**Step 1:Load Dataset**

**Step 2:Feature engineering :**

Convert the dataset to the appropriate format.

Verify missing data or any anomaly in the dataset.

Drop some columns that are not required.

In this step, we actually check for missing datasets and make a pandas data frame.

Later on divided the whole dataset into 2 dataframes, 1 for PD\_Yes and 2 for PD\_No.

**Step 3 :**

Find basic stats of the dataset like mean, standard deviation, maximum, minimum, etc. from each column.

**Step 4: Correlation**

Correlation between all columns from 2 data frames.

Correlation refers to a statistical measure that quantifies the relationship between two variables. It indicates how closely two variables are related or how they change together. Correlation is often used to determine the strength and direction of the relationship between variables in a dataset.

The correlation coefficient, commonly denoted as "r," is a value that ranges between -1 and +1. Here's what the correlation coefficient represents:

* A value of +1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases in a linear manner.
* A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases in a linear manner.
* A value of 0 indicates no correlation or a very weak relationship between the variables.

Correlation is widely used in data analysis and machine learning to understand the relationships between variables, identify important features, and make informed decisions based on the data.

We have used advance correlation techniques below are the correlation techniques we used.

1) **The Pearson correlation coefficient**, often simply referred to as Pearson's correlation or correlation coefficient, measures the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables.

The Pearson correlation coefficient is denoted by the symbol "r" and ranges between -1 and +1. Here's what different values of the correlation coefficient represent:

The Pearson correlation coefficient is calculated based on the covariance (a measure of how two variables vary together) and the standard deviations of the variables. It is sensitive to outliers and assumes that the relationship between the variables is linear.

The formula for calculating Pearson's correlation coefficient is:

r = (Σ (X - X̄)(Y - Ȳ)) / √((Σ (X - X̄)²) \* (Σ (Y - Ȳ)²))

where X and Y are the variables, X̄ and Ȳ are their means, and Σ denotes the sum.

The Pearson correlation coefficient is widely used in statistics, data analysis, and machine learning to measure the strength and direction of relationships between variables.

2**) Kendall rank correlation coefficient**

The Kendall rank correlation coefficient, often referred to as Kendall's tau, is a statistical measure that quantifies the strength and direction of the association between two variables. Unlike the Pearson correlation coefficient, Kendall's tau does not assume that the relationship between the variables is linear or normally distributed.

Kendall's tau is typically used when dealing with ranked or ordinal data, where the exact values of the variables may not be important, but their relative order is. It measures the similarity in the ranks of the paired observations.

The calculation of Kendall's tau involves counting the number of concordant and discordant pairs of observations and then normalizing it according to the total number of pairs. It is less sensitive to outliers compared to the Pearson correlation coefficient.

Kendall's tau is commonly used in various fields, including social sciences, economics, and environmental studies, where the data may not be continuous and exhibit a clear linear relationship.

3**) Spearmanr correlation coefficient**

The Spearman's rank correlation coefficient, often referred to as Spearman's rho or simply Spearman's correlation, is a statistical measure that quantifies the strength and direction of the monotonic relationship between two variables. Like Kendall's tau, Spearman's correlation is commonly used when dealing with ranked or ordinal data.

Spearman's correlation coefficient can also be used for continuous data, but it does not assume a linear relationship between the variables. Instead, it assesses how well the relationship between the variables can be described using a monotonic function.

To calculate Spearman's correlation coefficient, the ranks of the paired observations are used instead of their actual values. Ranks are assigned to each observation, and then the correlation is calculated using the same formula as for Pearson's correlation, but with the ranked values.

Spearman's correlation coefficient is commonly used when the relationship between variables is not clearly linear and may follow a different pattern.

4**) The point biserial correlation coefficient** is a statistical measure that quantifies the strength and direction of the relationship between a binary (dichotomous) variable and a continuous variable. It is a special case of the Pearson correlation coefficient that is used when one variable is binary and the other is continuous.

The point biserial correlation coefficient can take values between -1 and +1, where:

* A value of +1 indicates a perfect positive relationship. It means that as the binary variable increases, the continuous variable also increases.
* A value of -1 indicates a perfect negative relationship. It means that as the binary variable increases, the continuous variable decreases.
* A value of 0 indicates no relationship or a very weak relationship between the variables.

To calculate the point biserial correlation coefficient, the binary variable is typically coded as 0 and 1, where 0 represents one category and 1 represents the other category. The correlation is then calculated using the same formula as for the Pearson correlation coefficient.

So our goal was to find the specific columns/feature name that has the least/no relationship of data frame one with another.

**Step 5: identified the salient variables**

After applying the four correlation functions we took the average of them and based on one single average value we identified the salient variables (features)/columns.

Note :

This column 'UPDRS The Unified Parkinson’s Disease Rating Scale (UPDRS) score that isassigned to the subject by a physician via a medical examination to determine the severity and progression of Parkinson’s disease’ in both data frames this column has ties in the data so we drop this column from both of dataset