PARKINSONS DISEASE PREDICTION

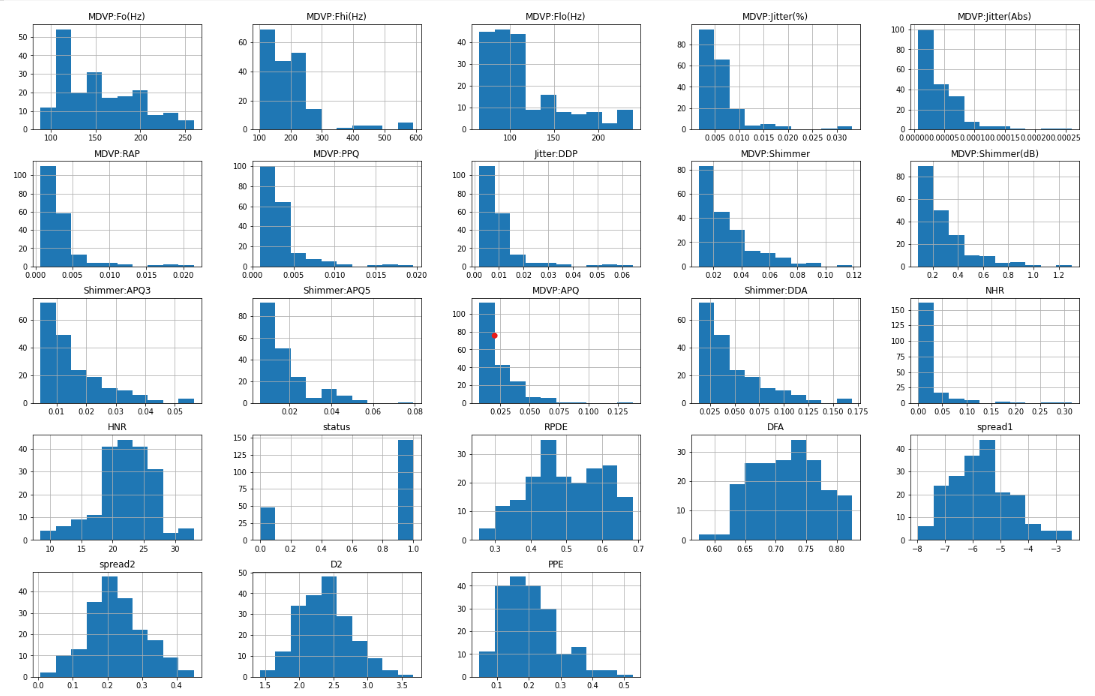
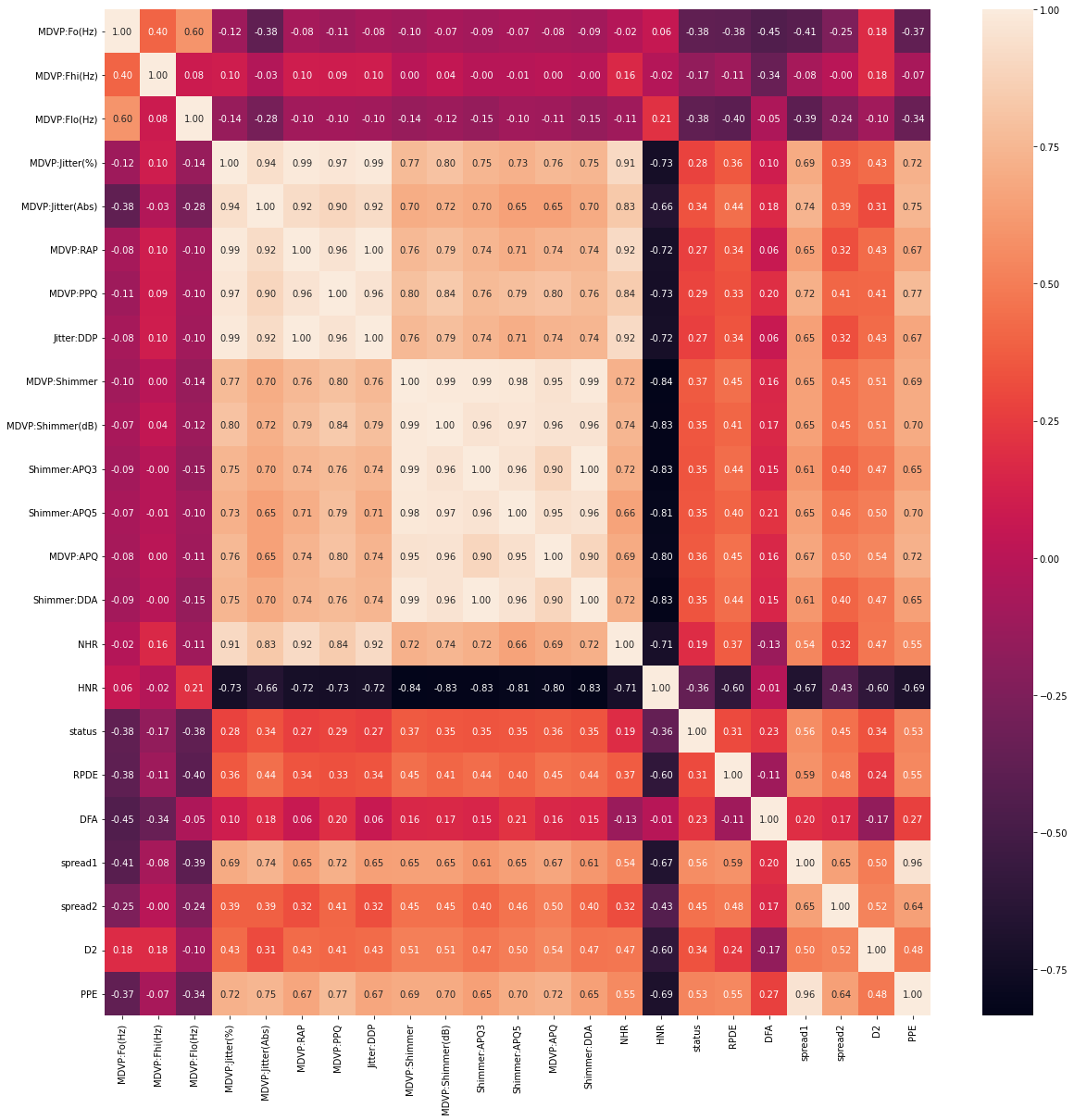
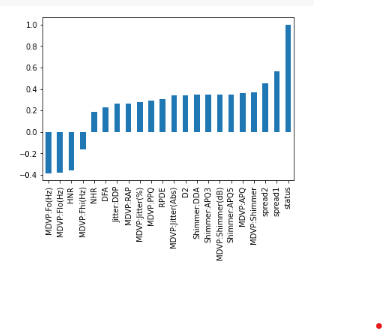
Parkinsons Disease is a nervous disorder which usually occurs in people, who are old and do not lead a healthy lifestyle. Some of the symptoms of Parkinsons disease include Tremors, inability to stand, etc.

**DATASET:**

The dataset regarding the details of individuals who might or might not have parkinsons disease is downloaded from kaggle repository. The dataset consists of 195 rows and 24 features, among which “status” is the dependent variable consisting of 2 labels; 1-The particular individual has Parkinsons disease, 0-The particular individual does not have parkinsons disease. All the other features are taken to be as independent features and the “name” column is dropped since; it is irrelevant to the analysis.

**ANALYSIS:**

Basic exploration of the dataset is done, and no missing values, no duplicate values are observed. Multiple histograms are plotted to graphically portray the numerical data distribution. Correlation of each independent variable with regards with to the dependent variable is also found using the corr() function.Correlation heatmaps are used to further understand the correlation between each independent variable and the dependent variable.



**ANALYSIS:**

The dependent variable(status) is found to have 147 – ‘1’ and 48 – ‘0’. The MinMaxScaler is used to scale the independent features and normalize them between the specified range of (-1,1). The splitting of dataset into train and test split is done in the ratio--80/20. And K-fold cross validation is done with 5 splits in order to compare the performance of K-fold and train and test split. Model building is the essential part of the data science project. In order to accurately predict if a particular individual has parksinsons disease or not, 7 well known and popular supervised machine learning algorithms are built,

1. Extreme Gradient Boosting
2. Adaptive Boosting
3. Logistic Regression
4. Naïve Bayes
5. Decision Tree
6. Support Vector Machine
7. K-Nearest neighbor

**Extreme Gradient Boosting:**

XGboost is a popular machine learning algorithm which can be used in both regression and classification problems. It works on the basis of sequential processing, where a weak learner’s mistakes are reduced until it becomes a strong learner. XGboost is also referred to as the “queen of machine learning algorithms.”

**ADAPTIVE Boosting:**

Adaboost falls under the same boosting technique, where we take a collection of weak learners and transform them into strong learners. Their main objective would be to identify the mistakes made by the previous model and rectifying them.

**Logistic Regression:**

Logistic regression is a very commonly used machine learning algorithm in the data science community. Logistic regression performs best when the predictor variable is a categorial variable and comes under a binary classification problem. It uses a S-shaped curve called sigmoid.

**Naïve Bayes:**

Naïve bayes is a supervised ml technique which works on the application of Bayes theorem. It is known for providing equal weightage and independence to all the features so that all the variables will equally contribute to predicting the dependent variable.

**Decision Tree:**

Decision tree is a popular machine learning algorithm, which is used for both classification and regression approaches. It is a tree like structure which makes decisions based on some questions derived from the class labels. Even though it can be used for both regression and classification problems, it is mostly used for classification problems.

**Support Vector Machine:**

Support vector machine is a supervised machine learning algorithm which can be used for both classification and regression problems. The main aim of SVM is to find the most optimal hyperplane which ideally splits the dataset into classes, which makes it easy to predict the dependent variable. It is also well known for its kernel trick for handling non-linear spaces.

**K-Nearest neighbors:**

K-nearest neighbors is a supervised machine learning algorithm known providing higher model accuracies. It is also known as a lazy learner algorithm. It is used for both classification and regression approaches.

Performance of Machine Learning models, when train-test split is used:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification Algorithms** | **Accuracy** | **Precision** | **Recall** |
| XGBOOST | 92.3% | 0.93 | 0.92 |
| ADABOOST | 92.3% | 0.96 | 0.92 |
| Support Vector Machine | 89.74% | 0.92 | 0.90 |
| Logistic Regression | 87.17% | 0.88 | 0.87 |
| Decision Tree Classifier | 87.17% | 0.87 | 0.87 |
| Naïve Bayes | 71.79% | 0.72 | 0.72 |
| K Nearest Neighbors | 94.8% | 0.95 | 0.95 |

Performance of Machine Learning models, when K-fold cross validation is used:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification Algorithms** | **Cross\_val score** | **Precision** | **Recall** |
| XGBOOST | 82.5% | 0.83 | 0.82 |
| ADABOOST | 79.4% | 0.79 | 0.81 |
| Support Vector Machine | 82.5% | 0.83 | 0.83 |
| Logistic Regression | 81.5% | 0.80 | 0.82 |
| Decision Tree Classifier | 77.4% | 0.76 | 0.78 |
| Naïve Bayes | 68.2% | 0.68 | 0.68 |
| K Nearest Neighbors | 64.6% | 0.64 | 0.64 |

**Conclusion:**

Therefore, we can conclude by saying that XGboost and adaboost performed the best when the train/test split with the ratio 80/20 was carried out. And models XGboost and Support vector machine performed the best, when K-fold cross validation was done. We can come to a conclusion that the above mentioned models are well and good and competent enough to be used to predict, if a particular individual has parkinsons disease or not.