Early Detection of Parkinson's Disease through Voice Analysis using Machine Learning

INTRODUCTION:

Parkinson's disease, a progressive neurodegenerative disorder, poses significant challenges for early detection. This project aims to create a non-invasive diagnostic tool using machine learning techniques applied to voice recordings. The objective is to identify individuals with Parkinson's disease at an early stage, enabling timely intervention and improving patient outcomes.

OBJECTIVES:

- **Exploring the Dataset:**
  - Uncover patterns, trends, and connections within the dataset.
  - Investigate the underlying relationships between voice features and Parkinson's disease.
- **Conduct Extensive Exploratory Data Analysis (EDA):**
  - Delve into bivariate relationships to reveal nuanced connections.
  - Extract valuable insights from the complex interplay of features and their impact on the target variable.
- **Preprocessing Steps:**
  - Identify and remove irrelevant features to streamline the dataset.
  - Implement strategies to handle missing values and outliers.
  - Encode categorical variables to ensure compatibility with machine learning models.
- **Model Building:**
  - Develop pipelines to scale models requiring normalization.
  - Implement classification models, including KNN, SVM, Decision Trees, and Random Forest, for accurate prediction of Parkinson's disease.
  - Emphasize achieving a high recall for class 1 to ensure thorough identification of individuals with Parkinson's disease.
- **Neural Network Architecture:**
  - Design and implement a neural network architecture to enhance the model's predictive capabilities.
- **Evaluate and Compare Model Performance:**

- Utilize precision, recall, and F1-score metrics to comprehensively assess the effectiveness of different models.
- Compare the performance of Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor models.

Data Scaling:

- Adoption of normalization techniques to address significant variations in measurement units.
- Consideration of skewness in feature distributions, necessitating appropriate scaling for improved model performance.

Crosstab Analysis:

Identification of Parkinson's disease prevalence within specific ranges of 'mdvp_fo_hz,' indicating a potential correlation.

Correlation Analysis:

Strong correlations observed among variables related to voice shimmer and jitter, aligning with the data description.

Observations:

Logistic Regression and SVC demonstrate better performance on the validation data, with accuracies ranging from 75% to 80%. A confusion matrix and classification report for the validation data using the Logistic Regression model are presented, providing insights into model performance.

- **Model Performance**: The implemented models, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor, consistently demonstrate high precision, recall, and F1-scores. This indicates the models' efficacy in accurately predicting individuals with Parkinson's disease based on voice recordings.
- **Feature Importance and Insights**: Through extensive exploratory data analysis and model evaluation, certain voice features emerge as critical indicators of Parkinson's disease. The understanding of feature importance provides valuable insights for both the medical and machine learning communities, contributing to a deeper comprehension of the disease's early markers.

CONCLUSION:

The project establishes a robust machine learning model using Logistic Regression, SVM, Decision Tree, Random Forest, and KNN for early Parkinson's disease detection through voice analysis. Achieving high precision, recall, and F1-scores, crucial voice features emerge as significant indicators, contributing valuable insights. The developed diagnostic tool holds promise for clinical integration, enhancing early screening and monitoring. This project is a noteworthy advancement in medical AI, potentially revolutionizing Parkinson's disease diagnosis and management efficiency.