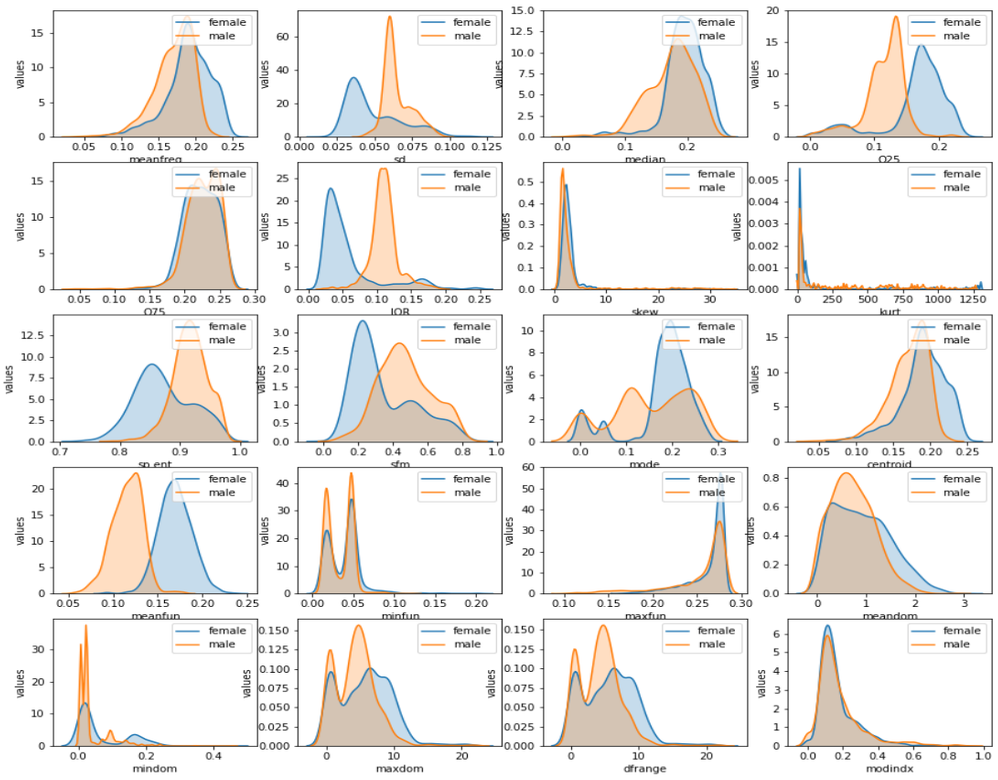
Gender Recognition by Voice and Speech Analysis using Machine Learning

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.

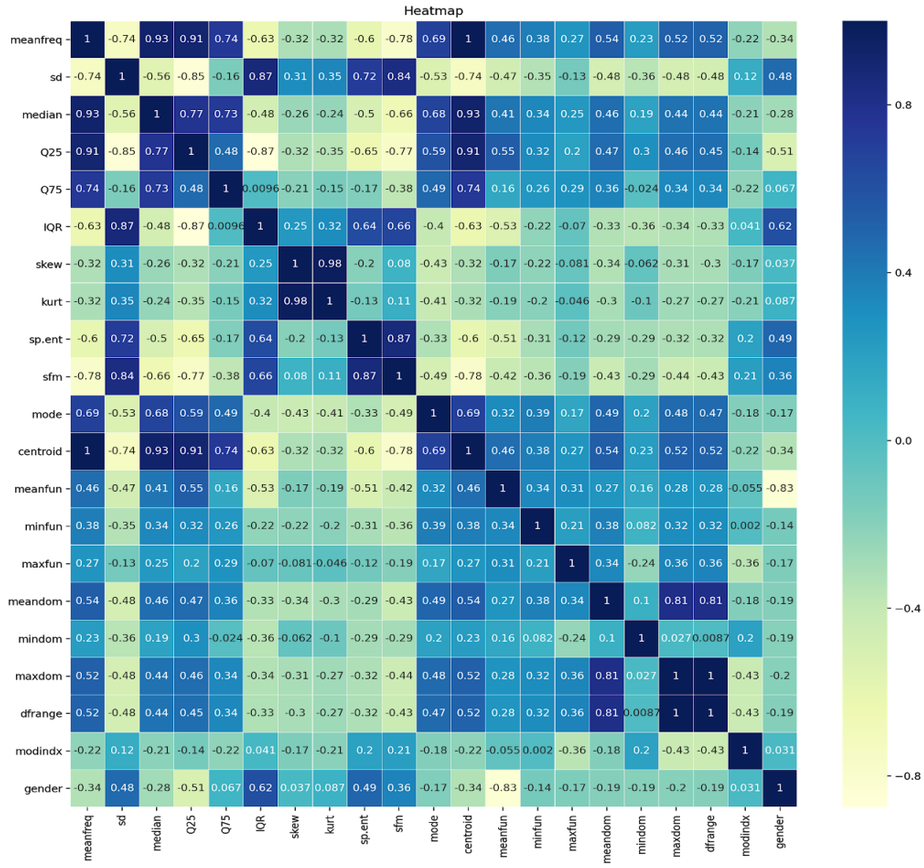
Problem: Apply Logistic Regression model which can predict the gender based on different acoustic properties of voice with high accuracy.

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 fair idea of the distribution for both the genders. To know the dependencies of different properties of voice on gender, the heatmap was drawn.



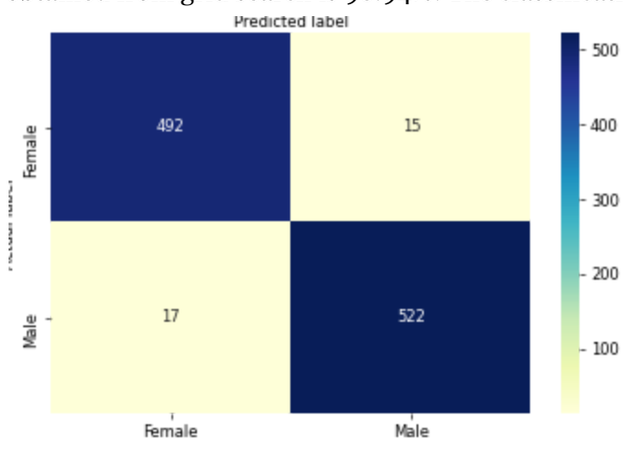
The heatmap shows that there is a high correlation in the data and multicollinearity can be a problem while analysing the data.

Method: To classify the data Logistic Regression model is fitted on the training data set and cross validated. The model is tested on a testing data set.

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



Model metrics/ Results:

Scores	Logistic Regression
Accuracy	96.94
Recall	96.85
Precision	97.21
F1 Score	96.94

Conclusion: The Logistic Regression model gave the **Accuracy of 96.94% with F1 Score 96.94%**.

Future Exploration: I would like to extend the project to apply different classification models on the data for example Decision Tree, Random Forest and Support Vector Machine to find the best model for this problem.