Parkinson's Disease Detection Project: Summary of Discussions and Findings

1. Project Overview

The primary goal of this project is to develop a classification model that predicts whether an individual has Parkinson's disease based on vocal features derived from voice recordings. This is achieved through a series of machine learning tasks including data processing, exploratory analysis, model training, and evaluation.

2. Data Exploration and Preprocessing

2.1 Dataset Overview

- The dataset contains vocal characteristics such as jitter, shimmer, fundamental frequencies, and noise-to-harmonic ratios, along with a target variable indicating Parkinson's status (1 = Parkinson's, 0 = Healthy).

2.2 Exploratory Data Analysis (EDA)

- Distribution Analysis: Several features (e.g., jitter, shimmer, fundamental frequencies) showed significant right-skewness. This implied that a few patients exhibited notably high vocal instability, typically associated with Parkinson's disease.

- Nonlinear Complexity Measures: Attributes such as RPDE, DFA, and PPE provided insights into the vocal irregularities of patients with Parkinson's.

3. Data Preprocessing

3.1 Log Transformation

- We applied log transformations to right-skewed features to make their distribution more normal. Some columns required additional transformations, but after review, it was decided not to apply repetitive log transformations, to avoid overprocessing the data.

3.2 Data Scaling

- StandardScaler was used to standardize the features to have a mean of zero and unit variance. This helped in ensuring that the models were not biased towards features with larger ranges.

3.3 Data Splitting and Handling Class Imbalance

- The dataset was split into training and test sets (80%/20%).

- SMOTE Oversampling: Synthetic Minority Oversampling Technique (SMOTE) was used to address class imbalance in the training dataset. Both oversampling and undersampling techniques were used to balance the data, and results were compared.

4. Model Training and Evaluation

4.1 Model Selection

- Logistic Regression and Random Forest models were selected for training. The performance of these models was assessed under different conditions: without sampling, with oversampling, and with undersampling.

- PCA (Principal Component Analysis) was applied to reduce dimensionality and further improve model performance by removing collinearity between features.

4.2 Performance Comparison

- Logistic Regression showed a notable improvement in accuracy when SMOTE was used, but Random Forest consistently outperformed Logistic Regression across all sampling techniques.

- Without Sampling: Logistic Regression had poor recall for minority class predictions, indicating that it struggled to correctly identify Parkinson's cases in the imbalanced dataset.

- With SMOTE Oversampling: Both models improved significantly, but Random Forest demonstrated higher precision, recall, and overall accuracy compared to Logistic Regression.

- Undersampling: The performance of Random Forest remained stable, while Logistic Regression's accuracy declined slightly, suggesting that Random Forest was more resilient to smaller sample sizes.

5. Feature Importance Analysis

- Random Forest was used to perform feature importance analysis, which indicated that nonlinear vocal features, such as RPDE, PPE, and DFA, were the most significant predictors of Parkinson's disease.

- Logistic Regression could not provide feature importance in the same way as Random Forest due to its linear nature. Instead, coefficient magnitudes were analyzed to infer feature significance.

6. Findings and Discussions

6.1 Key Insights

- Nonlinear Complexity Features: Measures such as RPDE, PPE, and DFA emerged as strong predictors of Parkinson's disease, indicating the importance of analyzing signal complexity in vocal data.

- Feature Transformations: Log transformation helped normalize the skewed features, improving model training, although further transformations were avoided to prevent overfitting or redundancy.

- SMOTE Effectiveness: Oversampling using SMOTE provided significant performance improvements, particularly for Logistic Regression, which was initially heavily affected by class imbalance.

6.2 Model Performance and PCA Impact

- PCA reduced the feature space and helped in improving computational efficiency. However, the accuracy improvements after PCA varied, with Random Forest showing only a marginal change while Logistic Regression exhibited more fluctuation.

7. Recommendations and Next Steps

7.1 Model Improvements

- Considering exploring additional ensemble models such as Gradient Boosting or XGBoost for potentially better performance.

- Further hyperparameter tuning, particularly for Random Forest, could help enhance model accuracy.

7.2 Deployment Plan

- Working on deploying the Random Forest model using Flask or FastAPI for setting up API endpoints. This model has demonstrated the best performance and reliability across different sampling techniques.

7.3 Interpretable Machine Learning

- To enhance transparency, incorporating SHAP or LIME for explaining individual predictions to users. This will help users understand which features contributed most to the model's predictions.