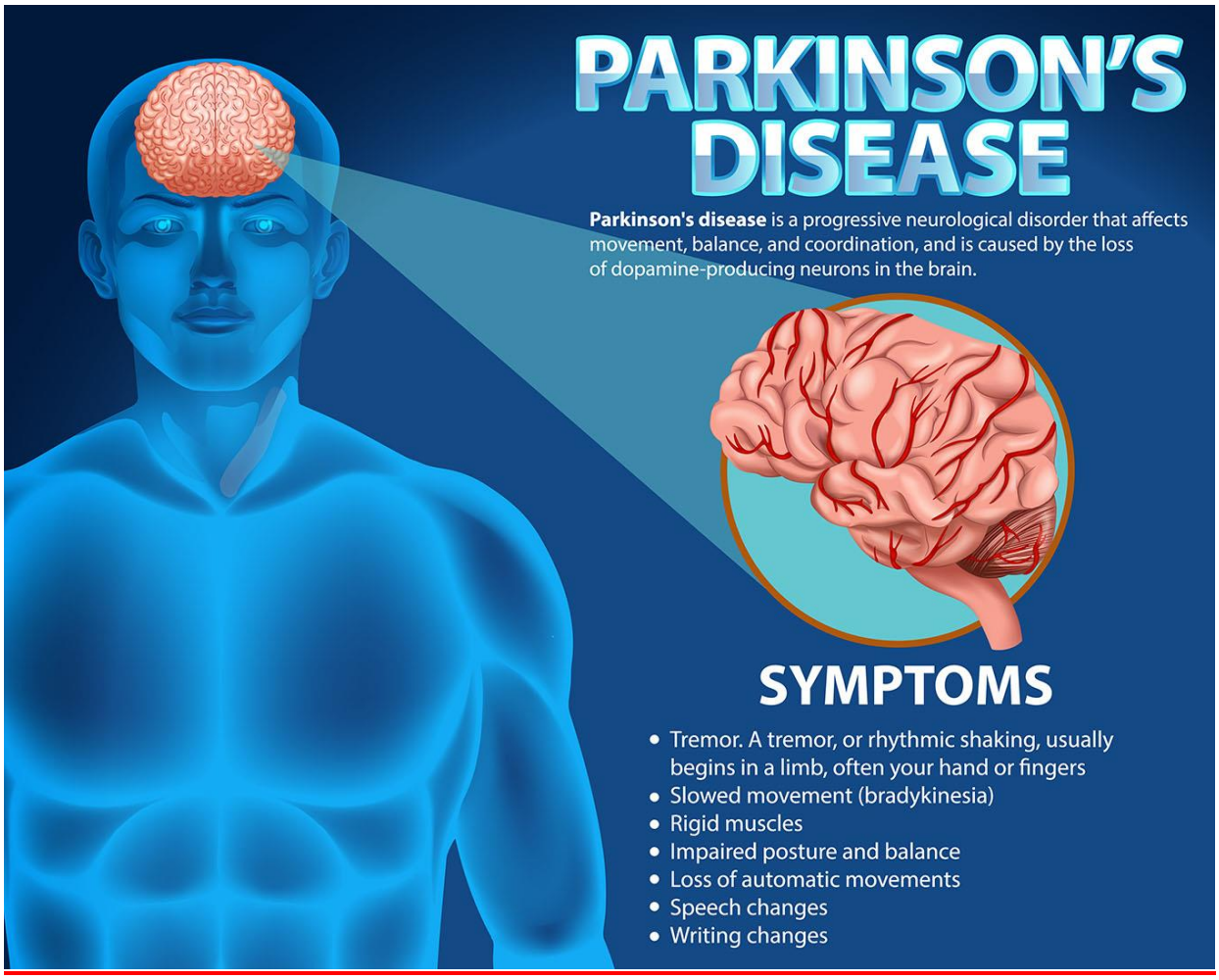
Parkinson's Disease



CS_10

NAME	ID
عمرمحمد سعيد محمد عبده	2022170282
عبدالله ابراهيم احمد ابراهيم	2022170220
على بسام المصرى	2022170259
عبدالرحمن حمدى عبدالحليم عمران	2022170217
مصطفى محمد مجدى مصطفى	2022170425
اسلام عمرو عبدالعزيز السيد	2022170060

Preprocessing:

- **Check nulls:** check if there any null values, only founded in 'EducationLevel' with percentage about 20% of the column. We decided to remove the column as all because if we fill this column with any replacement value like (mode) we saw that it may be lead to wrong information(even though we tried to take it in the model but there is no corr. between it and the target (UPDRS)) ,, also we checked the nulls in the remaining columns to be in shape like unknown or negative values but no nulls like this exist.

- **There exist 2 columns in shape of dictionary we split this 2 columns to many symptoms and medicalHistory features like (Tremor , FamilyHistoryParkinsons,...)**

- **Encoding:**
  - **One-hot encoding:** Apply one-hot encoding on the 'Ethnicity' column because if we make it using label encoding it will give an order meaning which is wrong

  - **Label encoding:** Apply Label encoding on the other categorical data and here order is no matter because all remaining features classified as (yes or no) so there is no order in binary classification to limit from the size of the columns that will be exist if we used one-hot encoder

- **Scaling & Normalization:** Scaling the data to put all the features on the same range, noticed that the continuous features were normally distributed but to be accurate we use log() transformations then applying standard scaling on it (giving more better result on modeling)

- **Outliers detection :** We measure the outliers on all numerical columns by measuring Q1,Q3 and IQR and we found that there is no outliers in our data

- **Convert formatting:** WeeklyPhysicalActivity column is presented in HH:MM format so we changed it to hours (int64)

- **Dropping columns :**
  ('PatientID','EducationLevel','DoctorInCharge')

  - For PatientID : there are unique identifier for database not helpful in modeling

  - For EducationLevel : as we mentioned before there exist many of nulls

  - For DoctorInCharge : all rows are the which is duplicated not helpful in modeling

Feature Engineering :

We extract new features from the existing features like

- Age with Symptoms

- BMI with Symptoms

- Combine Blood Pressure and Cholesterol , Creating Binary Count of Symptoms Present (Yes/No)

Mapping the existing features from numerical to categorical like

- BMI Categories :

Below 18.5 → Underweight

18.5 – 24.9 → Normal weight

25 – 29.9 → Overweight

30 and above → Obese

- FunctionalAssessment :

Below 5 → not good enough

Above 5 → very good

- UPDRS (target variable):

Below 49 → Not affected , 49 - 100 → Moderate , 100 and above → Severe

-CholesterolTotal:

Healthy Range: Total cholesterol should be below 200 mg/dL.

Levels between 200–239 mg/dL are borderline high, and above 240 mg/dL is high.

- Diet & Sleep Quality :

Below 5 → not good enough

Above 5 → very good

Feature Selection :

Techniques :

(RandomForestRegressor , SelectKBest , LinearRegression (abs coff. value))

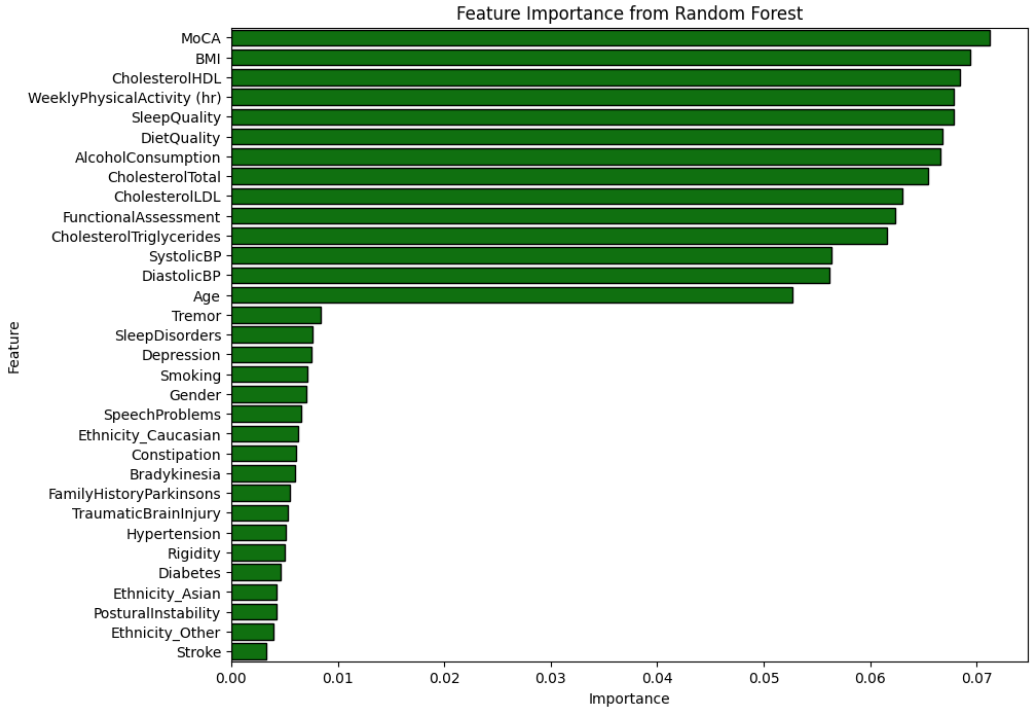
Select k best :

Top Selected Features:

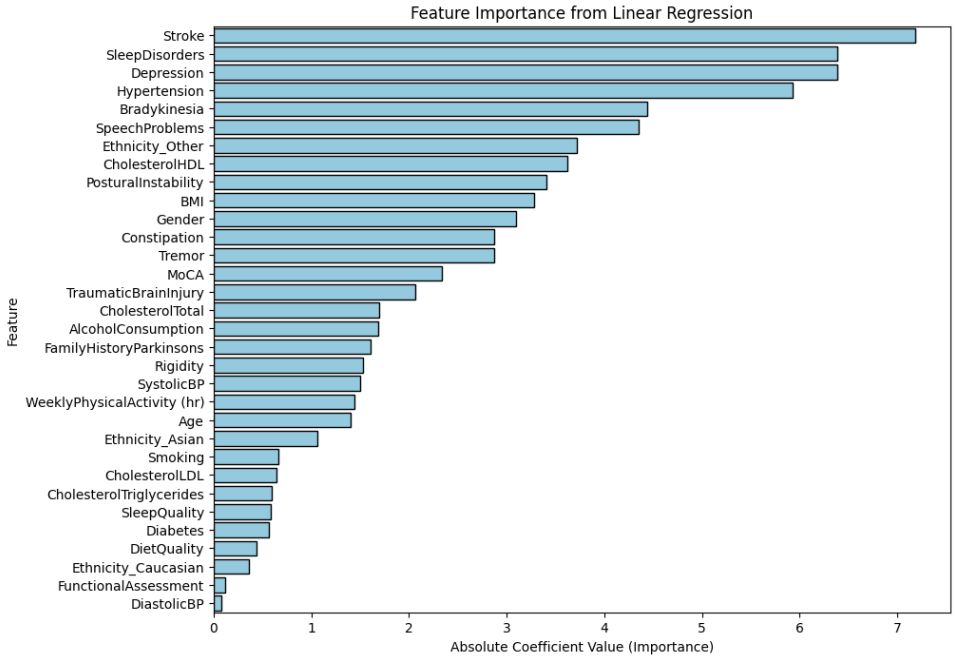
['Gender', 'BMI', 'DietQuality', 'SleepQuality', 'CholesterolTotal',

'CholesterolHDL', 'WeeklyPhysicalActivity (hr)']

RandomForest Regressor :



LinearRegression :



Analysis:

- Using *describe() & Kdeplot()* **functions noticed that most of the data are normally distributed and there is no skewness in data**

- Correlation:

Based on the correlation we see between data , we found that the correlation between features and target (UPDRS) is very small nearly to 0 even though the new features we extracted from the existing features also very small corr nearly to 0.02 to the target which is approximately equal to 0

but from our research on the Parkinson data we found that we have features in our data which is very important in the real world in predicting the percentage of existing this disease like

- presence of Tremor , FamilyHistoryParkinsons , Hypertension

Depression , Rigidity , SpeechProblems , SleepDisorders

This is the logically and medical POV but in this data and our models tells us that there is no any relationship between between these symptoms and our target (UPDRS)

Regression techniques used in the model:

- Linear regression
- Polynomial Regression
- Random forest Regression

	Linear Regression	Polynomial Regression	Random forest Regression
r2_score_train:	0.01	0.457	0.03
r2_score_test:	0.0019	-1.35	0.005
MSE	3395.1	8006.8	3381.8

Features used:

['WeeklyPhysicalActivity (hr)' , 'CholesterolHDL' , 'BMI' , 'AlcoholConsumption' , 'CholesterolTotal' , 'SleepQuality' , FunctionalAssessment' , ' CholesterolLDL' , DietQuality , CholesterolTriglycerides , SystolicBP , MoCA , DiastolicBP , Age]

Training data size: 70%

Test data size: 30%

Trying to change in splitting of data

Training data size: 80%

Test data size: 20%

But ALL results indicating **OVERFITTING** (training data R2 score = 0.8 while testing

R2 score = - 0.0019)

Further techniques used:

- **Hyperparametr Tunning:** By using GridSearchCV in random forest regressor to fit on data to get more better readings

- **Ridge and Lasso Regression :** To help in overfitting problem

- **Oversampling :** To increase number of rows but we canceled it because our problem in the low correlation between features and target , but oversampling used in balance data to avoid skewness

Conclusion :

This dataset performs very poorly in modeling. However, in real life, to diagnose whether a person has this disease or not, it is essential to conduct medical tests and gather some clinical information — and these types of information are actually present in the features of our dataset. Despite that, all modeling techniques we applied are still performing poorly.