Gender Recognition by Voice and Speech Analysis using Machine Learning

With increasing application of machine learning in nearly all day to day usability, voice analysis using ML is inevitably one of the most required one. Voice based security systems, recommendation systems, voice based devices are only a few straight use cases. Therefore, I am trying to design a model which can identify the important acoustic properties for a voice clip. Using the model, I should be able to predict and classify the voice to two genders i.e. male or female.

Problem: Find the best model which can predict the gender based on different acoustic properties of voice.

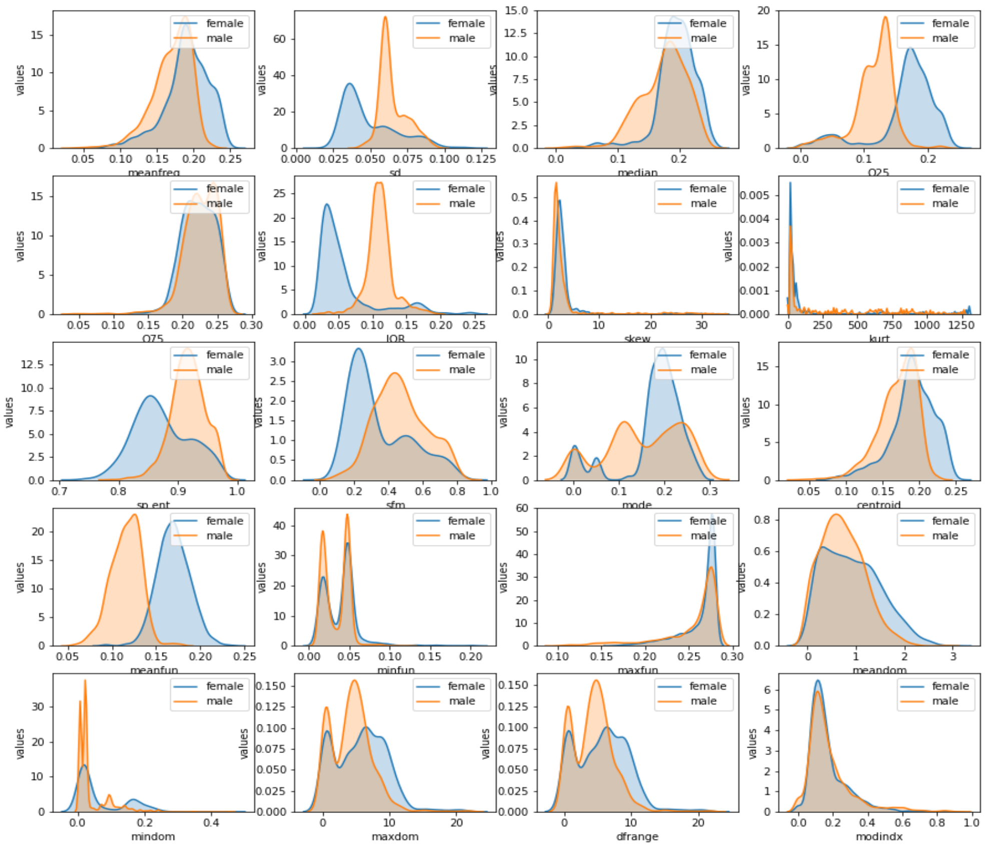
Data: The data set consists of **3168 samples**, generated on different acoustic properties of voice. All samples are mapped to the gender of the voice. There are 21 columns (20 columns for each feature and one label column for the classification of male or female). It was generated by converting sound waves for both the genders using the warbleR package in R. The data set is equally distributed for both male and female.

The data is available on Kaggle and can be downloaded from here. The available dataset is clean with no missing values.

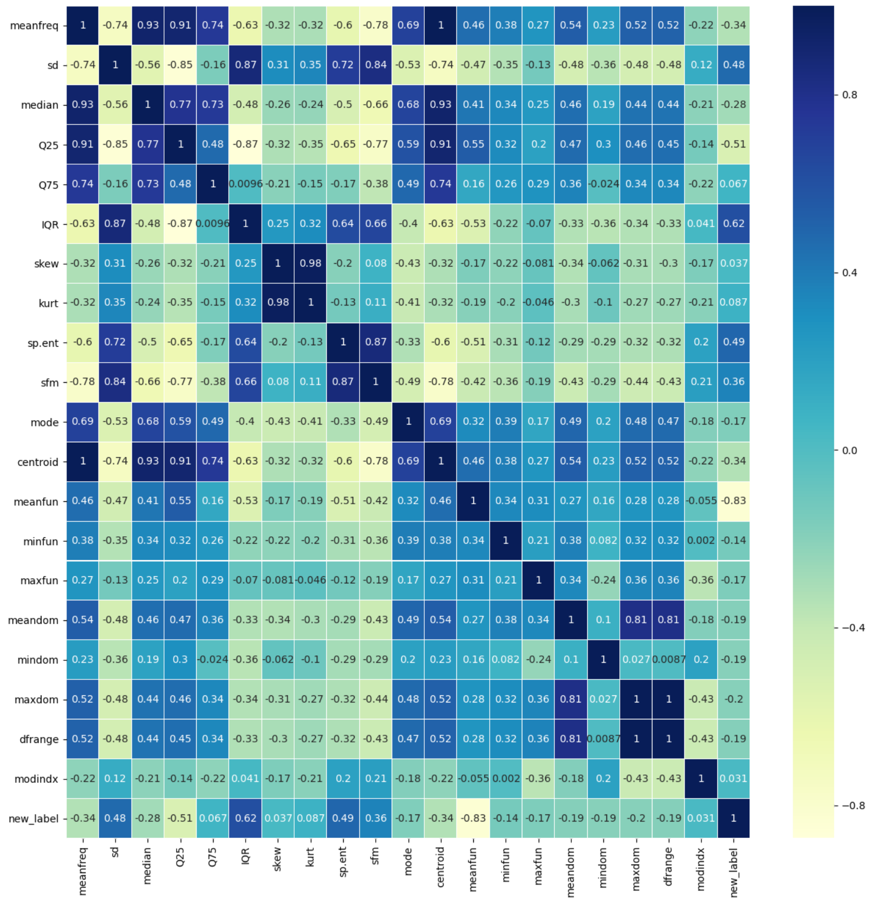
Exploration and Visualisation of data

EDA Notebook

Since all the columns are statistical computation of the frequency of the voice, there is a high correlation among most of the columns. Below **density plots** help to understand the distribution of different columns for the male and female genders.



Columns as ‘IQR’, ‘Q25’, ‘meanfun’, ‘sd’, ‘sp.ent’ are giving fairly differentiating male and female for column gender.
To know the dependencies of different properties of voice on gender, the **heatmap** was drawn.

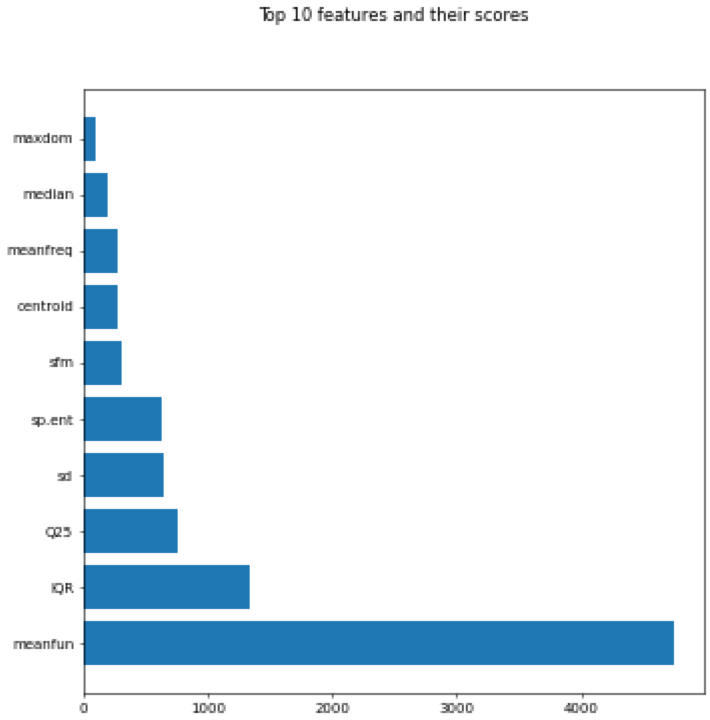


The heatmap shows that there is a **high correlation** in the data and multicollinearity can be a problem while analysing the data.

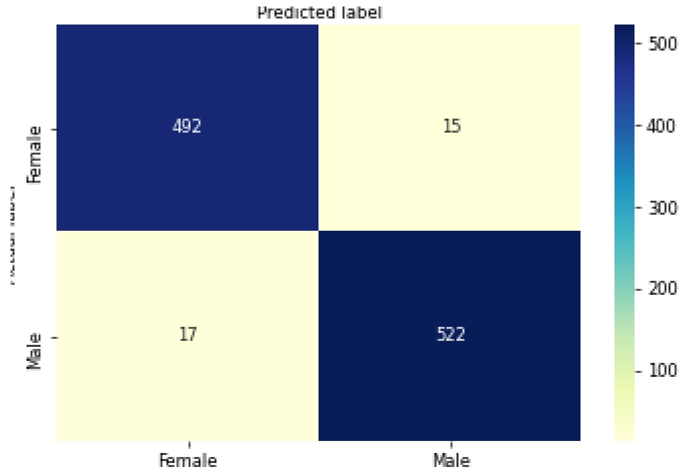
Method: To classify the data four approaches were investigated. Logistic Regression, Decision Tree, Random Forest and Support Vector Machine.

Modeling Modeling Notebook

Logistic Regression: I have applied the Logistic Regression model on the data. Since there are so many variables with high correlation as seen in the heatmap, I choose **5** variables which have high correlation with column gender and not significant correlation with each other. Variables are selected using SelectKBest method from scikit learn.The score of top 10 features is as follows.



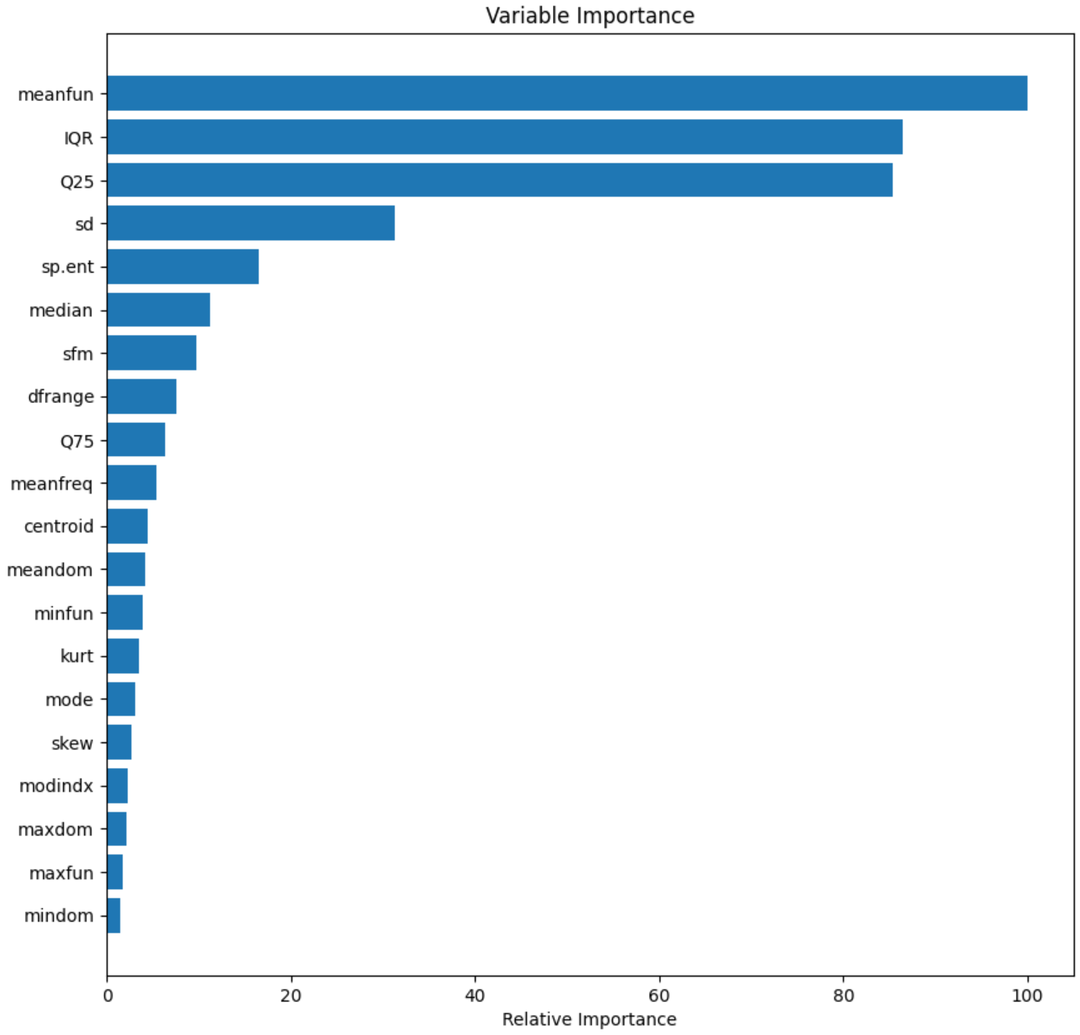
The grid search was done to find the best hyperparameters and the model was cross validated. The accuracy obtained from grid search is 96.94%. The classification metrics from the model is as follows



Decision Tree: The basic decision tree is fitted on the data with criterion "gini", max_depth 11 and random_state 1. The accuracy on test data is 96.56% and **accuracy on train data is 100%**. As 100% accuracy shows that the model was fitted too well on the train data that it's going to be a biased model for train data and could be less effective in predicting the new validation or test data sets in future.

To overcome the biases in the model, I have tuned the hyperparameters of the model using grid search. The best criteria came out to be 'Entropy' with max depth of the tree 5. Model **accuracy on train data using criteria Entropy and max depth of the tree 5 is 98.92%.. And on test data it is 96.85%** . With these parameters the model looks less biased and gives better accuracy than the basic model.

Random Forest: The basic random forest model gives the accuracy of 98.28%. To improve the model, I have explored the important features in the data. The feature importance chart is as follows:



Since only a few variables have high importance in the model. I conducted the grid search to tune the hyperparameters. The hyperparameters from grid search are as follows,

Parameter	Value
max_features	3
min_sample leaf	3
n_estimator	200

The accuracy on test data using these parameters is 97.90%.

Support Vector Machine: The basic SVM model has applied on the data with default parameters. The accuracy from the basic model is 97.90%. As so many factors can affect the SVM model, I tried to run the grid search and the accuracy after tuning the hyperparameters has reached to 98.18.

Model Evaluation Metrics/ Results:

Scores	Logistic Regression	Decision Tree	Random Forest	Support Vector Machine
Accuracy	96.94	96.85	97.90	98.18
Recall	96.85	96.29	97.59	98.14
Precision	97.21	97.56	98.32	98.33
F1 Score	96.94	96.85	97.90	98.18

Conclusion: All above applied methods Logistic Regression, Decision Tree, Random Forest and SVM have shown high accuracy which is >96% on this data but SVM model turns out to be the best model. The accuracy of the SVM model is highest which is 98.18%. Along with the highest F1 Score.

Future Exploration: I would like to extend the model to a recommendation system for recommending answers based on the questions asked by either male or female.