PARKINSON'S DISEASE PREDICTION USING SUPERVISED LEARNING METHODS

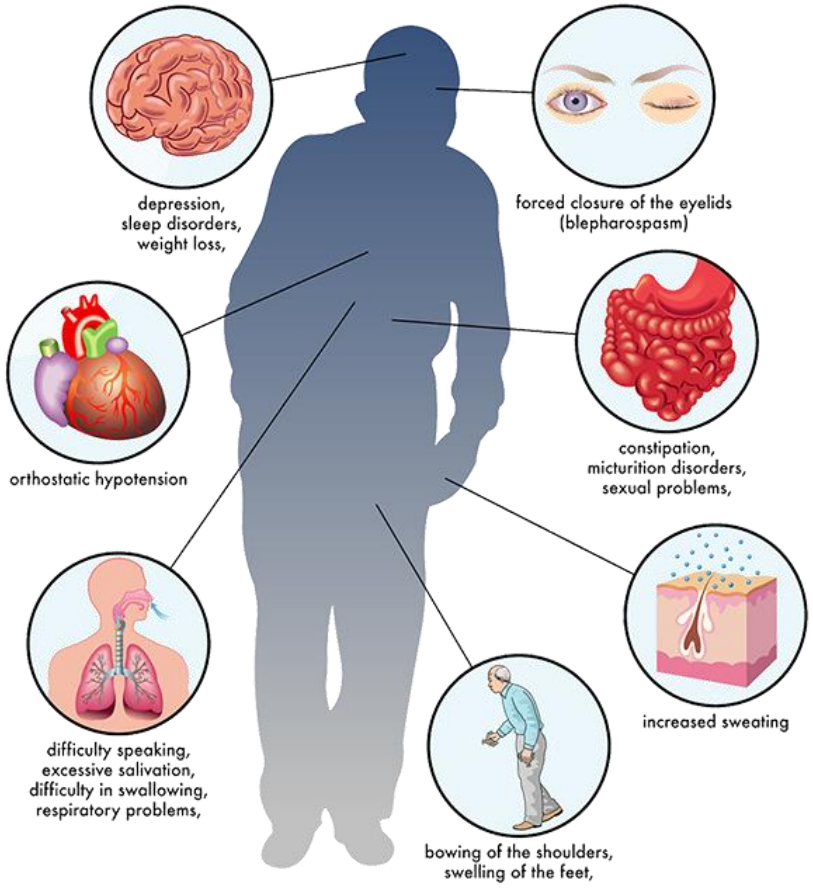
BY: ASMAA ALI (1910069)

PROBLEM STATEMENT

The aim of the project is to examine several classification techniques to classify the status of people to Health (1) or Parkinson's (0). We do this by applying supervised learning methods for Parkinson's disease prediction by interpreting the voice measures of the voice recordings of each individual saying only a single term.

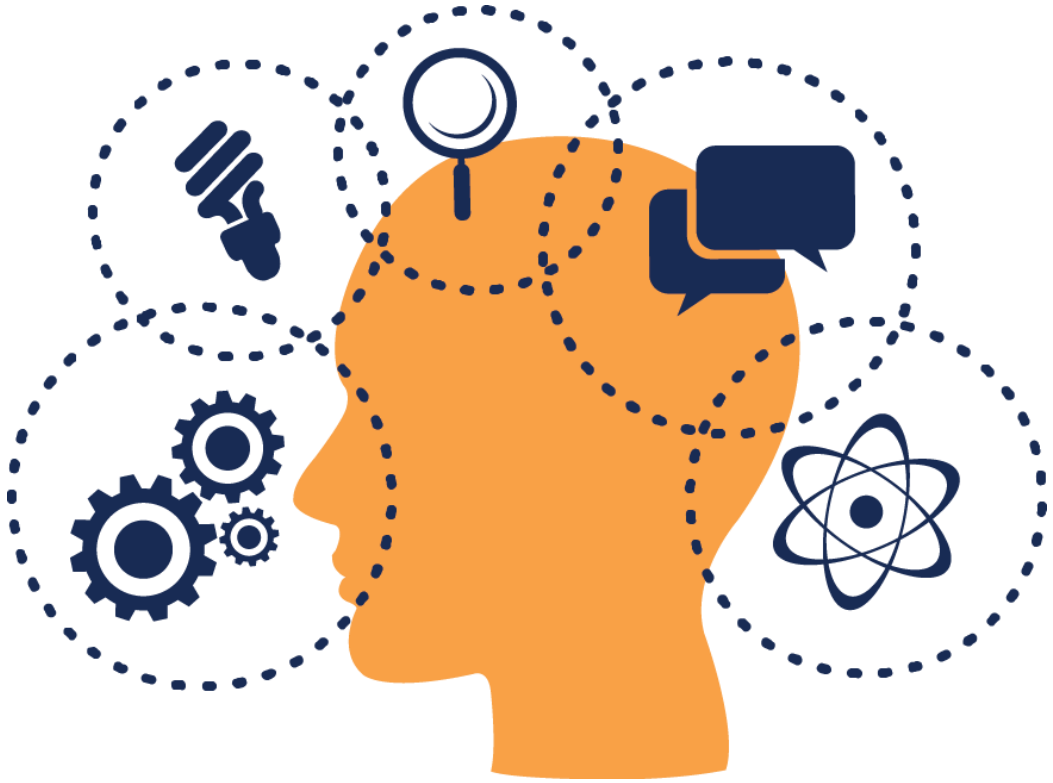
MOTIVATION

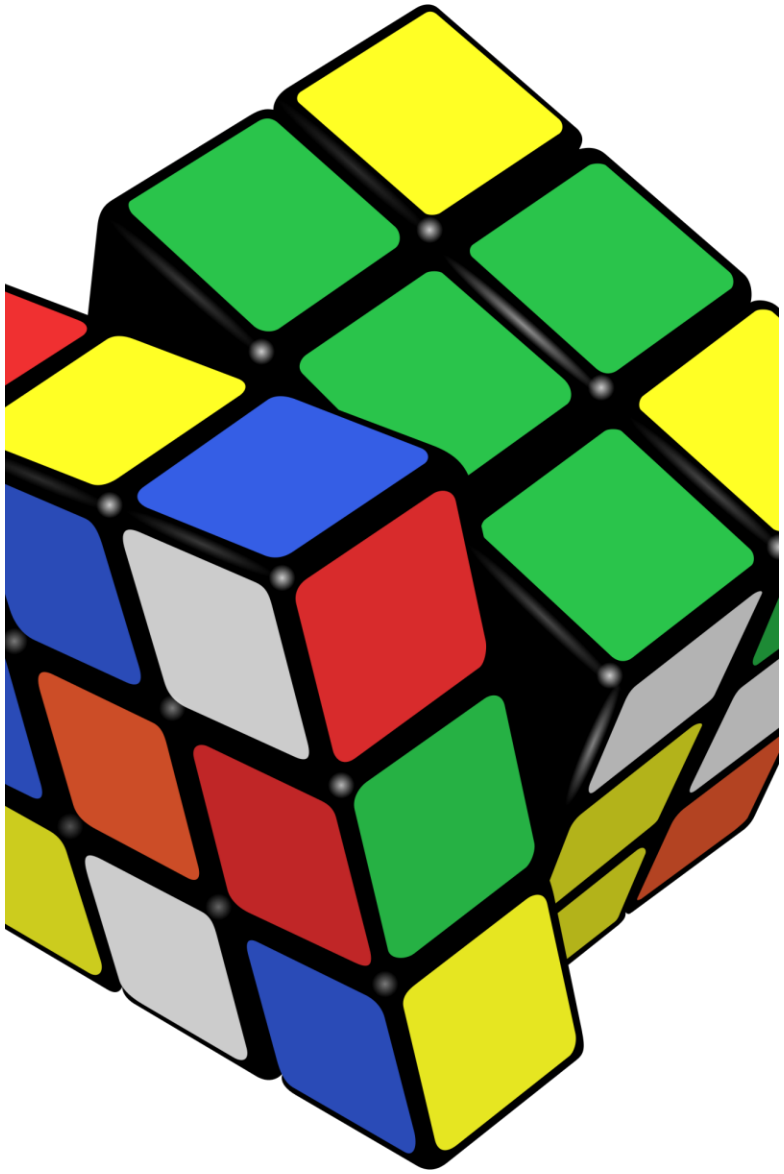
Parkinson's disease is the second most common age-related neurodegenerative disorder after Alzheimer's disease. An estimated seven to 10 million people worldwide have Parkinson's disease. Parkinson's disease is a progressive disorder of the central nervous system affecting movement and inducing tremors and stiffness. It has 5 stages to it. Early detection and diagnosis is important because the treatments for PD are more effective in the early stages of the disease. In addition, physical therapy and exercise, which greatly improve symptoms and delay progression of the disease, are much easier to perform in the early stages.



THE SOLUTION

1. Data Exploration and Visualization
2. Data Preprocessing
3. Evaluating and Tuning Linear Classifiers
4. Evaluating and Tuning Ensemble Methods
5. Model Selection and Finalization
6. Conclusions



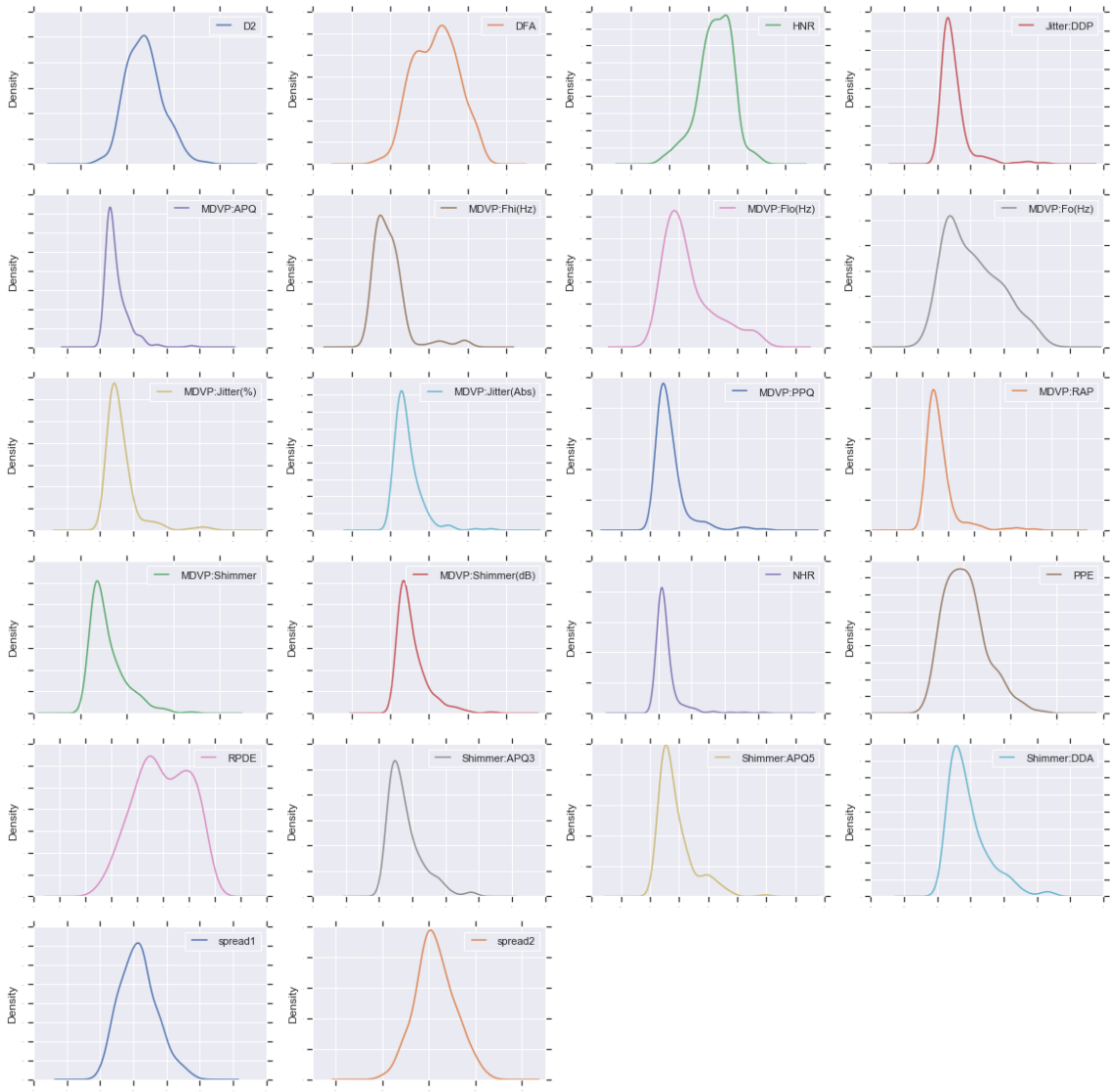


STRUCTURE OF DATA

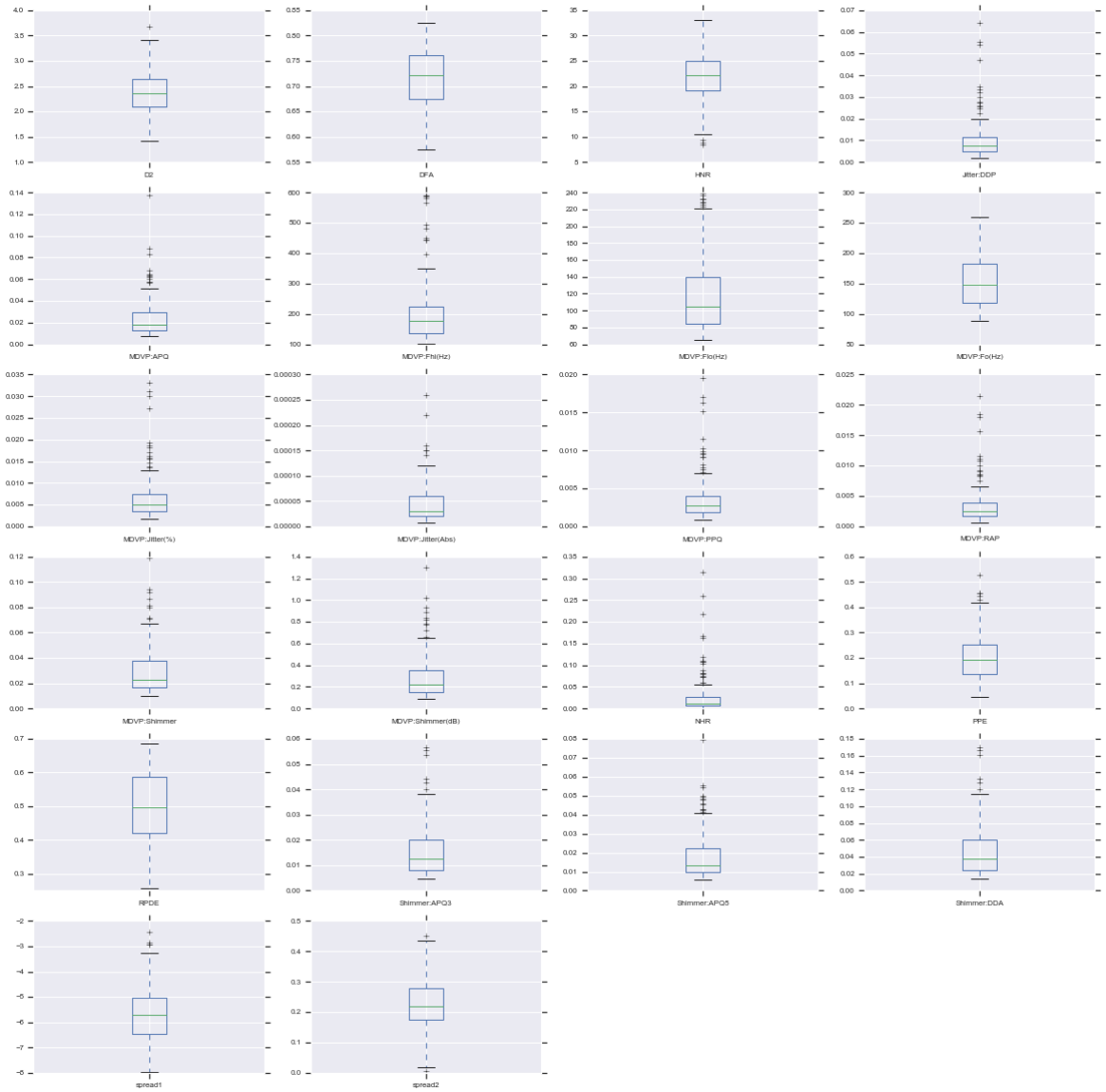
This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals. There are around six recordings per patient.

The base features in this data:

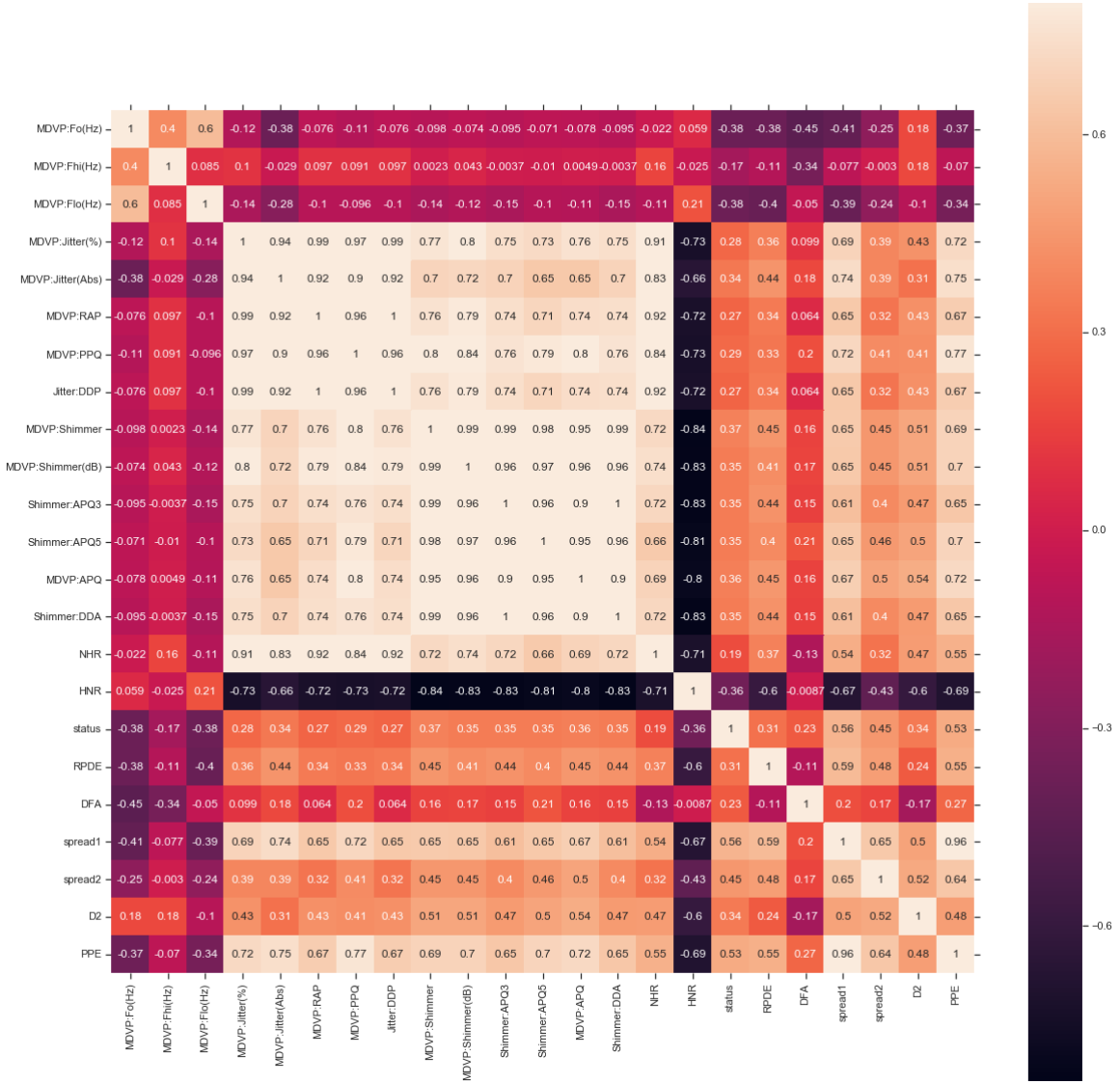
Name, MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, PPE



DATA VISUALIZATION



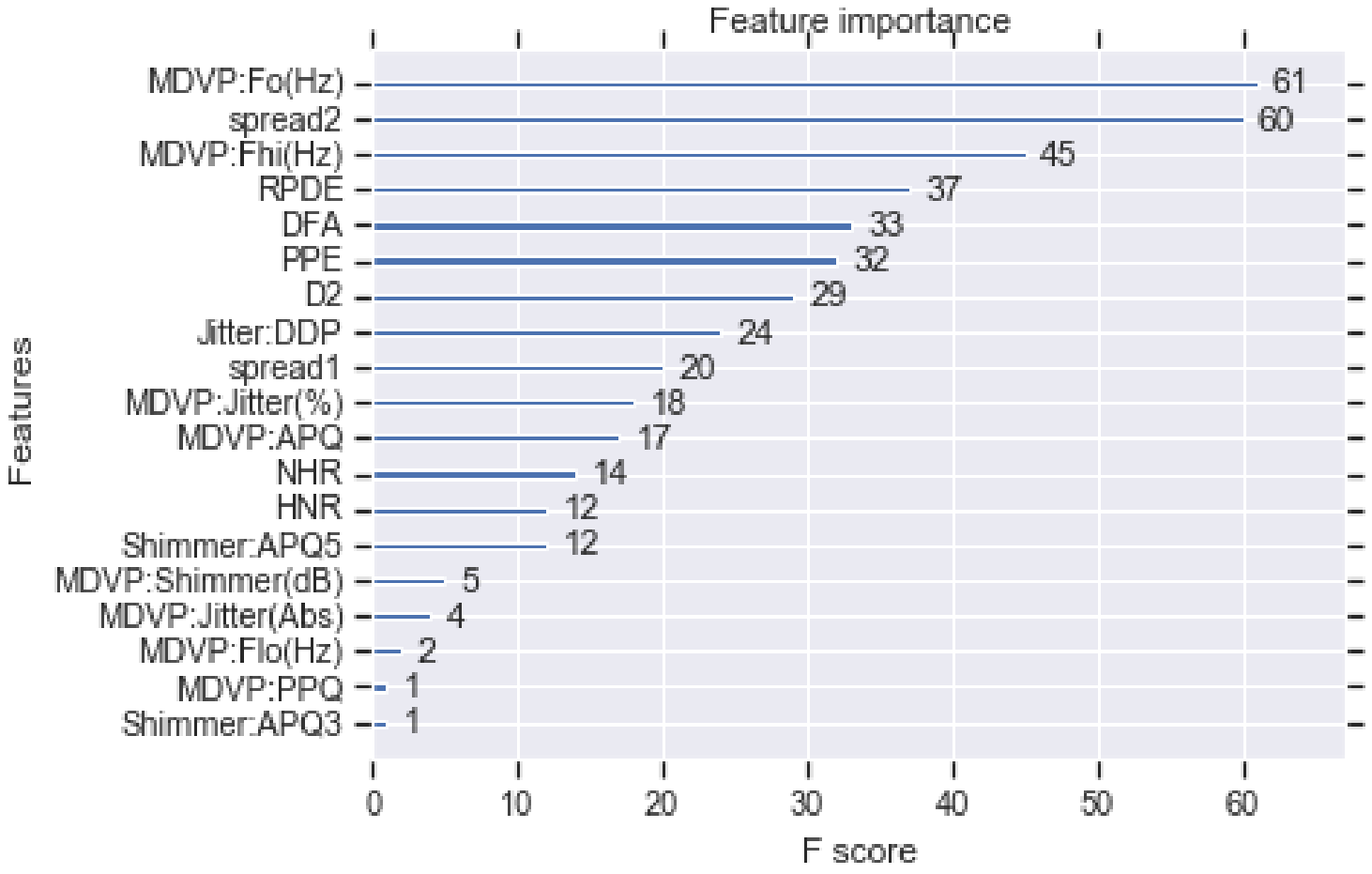
DATA VISUALIZATION

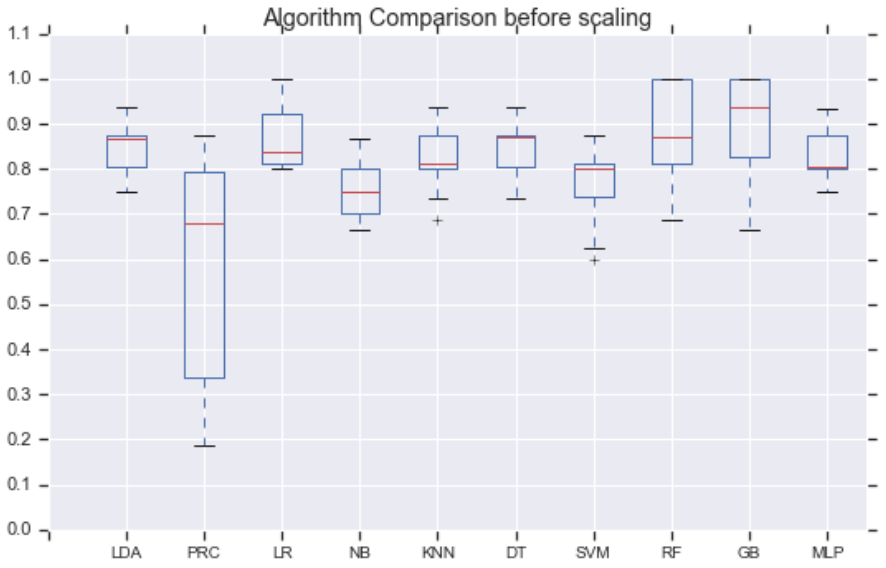


DATA VISUALIZATION

FEATURE SELECTION

- Removing the highly correlated features.
- Feature Importance and Feature Selection With XGBoost

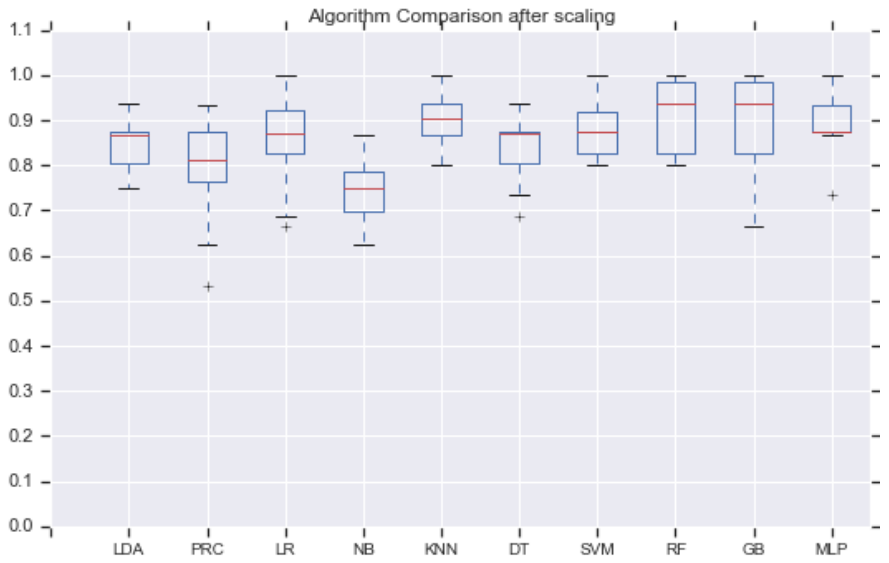




Algorithm	Accuracy before scaling	Accuracy after scaling
LDA	0.85	0.85
PRC	0.59	0.79
LR	0.87	0.86
NB	0.76	0.74
KNN	0.83	0.90
DT	0.85	0.84
SVM	0.76	0.88
RF	0.88	0.91
GB	0.90	0.89
MLP	0.83	0.90

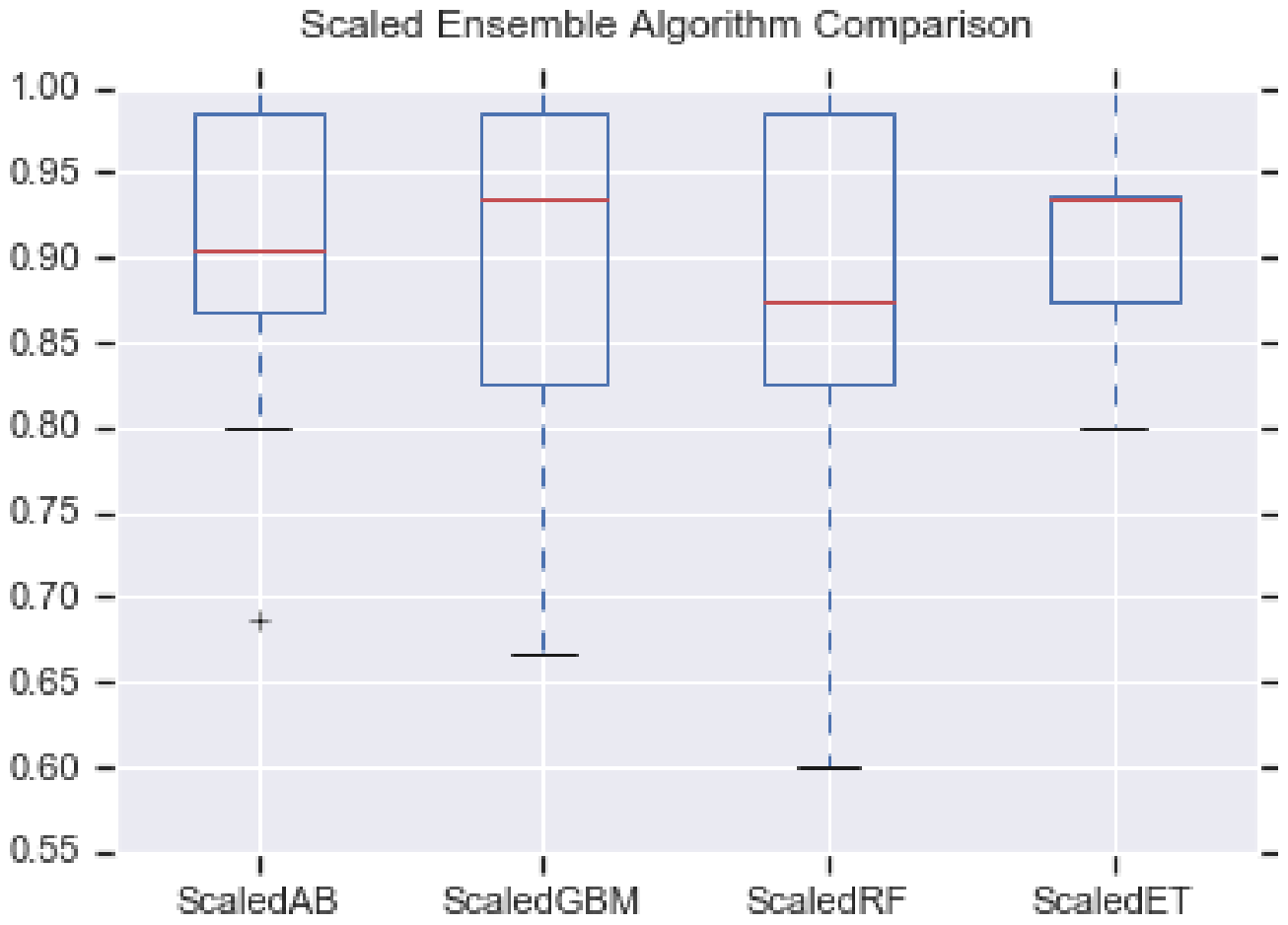
EVALUATING AND TUNING LINEAR CLASSIFIERS

- KNN Best: 0.974359 using {'n_neighbors': 1}
- GBM Best: 0.903846 using {'learning_rate': 0.15, 'n_estimators': 200}

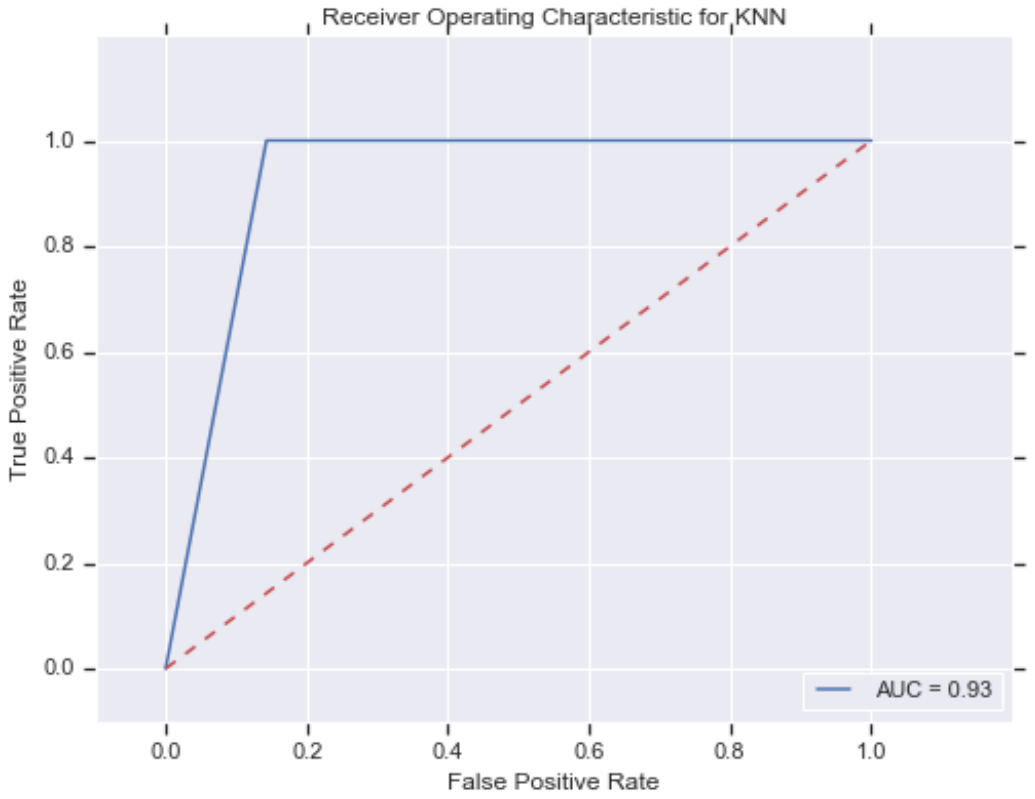


EVALUATING AND TUNING ENSMBLE METHODS

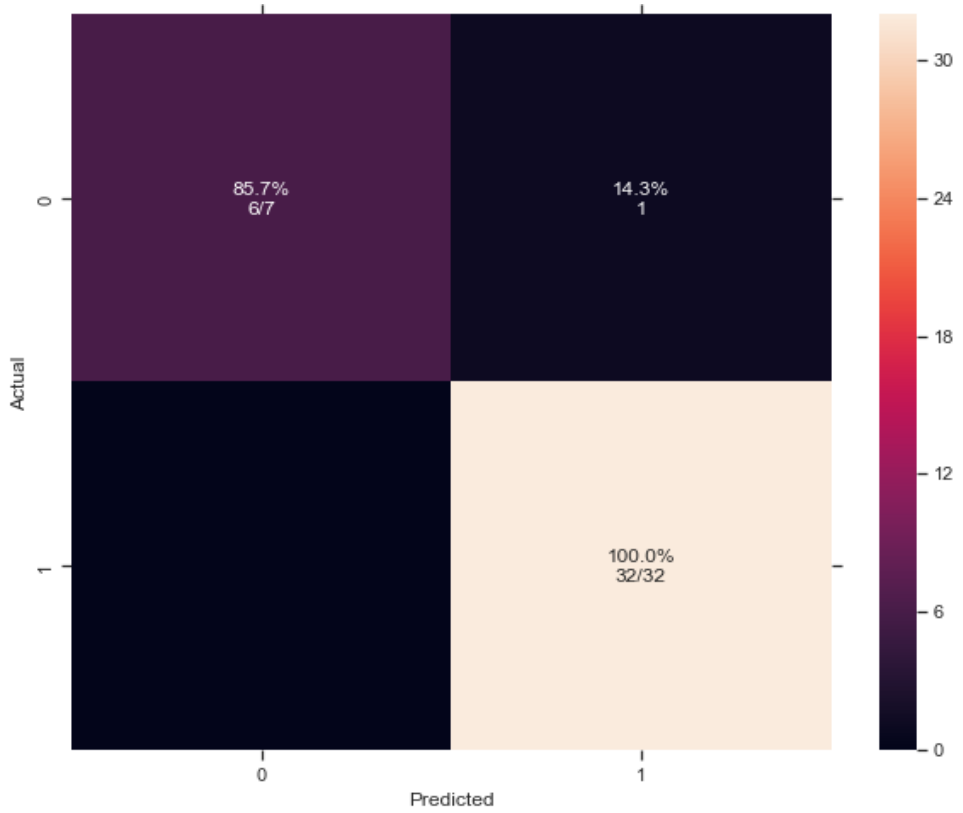
```
Best: 0.942308 using {'n_estimators': 100}

0.935897  with: {'n_estimators': 50}
0.942308  with: {'n_estimators': 100}
0.942308  with: {'n_estimators': 150}
0.935897  with: {'n_estimators': 200}
0.935897  with: {'n_estimators': 250}
0.935897  with: {'n_estimators': 300}
0.942308  with: {'n_estimators': 350}
0.935897  with: {'n_estimators': 400}
```



MODEL SELECTION AND FINALIZATION (KNN)



	precision	recall	f1-score	support
0	1.00	0.86	0.92	7
1	0.97	1.00	0.98	32
accuracy			0.97	39
macro avg	0.98	0.93	0.95	39
weighted avg	0.98	0.97	0.97	39



CONCLUSIONS

- Data Analysis (some skewed distributions and correlated attributes).
- Evaluate Algorithms (Gradient Boosting Classifier looked good).
- Evaluate Algorithms with Standardization (KNN and Gradient Boosting Classifier looked good).
- Algorithm Tuning (K=1 for KNN was best).
- Ensemble Methods (Extra Trees Classifier looked good).
- Tuning Ensemble Methods (getting the most from ExtraTreesClassifier).
- Finalize Model (use all training data and confirm using validation dataset on the selected algorithm (KNN)).

THANK YOU