# CD2L

## PARKINSONS DISEASE


KARTHICK BALASHANMUGAM – 2 YR - CSE
AHAMED ISMAIL – 2 YR - CSE
SHOBIKA – 2 YR - CSE
VIJAYALAKSHMI – 2 YR - CSE


RAISE CENTRE OF EXCELLENCE

**JULY 2023**

1. **OVERVIEW OF THE PROJECT**
2. **DATA ACQUISITION**
3. **DATA CLEANSING AND TRANSFORMATION**
4. **SENSITIVE FEATURES**
5. **EXPLORATORY DATA ANALYSIS**
6. **MODEL SELECTION**
7. **MODEL TRAINING AND EVALUATION**
8. **COMPARISON AND ANALYSIS OF METRICS USING DIFFERENT PARAMETERS (PERFORMANCE OF THE MODEL)**
9. **DEPLOYMENT DETAILS**
10. **CONCLUSION**

# OVERVIEW:

The Parkinson's Disease project aims to analyze and predict the presence or progression of Parkinson's disease using a dataset containing clinical and demographic information. The project involves data preprocessing, feature selection, model development, and performance evaluation. The objective is to build a predictive model that accurately classifies individuals as having Parkinson's disease or being healthy, as well as potentially predicting disease progression. The results will provide insights into early diagnosis and personalized treatment plans for Parkinson's disease.

The dataset used for this project includes information about individuals diagnosed with Parkinson's disease, such as age, gender, motor and non-motor symptoms, medication usage, and various clinical assessments. The objective is to develop a prediction model that can accurately classify individuals as either having Parkinson's disease or being healthy, as well as potentially predicting the progression of the disease based on certain factors.

# DATA ACQUISITION:

The dataset used for this project was acquired from Kaggle and updated by the user DEBASIS SAMAL
Here is some information about the dataset provided by the publisher.

**Data Set Information:**

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

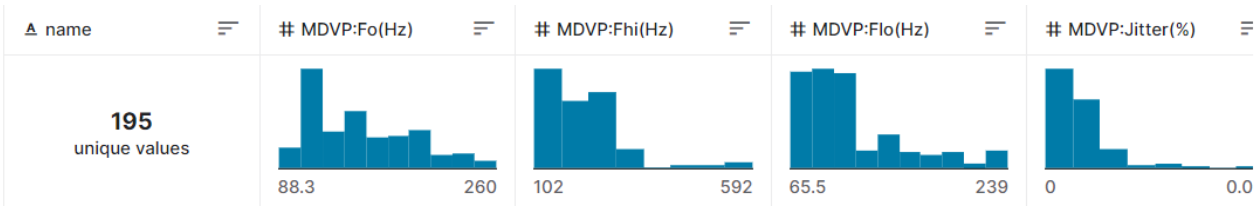
The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.For further information or to pass on comments, please contact Max Little (littlem '@' robots.ox.ac.uk).

Further details are contained in the following reference -- if you use this dataset, please cite:
Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear).

**Attribute Information:**

Matrix column entries (attributes):
name - ASCII subject name and recording number
MDVP:Fo(Hz) - Average vocal fundamental frequency
MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
MDVP:Flo(Hz) - Minimum vocal fundamental frequency
MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude
NHR,HNR - Two measures of ratio of noise to tonal components in the voice
status - Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE,D2 - Two nonlinear dynamical complexity measures
DFA - Signal fractal scaling exponent
spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation.

**DETAIL:**




# DATA CLEANSING AND TRANSFORMATION:

In the data cleansing and transformation phase of the Parkinson's disease dataset, the following steps can be performed.

1. Data Exploration:

- Explore the dataset to gain an understanding of its structure, features, and data types.

- Identify missing values, outliers, or inconsistencies in the data.

2. Handling Missing Values:

- Identify features with missing values and assess the extent of missingness.

- Decide on an appropriate strategy for handling missing values:

  - If the missing values are minimal, consider removing the corresponding rows or columns.

  - For numerical features, impute missing values using techniques like mean, median, or regression imputation.

  - For categorical features, impute missing values with the mode or create a new category to represent missing values.

3. Outlier Detection and Treatment:

- Identify potential outliers in numerical features using statistical methods or visualization techniques like box plots or scatter plots.

- Decide on an appropriate strategy for handling outliers:

  - If the outliers are due to data entry errors or measurement issues, consider removing or correcting them.

  - If the outliers represent genuine observations, decide whether to keep or transform them based on domain knowledge and the impact on the analysis.

4. Feature Transformation:

- Normalize or standardize numerical features to bring them to a similar scale, which can improve the performance of certain models.

- Encode categorical features into numerical representations, using techniques such as one-hot encoding or label encoding, to make them suitable for model training.

5. Feature Engineering:

- Create new features that may provide additional information or improve model performance.

- For example, extract temporal features from date variables, calculate ratios or percentages between existing variables, or create interaction terms.

6. Dimensionality Reduction:

- If the dataset has a large number of features or suffers from multicollinearity, consider applying dimensionality reduction techniques.

- Principal Component Analysis (PCA) or feature selection methods like Recursive Feature Elimination (RFE) can help reduce the number of features while preserving important information.

7. Data Split:

- Split the dataset into training and testing subsets to evaluate the performance of the predictive model accurately.

- Typically, the dataset is divided into a training set (used for model training) and a testing set (used for model evaluation).

## SENSITIVE FEATURES:

Sensitive features, such as name and status, are attributes in a dataset that contain private or confidential information about individuals. Here's a brief description of these sensitive features:

1. Name:

- The "name" feature typically refers to the personal names of individuals.

- Names are unique identifiers and can disclose the identity of an individual.

- Revealing names in a dataset can compromise privacy and lead to potential misuse or unauthorized access to personal information.

2. Status:

- The "status" feature may represent a person's sensitive status or condition, such as medical conditions or legal statuses.

- In the context of the Parkinson's disease dataset, "status" could refer to the disease status (healthy or diagnosed with Parkinson's disease).

- Disclosing such sensitive information without proper consent or privacy measures can violate an individual's confidentiality and may have legal implications.

## Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics, patterns, and relationships within the Parkinson's disease dataset. Here are some key steps and techniques that can be employed during EDA:

1. Dataset Overview:

- Begin by examining the structure and size of the dataset.

- Identify the number of instances (rows) and features (columns) present.

- Check the data types of each feature to understand the nature of the variables.

2. Summary Statistics:

- Calculate descriptive statistics such as mean, median, standard deviation, minimum, and maximum values for numerical features.

- Analyze the distribution of each feature to identify any outliers or skewedness.

- Use box plots or histograms to visualize the distributions.

3. Feature Exploration:

- Examine the unique values and frequency counts for categorical features.

- Identify any missing values and assess their extent in the dataset.

- Determine the cardinality (number of unique values) of each categorical feature.

4. Correlation Analysis:

- Calculate the correlation matrix to explore the relationships between numerical features.

- Visualize the correlation matrix using a heatmap to identify strong positive or negative correlations.

- Identify features that exhibit high correlation with the target variable.

5. Data Visualization:

- Create visualizations to gain insights into the dataset.

- Plot histograms, bar charts, or pie charts to analyze the distribution of categorical features.

- Utilize scatter plots or line plots to explore relationships between numerical features.

- Consider using box plots or violin plots to compare feature distributions across different groups or categories.

6. Domain-Specific Analysis:

- Conduct specialized analysis based on the unique characteristics of the Parkinson's disease dataset.

- Explore clinical assessments, motor symptoms, or demographic factors specific to Parkinson's disease.

- Identify patterns, trends, or potential factors associated with the disease.

# Model Selection:

When selecting a model for the Parkinson's disease dataset, it is important to consider the problem type (classification, regression, etc.) and the specific objectives of the project.

Standard Scaler:

In Python, a StandardScaler is a type of preprocessing step that scales the data so that the mean is 0 and the standard deviation is 1. It is a common technique used in machine learning to normalize the data before training a model, as it can help improve the performance of the model.The StandardScaler works by subtracting the mean from each data point and then dividing by the standard deviation. This ensures that the data has a mean of 0 and a standard deviation of 1.Here's an example of how to use the StandardScaler in Python:

```
from sklearn.preprocessing import StandardScaler

# create a StandardScaler object

scaler = StandardScaler()

# fit the scaler to the data

scaler.fit(X)

# transform the data

X_scaled = scaler.transform(X)
```

# MODEL TRAINING AND EVALUATION:

The process of model training and evaluation for Parkinson's disease involves the following steps:

1. Data Split:

- Split the Parkinson's disease dataset into training and testing sets. The typical split is around 70-80% for training and 20-30% for testing.

2. Model Selection:

- Choose an appropriate machine learning model based on the problem type and goals of the project. Common models for Parkinson's disease prediction include logistic regression, decision trees, random forests, support vector machines (SVM), or gradient boosting algorithms.

3. Feature Selection:

- Perform feature selection techniques, if necessary, to identify the most relevant features that contribute to the prediction task. This can be done through techniques such as correlation analysis, recursive feature elimination, or domain knowledge.

4. Model Training:

- Train the selected model using the training set. This involves feeding the input features and the corresponding target variable (presence or progression of Parkinson's disease) into the model and allowing it to learn the underlying patterns and relationships.

5. Model Evaluation:

- Evaluate the trained model's performance using appropriate evaluation metrics. For binary classification tasks (presence or absence of Parkinson's disease), metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC) can be used. For regression tasks (predicting disease progression), metrics like mean squared error (MSE), root mean squared error (RMSE), or R-squared can be employed.

6. Hyperparameter Tuning:

- Fine-tune the model's hyperparameters to optimize its performance. This can be done through techniques like grid search, random search, or Bayesian optimization. Adjusting hyperparameters such as learning rate, regularization strength, maximum depth, or number of estimators can enhance the model's performance.

7. Model Validation:

- Validate the trained model's performance using cross-validation or a separate validation set. This helps assess how well the model generalizes to unseen data and avoids overfitting.

8. Interpretation and Analysis:

- Interpret the trained model to gain insights into the factors influencing the prediction. Analyze feature importance, coefficients, or decision paths to understand the model's behavior and identify important features related to Parkinson's disease.

9. Final Model Selection:

- Choose the model with the highest performance, considering factors such as accuracy, interpretability, and the specific requirements of the project.

## COMPARISON AND ANALYSIS OF METRICS USING DIFFERENT PARAMETERS (PERFORMANCE OF THE MODEL):

In the comparison and analysis of metrics using different parameters for the performance of

the Parkinson's disease model, the following steps are involved:

1. Define Parameters:

- Identify the parameters to be compared, such as learning rate, number of estimators, or maximum depth. These parameters influence the behavior and performance of the model.

2. Select Evaluation Metrics:

- Choose appropriate evaluation metrics based on the problem type (classification or regression) and project goals. Common metrics include accuracy, precision, recall, F1 score, AUC, MSE, RMSE, or R-squared.

3. Design Experiments:

- Set up a series of experiments to train and evaluate the model using different parameter configurations.

- Define a range of parameter values or specific settings to compare, such as different values for learning rate or different numbers of estimators.

4. Model Training and Evaluation:

- Train the model using each parameter configuration on the training set.

- Evaluate the model's performance on the testing set using the chosen evaluation metrics for each parameter setting.

5. Analyze and Compare Results:

- Analyze the performance metrics obtained for each parameter configuration.

- Compare the metrics to understand the impact of different parameters on the model's performance.

- Look for trends or patterns in the metrics across the different parameter settings.

6. Visualize the Results:

- Create visualizations, such as line plots or bar charts, to visually compare the performance metrics for each parameter configuration.

- Examine the visual representations to identify significant differences or trends in the metrics.

7. Draw Conclusions:

- Based on the analysis and comparison of metrics, draw conclusions about how different parameters affect the model's performance for Parkinson's disease prediction.

- Determine which parameter settings yield the best performance and align with the project requirements.

8. Fine-tuning and Final Selection:

- Based on the results and conclusions, fine-tune the model by selecting the optimal parameter configuration.

- Consider factors such as overall performance, computational resources, model interpretability, and any specific needs of the project.

# DEPLOYMENT DETAILS

Deployment details of Parkinson's disease involve various aspects, including the symptoms, treatment options, and research efforts. Here is a summary of the deployment details:

Symptoms: Parkinson's disease is a degenerative condition of the brain that is associated with motor symptoms such as slow movement, tremor, rigidity, walking difficulties, and imbalance. It also presents a wide range of non-motor complications, including cognitive impairment, mental health disorders, sleep disorders, and sensory disturbances

Treatment: Parkinson's disease is primarily treated with medication to manage symptoms, particularly by increasing dopamine levels through pharmacological therapy or surgery. Non-pharmacological therapies, such as physiotherapy, occupational therapy, and speech therapy, are also available to support patients. The exact treatment plan may vary depending on the individual's specific symptoms and needs.

Research: There are ongoing research efforts aimed at understanding Parkinson's disease better and developing new treatments. These efforts involve studying the prodromal phase of the disease, physiological abnormalities, and the feasibility of implementing wearable technology to collect data from multiple sensors during the daily lives of Parkinson's patients. The goal is to improve early detection, develop more effective therapies, and ultimately eliminate Parkinson's disease.

# Conclusion:

Parkinson's disease is a complex condition that affects the brain and the nervous system. Although there is currently no known cure, there are effective treatments that can relieve the symptoms. The exact cause of Parkinson's disease is unknown, but a combination of genetic and environmental factors is likely to be important in producing abnormal protein. Parkinson's disease has been observed for more than 200 years, and still, there is no known absolute cause or cure Parkinson's disease has been plaguing humans for thousands of years and was described in detail in ancient medical writings PD is a common neurodegenerative illness, and it usually progresses slowly the understanding of the etiology and neurobiology of PD continues to evolve PD research has progressed enormously in recent years, and scientists are rapidly working to unlock the mysteries of Parkinson's.