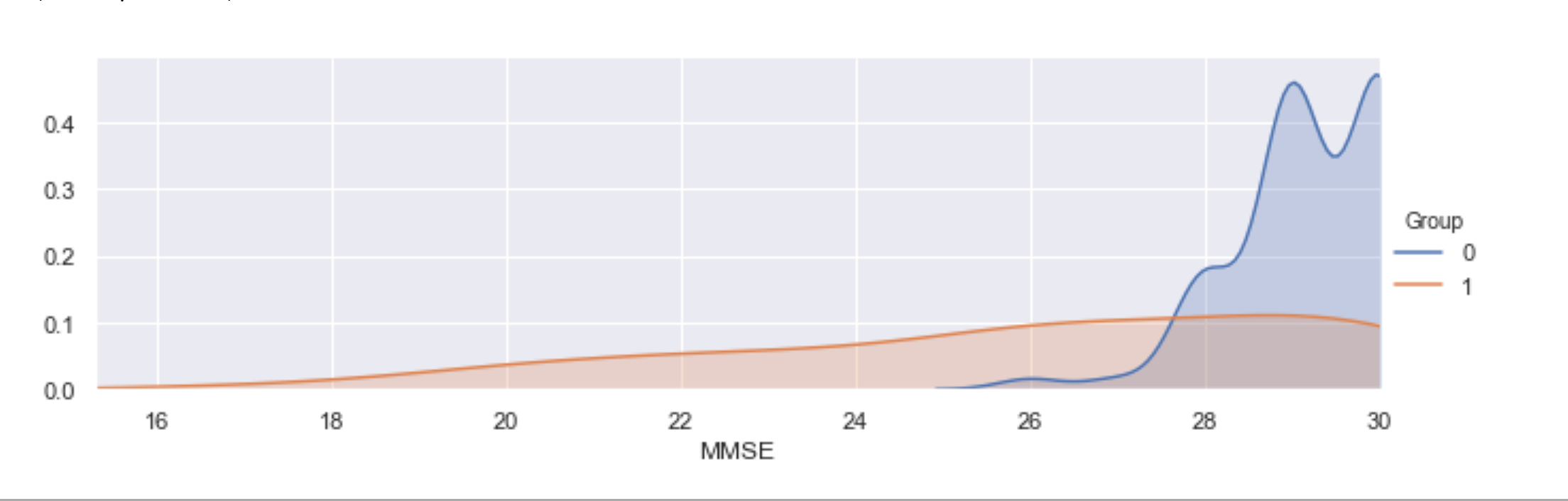
1. One can try dropping the columns which have most number of the null values in the dataset. These columns can be irrelevant and may have less significance to the data.
2. If there are a few number of null values in the column, you try removing by replacing the null value with the mean, median or mode etc. That depends on the values in the columns.
3. If there are any categorical values in the data set, you can replace these values by using one hot encoding method. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

Once we are sure that all the null values are taken care of we can start with the further analysis.

Here I have dropped multiple visits to avoid confusion and only considered the 1st visit. I have also converted the string values into the binary or number equivalent format to get the process done.

The following is the visualization of the men to female ratio:

From the above graph we can see that there are almost equal numbers of demented men and women. On the other side we can see that, the number of nondemented females are low compared to men. This shows that men are more likely suffer from dementia than women.



The above is graphical comparison of the Mini Mental Status Exam (MMSE) which has the maximum score of 30. The above graph shows the Nondemented group got much more higher MMSE scores than Demented group.

From the several graphs I have concluded that, Nondemented group has higher brain volume ratio than Demented group. This is assumed to be because the diseases affect the brain to be shrinking its tissue.

Other than this, I came with the following conclusion from the graphs:

1)Men are more likely with demented, an Alzheimer's Disease, than Women.

2)Demented patients were less educated in terms of years of education.

3)Nondemented group has higher brain volume than Demented group.

4)Higher concentration of 70-80 years old in Demented group than those in the nondemented patients.

## MODELS:

## Here I have created total 6 models which are as follows along with their result:

## SVM:

## A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

## In this model, I have iterated the values for the parameter c and kernel. I have used cross validation to get the accuracy score. I have rebuilt the model with the best parameters to get the best score. After successfully building the model, SVM gave the accuracy of 81.5%.

## DECISION TREE:

## A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

## In this model, I have iterated the values for maximum deapth. I have used cross validation to get the accuracy score. I have rebuilt the model with the best parameters to get the best score. After successfully building the model, Decision tree gave the accuracy of 81.57%.

## RANDOM FOREST:

## Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

## The method to build the model for this is same as the previous ones. But here I have received accuracy of 86.84%.

## K-MEANS CLUSTERING:

## K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space.

## To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

## In the K-Means clustering I have selected number of clusters as 3. Which means in the output there should be 3 clusters formed. After building the model correctly we get the following output:

## 

## In the figure, we can see that there are 3 clusters formed properly after applying the k-means classification. This model gives us average accuracy of 80% which is considered to be good.

## KNN:

## K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

## In this model the data is split into train and test. The data is trained in the KNeighborsclassifier and the model is evaluated for the prediction.

## Here I have received average 70% of the accuracy.

## NAÏVE BAYES:

## A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis. These classifiers are widely used for machine learning because they are simple to implement.

## The model is generated using Bernoulli. The test accuracy turns out to be 91.11%.

## BEST MODEL:

## After observing the models, I have found that the Naïve Bayes model gets the most accuracy. Which shows that the it is effeicient and makes more important because it is easy to implement on any data set.

## CONCLUSION:

## After going through the all of the analysis we get to see the accuracy or precision of different models. The accuracy is as followed:

## SVM = 81.5%

## Decision Tree = 81.57%

## Random Forest = 86.84%

## K-Means Cluster = 80.00%

## Naive Bayes = 91.11%

## Hence, I can conclude that the Naive Bayes model is the best model for this data set. It is very simple and efficient compared to all other models.

## ACKNOWLEDGEMENT:

## We would like to thank Prof. Nick Brown and Prabhu Subramanian for their constant support and guidance. They have helped us to achieve our goal.

## REFERENCES:

## 1.The reference dataset was from the OASIS open dataset website http://www.oasis-brains.org/ and used for the analysis is https://www.kaggle.com/vaibhavmathur96/detecting-early-alzheimer-s.

## 2.All these data sets are made up of data from the US government. Data from the World Factbook is public domain. The website says, "The World Factbook is in the public domain and may be used freely by anyone at any time without seeking permission." https://www.cia.gov/library/publications/the-world-factbook/docs/faqs.html

## 3.https://docs.anaconda.com/anaconda/packages/pkg-docs/ - Used for python packages.

## 4.Google was used for understanding the graphs and explaination-www.google.com

## 5.Udemy and Edureka videos are referred for better understanding

## 6.Github

## 7.K-Means reference: https://www.kaggle.com/andyxie/k-means-clustering-implementation-in-python

## 8.KNN reference: https://www.youtube.com/watch?v=s-9Qqpv2hTY

## 9.Naive Bayes reference: https://www.youtube.com/watch?v=99MN-rl8jGY

## 10.https://en.wikipedia.org/wiki/Precision\_and\_recall

## 11.http://scikitlearn.org/stable/modules/preprocessing.html#imputation