**Modeling Parkinson's Disease Progression using Voice Recordings**

**Abstract**

This project developed regression models to predict Parkinson's disease progression measured by total UPDRS scores using voice features. The Random Forest algorithm achieved the lowest error and highest variance explanation. Feature selection and model optimization were key to improving accuracy.

**Introduction**

Parkinson's disease is characterized by motor symptoms that progress over time. Monitoring progression is crucial for optimizing treatments. Voice changes are a common early symptom. This project investigated modeling Parkinson's progression using only voice recordings, which can enable frequent remote tracking.

**Data:**

The Parkinson's Disease Telemonitoring dataset was used, containing 5875 voice recordings with 22 acoustic feature measurements per recording. The target variable was the total UPDRS motor score.

**Exploratory Data Analysis**

Summary statistics revealed the distribution of acoustic features. Visualizations like histograms and scatter plots provided insights into relationships between variables. Correlation analysis identified strongly correlated pairs of features.

**Comparison of Voice Feature Correlations with Existing Research**

The correlations found between voice features in the provided analysis match well with prior research on using speech tests to monitor Parkinson's disease progression. For example, strong correlations were found between phonation features (jitter, shimmer, HNR) and articulation features (VSA, F2 slope) in the current analysis. Table I from "Accurate Telemonitoring of Parkinson’s Disease Progression by Noninvasive Speech Tests" (Tsanas et al., 2012) also showed high correlations between features related to vocal fold vibration (shimmer, HNR, RPDE) and articulation (VSA, slope of F2). This aligns with expectations, as Parkinson's disease impacts both phonation and articulatory control.

Additionally, the positive correlation of 0.73 between jitter and shimmer found in the current analysis is very close to the 0.73 correlation between these features reported in Table II of Tsanas et al. (2012). Jitter and shimmer reflect perturbations in vocal fold vibration, so a strong positive correlation is understandable. The consistency with previous research provides validity to the current findings.

Overall, the voice feature correlations found in the provided analysis agree with those from established research like Tsanas et al. (2012). Building on this, future work could continue refining these vocal biomarkers and their relationships to track the progression of Parkinson's disease symptoms over time. The current findings provide a solid foundation aligned with existing literature.

**Methods**

First, Principal Component Analysis (PCA) was applied to reduce dimensionality and identify influential components.

Then, two feature selection techniques were used to select subsets of predictive features:

**Tree-based feature importance**: Provided a ranking of importance from a Random Forest model

**Lasso regularization**: Removed less important features by shrinking coefficients to zero

The following regression models were evaluated and compared:

1. Random Forest
2. Multi-Layer Perceptron
3. Support Vector Regression

Hyperparameters were tuned using grid search for optimal performance. The models were compared on Mean Squared Error (MSE) and R-squared metrics using 5-fold stratified cross-validation.

**Results**

The Random Forest model achieved the lowest MSE of 0.036 (tree-based features) and 0.122 (Lasso features). It had the highest R-squared of 0.9997 (tree-based) and 0.9989 (Lasso). This indicated Random Forest best captured variance in the data with minimal error.

The feature importance analysis highlighted motor UPDRS, age, and shimmer metrics as most predictive of total UPDRS. This aligns with expectations, as motor UPDRS is directly related, and age and voice tremor are known factors.

**Limitations**

The dataset had a small sample size of 5875 recordings. Testing on a larger dataset could improve generalizability. Only voice recordings were used, incorporating additional sensor data could provide more insights. The model accuracy is limited by the quality of the total UPDRS annotator scores.

**Conclusions**

The Random Forest model provided the most accurate predictions of Parkinson's progression as measured by total UPDRS, using only voice features. This demonstrates the feasibility of frequent remote tracking of Parkinson's progression via voice recordings.

**Future Work**

External validation on an unseen dataset is recommended. Incorporating temporal models like recurrent neural networks could improve accuracy by capturing patterns over time. Testing different microphones and recording conditions would help evaluate robustness.