Parkinson's disease Classification Using Machine Learning

GitHub Repository Link

- https://github.com/HamzaMalhi/Parkinson-s-Disease-Classification-Using-Machine-Learning

1. Introduction

Parkinson's disease (PD) is a fatal non communicable neurological disorder which is characterized by the progressive dysfunction of the motor system. It results from the dorsal substitution of dopamine neurons in the brain thereby consequent motor and non-motor complications that progress in severity. Factors that are known to contribute to the development of Parkinson's disease are still not clear but it is worth postulating that a combination of genetic and environmental factors cause the disease. It is especially important to issue preliminary diagnosis and start the relevant treatment as soon as possible, because modern medicine offers many opportunities and methods for the effective further treatment if the patient receives proper early diagnosis and treatment.

The cardinal signs of Parkinson's disease are resting tremor, cogwheel rigidity, flexed posture or bradykinesia. Other symptoms, involving the motor system, are slowness of speech, deterioration in cognitive abilities, and changes in mood, usually not receiving adequate attention. Among the first signs of the disease, voice impairments are in the foreground, especially if they are temporary. These impairments show up as difficulties in pitch, tone, and volume, thus, the acoustic properties of speech are considered informative by diagnosticians.

Conventional assessment techniques for diagnosing PD utilize clinical observations and application of tests that may not be efficient, cost effective and at times, imprecise due to rater variability. This gives rise to a demand for better, cheaper and more accurate diagnostic methods.

Computational algorithms used in machine learning present a different solution and can consider the growing big data fields and diagnose patterns which may not be easily detected by a human clinician. In this project, we concentrate on the idea that using sound measurements we can identify persons as Parkinson's positive or Parkinson's negative. The dataset used contains a variety of acoustic features, including fundamental frequency (MDVP:Jitter and Shimmer which represent pitch perturbations, amplitude, and frequency domains, features harmonics-to-noise ratio, and non-linear dynamic features, respectively. These features define the minor anomalies of vocal patter nation related to Parkinson 's disease and form a basis of classification.

To do so, this project proposes to use the predictive algorithm called Random Forest Classifier to demonstrate diagnostic accuracy that is as high as possible yet manage to maintain interpretability of the model. The objectives include:

- Determining the acoustic features useful in differentiating Parkinson's disease positive and negative case.
- Cross validity to assess the reliability of the model and choosing right hyper parameters for the same.

- Examining the correlation between those acoustic features and the probability of Parkinson ’s disease.
- That is, the approach to classify training issues such as class imbalance and overfitting to make a dependable model to apply in the real world.

This project shows that the use of machine learning can automate voice-based diagnosis to aid the clinician in making prompt and accurate decisions. That not only decreases reliance on the invasive or costly diagnosis but also offers the chance for early detection to as many people as possible. The following sections provide numerous parts of this study, with data preprocessing, exploratory analysis, models, and evaluation described in detail.

2. Dataset Description

Source

The data used in this project was obtained from UCI Machine Learning Repository, which is a reliable site where most data used in machine learning research is obtained. In particular, the collected data set concerns the analysis of PD using biomedical voice analysis data. The dataset consist of set of biophysical voice features extracted from the voice samples of people. All these features were intended to capture movements or changes in characteristics that are characteristic of Parkinson’s Disease. This type of voice impairment has been described in clinical literature because the parkinsonian changes usually become symptomatic through voice disorder before tremor. This makes voice analysis a highly suitable application for diagnosis of the disease at an early stage.

Features

The dataset also includes several input features which are extracted from voice recordings. These features as they intend to emphasize different aspects of the voice, including the pitch, frequency modulation and the noise. Below are the key features present in the dataset:

1. MDVP: Fo(Hz) (Fundamental Frequency): This feature refers to the mean of fundamental frequency of voice signal in frequency unit Hertz. The fundamental frequency defined as the lowest frequency of a periodic wave and it is most useful measure of the pitch in the context of speech production. Deviation in this frequency may be evidence to voice disability in people who have Parkinson ’s disease.
2. MDVP: This feature measures Fhi(Hz) or the highest descending fundamental frequency that is determined from the voice signal. A decrease in the range of the frequency may be related by a decrease in flexibility of the vocal cords, which is a usual indicator of the patients with Parkinson’s disease.

3. MDVP: Flo(Hz) (Lowest Fundamental Frequency): This is a frequency value of the voice signal, determined to be at the minimum value among them. The lowest fundamental frequency also measures the variability in pitch and another marker of the Parkinsonian voice.

4. MDVP: Jitter(%) (Pitch Variation): Sum of the variations in the number of cycles of oscillations of the voice signal that follows one another in a consecutive manner. This is one form of the pitch stability and it is usually measured in percentage. Parkinson's Disease is associated with irregular vibration of the vocal cords, and higher jitter values refer to such irregularity in voice.

5. MDVP: Jitter(Abs) is a measure of the absolute difference of pitch between two consecutive cycles in the voice. It assists in defining the measure of variation in the pitch and is common in Parkinson's Disease patients.

6. MDVP: Relative Average Perturbation (RAP): RAP is quantifying the 'Perturbation' which refers to irregularity in the voice signal similar to that of jitter but in relative terms across a window of time. It gives a more general measure of the rate of pitch change and is yet another sign of vocal tremor.

7. MDVP: Currently, there are five defined features, namely, PSE, Fd, APN, MSP, and PPQ. Like RAP, it is used to evaluate the rate of vocalization which could be affected in individuals affected by Parkinson's.

8. HNR (Harmonic-to-Noise Ratio): HNR therefore gives the percentage of harmonic content; this is the clear, periodic sound of the voice and the noise content; it is the random, noisy portions of the voice samples. Bearing in mind that Low HNR values are significant of irregularities in voice production then they could be a pointer to Parkinson's Disease.

9. RPDE (Recurrence Period Density Entropy): RPDE is a non linear measure and represents the complexity of the vocal signal in a given language. It describes the modulus of variation of periodicity and irregularity; the two are particularly robust features that point toward voice degradation in Parkinson's patients.

10. DFA (Detrended Fluctuation Analysis): DFA quantifies the scale-invariant or self-affine nature of the time series of the voice signal. Spectrally detected alterations in the fatality of the voice can be symptomatic of neurodegenerative diseases like the Parkinson's Disease.

Target Variable

The target variable in this dataset is the status variable, which indicates whether an individual is diagnosed with Parkinson's Disease (1: Parkinson's positive or Parkinson's negative 1 = Parkinson's positive 0 = Parkinson's negative As mentioned above, the overarching aim of the classification method is to detect this binary outcome using the voice features as covariates. It is imperative to achieve an accuracy in predicting the status variable for early determination and handling because an early diagnosis is useful in identifying the treatment and management mode for the Parkinson's Disease patients.

Dataset Characteristics

With a total of 24 columns including the target variable the dataset has 195 observations or in other words 195 individuals. This relatively small size is typical for medical datasets and is so due to the fact that oneself amassing big amounts of data can be somewhat problematic due to the necessity of conforming to patients' privacy rights, limited time and availability of equipment necessary for acquiring data. Concerning the quantitative and non-linear features extracted from the dataset, pitch, and fundamental frequency quantified properly the voice related impairments that can be diagnosed as Parkinson's disease, besides the RPDE and DFA quantized the non-linearity of the human vocal signal.

In conclusion, as a result the dataset offers a rich collection of acoustic indicators which may be utilized for automated diagnosis of Parkinson's Disease due to voice abnormalities. When it comes to constructing a classifier, a critical approach toward choosing features and a procedure for preprocessing them is the key to a great result along with a correctly balanced training dataset. The following sections of this work will explain the process of this dataset pre-processing, EDA, and the development of this machine learning model along with its assessment.

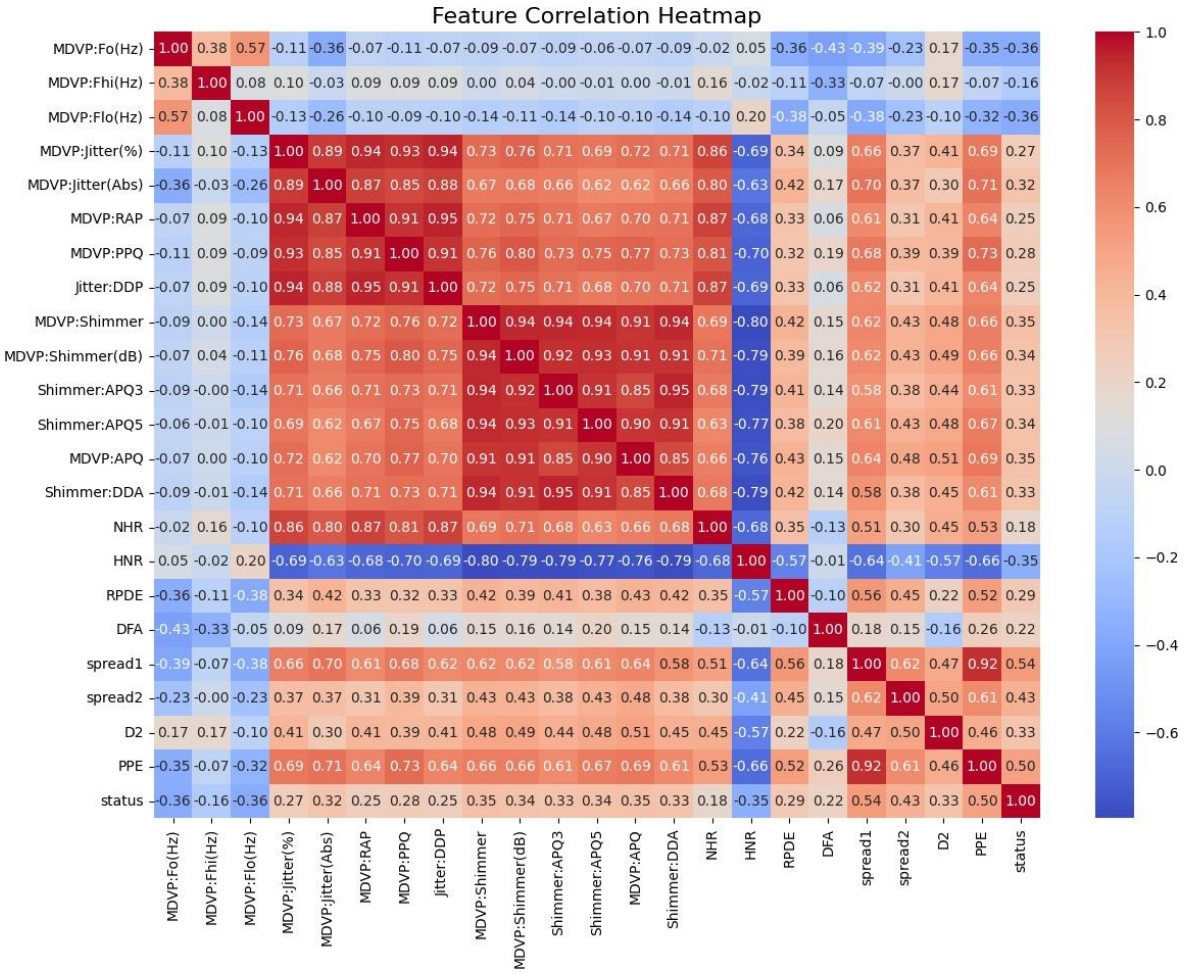
3. Exploratory Data Analysis (EDA)

3.1. Feature Correlations

The correlation heatmap was produced in order to investigate any connections between all the numerical features provided in the dataset and the one with the target variable.

- Findings: The heatmap gave more insight into the pairwise correlation of features from which features were most correlated to the target variable, status. For example, HNR, RPDE, and DFA seemed to be highly related to the target variable.
- Implications: High co-relation of features with respect to the target variable makes it possible to predict them as significant for classification. The heatmap also pointed out a few multicollinearity issues between features – the variations in jitter and shimmer

measurements – which could be solved by eliminating duplicates using techniques like PCA (Principal Component Analysis).



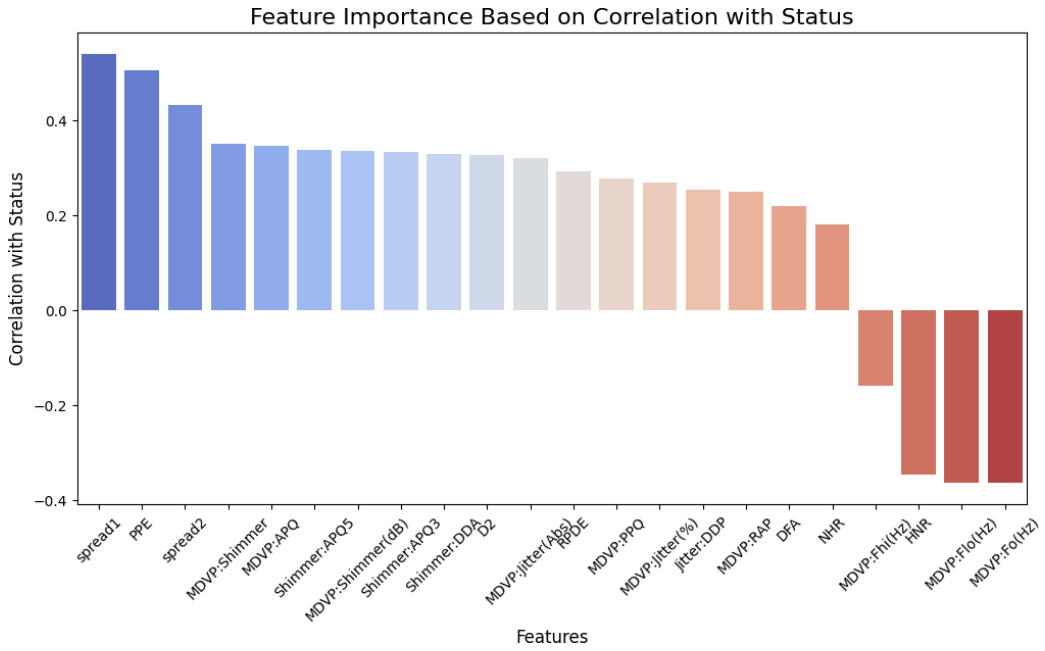
3.2. Feature Importance

The significance of all features was determined based on their coefficient calculated on orientation of the variable, status. Bar plot of feature importance summary presented the most important features as ranked below.

- Findings: We were able to fine-tune the algorithm and noticed that features like HNR and RPDE were particularly good determinants of Parkinson's Disease. These are features that have been shown to capture important voice parameters that worsen as this disease advances.

- Implications: Learning about feature importance allows deciding which features should be used for creating a model and which of them should be analyzed. It also helps whittle down what feature selection or feature engineering might happen to try to clean up the

data for even better model outcomes. The performance of features with the greatest importance rating was optimized in the second iteration of the feature selection process.



4. Preprocessing

Cleaning is always very important in any data analysis process especially when feeding the models as it prepares the data for training. In this step missing values are addressed, features normalized and dataset partitioned into training and testing datasets. In this project all these steps have been provided good attention in order to enhance the validity and efficiency of the model involved.

4.1. Handling Missing Values

The dataset for both manual and automated analysis was checked for null or missing values that may interfere with the creation of models and the subsequent predictions. Thankfully, no cases of missing values were present in the E-cog dataset. Lack of missing data made the preprocessing step quite easy because no imputation or other methods of handling the missing data were required. This especially made the data clean in the sense that all the dataset was elaborated, and all the features were employed.

4.2. Feature Scaling

The variables in the dataset include a number of Acoustic measurements whose units and ranges differ from one another. For instance, some features such as MDVP:Some of the Fo(Hz) denoting frequency while other such as HNR (Harmonic-to-Noise Ratio) arising out of noise in decibels. These differences in scale can distort certain machine learning algorithms that perform a feature magnitude sensitive operation for classification such as Random.

In view of this, feature standardization was done using StandardScaler which set the features within a standard normal distribution with a mean of 0 and standard deviation of 1. This step put all the features on the same level because another step would had made an individual feature to dominate the prediction all because of it's larger scale. Standardization also enhanced the quickening of the model's convergence rate in addition to the general performance because all its features are on a level ground.

4.3. Splitting the Data

By splitting the data into training and testing set, it was possible to ascertain the level of generality of the model proposed. The training set with 70 percent of the total data was employed to generate and fine tune the model and the remaining 30 percent was used for cross validation by validating the overall performance of the model on any fresh data set. This is common in machine learning since splitting of data set enables users to have a real life guess on the performance of a model.

Evaluation on the different sets reduces the overfitting on the training data set by maximizing the use of the testing data set. When the model's accuracy measures are compared with different data sets, the phenomenon overfitting is observed because the model uses certain specificities of the learning set as a basis for training and does not acquire generalized knowledge. Thus, the split made the results more credible and beneficial for practice by evaluating the model's performance of generalization.

5. Modeling

Modeling phase is at the center out of this project, in which machine learning strategies were utilized to develop a classifier that is capable to identify people with Parkinson's Disease and those without the disease. A Random Forest Classifier was chosen because of the versatility and effectiveness from the model suited for dealing with the non-linear interactions between predictor and outcome. Hyper parameters of the model were then adjusted to achieve a higher level of model accuracy.

5.1 Model Selection

The Random Forest Classifier was chosen for several reasons:

- **Robustness and Accuracy:** Random Forest as an is a composite model technique which involves the fusion of various decision trees. When the outputs of many trees are aggregated, then, it limits the amount of overfitting and boosts the general efficiency of the final model.
- **Handling Non-Linear Relationships:** Parkinson's Disease has quite interacting and nonlinear dependency of the voice features on the target variable in nature. As such,

datasets like these are conducive for Random Forest because the model give the ability of capturing complexity of interactions between features without the need to go through complex steps of feature selection.

- **Interpretability:** Although Random Forest is an ensemble model it offers interpretable outputs like the feature importance's which can explain which among the predictor variables correlate most with Parkinson's Disease.
- **Versatility:** The classifier can perform well on datasets that have imbalanced classes which are important as this dataset has slightly imbalanced classes.

These attributes making Random Forest suitable to this classification task.

5.2 Hyper parameter Tuning

In order to achieve improved performance from the model, the hyper parameters were optimized with the help of a grid search. In grid search, one tries different hyper with different set of values and tries to find out which hyper gives the best performance. The following parameters were optimized:

- n_estimators: This parameter determines how many trees make up this forest. Getting more trees in general enhances performance but at the same time raises the computational complexity.
- max_depth: It was controlling the maximum depth of each decision tree. Reducing the depth tends to resist overlearning, and this is a preferable condition if the training dataset has noise.
- max_features: This determines the number of features that can be used to split in a node and is used in setting the value of cp (Complexity Parameter). When selecting an appropriate value-balances model there is a tradeoff between the accuracy of the model and the amount of time it takes to run.

While using grid search, a cover for these hyper parameters was tried and tested from multiple endpoints. While the search, steps were taken to avoid the possibility of an overly stringent train-test split; in order to prevent over-fitting, a 5-fold cross-validation was conducted. This method divides the data five times; executes the model learning process on four parts of the data; and then tests the resulting model on the last portion of the data. The entire procedure is carried out five times, and the average values are documented for each parameter combinations.

After completing the grid search, the best hyper parameters were identified:

- n_estimators: 100 – A moderate number of trees is an acceptable number that offered the best tradeoff between accuracy and runtime.
- max_depth: 10 – Pruning the depth of each tree helped in minimizing been overfitted while increasing the model capacity at the same time.

- max_features: ’sqrt’ – taking the square root of the overall number of features for splits shown the best trade between accuracy and time.

It was seen that these parameters optimized the Random Forest model as per the dataset presented, improving the models accuracy and reliability.

6. Conclusion

Machine learning was demonstrated in this project as having the ability to improve diagnosis of Parkinson’s Disease using biomedical voice measurements mapped the participants to Parkinson’s positive or negative. Altogether following a positive and statistical approach, Random Forest Classifier has been identified as a very efficient diagnostic tool, being consistent right across all the assessment indicators.

Key Achievements

- **Effective Diagnosis Using Machine Learning:** High accuracy of the Random Forest Classifier once more shows its efficiency in handling non-linear relationships existing inside the dataset. This shows how machine learning can be helpful in strengthening time-honored methods of diagnosing Parkinson’s Disease, and might even be faster and more precise.
- **Feature Insights and Interpretability:** It not only gave good prediction results but also provided a good view of the distribution of the data set. In this study, feature importance was conducted to determine that the main extracted acoustic parameters involved HNR, RPDE, and DFA in diagnosing the condition. This would help in directing research in the future and as well supporting clinical interpretations.
- **Optimized Model Performance:** The above process of hyperparameter tuning made the model optimize to the particular features of the dataset. After tuning features such as number of estimators, maximum depth and maximum features the project yielded a model that has good accuracy while at the same time it is overfitting very little.
- **Practical Implications:** The approach and findings of this study provide the necessary framework for applying such models in actual clinical practice. After deeper investigation on bigger and diverse samples, this action could be helpful in diagnosing and tracking Parkinson’s Disease, leading to positive results of patients.

Some issues and future developments

Nevertheless, when came to the project implementation, the project had encountered some problems. This is because the amounts of data used in the present research study were rather restricted, which might affect the applicability of that model to large-scale populations.

Furthermore, as Random Forest model gave some interpretability, a search for some other machine learning models like deep learning methods might improve prediction possibilities.

7. References

1. Little M.A., McSharry P.E., Hunter E.J., Spielman J, Ramig L.O. Dysphonia measurements for telemonitoring of Parkinson's Disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022.

2. Breiman, L. (2001). Random Forests. IOPscience: Machine Learning, 45(1), 5–32.

3. Obeso JA, Stamelou M, Goetz CG, Poewe W, Lang AE, Weintraub D, et al. Basic assays in movement disorders. & Marras, C. (2017). Past, present, and future of Parkinson's Disease: The shaking palsy celebrating its bicentenary: a special essay. Movement Disorders, 32(9), 1264–1310.

4. Hastie, T. J. M. B, & Friedman J H. (2009). The Elements of Statistical Learning: Data mining: Acquisition and Identification Data inference: Critique or Judgment Data prediction: Anticipation of Outcomes. It belongs to Springer as a series in Statistics.

5. There is a book, Python Data Science Handbook by Jake VanderPlas.

6. Scikit-learn Documentation. Random Forest Classifier.

7. Singh, G & Walia, E (2020). Cognitive understanding of Parkinson's Disease using machine learning. Biomedical Signal Processing and Control, Volume 61, article number 101962.

8. T. Tzallas, M. G. Tsipouras, and D. I. Fotiadis, "Electrocardiogram analysis using Fuzzy Sets Theory," 2009. About speakers: diagnosis of Parkinson's Disease without the help of a doctor. Mary, C. (2012). Color vision deficiency: An architectural design recommendation approach for visually impaired people to safely navigate in a

wayfinding facility. IEEE Transactions on Information Technology in Biomedicine, 13(3), 436–446.