**Classification of dysphonic Voice Patterns of Patients with Parkinson's Disease**

This study aims to train a model that classifies voices of people with Parkinson’s disorder from those without Parkinson’s disorder by analysing dysphonic voice patterns.

Dysphonia refers to having an abnormal voice. It is also known as hoarseness. Dysphonia has many causes. Idiopathic Parkinson's disease (IPD) is known as a chronic neurodegenerative disorder that may lead to producing dysphonic voice due to probable neurogenic interruptions in the laryngeal nerve paths. It is reported that about 70% to 80% of IPD patients would suffer from dysphonia or other phonatory disorders, with the symptoms of decreased variation, roughness, increased asthenia, dysarthria, or voice tremor.

The dataset set used for this study was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders. The dataset contains of a range of biomedical voice measurements from 31 people, 23 with IPD. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals.

It can be found at https://archive.ics.uci.edu/ml/datasets/Parkinsons

To achieve the aim of this study, the dataset was loaded and explored for irregularities. It was observed that the dataset was unbalanced. There were significantly more voice patterns of patients with IPD than healthy subjects. A logistic regression model was ran as a baseline model. The accuracy came out low as expected. Then the features were checked for correlation. The highly correlated features were dropped to improve the classification performance.

Several models were run on the reduced dataset. Then the dataset was balanced out using SMOTE method for improved performance. Then the models were fitted again, and this led to more improved performance.

Ensemble methods were then employed. Adaboost, gradient boost with hyperparameters tuning were applied. The adaboost model showed overfitting and the gradient boost performed worse upon parameter tuning. The bagging method used was Random Forest. The parameters were tuned by first running a random search, then it was narrowed down using grid search. The model showed high accuracy. Although it tends to overfit the training set, the test data also performs well.

In conclusion, the bagging ensemble method indeed showed the high accuracy just as mentioned in the paper. However, the models were all overfitted. Deep learning would be a good option to resolve this.