Diagnosing Parkinson Disease using Voice Sample Data Analysis Report.

Assignment 3 group project report

FOUNDATION OF DATA SCIENCE

Course lecturer: John Denison

Close-up of a person's hands

Description automatically generated

Group 82:

Kritica Shakya

Neha Pandey

Shatabdi Shah

Sandeep Baniya

Contents

[Table of figure 2](#_Toc147413337)

[Introduction 3](#_Toc147413338)

[Background 3](#_Toc147413339)

[Project objectives 3](#_Toc147413340)

[Datasets 3](#_Toc147413341)

[Libraries 3](#_Toc147413342)

[Data loading and Preprocessing 4](#_Toc147413343)

[Dataset loading and initial exploration. 4](#_Toc147413344)

[Handling missing values 4](#_Toc147413345)

[Exploratory Data Analysis (EDA) 4](#_Toc147413346)

[Descriptive data analysis 4](#_Toc147413347)

[Data visualizations 5](#_Toc147413348)

[Feature selection and engineering 8](#_Toc147413349)

[Linear regression modelling 9](#_Toc147413350)

[Performance analysis 11](#_Toc147413351)

[Optimization 12](#_Toc147413352)

[Rescaling 12](#_Toc147413353)

[Log transformation to selected feature. 15](#_Toc147413354)

[Gaussian 17](#_Toc147413355)

[Non-linear transformation 19](#_Toc147413356)

[Evaluation 22](#_Toc147413357)

[Discussion and Limitations 22](#_Toc147413358)

[Discussion 22](#_Toc147413359)

[Contribution to knowledge 22](#_Toc147413360)

[Limitation 22](#_Toc147413361)

[Brief descriptions of individual contributions 23](#_Toc147413362)

[Conclusion 23](#_Toc147413363)

## Table of figure

[Figure 1 pairplot 6](#_Toc147413299)

[Figure 2 heatmap 7](#_Toc147413300)

[Figure 3 boxplot 8](#_Toc147413301)

[Figure 4 old jitter 16](#_Toc147413302)

[Figure 5 new jitter 16](#_Toc147413303)

# Introduction

## Background

The American National Institute of Neurological Disorders and Stroke (NINDS) states that the Parkinson's disease (PD) is a condition of movement in human neurological system. This chronic illness in a human suffers for various kinds of system like tremors, stiffness, bradykinesia (slow movement), and postural instability Highly influential individuals such as legendary boxer Muhammad Ali, former U.S. President George H.W. Bush, actor Michael J. Fox of "Back to the Future" fame, heavy metal legend Ozzy Osbourne, and the late Pope John Paul II have all battled with PD. Despite suffering with PD, they shined in their own fields.

## Project objectives

This group assessment project is focused on the investigation of acoustic measurement data derived from voice samples of Parkinson's disease patients as well as, if appropriate, healthy persons. PPD infected individuals were examined by the professionals and these professionals used the Unified Parkinson's condition Rating Scale (UPDRS) to find out the severity and the course of the condition for further discovery about this disease.

The main objective of this project is to use various technology and data analytic methods to further help in establishment of a precise approach to detect Parkinson's disease. Our main aim from this project is to predict the motor and the total UPDRS scores on the dataset of individuals with Parkinson disease.

# Datasets

There were given two datasets in this project:

po1\_data.txt which contained 40 study subjects which had 20 individuals with Parkinson disease (PPD) and 20 were healthy people. The dataset contains the voice samples which contain the data of the different measurement categories of individuals. We used this dataset to determine a set of salient features that could distinguish people with PD to healthy.

Po2\_dataset.csv which contains a set of voice measurements based on 42 people with Parkinson disease. In the dataset we have a list of variables which indicates the Parkinson disease symptoms. We have data on categories such as recurrence Period Density Entropy, Detrended Fluctuation Analysis and Pitch period Entropy.

# Libraries

we have used Libraries such as:

* Pandas: used for data manipulation and analysis in the data frame
* Matplotlib: used for plots and charts.
* Seaborn: used for data visualization.
* Numpy: used for working with numerical datas
* Math: used for mathematical functions
* Train\_test\_split: for splitting the dataset into training and testing set.
* Scikit-learn linear regression used for evaluating model performance.
* Scikit’s-learns metrics used for evaluation
* Statsmodel for detailed statistical analysis
* Standard scaler for standardizing features
* Power transformer of yeo-johnson used to make feature distribution gaussian like.

# Data loading and Preprocessing

## Dataset loading and initial exploration.

The po2\_dataset.csv was loaded into the data frame using the pandas. DF HEAD AND INFO

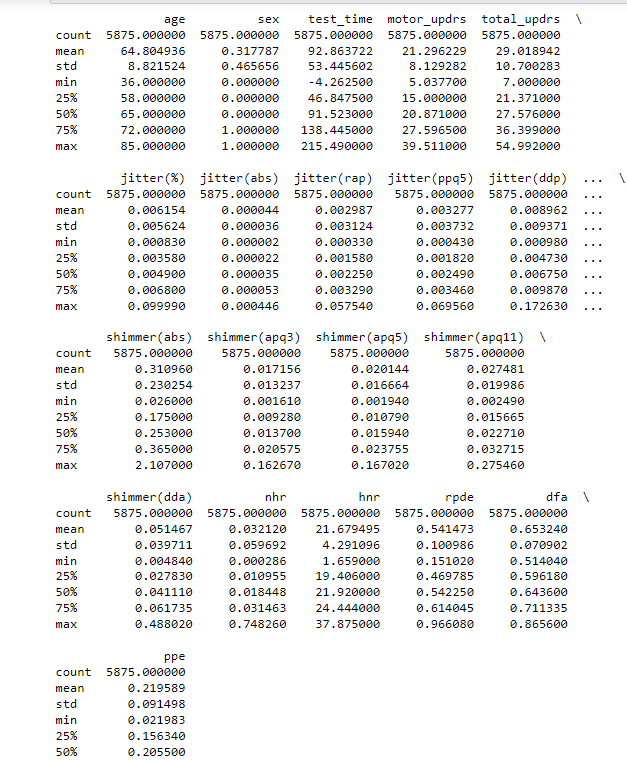
## Handling missing values

In the dataset, we have used df.isnull().sum() to identify the missing data. It calculates the total number of missing data for each column in the data frame and stores it in the missing data and displays the result of it. This is the widely used approach to identify missing values in a dataset. Identifying and addressing missing values is a very important step in the preprocessing to maintain accuracy for further data analysis.

# Exploratory Data Analysis (EDA)

## Descriptive data analysis

In this project we have done a descriptive analysis of the dataset categories of the sample which is a common practice in data analysis for further analysis. The code calculates the descriptive statistic using the describe () method and displays it. The descriptive statistics displays measures of central tendency eg. Mean, median and mode and measures of dispersion which helps us get insights on the dataset distribution, characteristics, issues, outliers. The below figure is a descriptive data analysis of the po2 dataset.



## Data visualizations

1. Pairplot:

To visualize the relationship between the data pair plot is generated. Using the seaborn the code contains some data visualization and exploitation. This helps to provide relationships between two variables, correlations and patterns and guide further decision and analysis too. To visualize relationship between the data pair plot is generated where the variables include ‘age’, ‘motor\_updrs’, 'total\_updrs', 'jitter (%)', 'shimmer (%)', and 'nhrand.

A collage of blue and black graphics

Description automatically generated A collage of blue dots

Description automatically generated

Figure pairplot

1. Heatmap:

To visualize the total UPDRS and motor UPDRS Heatmaps are generated which help to give a proper overview of the data. The correlation between age and motor\_updrs seems to have strong positive correlation indicating age increases, motor\_updrs also increases. Similarly, jitter(abs) and shimmer(abs) have strong correlation whereas between shimmer(apq111) and nhr it has strong negative correlation which mean when nhr increase, shimmer decrease and vice versa.

A graph of a graph

Description automatically generated with medium confidence

Figure heatmap

1. Boxplot

To visualize the total UPDRS and motor UPDRS Box plots are generated which help to give a proper overview of the data. The box plot depicts that motor and total UPDRS score differ in Parkinson disease and since, IQRs of the box plots are similar to each other it suggests that there is degree of variance in motor and total UPDRS scores.

A diagram of a box plot

Description automatically generated with medium confidence

Figure boxplot

# Feature selection and engineering

For Feature selection and engineering, we choose these selected features (age, sex, jitter%, shimmer%, nhr) and target variables were motor\_updrs and total\_updrs to improve the model performances.

Feature selection is done to select a particular feature which is the suitable one to improve the model performance of the model. For feature engineering we have applied the yeo-Johnson transformation to the selected data to help improve the linearity assumption required for regression modelling. To address non-Gaussian distributions implementing yeo-Johnson transformation first PowerTransformer is imported which is useful for transforming data when dealing with non-normally or skewed variable. Then columns for transformation are selected and PowerTransformer is initialized. Using fit\_transform method fit and transform the data and display the row to check transformed data.

The feature selected have undergone the transformation to address the non-gaussian distribution for it to use while linear regression modelling. We can predict the motor and total UPDRS based on the selected variables.

A table with numbers and text

Description automatically generated

# Linear regression modelling

in the po2\_dataset.csv, multiple linear regression modelling is done to predict the motor UPDRS and total UPDRS score as well as to evaluate data and to have the better understanding of the factors that influence Parkinson disease.

From response variables (y) separating explanatory variables (x). Here, predictor variables are y total and y motor, which are used to predict motor UPDRS and total UPDRS. Splitting the dataset using train test split function into training and test sets. Here, 60% of data is used for training and 40% of data is used for testing. To predict ‘motor UPDRS’ and ‘total UPDRS’ linear regression ‘model-motor’ and ‘model total’ are created respectively. Fit method is used to fit the models to train linear regression using training data. On the test data, model motor and model total, two trained models are used to make predictions. Printing intercept and c coefficients where intercept is y-intercept and coefficients are weight assigned. For ‘motor UPDRS’ and ‘total UPDRS’ metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Squared Error (RMSE), adjusted R-squared and R-squared (R²). Then, the number of predictors is calculated. For both motor UPDRS and total UPDRS adjusted value is calculated. Performance metrices MAE, MSE, RMSE, R-squared and Adjusted R-squared for motor UPDRS and total UPDRS are displayed. In the training set the mean of 'motor\_updrs' and 'total\_updrs is calculated and their mean values are replicated multiple times. The predicted value is displayed next to the actual value and provide additional performance focusing the baseline metrics and display the values.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A white background with black and red dots

Description automatically generated

## Performance analysis

We have performed an analysis on the linear regression above on various scenarios:

* 50% training, 50% test.
* 60% training, 40% test.
* 70% training, 30% test.
* 80% training, 20% test.

From the response variables (y) separate explanatory variables (X) and defining the various test and training ratio. Using the split method () to split the data set into training and test set. Linear regression models for both motor UPDRS and total UPDRS are created and predicted the value on the test set. For motor UPDRS and total UPDRS calculating performance metrics MAE, MSE, RMSE and R-squared (R²). Analyzing and interpreting the performance metrics. For motor UPDRS and total UPDRS adjusted R-squared is calculated and displayed.

A screenshot of a computer

Description automatically generatedA white background with black text

Description automatically generated

# Optimization

## Rescaling

The summary covers two linear regression models—one predicting Motor UPDRS scores and the other Total UPDRS scores. We've used -score standardization to the explanatory variables to build and evaluate the linear model for motor and total UPDRS. R-squared values show how much of the score variability the models explain (around 24.6% and 26.6% for each score). Adjusted R-squared considers predictor count. The models are statistically significant with low p-values and have coefficients revealing variable impacts and precise standard errors. Diagnostic stats like Omnibus and Jarque-Bera check model assumptions, while the Condition Number spots multicollinearity. These insights help assess model effectiveness, significance, and areas for improvement.

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer

Description automatically generated

## Log transformation to selected feature.

The summary presents a linear regression model using "Absolute jitter" as the independent variable and another undisclosed variable as the dependent one. The R-squared value is an unusual perfect 1.000, indicating that "Absolute jitter" explains the entire variance in the dependent variable, raising concerns of overfitting. The F-statistic is remarkably high, implying strong predictiveness. However, diagnostic statistics suggest non-normally distributed residuals. Overall, the model appears to be overfitting, demanding further scrutiny and evaluation for practicality and data suitability.

A graph with a line

Description automatically generated

Figure old jitter

A graph of jigsaw puzzle

Description automatically generated

Figure new jitter

A screenshot of a computer screen

Description automatically generated

## Gaussian

The output summarizes two linear regression models that is the "Initial Linear Regression Model" and the "Linear Regression Model with Transformed Variables." In the first model, it explains around 77.8% of the variance in "ppe" with strong overall significance. The second model, with some variable transformations, slightly improves the explanation (79.8%) while still being highly significant. Both models look promising, but we should do additional checks for things like multicollinearity and the normality of residuals to ensure their reliability. Plus, it's crucial to have domain-specific knowledge for understanding the real-world impact of the coefficients.

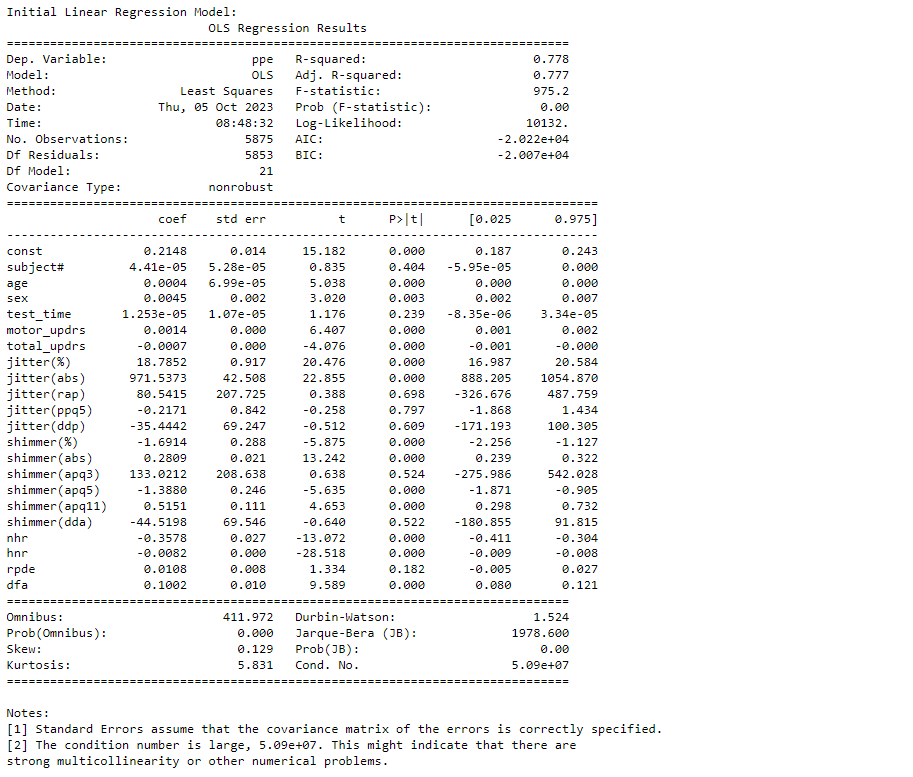
A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

## Non-linear transformation

The shown result represents two linear regression models: the "Initial Linear Regression Model" explains about 77.8% of the variance in "PPE" with good fit but potential issues in assumptions, while the "Linear Regression Model with Transformed Variable" boasts a perfect R-squared value of 1.000, indicating a strong fit but raising concerns about overfitting. To ensure the reliability of these models, further investigation and alternative approaches may be necessary.

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer

Description automatically generated

# Evaluation

As we evaluate the predictions and output of the UPDRS scores from the multiple linear regression model, we can predict that:

For the motor UPDRS score of the multiple linear regression in the po2\_dataset displayed MAE: 6.51, MSE: 58.48, RMSE: 7.64, R2: 0.08, Adjusted R2: 0.083. As we can see the R-squared score is low which suggests that the model isn’t quite suitable at prediction of the motor UPDRS score and independent variable.

For the total UPDRS score of multiple linear regression in the po2\_dataset displayed MAE:8.18, MSE: 96.90, RMSE: 9.84, R2: 0.12, Adjusted R2: 0.11.

We have provided some baseline predictions for motor and total UPDRS. The baseline prediction is done by the representation of mean which is a reference point to evaluate the multiple linear regression model.

# Discussion and Limitations

## Discussion

For this project, the main objective was to predict the motor and the total UPDRS scores of people with Parkinson disease. We implemented a multiple linear regression analysis to analyze the relationship between independent variables and the UPDRS scores of motor and total. The predictions on our model help in assisting the health workers to monitor and analyze the patients of Parkinson disease.

Our multiple linear regression modelling shows an R-squared value of 0.0855 for motor UPDRS and 0.1218 for total UPDRS which indicates the existence of linear relation between the selected variables in the feature selection and the UPDRS. For motor UPDRS and total UPDRS prediction the code displays the MSE, and R squared values which provide indication of model performance. Using the seaborn the code contains some data visualization and exploitation. This helps to provide relationships, correlations and patterns and guide further decision and analysis too. To visualize relationship between the data pair plot is generated where the variables include ‘age’, ‘motor\_updrs’, 'total\_updrs', 'jitter (%)', 'shimmer (%)', and 'nhrand.

## Contribution to knowledge

This project for the diagonising Parkinson patients using voice sample analysis contributes to the healthcare assessments by finding out their relationships between the features and the scores. We have room for improvements in this project, but this offers an initial point to go into research for more predictive methods. This prediction can aid in the health sector to plan and evaluate the disease.

## Limitation

1. For analysis on this project there were limitation by the dataset which could impact the accuracy of our model.
2. Prediction power is limited as we have low R-squared values in our model.
3. Limited features were used which influenced the UPDRS score.
4. May need broader range of factors.
5. Linearity assumption: The code assumes that there is the linear relationship between motor UPDRS and total UPDRS score but linear regression model may not capture it accurately if actual relationship is non-linear.

# Brief descriptions of individual contributions

Kritica Shakya: For this project, I did the coding part for task 1. My responsibility here was coding. I coded the multiple linear regression and further optimize for more information. I also did some contribution on documentation as well.

Neha Pandey: I devoted my time to research and went through the provided dataset in po2\_data.csv file and analysis the performance of linear regression made in task1. Additionally, I be partly responsible for data preparation and collection and supported the team.

Shatabdi Shah: The main contribution from my part was completing task 3. I also contributed to the preparation of the documentation file. I updated the linear regression and reported the performance gain to add more insight.

Sandeep Baniya: I took the responsibility for the completion of task 4 as well as analyzing the minute details in preparation parts and analyze the results carefully. I helped my team find out the required documents and parts for the analysis.

# Conclusion

As a result of our efforts to forecast the motor and overall UPDRS scores in Parkinson's disease patients, we have gained insightful knowledge and seen the potential of regression modeling. It's vital to recognize the intrinsic complexity of these scores, which can be affected by a variety of factors, even though we have reached a respectable level of predictive accuracy. In this field, model complexity, data quality, ethical issues, and the requirement for clinical interpretation have all been underlined by our experience. In terms of managing Parkinson's disease, this effort offers a solid foundation for further study and the improvement of predictive models, with the potential to enhance patient care. As R's value of log transformation is superior to other initial techniques, we can claim that this approach is the best for the project.

We can conclude that while the linear regression model done in this project gives prediction on motor UPDRS and Total UPDRS, in terms of accuracy we can still improve it. The yeo-transformation which was applied for target variables had a positive effect on the performance of the model.