Parkinson Disease Detection App: Updated Business Requirements

1. Data Pre-Processing & Cleaning

1.1 Load and Review Data

- Load the dataset and review its structure, focusing on columns, data types, and the distribution of the target variable. Verify the alignment between Kaggle and UCI datasets to ensure the dataset is suitable for analysis.

1.2 Understand Target Variable Distribution

- Analyze the target variable distribution to understand if the dataset is balanced between Parkinson's and healthy cases. Flag any imbalance for further analysis.

1.3 Handle Missing or Outlier Values

- Handle missing values using suitable imputation strategies, such as the mean or median. Examine any unusual outliers to identify data entry errors or domain-specific anomalies.

1.4 Data Transformation

- Apply scaling or normalization, such as using StandardScaler or MinMaxScaler, to handle models sensitive to scale differences.

2. Exploratory Data Analysis (EDA)

2.1 Define EDA Goals

- Establish goals for understanding feature relationships with Parkinson's status, such as analyzing frequency, amplitude, and noise metrics.

2.2 Analyze Features

- Examine vocal frequency metrics (e.g., MDVP(Hz), MDVP(Hz), MDVP(Hz)) using visualizations like histograms and box plots to identify differences between Parkinson's and healthy cases.

2.3 Study Noise Variation Metrics

- Analyze noise variation metrics (e.g., Jitter, Shimmer, NHR, HNR) to explore the correlation with Parkinson's presence. Use visual tools like heatmaps if applicable.

2.4 Examine Nonlinear Complexity Measures

- Explore nonlinear complexity measures (e.g., RPDE, DFA, spread variables) to understand vocal irregularities in Parkinson's patients.

2.5 Summary Insights

- Summarize key observations from EDA to inform subsequent feature engineering and model selection steps.

3. Feature Engineering

3.1 Feature Transformation or Interaction

- Create interaction features based on EDA insights (e.g., combining amplitude and noise ratios).

3.2 Dimensionality Reduction

- Apply PCA to reduce dimensionality if there are correlated features, streamlining data while retaining critical information.

4. Data Splitting

4.1 Train-Test Split

- Perform an 80/20 split for training and testing, ensuring that the test set accurately represents the entire dataset.

5. Handling Class Imbalance

5.1 Oversampling and Undersampling Techniques

- Use SMOTE to handle class imbalance in the training set, and apply undersampling if needed for comparison purposes.

6. Training Machine Learning Models

6.1 Model Selection and Metrics

- Use models such as Logistic Regression and Random Forest to predict Parkinson's status, comparing results using metrics like accuracy, recall, F1-score, and AUC-ROC.

6.2 Model Training and Evaluation

- Train models on preprocessed and balanced data, and perform hyperparameter tuning using techniques like GridSearchCV to optimize performance.

7. Feature Importance Analysis

7.1 Identify Significant Predictors

- Use feature importance measures, such as Random Forest feature importance, to identify significant predictors contributing to the model's predictions.

7.2 Summary of Predictive Features

- Summarize the most influential features for Parkinson's detection to provide transparency in the model's decision-making process.