Detection of Parkinson's Disease using Machine Learning Algorithms

Vaidehi Mungekar
Red ID: 824666863
vmungekar8240@sdsu.edu

Kalpitha Narayana Moorthy
Red ID: 825921428
knarayanamoort7305@sdsu.edu

San Diego State University, California

Abstract

Parkinson's disease (PD) is a degenerative, progressive disease that causes a variety of motor and cognitive symptoms. Its diagnosis is difficult because the symptoms are close to those of other diseases such as normal aging and critical tremor. Patients suffering with Parkinson's disease, do not receive the best possible treatment, resulting in disability and rising societal costs. Although Parkinson's disease cannot be cured, early diagnosis and treatment can greatly improve symptoms and overall quality of life. This project provides an unbiased approach of detecting Parkinson's disease by applying machine learning using five different feature selection methods. We have used a dataset on Kaggle to detect Parkinson's disease in 195 individuals.

1. Introduction

The term parkinsonism refers to a symptom complex that includes resting tremor, bradykinesia, and muscle rigidity, and is used to characterize the motor features of Parkinson's disease. The number of people suffering from Parkinson's disease has risen rapidly around the world, especially in Asian developing countries. In community-based surveys of patients taking antiparkinsonian medication, diagnostic precision for the condition and other types of parkinsonism was investigated. Only 74 percent of patients had a diagnosis of parkinsonism, while 53 percent had clinically probable Parkinson's disease. It affects more than 1% of the population over the age of 60, with both motor and non-motor symptoms that worsen as the disease progresses. Since it cannot be healed, treatment methods concentrate on reducing the effects of Parkinson's disease. Early PD intervention can delay symptom progression and increase overall quality of life in the long run, according to research.

In this project, we have used the Parkinson's disease dataset obtained from Kaggle. The dataset consists of 195 individual's data and 23 characteristics to identify Parkinson's disease. We have used five different feature selection methods: Correlation Coefficient, Information Gain, Forward Feature Selection, Backward Feature Selection and Embedded method to determine relevant features useful to accurately predict the disease. We have applied various Machine Learning algorithms on a dataset to identify patients having Parkinson's disease. Depending on the performance of algorithms (accuracies of algorithms), we have identified the best feature selection methods and Machine Learning algorithm that results in identifying patients with Parkinson's disease accurately.

2. Workflow

The set of steps we followed while working on the project are as follows:

2.1 Analysis and Extraction

When we started working on the project to identify patients who have Parkinson's disease, we analyzed many datasets with various features. Among all the datasets that

we studied; we found a dataset with the range of biomedical voice measurements to be interesting to work on as that had various characteristics. A lot of research showed biomedical voice measurements to be important criteria to identify the disease. After careful evaluation, we decided to work on a dataset on Kaggle having a total 195 individual's data with 23 biomedical voice measurements. The dataset we referred to is: https://www.kaggle.com/wajidsaw/detection-of-parkinson-disease. The original dataset was from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/parkinsons. We read the data from the CSV file and used Python to perform analysis on the dataset.

2.2 Dataset Description

The dataset consists of 23 following biomedical voice measurements:

MDVP:F0(Hz):Average vocal fundamental frequency

MDVP:Fhi(Hz):Maximum vocal fundamental frequency

MDVP:Flo(Hz):Minimum vocal fundamental frequency

MDVP:Jitter(%):Multidimensional Voice Program (MDVP) jitter in percentage

MDVP:Jitter(Abs):Multidimensional Voice Program (MDVP) absolute jitter in ms

MDVP:RAP: Multidimensional Voice Program (MDVP) relative amplitude perturbation

MDVP:PPQ:Multidimensional Voice Program (MDVP) five-point period perturbation quotient

Jitter:DDP:Average absolute difference of differences between jitter cycles

MDVP:Shimmer: Multidimensional Voice Program (MDVP) local shimmer

MDVP:Shimmer(dB): Multidimensional Voice Program (MDVP) local shimmer in dB

Shimmer:APQ3: Three-point amplitude perturbation quotient

Shimmer:APQ5: Five-point amplitude perturbation quotient

MDVP:APQ11: Multidimensional Voice Program (MDVP) 11-point amplitude perturbation quotient

Shimmer:DDA:Average absolute differences between the amplitudes of consecutive periods

NHR:Noise-to-harmonics ratio

HNR:Harmonics-to-noise ratio

RPDE:Recurrence period density entropy measure

DFA:Signal fractal scaling exponent of detrended fluctuation analysis

Spread1:Two nonlinear measures of fundamental

Spread2:Frequency variation

D2:Correlation dimension

PPE:Pitch period entropy

Status:Detection of Parkinson's Disease

Initially, these terminologies were difficult to understand but as we progressed we became familiar with these features.

2.3 Importing libraries and loading the dataset

From loading the CSV file, plotting the visualizations, performing operations on datasets, choosing various feature selection techniques, implementing machine learning algorithms to designing models to calculate the accuracies of algorithms, we used different libraries which are: os, numpy, seaborn, matplotlib, pandas, scikit learn and many more functions and modules in these main libraries (scatter_matrix, GradientBoostingClassifier, LinearRegression, LogisticRegression, KneighborsClassifier, StandardScaler, SequentialFeatureSelector, Support Vector Classifier, DecisionTreeClassifier, Gaussian Naïve Bayes, GridSearchCV, AdaBoostClassifier, confusion_matrix and accuracy_score RandomForestClassifier, ExtraTreesClassifier,)

2.4 Descriptive Statistics

We tried to analyze the dataset first in order to identify the data values in it for various features. We identified data types for each feature and provided descriptive statistics to identify mean, minimum, maximum, percentiles, count and standard deviation.

2.5 Data Cleaning

We have used specific methods to clean our dataset. We checked if any of the rows have null values and whether they need to be dropped. However, the data was clean, and we were not required to make any column datatype or row changes for the data. Only for the purpose of ease, we made changes to start the row from index 1.

2.6 Visualization

For the obtained descriptive statistics, we have plotted two different types of visualizations. One is Histogram and the other one is Scatter plots for all columns to show the distribution of the data.

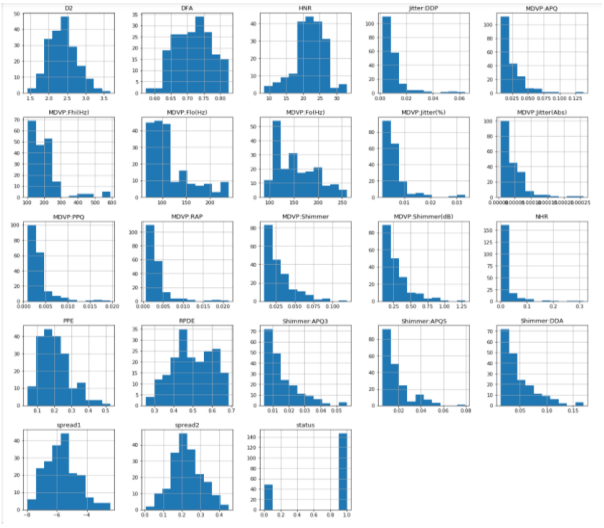

Figure 1. Histogram of Features

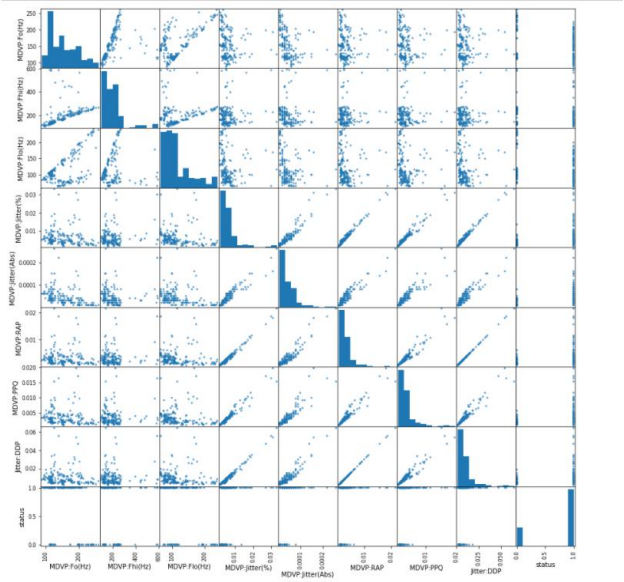

Figure 2. Scatterplot of Features

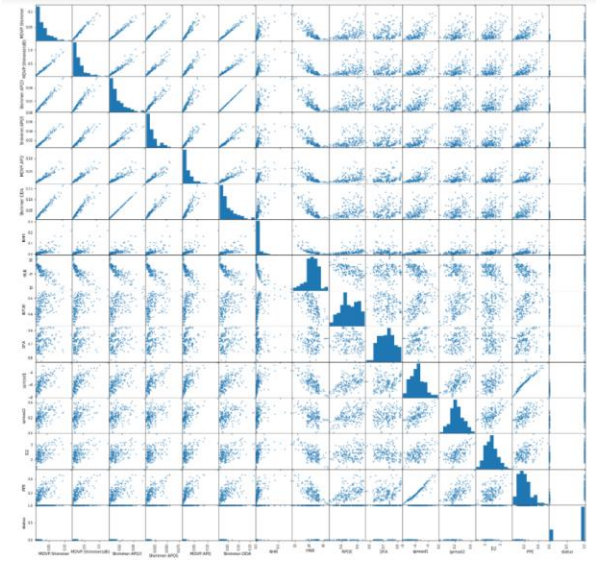

Figure 3. Scatter plot of Features

2.7 Splitting the dataset into features and result column

We split the dataset into features and result column. 22 Features were included, and the Patient name column was dropped among the features. In the result column, we have the status of the individual: 1 being the person having Parkinson's disease and 0 being a healthy person. Among these 22 features, we have selected few features based on specific feature selection methods.

2.8 Feature Selection Methods

Feature selection allows the machine learning algorithm to learn more quickly. It simplifies a model's complexity and makes it easier to understand. When the right subset is selected, a model's accuracy increases. We have used the following five feature selection methods:

2.8.1 Correlation Coefficient

High correlation features are more linearly dependent and therefore have almost the same effect on the dependent variable. When two features have a high correlation, one or all of them can be dropped. Considering the remaining features, we can find correlation of those with result column and generate training and testing datasets.

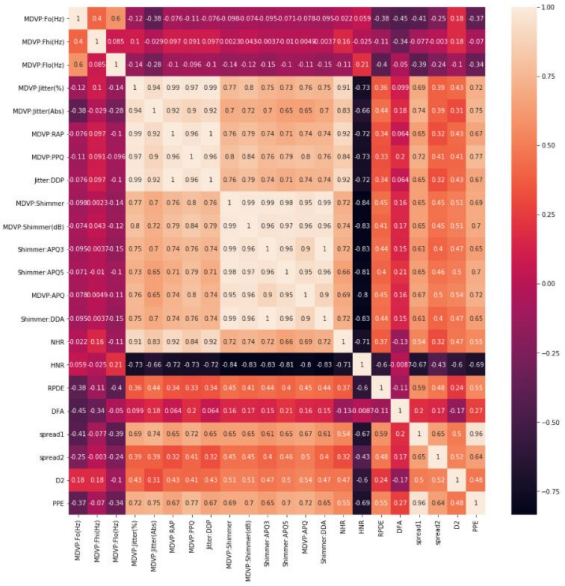

Figure 4. Correlation Matrix of Features

As the features MDVP:Shimmer(dB), NHR, Shimmer:APQ5, MDVP:APQ, Jitter:DDP, Shimmer:DDA, MDVP:RAP, MDVP:PPQ, Shimmer:APQ3, MDVP:Jitter(Abs) and PPE are highly correlated among themselves, we dropped them and used remaining features using this method.

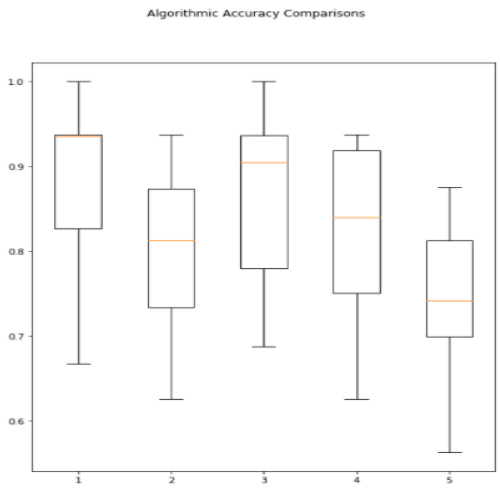

Figure 5. Boxplot for Correlation Coefficient

2.8.2 Information gain

By comparing the gain of each variable in the sense of the target variable, information gain can also be used for feature selection. The estimate is referred to as shared knowledge between the two random variables in this slightly different use.

Based on the information gain graph displayed, we chose columns having value greater than 0.10 as features: MDVP: Fo(HZ), MDVP: Fhi(HZ), MDVP: Flo(HZ), MDVP: Jitter(%), MDVP: Jitter(Abs), NHR, HNR, MDVP:Shimmer, DVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, spread1, spread2, D2, PPE

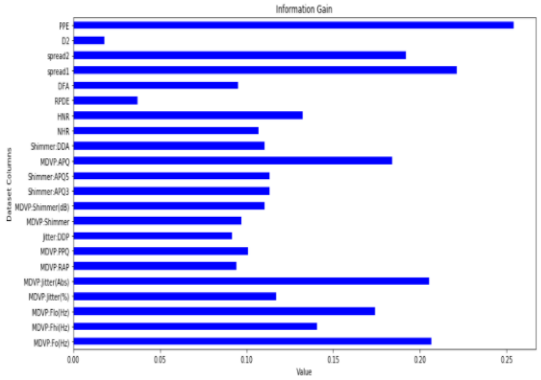

Figure 6. Information Gain Graph

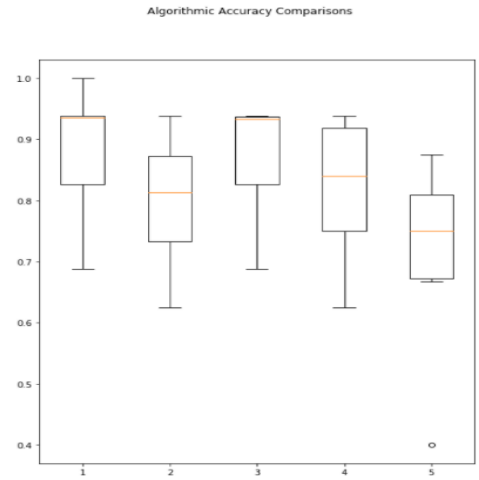

Figure 7. Boxplot for Information Gain

2.8.3 Forward Feature Selection

In a greedy manner, the Sequential Feature Selector adds (forward selection) or subtracts (backward selection) features to form a feature subset. Based on the cross-validation score of an estimator, this estimator chooses the best function to add or remove at each point. It usually involves two steps. First, the best single feature is selected. Next, using one of the remaining features and this best feature, pairs of features are created, and the best pair is chosen. We used logistic regression here. Based on the evaluation, the best 11 selected features are: MDVP:Fo(Hz), MDVP:Jitter(%), DVP:RAP, MDVP:PPQ, Jitter:DDP, NHR, HNR, RPDE, spread1, spread2, D2.

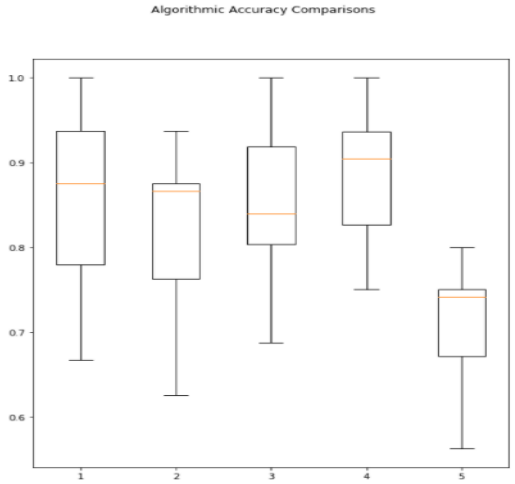

Figure 8. Boxplot for Forward Selection

2.8.4 Backward Feature Selection

Backward selection begins with all the dataset's features. It then runs a model and calculates a p-value for each function associated with the model's t-test or F-test. The function with the highest non-significant p-value will be removed from the model, and the process will begin all over again. We have opted for Logistic regression and Decision tree classifier to implement this. Based on the

evaluation, the best 11 selected features are: MDVP:Flo(Hz), MDVP:Jitter(%), PPE, DFA, MDVP:RAP, Jitter:DDP, DVP:Shimmer(dB), NHR, RPDE, spread1, spread2

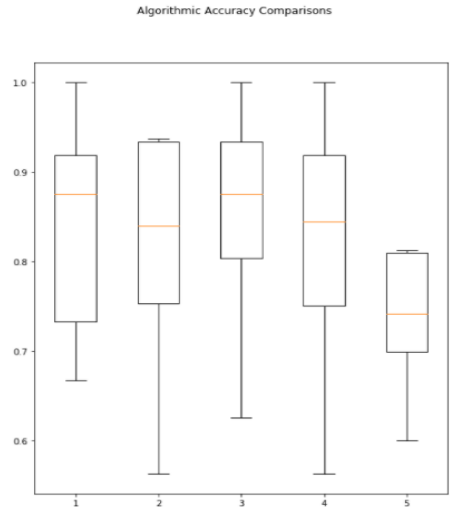

Figure 9. Boxplot for Backward Selection

2.8.5 Embedded method

A learning algorithm uses its own variable selection method to simultaneously perform feature selection and classification/regression. Filter and wrapper methods are combined in embedded methods. Algorithms with built-in feature selection methods are used to implement it.

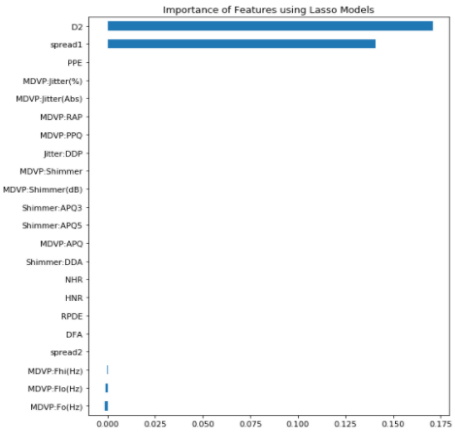

Figure 10. Importance of Features

LASSO and RIDGE regression are two common examples of these approaches, both of which have built-in penalization functions to minimize overfitting. To perform the Embedded method, we have used Logistic regression. Based on the evaluation, the best 5 selected features are: MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), spread1, D2

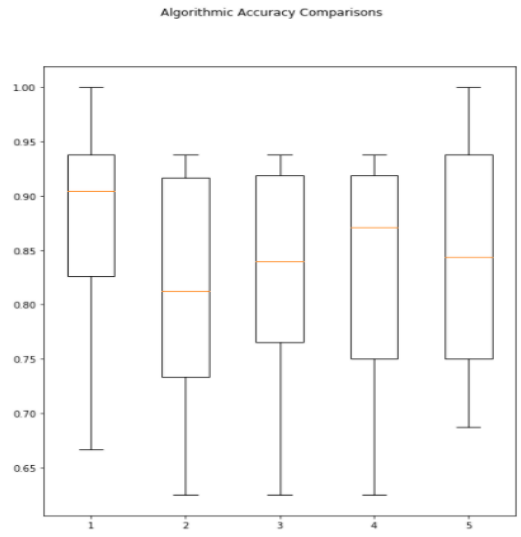

Figure 11. Boxplot for Embedded Selection

3. Algorithms

3.1 Logistic regression

Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable in its most simple form, though there are several more complex extensions. Logistic regression is a technique for estimating the parameters of a logistic model in regression analysis.

3.2 Support vector machine

Support vector machines are a class of supervised learning methods for classification, regression, and identification of outliers. Support Vector Machines have the following advantages: They are effective in high-

dimensional spaces. When the number of dimensions exceeds the number of samples, the method is still accurate.

3.3 Decision Tree

Decision Trees are a form of supervised machine learning in which the data is continuously split according to a parameter. Two entities, decision nodes and leaves, can be used to illustrate the tree.

3.4 K Nearest Neighbors

The K-Nearest Neighbors algorithm is based on the Supervised Learning methodology. The KNN algorithm assumes that the new case/data and existing cases are identical and places the new case in the category that is most similar to the existing categories.

3.5 Naive Bayes

A naive classifier model makes a prediction without any complexity, usually with a random or constant outcome. Such models are naive in that they make no predictions based on domain knowledge or learning.

4. Results

The accuracies of the algorithms based on various feature selection methods without feature scaling are as follows:

Feature Selection / Algorithm	Correlation Coefficient	Information Gain	Forward Feature Selection	Backward Feature Selection	Embedded Feature Selection
Logistic Regression	0.871667	0.878333	0.859583	0.840417	0.865417
Support Vector Machine	0.795417	0.795417	0.821667	0.815417	0.801667
Decision Tree	0.842083	0.866667	0.872500	0.841250	0.820417
K Nearest Neighbor	0.821667	0.821667	0.865833	0.821667	0.827917
Naïve Bayes	0.750000	0.722083	0.705000	0.742917	0.845417

The accuracies of the algorithms based on various feature selection methods with feature scaling are as follows:

Feature Selection / Algorithm	Correlation Coefficient	Information Gain	Forward Feature Selection	Backward Feature Selection	Embedded Feature Selection
Ada Boost Classifier	0.852917	0.853750	0.840417	0.834583	0.859167
Gradient Boosting Classifier	0.885417	0.905000	0.898333	0.904167	0.872500
Random Forest Classifier	0.891250	0.891667	0.865833	0.865417	0.878333
Extra Trees Classifier	0.929583	0.917083	0.904583	0.892083	0.878750

Based on these observations, Logistic Regression is the best algorithm without feature scaling and Extra Tree Classifier is the best algorithm with feature scaling.

5. Major Challenges

For selecting relevant features, various techniques are available apart from the techniques used in the project such as Exhaustive Feature Selection, Fisher's Score, Variance Threshold and Mean Absolute Difference. Choosing a proper feature selection method was a challenge in the project. Understanding factors that cause Parkinson's disease was a major challenge. Several other illnesses have similar symptoms; therefore misdiagnosis may occur.

6. Related Work

Diverse research is going on for detection of Parkinson's disease including:

- Using image processing (Single Photon Emission Tomography) and artificial neural network (ANN)
- Using Deep Neural Networks on rapid eye movement, Cerebrospinal fluid data, and dopaminergic imaging markers

- Using the genetic algorithm and SVM classifier for early detection

7. Conclusion

The main aim of this project is to detect Parkinson's disease using machine learning algorithms. To achieve accurate detection, we have used five different strategies. Also, we have examined the results using testing dataset of Parkinson's patients dataset. Detecting Parkinson's disease using biomedical voice measurement includes selecting relevant features as an important criterion for an algorithm to provide better accuracy.

8. References

[1]
https://scikit-learn.org/stable
https://jnnp.bmj.com/content/79/4/368
[2]
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517533/
[3]
https://www.sciencedirect.com/science/article/abs/pii/S0306987719314148
[4]
https://www.frontiersin.org/articles/10.3389/fict.2019.00010/full
[5]
https://ieeexplore.ieee.org/document/9087433
https://ieeexplore.ieee.org/abstract/document/8615607
[6]
https://ieeexplore.ieee.org/document/8284216
https://en.wikipedia.org/wiki/Machine_learning