UPSCHOOL MACHINE LEARNING & DEEP LEARNING PROGRAM
IN PARTNERSHIP WITH GOOGLE DEVELOPERS

Prediction of The Different Progressive Levels of Alzheimer's Disease with MRI Data

Data Science Project Report

Eda AYDIN

Table of Contents

A. Business Understanding – Project Objective

This is an optional model development project on a real dataset related to predicting the different progressive levels of Alzheimer's disease (AD). The student is expected to use machine learning algorithms, and the TensorFlow library for the modeling process and will be asked to submit predicted labels for a test dataset by which their score be evaluated objectively.

This project is included in the *UpSchool – Google Developers Machine Learning – Deep Learning Program.*

In this project, it is necessary to provide an end-to-end data science model to determine the level of Alzheimer's disease. Levels are categorized in order from low to high: 0,1,2,3. (these are the progressive levels of Alzheimer's disease).

B. Data Understanding

B.1 Data Short Information

The dataset is provided by the NACC – Uniform Data Set by the University of Washington. In addition, the Researchers Data Dictionary (NACC Uniform Data Set - Researcher Data Dictionary, 2015) pdf file prepared by the researchers of the same university is used to determine the terms in the dataset. In addition, the Genetic Data Dictionary (Researchers Data Dictionary - Genetic Data (RDD-Gen), 2015) and Imaging Data Dictionary (Researchers Data Dictionary Imaging Data, 2015) pdf files of the same university are used to determine MRI terms.

The train dataset size should be 70%, the validation dataset size 15% as well as the test size 15%. The target metric will be F1-Score.

B.2 Types of Libraries

The code is written in Python and it is aimed to use NumPy, and Pandas libraries in this data analysis part, Matplotlib in the data visualization part, sci-kit-learn, Keras, and TensorFlow in the modeling part.

B. 3 Data Preprocessing

The data equals (1354, 186). The first 5 rows of the dataset look like this.

	SEX	EDUC	MARISTAT	INDEPEND	RESIDENC	NACCFAM	ANYMEDS	SMOKYRS	NACCTBI	DIABETES	ALCOHOL	HXHYPER	HYPERCHO	HXSTROKE	FOCLSIGN	HACHIN	CDRGLOB	DEL	HALL	AGIT	DEPD	ANX	ELAT
0	2	18	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
1	1	11	1	1	1	1	1	0	0	0	0	1	1	0	0	1	2	0	0	1	0	0	0
2	2	16	1	1	1	1	1	10	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
3	1	14	1	1	1	0	1	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0
4	1	16	1	1	2	0	1	50	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0

Figure 1 - The first 5 rows of the dataset

It is observed that there is no missing data in the dataset in the first place. By determining different quantile intervals, it is observed how the data set takes values in these intervals. The main reason for doing this is to determine outliers or if there is missing data between the intervals

B.4 Variable Type and Data Structure Consistency

In the model phase, categorical, numerical, cardinal, and nominal columns are determined to approach all kinds of data more accurately. A total of 27 categorical features, 159 numerical features, and 27 nominal features are determined.

B.5 Building the Target Variable (Regression or Classification)

To understand what kind of model should be built, visualization is done to better understand the CDRGLOB feature, which is determined as the target feature. As a result, it is determined that modeling with classification is required and there is an unbalanced data set.

C. Data Analysis

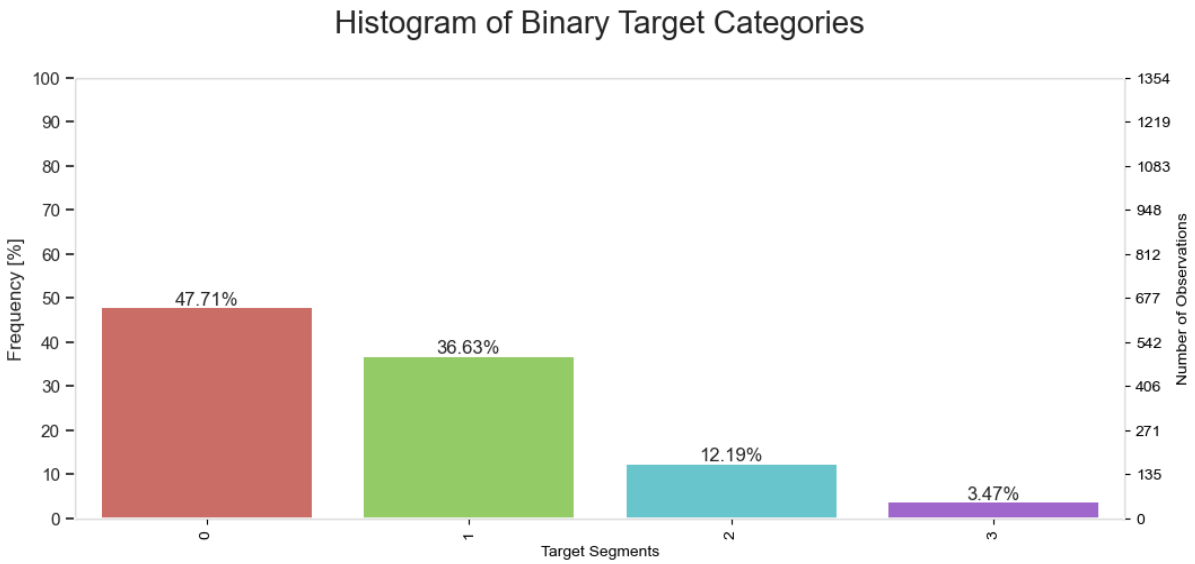

Figure 2 - Histogram of Target Variable (CDRGLOB)

C.1 Preparatory Data Analysis (PDA)

Before analyzing the data, drop operation is not performed because there is no feature with more than 80% null value, but the data gives detailed results as a result of correlation.

To detect outliers for each feature, the 1st and 3rd quartiles are set to 0.25 and 0.75, and then the values above or below these values are adjusted to change with the threshold values.

It is checked whether there is a NaN value for each feature in the data set and the result is obtained.

Among the categorical features, it is determined that there are 6 features that need to be one-hot encoding, and this process is done by considering the bias situation.

C.2 Exploratory Data Analysis

C.2.1 Categorical Variables

In the data visualization part, firstly, visualization is made for each categorical feature using a bar chart. As a result of this analysis, the following comments are obtained.

- Looking at the gender distribution of the patients in the data, it is observed that the male population was more.
- More than 87% of patients in the data have the following characteristics:
    - Living alone
    - Using drugs,
    - Alcohol consumption,
    - No history of traumatic brain injury, no diabetes, no focal neurological symptoms
- Looking at the relationship between the categorical features and the target feature (CDRGLOB), the following features are those that cause serious imairment in Alzheimer's disease:
    - Completely dependent
    - Staying in a hospital or nursing home,
    - Eating disorder, increased nighttime behavior, motor disorder, irritability, instability, apathy, anxiety, depression, hallucinations in the last month,
    - Being a drug user

C.2.2 Numerical Variables

Visualization is done for each numerical feature using a histogram. As a result of this analysis, the following comments are obtained.

Considering the relationship between the numerical features and the target feature (CDRGLOB), the following features are the ones that cause severe impairment in Alzheimer's disease.

- Total years of smoking
- Patient's age at the first visit,

- Total brain white matter hyperintense volume[1]
- Total cerebrospinal fluid volume[2]
- Segmented left / right lateral ventricle volume (cc)[3]
- Segmented total third ventricle volume (cc)

C.2.3 Multivariate Plots

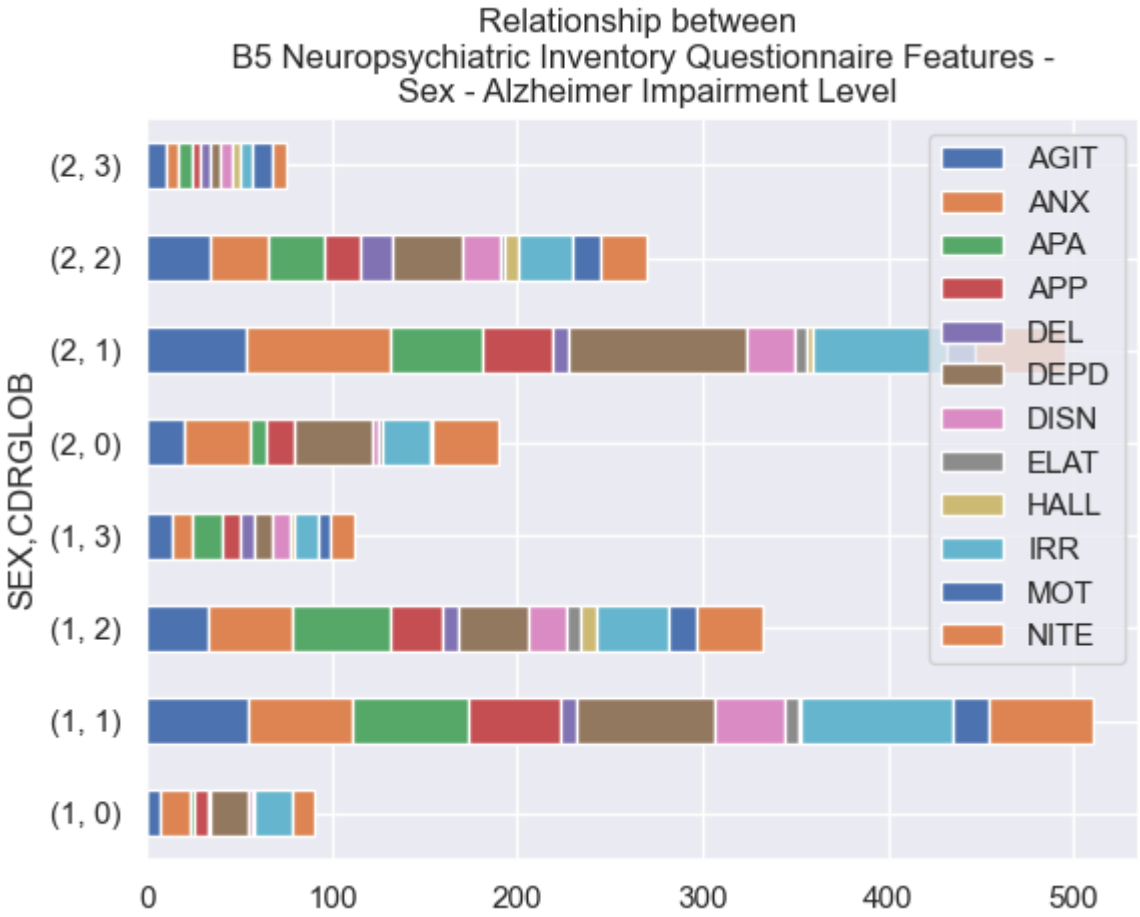

Figure 3 - Relationship between B5 Neuropsychiatric Inventory Questionnaire Features- Sex- Alzheimer Impairment Level

---

[1] White matter hyperintensities (WMHs) are lesions in the brain that show up as areas of increased brightness when visualized by T2-weighted magnetic resonance imaging (MRI). WMHs are also referred to as Leukoaraisosis and are often found in CTs or MRIs of older patients. WMHs are significantly associated with an increased risk of stroke, and dementia. (Dr. Sanil Rege, 2020)

[2] Cerebrospinal fluid (CSF) is a clear, colorless, watery fluid that flows in and around your brain and spinal cord. Your brain and spinal cord make up your central nervous system. It controls and coordinates everything you do, including your ability to move, breath, see think, and more. (Cerebrospinal Fluid (CSF) Analysis)

[3] The right and left lateral ventricles are structures within the brain that contain cerebrospinal fluid, a clear, watery fluid that provides cushioning for the brain while also helping to circulate nutrients and remove waste. (Lateral Ventricles, 2015)

These are the results from this chart.

- Aggression, depression, anxiety disorder and irritability in the last month affect the result as questionable impairment in both men and women.
    - (2,1) → (Man, Mild imparment)
    - (1,1) → (Woman, Mild impairment
- In the range of high symptoms of Alzheimer's problem (1), it indicates a high importance of apathy, irritability, and anxiety disorder.
- When the range from moderate impairment (2) to severe impairment (3) is examined, a decrease is observed in the number of people affected by aggression, depression and anxiety disorder in the last month.

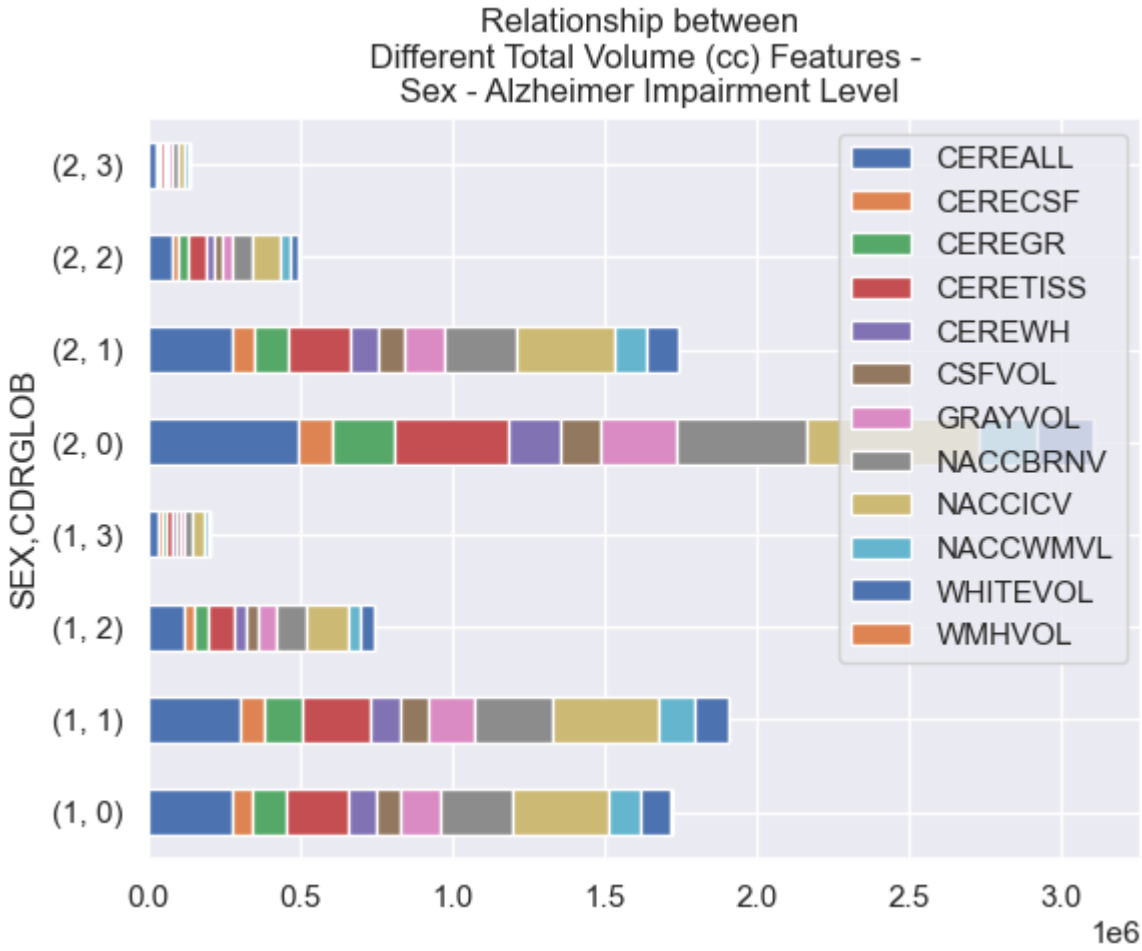

Figure 4 - Relationship between Different Total Volume (cc) Features- Sex- Alzheimer Impairment Level

These are the results from this chart.

- In male patients with severe impairment of Alzheimer's, total cerebrum cranial volume (CEREALL), total cerebrum brain volume (CERETISS), total cerebrospinal fluid volume (CERECSF) and total intracranial volume

(NACCICV) are observed to be higher. In women, it is observed that the effect occurs at the same rate when there is no impairment.

- WHITEVOL (total brain white matter volume (cc), NACCICV (total intracranial volume (cc)), NACCBRNV (total brain volume (cc)), CERETISS (total cerebrum brain volume (cc))) While there is a regular increase in male patients, it is observed that these characteristics decrease as the level of impairment increases in female patients.

C.3 Confirmatory Data Analysis

C.3.1 Chi-Square test for Nominal Features

Chi-Square test is performed on all nominal properties to see if there is a high correlation. In conclusion, since all other features except ANYMEDS have p values <0.05, it can be concluded that these features are related to the CDRGLOB target and should be included in your model for training.

C.3.2 ANOVA Test for Numerical Features

The significance value is accepted as 0.05. If the p-value is less than 0.05, we assume and assert that there are significant differences in the mean of the groups formed by each level of the categorical data. That is, we reject the NULL hypothesis.

In this case, it can be concluded that the numerical features of SMOKYRS, NACCICV, LPARSOPM should not be included in the model since their p value is greater than 0.05.

D. Modeling

D.1 Data Splitting – SMOTE Oversampling

The data set was divided into 70% train dataset, 15% validation dataset, and 15% test dataset. As a result, SMOTE Oversampling is performed because the train data set had an unbalanced data set in the CDRGLOB feature. In this way, the values of 4 different classes in y_train increased to 449[4].

[4] The reason for being 449 is that before the oversampling procedure, the total number of patients with a class 0 result is 449.

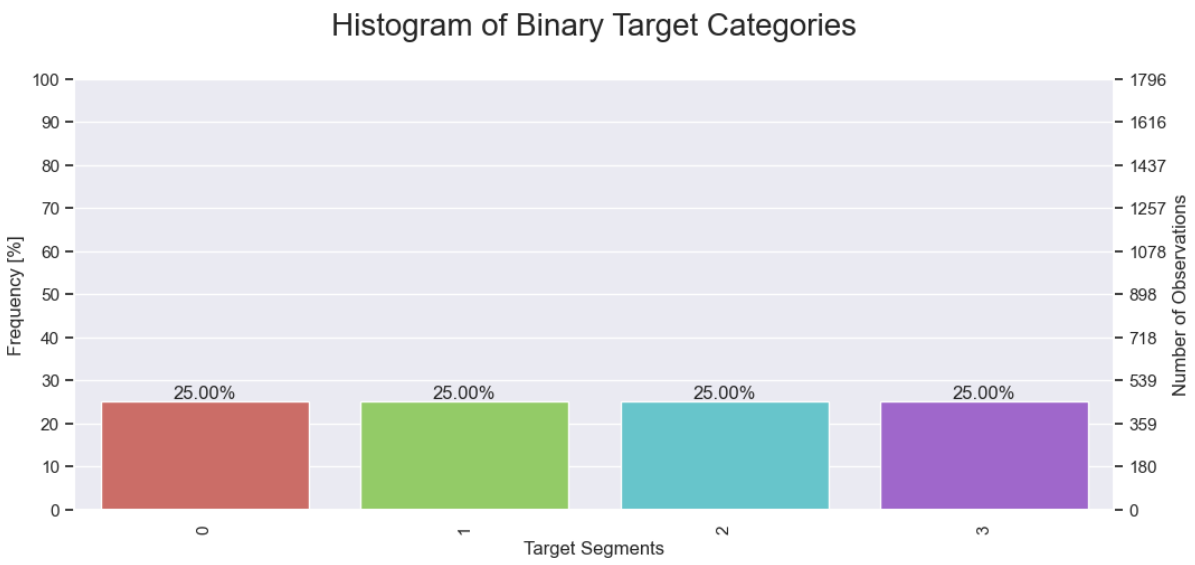

Figure 5 - Histogram of Train Dataset Target Categories

D.2 Data Transformation

MinMaxScaler process is applied to train dataset, validation dataset, and test dataset to set up the model properly. (MinMaxScaling process is important for neural networks modeling.)

D.3 Baseline Model

In a basic model, the accuracy value of the data set is calculated by establishing Logistic Regression and Random Forest Classifier models, without the operations done so far. The result is a value between 61% and 65%. It is aimed to exceed these values with operations such as SMOTE Oversampling, Hyperparameter Tuning Optimization, and Neural Networks modeling.

D.4 Classifier Modeling – Tunned Modeling

By establishing 6 different models and in addition to them, the hyperparameter tuning optimization process is applied.

- Logistic Regression Classifier
- Ensembled Classifiers
  - Random Forest Classifier
  - XGBoost
  - LightGBM
  - CatBoost
- Neural Network Classifiers
  - Multi-Layer Perceptron
  - Neural Networks

E. Evaluation

The result of the loss graph obtained from the neural network model is as follows.

Figure 6 - Loss Graphs of Neural Networks Model

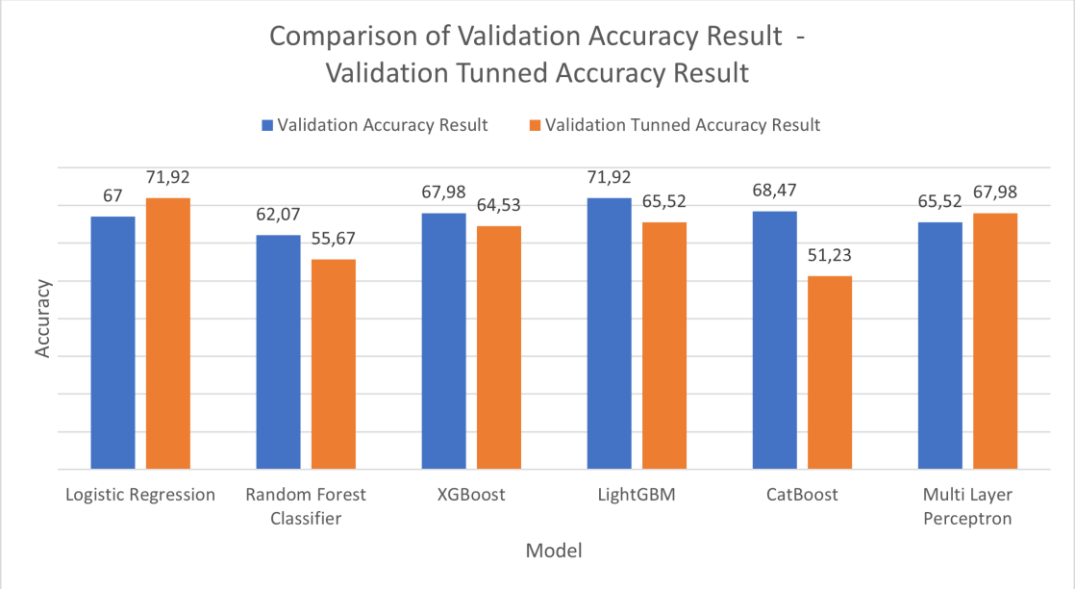

Figure 7- Compression of Validation Accuracy Result- Validation Tunned Accuracy Result

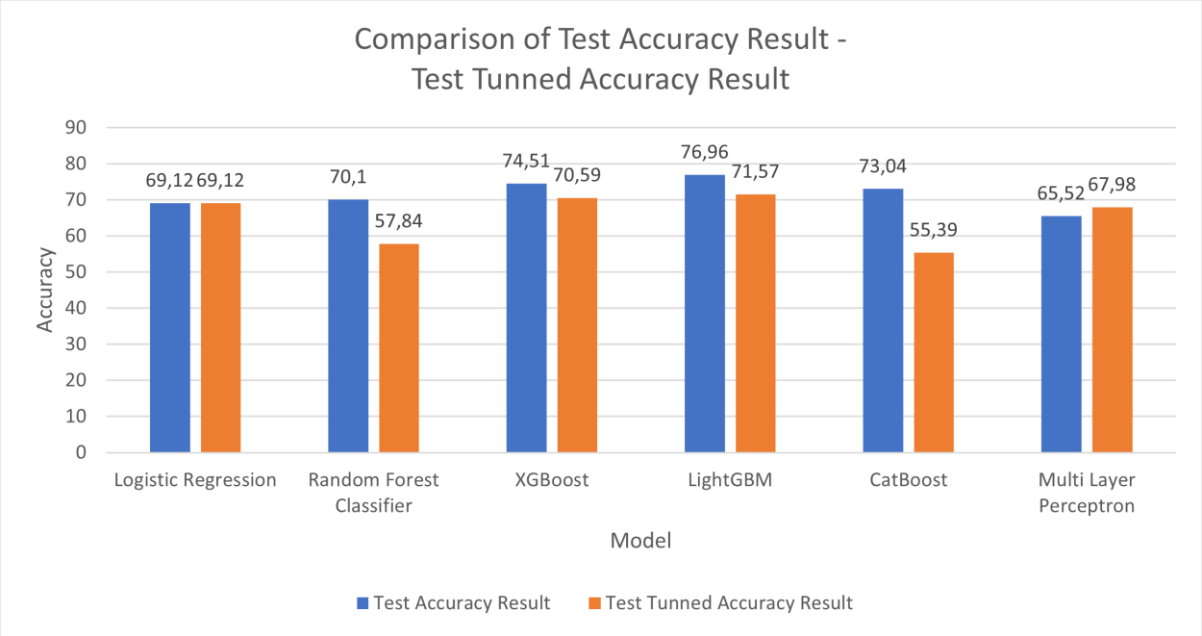

Figure 8 - Compression of Test Accuracy Result- Test Tunned Accuracy Result

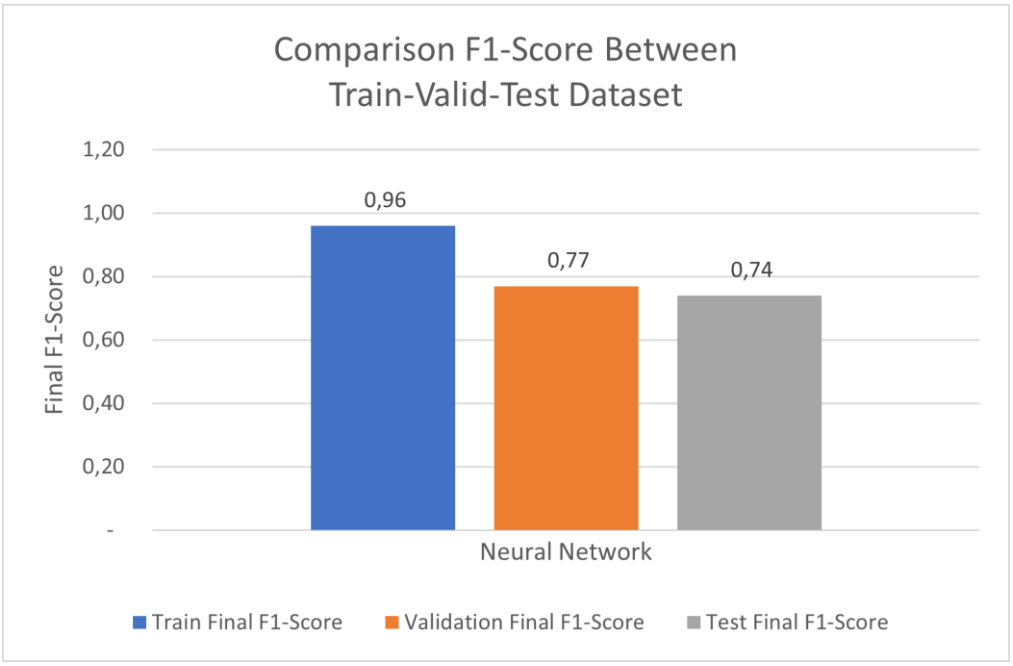

Figure 9 – Compression F1-Score Between Train-Valid-Test Dataset

References

Cerebrospinal Fluid (CSF) Analysis. (n.d.). Retrieved from MedlinePlus - Trusted Health Information for You: https://medlineplus.gov/lab-tests/cerebrospinal-fluid-csf-analysis/#:~:text=Cerebrospinal%20fluid%20(CSF)%20is%20a,%2C%20see%20think%2C%20and%20more.

Dr. Sanil Rege, D. J. (2020, 10 2). *White Matter Hyperintensities on MRI - Coincidental Finding or Something Sinister?* Retrieved from PsychScene Hub: https://psychscenehub.com/psychinsights/white-matter-hyperintensities-mri/

Lateral Ventricles. (2015, 04 15). Retrieved from HealthLine: https://www.healthline.com/human-body-maps/lateral-ventricles#1

NACC Uniform Data Set - Researcher Data Dictionary. (2015, March). Retrieved from National Alzheimer's Coordinating Center: https://github.com/edaaydinea/OP2-Prediction-of-the-Different-Progressive-Levels-of-Alzheimer-s-Disease-with-MRI-data/blob/main/additional%20resources/uds3-rdd.pdf

Researchers Data Dictionary - Genetic Data (RDD-Gen). (2015). Retrieved from National Alzheimer's Coordinating Center: https://github.com/edaaydinea/OP2-Prediction-of-the-Different-Progressive-Levels-of-Alzheimer-s-Disease-with-MRI-data/blob/main/additional%20resources/rdd-genetic-data.pdf

Researchers Data Dictionary Imaging Data. (2015). Retrieved from National Alzheimer's Coordinating Center: https://github.com/edaaydinea/OP2-Prediction-of-the-Different-Progressive-Levels-of-Alzheimer-s-Disease-with-MRI-data/blob/main/additional%20resources/rdd-imaging.pdf