Parkinson's Disease Progression Prediction

PROGRESS REPORT OF THE TERM PROJECT

Submitted by:

Kushagra Gupta-01114811621

BACHELOR OF TECHNOLOGY
IN

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Under the Guidance

of

Dr. Tripti Lamba

(Associate Prof., AIML)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Maharaja Agrasen Institute of Technology,
PSP area, Sector – 22, Rohini, New D elhi – 110085
(Affiliated to Guru Gobind Singh Indraprastha, New Delhi)

(JUNE 2023)

Introduction

Parkinson's disease (PD) is a disabling brain disorder that affects movements, cognition, sleep, and other normal functions. Unfortunately, there is no current cure—and the disease worsens over time. It's estimated that by 2037, 1.6 million people in the U.S. will have Parkinson's disease, at an economic cost approaching $80 billion. Research indicates that protein or peptide abnormalities play a key role in the onset and worsening of this disease. Gaining a better understanding of this—with the help of data science—could provide important clues for the development of new pharmacotherapies to slow the progression or cure Parkinson's disease.

Current efforts have resulted in complex clinical and neurobiological data on over 10,000 subjects for broad sharing with the research community. A number of important findings have been published

Problem Statement

we are looking at the Unified Parkinson's Disease Rating Scale (UPDRS) that was revised by the Movement Disorder Society (MDS) in 2008. This new scale - the MDS-UPDRS (which we will refer to within this EDA as simply the UPDRS) - consists of 4 separate parts. Each part consists of a questionnaire that rates signs or symptoms of Parkinson's Disease (PD). According to Holden et al (2018), the individual parts consist of:

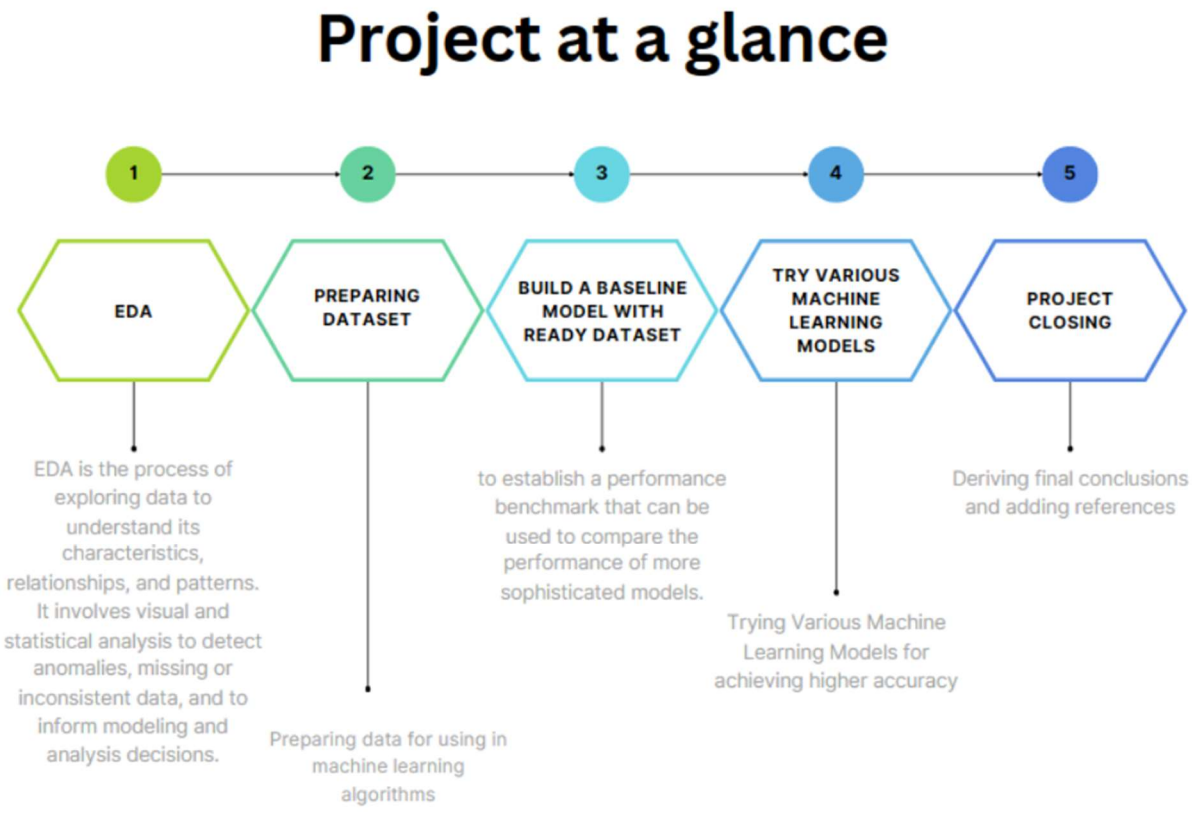
Part I - Non-Motor Aspects of Experiences of Daily Living

Part II - Motor Aspects of Experiences of Daily Living

Part III - Motor Examination

Part IV - Motor Complications

Questions within each part are scored on a 5 point scale ranging in values from 0 (normal) to 4 (most severe impairment). The maximum score that a patient may be assigned is 272 points. The challenge in this competition is to predict the UPDRS scores for parts 1 - 4 for each month that the patient had a visit and evaluation with a physician.



Objective

Our Objectives for the Project :

- The goal of the project is to predict the course of Parkinson's disease and to find to the best suitable Machine Learning Algorithm to predict the course of Parkinson's Disease.

Technology Used

1. Python: a popular programming language used for developing the system.
2. Web Scrapping Tools: Given Dataset from Kaggle
3. Machine learning algorithms: used for training of the Machine learning model

Methodology

Our Objectives for the Project :

- Phase 1- Exploratory data analysis for prediction and model training.
- Phase 2-Building simple baseline model with train dataset.
- Phase 3- Testing various Machine learning model and achieving high score. Our training dataset consists of protein levels, peptide level and clinical report

About the Data Set

Dataset	Size on Disk	Size in Memory
train_clinical_data	73 KB	159.73 KB
train_peptides	49 MB	44.94 MB
train_proteins	7.5 KB	8.88 MB
supplemental_clinical_data	75 KB	135.80 KB

As we can see, memory pressure isn't too bad, nor is disk space storage. This means that we won't likely have too many problems with various machine learning features.

Files

train_peptides.csv Mass spectrometry data at the peptide level. Peptides are the component subunits of proteins.

- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `UniProt` - The UniProt ID code for the associated protein. There are often several peptides per protein.
- `Peptide` - The sequence of amino acids included in the peptide. See this table for the relevant codes. Some rare annotations may not be included in the table. The test set may include peptides not found in the train set.
- `PeptideAbundance` - The frequency of the amino acid in the sample.

train_proteins.csv Protein expression frequencies aggregated from the peptide level data.

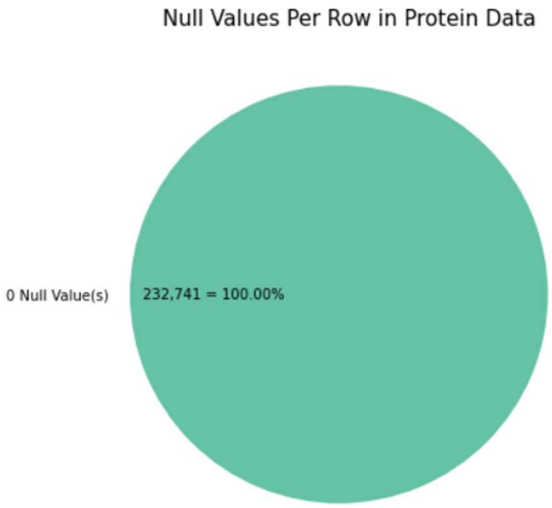
- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `UniProt` - The UniProt ID code for the associated protein. There are often several peptides per protein. The test set may include proteins not found in the train set.
- `NPX` - Normalized protein expression. The frequency of the protein's occurrence in the sample. May not have a 1:1 relationship with the component peptides as some proteins contain repeated copies of a given peptide.

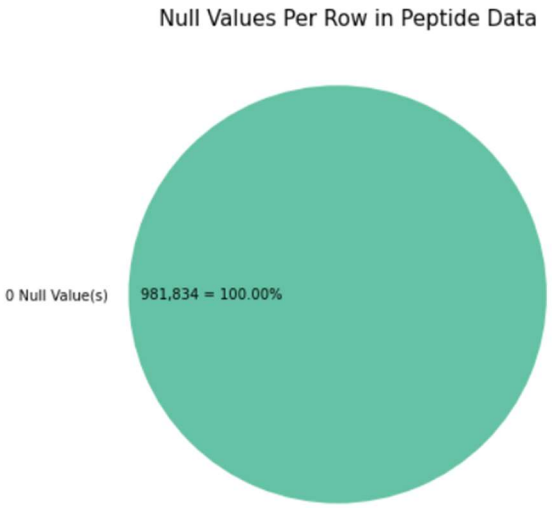
train_clinical_data.csv

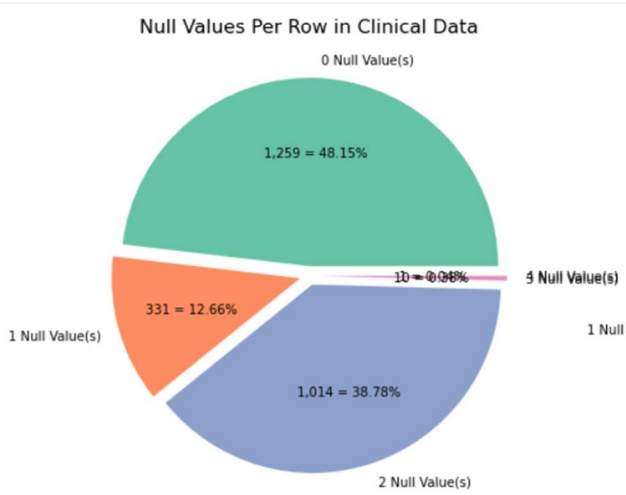
- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `updrs_[1-4]` - The patient's score for part N of the Unified Parkinson's Disease Rating Scale. Higher numbers indicate more severe symptoms. Each sub-section covers a distinct category of symptoms, such as mood and behavior for Part 1 and motor functions for Part 3.
- `upd23b_clinical_state_on_medication` - Whether or not the patient was taking medication such as Levodopa during the UPDRS assessment. Expected to mainly affect the scores for Part 3 (motor function). These medications wear off fairly quickly (on the order of one day) so it's common for patients to take the motor function exam twice in a single month, both with and without medication.

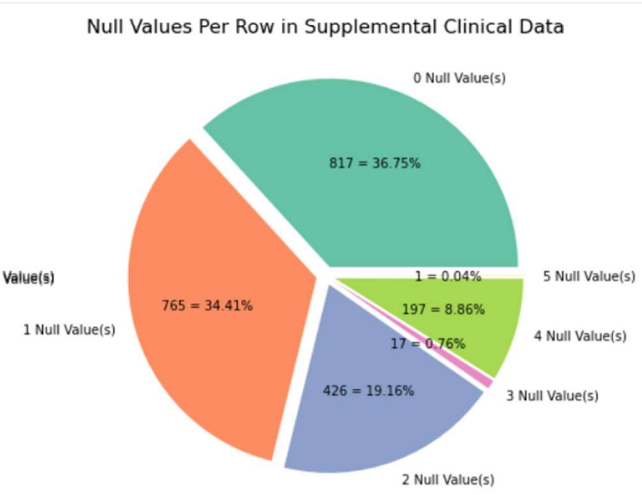
Observation and Analysis:

Exploring the issue of missing values in the dataset to see if there are systemic problems with data representation.









Null Values Per Row in Protein Data

Null Values Per Row in Peptide Data

Making Observations Based on Our Analysis

Rows with 1 Null Value

When a row contains only one null value, it is typically in the feature "upd23b_clinical_state_on_medication," which should have a value of either "On" or "Off." A null value in this field can be interpreted in two ways: the patient may have been off medication at the time of assessment, or the assessment may have failed to capture the patient's medication status. There are also 7 null values in the "updrs_3" column and 21 null values in the "updrs_4" column. According to Goetz et al (2008), "updrs_3" assesses motor function and "updrs_4" assesses motor complications, both with a minimum score of 0. A null value in either column suggests that the assessment was not performed, which is important because a score of 0 indicates that the patient was assessed and had normal responses.

Rows with 2 Null Values

When a row has two null values, they usually correspond to updrs_4 and upd23b_clinical_state_on_medication. As valid responses for the medication status field are limited to On or Off, a null value in this column is of interest as it is unclear whether the medication status was not recorded or if the patient was not taking any medication. The majority of other null values are found in the updrs_4 field, which assesses motor complications, while updrs_3 and updrs_2 have few null values. However, a null value in updrs_3 cannot be assumed to be a score of 0 as this indicates normal function. Similarly, null values in updrs_2 may indicate that the assessment was not performed.

Rows with 3 Null Values

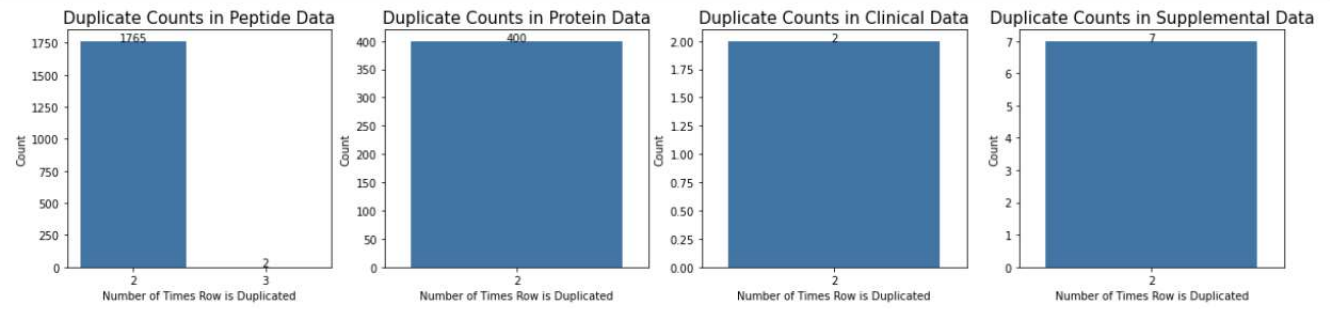
In 10 cases, rows have 3 null values, which correspond to missing information in updrs_3, updrs_4, and upd23b_clinical_state_on_medication. It should be noted that the missing values cannot be considered as 0 values.

Rows with 4 Null Values

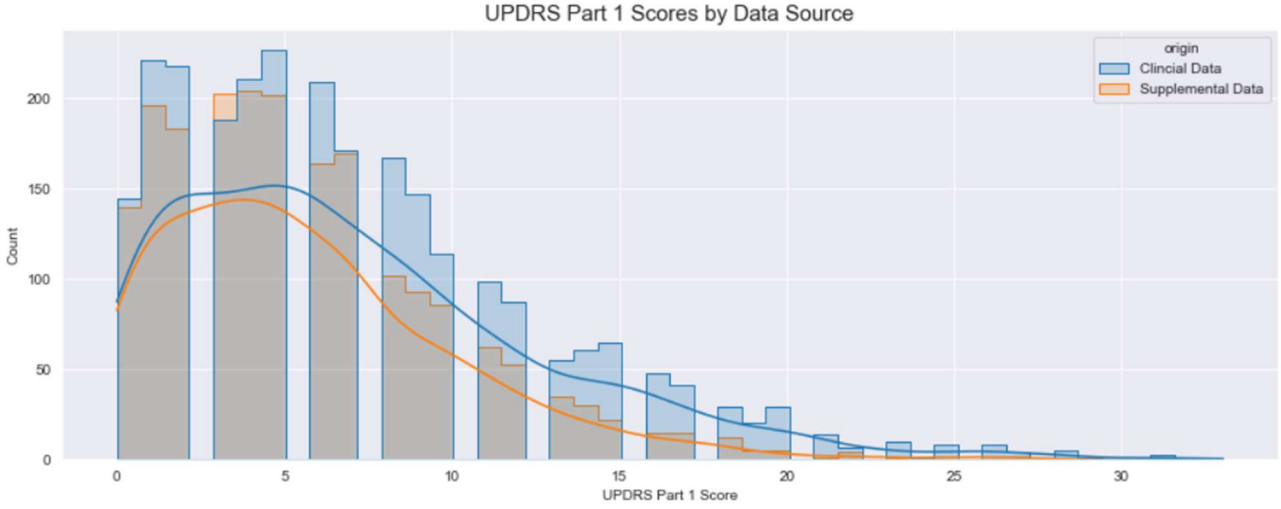
There is only one row in the dataset where four features contain null values. This suggests that only the UPDRS part 3 assessment was conducted during the visit, as the other features with null values correspond to motor complications and medication status. It is important to note that null values cannot be assumed to have a value of 0, as a score of 0 indicates normal function.

It is important to examine the supplementary information for any null values.

Handling Duplicate data

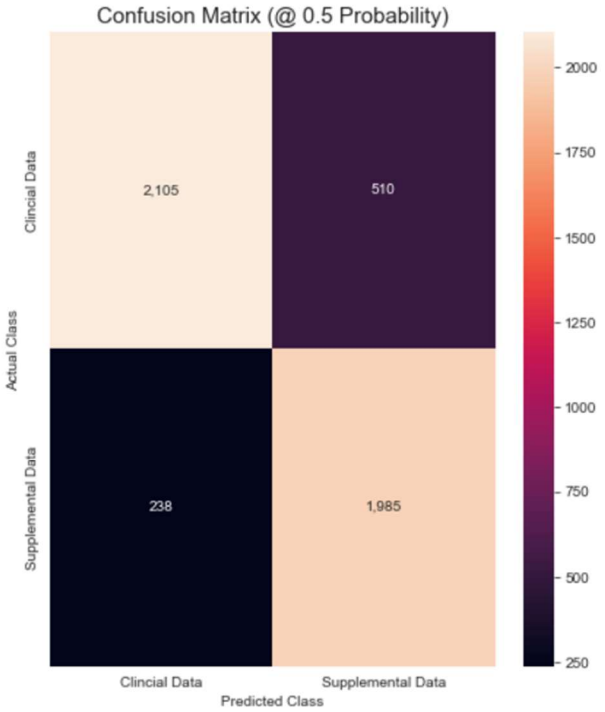


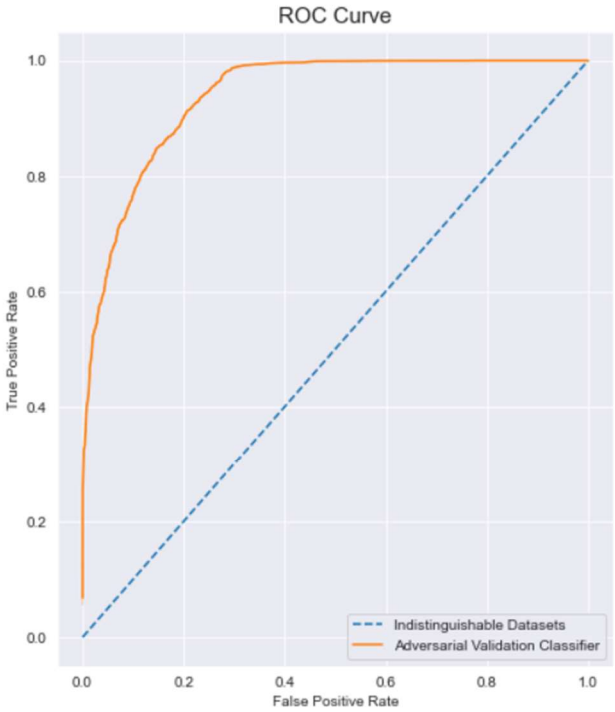
Analyzing UPRDS Score



Advertial Validation

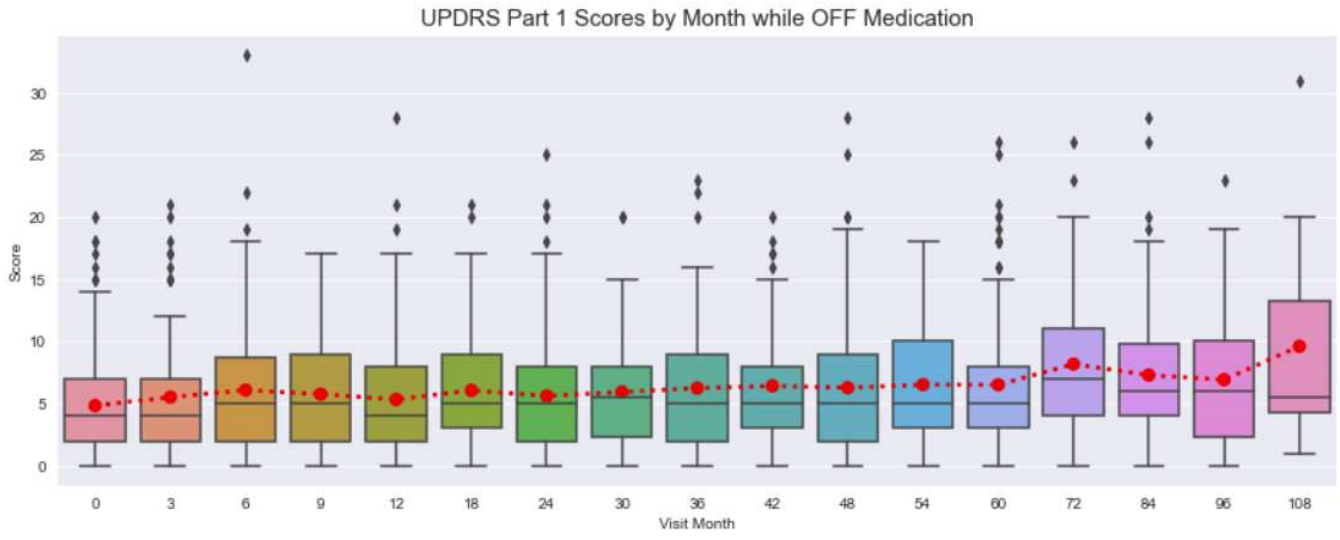
```
--> Overall results for out of fold predictions
: AUC ROC = 0.9389806206451068
```





We have trained a LightGBM classifier using clinical data and supplemental clinical data, and performs adversarial validation to check whether the two datasets are similar or not. It concatenates the two datasets, adds a label to identify the origin of each observation, and then performs 5-fold cross-validation. During each fold, it trains the model on the training set and evaluates it on the validation set. It uses early stopping and log evaluation callbacks to stop the training process when there is no improvement in the validation AUC ROC score for 50 consecutive iterations. It then computes the out-of-fold predictions and probabilities for each observation and stores them in arrays. Finally, it calculates the overall AUC ROC score, generates a confusion matrix, and plots the ROC curve.

Performing Statistical Analysis



Conclusion

Once the model is trained and validated, it can be used to predict the progression of the Parkinson's Disease with the help of model developed. Overall, Machine learning models have shown great potential in detecting Parkinson's Disease accurately and efficiently, which can aid in early diagnosis and treatment of the disease, leading to improved patient outcomes.