Kin-Keepers.com

Machine Learning Projects

DESIGN OF A MACHINE LEARNING BASED MODEL TO TRANSLATE FOR APHASIA AND/OR ALZHEIMER'S DISEASES

Researchers

Mr. Sudi Murindanyi

Mr. Sulaiman Kagumire

Volume One

Dementia Prediction

DISCLAIMER: This document is strictly private, confidential and personal to its recipients and should not be copied, distributed or reproduced in whole or in part, nor passed to any third party**.**

Abstract

Dementia being a major cause of creating dependency among aged people also has an inevitable impact on people suffering from it and the families around them. Since the symptoms are gradual and may overlap, diagnosing dementia and identifying its type is risky. The main purpose is to develop a machine learning-based method to early diagnose dementia using the dataset obtained from OASIS. Classification algorithms such as GaussianNB, K-Nearest Neighbors, Random Forest, XGBoost, LGBMClassifier, CatBoost, and GradientBoosting are used to find accuracy, recall, precision and confusion matrix. Implementation of the following algorithms provides accuracy in the range of 80 to 90 percent. KNN gives out accuracy of 81%, GaussianNB give an accuracy of 83%, Random Forest gives an accuracy of 89.3%, and XGBoost gives 88.4%, and LGBMClassifier gives 88%. Random Forest performs better than all other models considered.

Contents

Chapter one: Introduction

Dementia is a devastating illness that results in gradual loss of memory and other cognitive ability which is mostly identified in people more than 60 age groups, but this does not mean young people are not affected by it. Various types of dementia include Alzheimer’s, Lewy Body Disease, Frontotemporal dementia, Vascular dementia and more.

Alzheimer's disease (AD) is the most common type of dementia disease involving degeneration of the brain which is irreversible and gradually ends up with the complete brain failure. According to the statistics of Alzheimer's Association, AD accounts for 60–80% of the dementia cases [1]. In 2006, there were 26.6 million sufferers worldwide, and is expected to double by 2030 and triple by 2050 as projected by world health organization. AD is predicted to affect 1 in 85 people globally by 2050, and at least 43% of prevalent cases need a high level of care [2]. Aging and other factors increase the possibility of neuron degeneration and can lead to AD. As the world is evolving into an aging society, the burdens and impacts caused by AD on families and the society will be increasingly pronounced.

Studies have shown that AD is influenced by several factors such as age, education and socio-economic status. In normal aging, whole-brain volume decline begins in early adult hood and accelerates in advanced aging [3] [4] [5] [6] . Preferential volume loss of gray matter [7] and regionally specific thinning of the cortex are also noted [8]. Level of education, sex, socioeconomic status, and cardiovascular health have been identified as contributing factors in volume decline in advanced aging, suggesting that subclinical health conditions contribute to age related changes in brain structure [9]. Individuals with clinically diagnosed AD show substantially reduced over all brain volumes relative to age matched peers as well as regional volume loss that has been well documented in the hippocampal formation, among other regions [10]. Various biological and neuropsychological studies discover that AD can be predicted at its early stage and useful to take treatment in an efficient direction It starts from a specific subcortical region and increases to the cortical mantle with the passage of time. The most common effect of AD is memory loss and slows down the ability to do any task. It is found that MCI, a highly heterogeneous phenotypic spectrum, has very less considerable memory deficits than AD. These MCI may convert to AD. It was discovered that 10%-15% MCI patients converted to AD within a short span of time. So, MCI needs to be taken care with special attention in order to stabilize the chance of AD [11]. The development of AD can be predicted several years before, which is helpful in controlling the progress of AD.

MRI provides multi-mode information for the brain's structure and function. MRI works successfully in distinguishing healthy people with AD survivors. MRI results can identify the sMCI (stable MCI) and pMCI (progressive MCI) [11]. Neuroimaging techniques are progressing very fast that makes it difficult to integrate large scale high dimensional multimodal neuroimaging data. Thus, computer aided machine learning approaches are adopted for integrative analysis [12].

In this report, machine learning techniques are applied on OASIS dataset that has the data of demented, non-demented individuals. Machine learning has been implemented in various fields for the betterment of analyses and results, the health sector is the major area where the contribution of machine learning is found to be most useful. The main aim is to build a machine learning model that can diagnose dementia with better accuracy. In this report, KNN, Random Forest, XGBoost, GaussianNB, and Light GBM are implemented.

Chapter two: Exploratory Data Analysis

The Dataset

The ML models were evaluated on data obtained from the Open Access Series of Imaging Studies (OASIS). OASIS is a series of neuroimaging data sets that is publicly available for study and analysis.

The dataset used consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects (labelled converted) were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

The features included in the dataset are explained in ***Table 1***. Additional details of the directory structure, file naming scheme, and image characteristics can be found at http://www.oasis-brains.org/longitudinal_facts.html.

Table 1: Features and description

Feature	**Description**
Group	Class Label (Non-demented, Demented, Converted)
CDR	Clinical Dementia Rating [0=no dementia, 0.5=very mild AD 1=mild AD, 2=moderate AD]
M/F	Gender of subject (M or F)
Age	Age of subject
EDUC	Education level of test

SES	Socio-Economic Status, as assessed by the Hollingshead index of social position [(1(highest) to 5 (lowest)]
MMSE	Mini Mental Status Examination score [0(worst) to 30(best)]
eTIV	Estimated total intracranial volume
nWBV	Normalized whole-brain volume, expressed as a percent of all voxels in the atlas-masked image that are labelled as gray or white by the automated tissue segmentation process
ASF	Atlas scale Factor, scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix)

Data Analysis and Feature Engineering

Data Analysis (DA) is understanding the datasets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. EDA, is essentially a type of storytelling for statisticians. It allows us to uncover patterns and insights, often with visual methods, within data.

Feature Engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. This involves techniques like Label encoding, imputation, feature selection, and feature extraction. The features in your data will directly influence the predictive models you use and the results you can achieve. You can say that: the better the features that you prepare and choose, the better the results you will achieve.

The dataset has a shape of 373 rows and 10 columns (373 X 10) including the target label, ‘Group’. Features ‘Group’ and ‘M/F’ consist categorical data, hence calling for Label Encoding. Features ‘MMSE’ and ‘SES’ have missing values, hence calling for imputation using the mean value for each feature. As shown in *figure 1* below, Dementia is more common in men than women and more women turned out to have Dementia on subsequent visits.

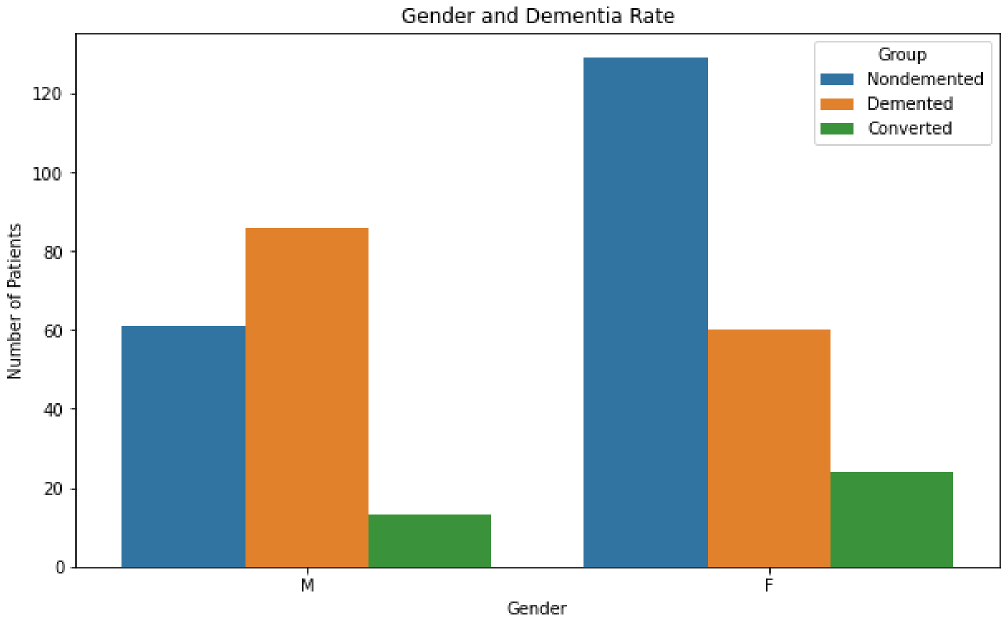


Figure 1: Gender and Dementia rate

The kdeplot in ***figure 2*** shows that the mean age in all the three classes, is 79 for Converted, 76 for Demented, and 77 for non-demented. KDE Plot (Kernel Density Estimate plot) is used for visualizing the Probability Density of a continuous variable. It depicts the probability density at different values in a continuous variable. Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. This also shows that patients age is normally distributed with minimum of 60 and maximum of 98.

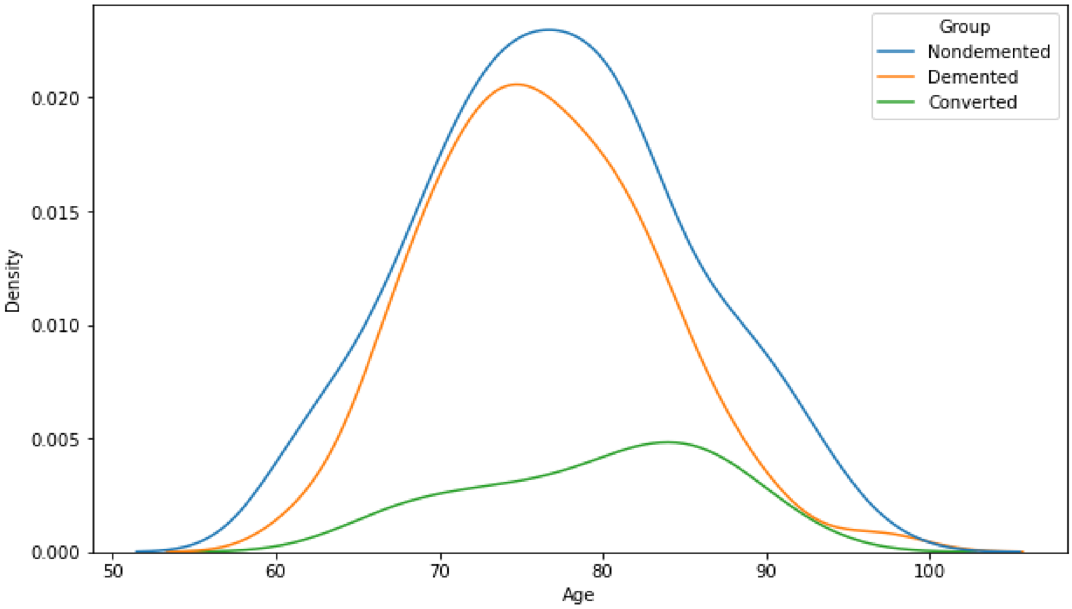


Figure 2: Probability density and Group

Dementia is observed as per age and discovered that the chance of dementia is more at an age of 70-80 and of course age plays a significant role. ***Figure 3*** illustrates a violin plot of Age and Clinical Dementia ratings. Violin plots use kernel density estimation for displaying underlying distribution

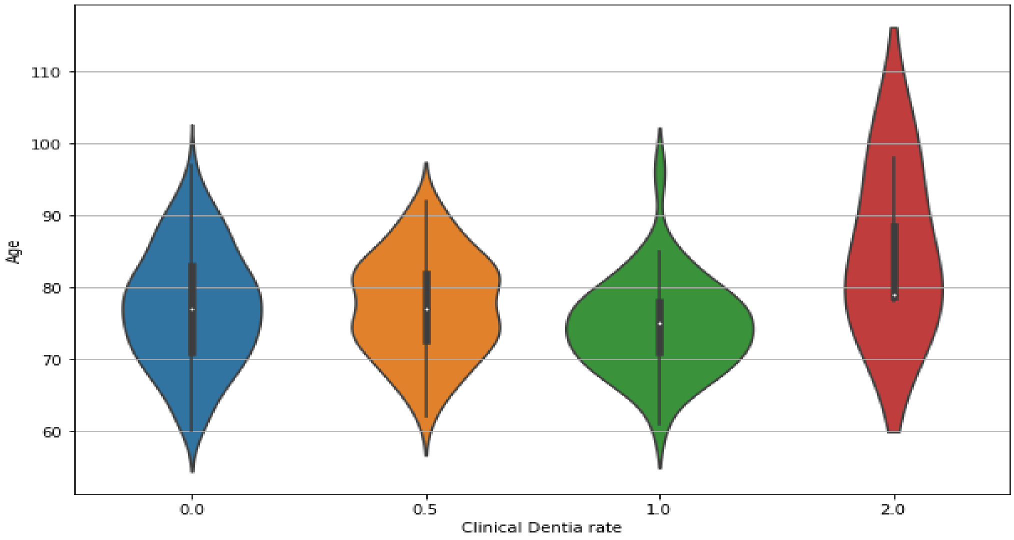


Figure 3: Age and Clinical Dementia rate

The scatterplot shown in ***figure 4*** illustrates the relationship between the brain volume and Age. We notice that there's a high negative correlation between Age and 'nWBV' because normalized whole brain volume reduces with increase in age. We also notice that more patients with Dementia tend to have a low brain volume compared to the non-demented.

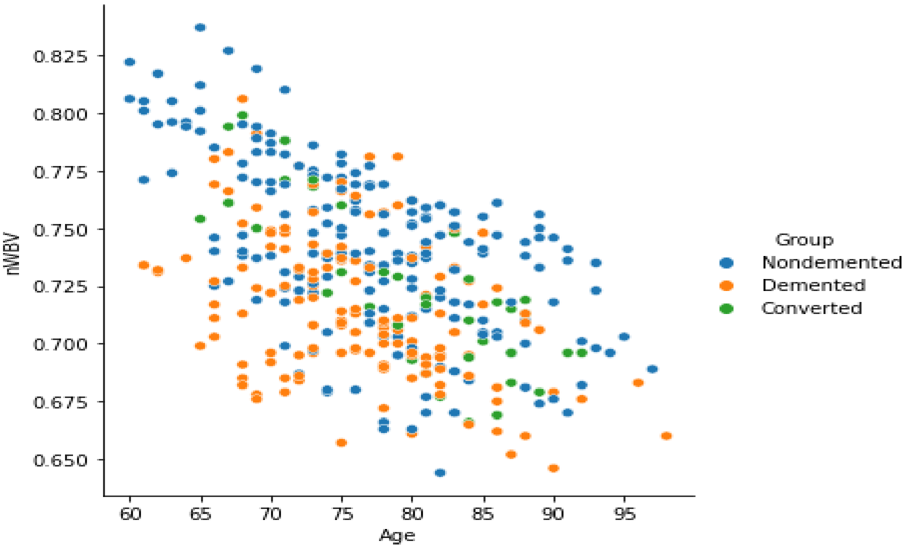
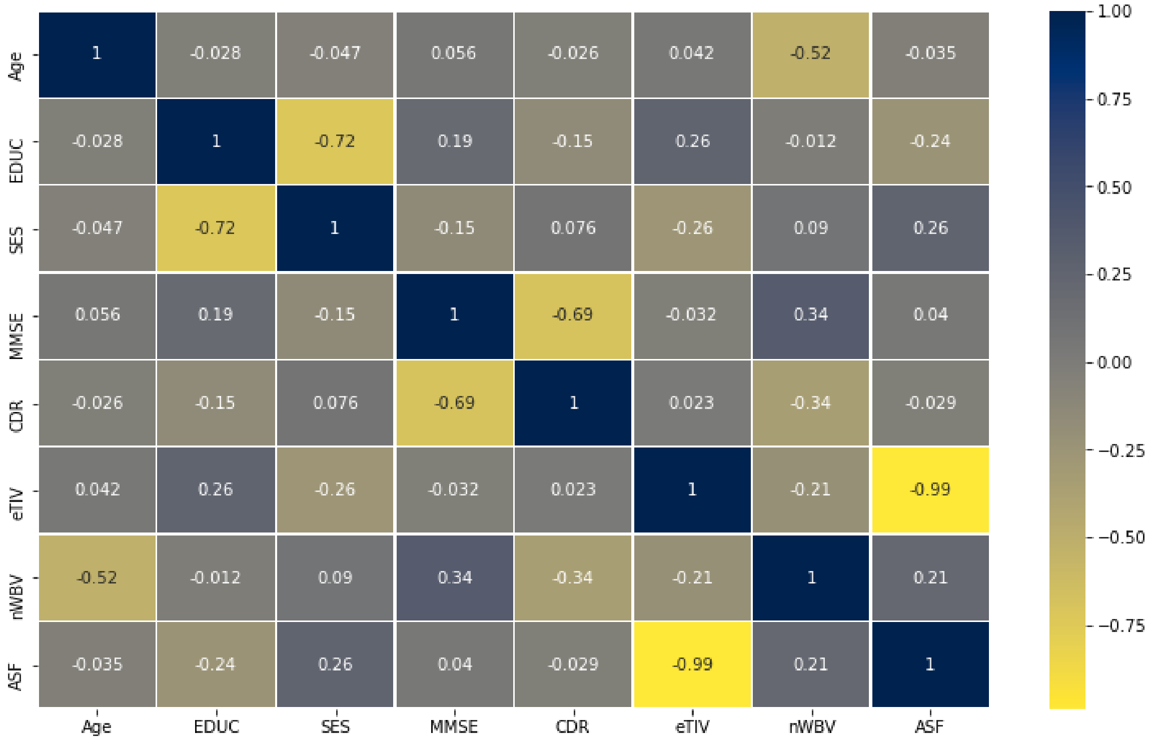


Figure 4: Brain volume and Age

The correlation matrix in ***figure 5*** shows the correlation coefficients that indicate the strength of the linear relationship between two different features. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. However, when interpreting correlation, it's important to remember that just because two variables are correlated, it does not mean that one causes the other.



	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
Age	1	-0.028	-0.047	0.056	-0.026	0.042	-0.52	-0.035
EDUC	-0.028	1	-0.72	0.19	-0.15	0.26	-0.012	-0.24
SES	-0.047	-0.72	1	-0.15	0.076	-0.26	0.09	0.26
MMSE	0.056	0.19	-0.15	1	-0.69	-0.032	0.34	0.04
CDR	-0.026	-0.15	0.076	-0.69	1	0.023	-0.34	-0.029
eTIV	0.042	0.26	-0.26	-0.032	0.023	1	-0.21	-0.99
nWBV	-0.52	-0.012	0.09	0.34	-0.34	-0.21	1	0.21
ASF	-0.035	-0.24	0.26	0.04	-0.029	-0.99	0.21	1

Figure 5: Correlation coefficients between features

We notice a high negative linear relationship between socio-economic status and level of education. This makes senses because the lower the economic status (5 is the lowest), the lower the level of education. ‘ASF’ and ‘eTIV’ are also highly corelated negatively. Clinical Dementia ratings and Mini mental examination score are other two highly correlated features, which also makes sense because patients with high scores in the MMSE are actually non-demented compared to demented patients who performed poorly in this examination.

Chapter four: Modeling and Evaluation

To build an early prediction model of AD dementia based on longitudinal data, we first train the dataset to learn compact representation and encode the dynamics of longitudinal measures for each subject. This work contains hyper-parametric classifiers from Naïve Bayes, XGBoost, Random Forest models, Light gradient boosting framework machine LGBM, and K-nearest Neighbors.

XGBoost is a special type of Ensemble Learning technique which perform combination of various weak learning and strong learning and focus on its predecessor's faults to improve its accuracy. Bagging: The accuracy of classification and regression tree can be improved through bagging and also known as ensemble method.

Random Forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

Light BGM extends the gradient boosting algorithm by adding a type of automatic feature selection as well as focusing on boosting examples with larger gradients. This can result in a dramatic speedup of training and improved predictive performance.

K-nearest neighbors is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

Naive Bayes classifiers works on Bayes' theorem of probability and produces very accurate classification output with a reduced training time when compared to conventional supervised or unsupervised learning algorithms.

Metrics

Table 2: Models results

Classifiers	**Accuracy**
XGBoost	88.4
Random Forest	**89.3**
KNN	81.2
GaussianNB	83.0
Light BGM	87.5

Table 1 shows accuracies from each classifier models used. Random Forest performs better than others.

The confusion matric below shows the True Positives, True Negatives, False Positives, and False Negatives for the Random Forest Classifier. A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

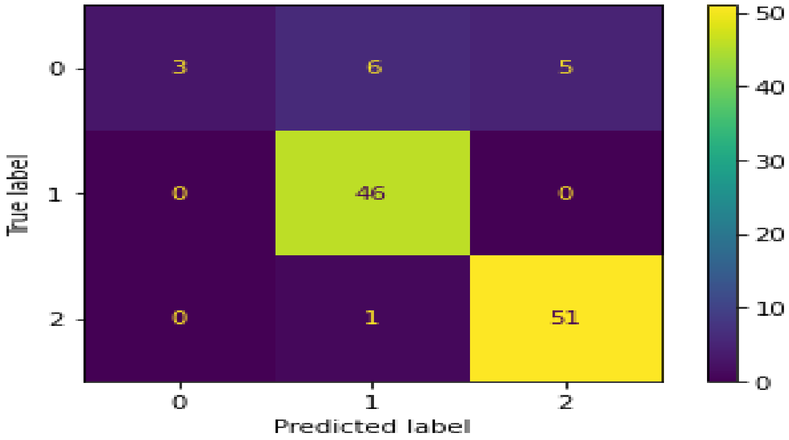


Figure 6: Confusion matric

The overall accuracy, recall, precision, and f1_score for each model is shown in the bar plot below.

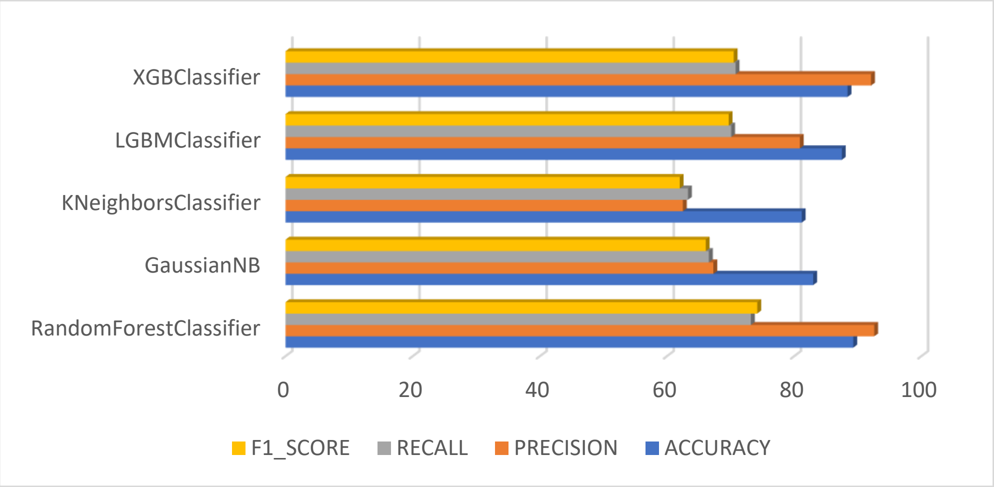


Figure 7: Models results

Chapter five: Conclusion, Next Task, and Challenge

Conclusion

This effort was to experiment with Alzheimer's data with machine learning algorithms to early diagnose dementia. We have investigated the longitudinal MRI images consisting of neuroimaging test images in the OASIS dataset. 8 out of 15 features were considered for data modeling and predicting the disease. The EDA conducted on the dataset reveals that the age is the most dependent feature for Dementia. This form of diagnosis aims to save time for neurologists and patients to get the correct diagnosis at the right time.

Next Task

The main objective of the project is to design a machine learning model that can bridge the gap between an individual with aphasia and/or Alzheimer’s diseases and the receiver. We divided our project into different tasks to help us understand and do the project in a good way. We started by doing Dementia Prediction using the OASIS dataset, and the results helped us understand Dementia, its features, and the model we developed that can be about to predict if a patient has Dementia or not. Our next step will be, using sensors data to classifier different levels of Alzheimer’s or analyze the downloaded dataset to produce Text Corpus and Audio Corpus for our Audio to Text model.

References

[1] "Alzheimer's and Dementia," 21 July 2021. [Online]. Available: https://www.alz.org/alzheimers-dementia/what-is-alzheimers.

[2] J. Elizabeth, Z.-G. Kathryn and H. Michael Arrighi, "Forecasting the global prevalence and burden of Alzheimer’s disease," *Alzheimer's & Dementia,* pp. 186-191, 24 August 2017.

[3] D. Charles, M. Joseph, H. Daniell, H. John, T. Mats and A. Rhoda, "Measures of brain morphology and infarction in the framingham heart study: establishing what is normal," *Neurobiology of Aging,* pp. 491-510, February 2015.

[4] A. F. Fotenos, "Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD," *Neurology,* April 2005.

[5] S. Elizabeth, T. Paul, L. Christiana, W. Suzanne, K. Eric and T. W. Arthur, "Longitudinal mapping of cortical thickness and brain growth in normal children".

[6] D. Sarah, "A review of the biological bases of ADHD: what have we learned from imaging studies?," *Ment Retard Dev Disabil Res Rev,* pp. 184-95, September 2003.

[7] N. Raz, A. Williamson, F. Gunning-Dixon and D. Head, "Neuroanatomical and cognitive correlates of adult age differences in acquisition of a perceptual-motor skill," in *Microscope Research and Technique*, pp. 85-93.

[8] D. H. Salat, "Thinning of the Cerebral Cortex in Aging," *Cerebral Cortex,* p. 721–730, 1 July 2004.

[9] F. Anthony, A. Z. Snyder, L. E. Girton, J. C. Morris and R. L. Buckner, "Brain volume decline in aging: evidence for a relation between socioeconomic status, preclinical Alzheimer disease, and reserve," *Arch Neurol,* pp. 113-120, 2008.

[10] M. A. Nicole, A. Yang, B.-H. Lori, D. Jimit, E. Guray and F. Luigi, "Sex differences in brain aging and predictors of neurodegeneration in cognitively healthy older adults," *Neurobiol Aging,* pp. 146-156, 20 5 2019.

[11] L. Johann, K. Sandra, L. Claus, M. Doris, K. Stefanie, P. Gisela, D.-B. Peter, P. Walter and A. Eduard, "Awareness of memory deficits in subjective cognitive decline, mild cognitive impairment, Alzheimer's disease and Parkinson's disease," *Int Psychogeriatr,* pp. 357-366, 2015.

[12] L. Siqi, L. Sidong, C. Weidong, P. Sonia, K. Ron and F. Dagan, "Early diagnosis of Alzheimer's disease with deep learning," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, Beijing, China(362), 2014.