Gender Classification By Voice Using Machine Learning

Kashika Puri
Department of Electrical and Computer Engineering
Queen's University
Kingston, Canada
17kp35@queensu.ca

***Abstract*— Gender classification is one of the major problems in field of speech analysis now-a-days. Identification of gender from acoustic properties of voice i.e. mean, median, frequency etc. is the highly important. Machine learning is used to solve this problem because it gives promising results for classification techniques. There are several algorithms that can be used to predict the gender using acoustic properties. In my project, I am evaluating classifiers using 6 different machine learning algorithms. These algorithms include K-Nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), Support Vector Machine Using Poly kernel and Gradient Boosting (GB). The main parameter involved is the accuracy obtained using all these classifiers. I am trying to assess the accuracy, recall, precision and F1 Score obtained after predicting on test data for all these classifiers and finally the best fit model will be generated for gender classification of acoustic data.**

Keywords—Machine learning, Gender classification, Acoustic attributes

I. INTRODUCTION

Speech signal is the most common means of communication for human beings. Dimorphism is the property of voice that is highly observed in human beings. Intonation, speech rate, and duration are certain characteristics that distinguish human voices, mainly male and female voices [1]. The task to identify a human's gender by voice is very easy when a human identifies it. On the other hand, it becomes difficult for a computer to identify whether the voice is of male or female. A human has instinct to identify the difference between the voice of male and female but when it comes to computer we need to train and make it learn by providing training data and various algorithms.

The aim of this project is to create a classifier which identifies the gender based on acoustic attributes of voice. The data set split into train and test data and model is build using train data. This model is trained using various machine learning algorithms which includes K-Nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), Support Vector machine Using Poly kernel and Gradient Boosting (GB). The confusion matrices are created and results of different models are compared with each other. The model with the highest accuracy is considered as the best model. There are various applications where gender recognition can be useful. Some of these include:

- for further identification of human sounds like male laughing, female singing
- categorizing audios/videos by adding tags and simplifying and reducing search space
- automatic salutations
- can help personal assistants like Siri, Google Assistant to answer the question
- with female generic or male generic results.

Tremendous work is already done in the field of gender identification. Becker [2] used a frequency-based baseline model, logistic regression model [3], classification and regression tree (CART) model [4], random forest model [5], boosted tree model [6], Support Vector Machine (SVM) model [7], XGBoost model [8], stacked model [9] for recognition of voices. Ali, Islam and Hossain have worked with frequency at maximum power in the system, power spectrum density and other frequency domain features for gender recognition [10]. Alías, Socoró and Sevillano worked on perceptual and physical features [11]. Subramanian worked in the field of classification of audio signals [12]. Anemüller and Kollmeier discussed about background on real acoustics and worked with perceptual features for speech detection [13]. Richard, Sundaram and Narayanan have worked on audio indexing and classification using perceptual features [14]. B. Moghaddam, and M.H. Yang used Support Vector Machine for visual gender classification. In their work, they have compared the results of SVM over other classification models which include Linear, Quadratic, Fisher Linear Discriminant, K Nearest-Neighbor. They also proved that SVM performs comparatively better than other classifiers [15]. Harb and Chen created their own set of features which consists of 20 spectral coefficients. For the construction of the classifier, they used Neural Network [16]. Meinedo and Trancoso even combined acoustic as well prosodic features for the purpose age and gender classification [17].

II. DATASET

The database which I am using for recognizing gender based on audio events is retrieved from Kaggle (Gender Recognition by Voice) [18]. It was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice

samples of male and female speakers and these voice samples are collected from the Harvard-Haskins Database of Regularly-Timed Speech, Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University, VoxForge Speech Corpus and Festvox CMU_ARCTIC Speech Database at Carnegie Mellon University [18]. The voice samples were pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz and the CSV file with 21 columns (20 columns for each feature and one label column for the classification of male or female) was created. I am using this CSV file as dataset which contains 1584 Male records and 1584 Female records along with acoustic properties. Each record has different acoustic properties and the acoustic properties are shown in Table I as follows.

TABLE I. ACOUSTIC PROPERTIES OF VOICE

Acoustic properties	
Properties	***Description***
meanfreq	mean frequency (in kHz)
sd	standard deviation of frequency
median	median frequency (in kHz)
Q25	first quantile (in kHz)
Q75	third quantile (in kHz)
IQR	interquantile range (in kHz)
skew	skewness
kurt	kurtosis
sp.ent	spectral entropy
sfm	spectral flatness
mode	mode frequency
centroid	frequency centroid
meanfun	average of fundamental frequency measured across acoustic signal
minfun	minimum fundamental frequency measured across acoustic signal
maxfun	maximum fundamental frequency measured across acoustic signal
meandom	average of dominant frequency measured across acoustic signal
mindom	minimum of dominant frequency measured across acoustic signal
maxdom	maximum of dominant frequency measured across acoustic signal
dfrange	range of dominant frequency measured across acoustic signal
modindx	modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
label	male or female

III. METHODOLOGY

The project is based on a classification problem as the output is based on classifying gender into male or female based on voice. The goal is to compare accuracy, precision, recall and F1 score of various models and suggest the best model that can be used for gender recognition by voice. While dealing with the task of classifying male and female voices i.e. identifying the gender of the person speaking, the job is typically divided into two main phases namely preprocessing and classification. The general block diagram of gender classification model using audio signal is given in Fig. 1 as follows.

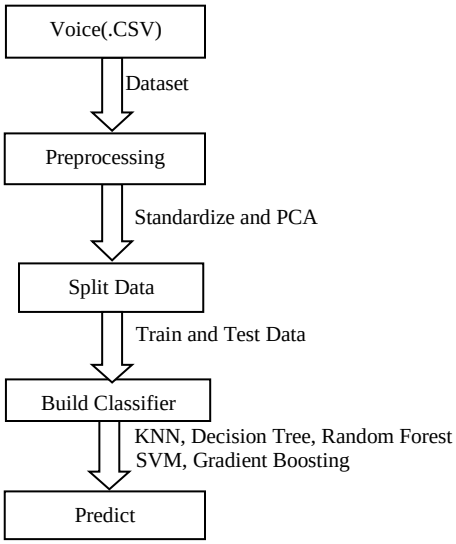

Fig. 1. Work flow to recognize gender using voice

A. *Preprocessing Techniques*

The dataset is already preprocessed. But still preprocessing techniques are performed to achieve the maximum accuracy. The preprocessing techniques used in my project are as follows:

1) *Missing values*

Handling missing values is first and the prior step towards building the model. Missing values occur in data due to various reasons, such as problems occurred during extraction or data collection process. So, I checked for the missing values in the dataset. There are no missing values found. If there were any missing values, I would have deleted the the missing values using list wise method. In list wise method, the complete row which contains missing value is erased. After checking for the missing value, the next and most important step is to encode the label column in the dataset. Label column contains the output of the dataset which can either be male or female. The output male is encoded as 1 and the output female is encoded as 0. The reason for this encoding is that , further a classification model is to built up and it will be easy if the output column contains two classes 1 or 0 instead of male or female.

2) *Standardization*

Standardization is a technique used to remove mean or to scale the variance. If the individual features of the data do not look like standard normally distributed data i.e. zero mean and unit variance then fitting the model will not be accurate. It is

always required to create the best fit model using machine learning algorithms. Standardization is basically subtracting the mean value of each feature, then scaling it by dividing the respective feature by its standard deviation [19]. I am performing this technique to centralize the data and it is implemented using sklearn library.

3) Principal Component Analysis

Principal component analysis (PCA) is a technique to reduce the learning space as it reduces a large number of correlated variables into a smaller number of uncorrelated variables. It decreases the diminesions of dataset. The aim of this technique is to reduce those features which are uncoorelated with the output or highly coorelated with other features. From those highly coorelated features, one feature is selected. In my dataset, at first I performed PCA for reducing 20 features. As, the dataset used is already preprocessed, so performing this technique does not prove to be much efficient.

B. Classification Techniques:

After performing the prepreprocessing techniques, the dataset is ready to be processed. For performing classification, the dataset is divided into two parts. 75% of dataset is in train data and 25% of dataset is in test data. There should always be more data in training data as compared to the test data set. It is so because more data is required to train the model as compared to test the model. Both training and testing data are further categorized into two parts. One is X which includes all the independent predictors and the other is Y which includes the output i.e. the label column which tells whether the voice is of male or female for the coressponding values of X. The following are the types of classifiers:

1) K Nearest Neighbors Classifier

K Nearest Neighbors (KNN) is the intuitive machine learning algorithm. KNN. It is an unsupervised technique which predicts the output while searching for the nearest and most similar K records from the dataset. To implement this technique, KNeighborsClassifier from sklearn library is used. The fitted model i.e. classifier is build using the training set (including both X and Y). Prediction is done on this model for the test sample. Finally, the predicted output is compared with the actual output and the accuracy, recall, precison and F1 score are calculated.

2) Decision Tree Classifier

Decision Tree (DT) is a non-parametric supervised learning technique used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A condition is checked at the root node and based on the output an appropriate branch is selected. It can further lead to another internal node where a new test condition is checked, or to a leaf node where class is assigned to the record [21]. This classifier is applied with the tree function in sklearn library. The model is trained with the help of X and Y values of train dataset. Then this model is predicted on X values of test dataset and the corresponding Y values are obtained. These Y values are compared with actual Y values e the accuracy, recall, precision and F1 score are recorded.

3) Random Forest Classifier

Random forest (RF) is an robust ensemble classifier which consists a number of decision trees. The output is obtained by the collective decision of output from individual trees. The root node is split using features in the dataset and each node is further divided using other features. I implemented it using Random Forest Classfier in sklearn library. The model is fitted using the X and Y of train dataset. Prediction of this model is done on X of test dataset and the output is predicted. This output belongs to class 0 or 1 which means the voice belongs to either male or female. The output is compared with Y of testing dataset and the accuracy, recall, precison and F1 score are obtained.

4) Support Vector Machine Classifier

Support vector machine (SVM) is a binary classifier which separates the data into two different classes at the same time. The decision boundary is build such that the distance between hyperplane and the training data is maximum. This is done to minimize the noice because if hyper plane passes close to the points, it will be noise sensitive. After determining the best decision boundary, it is easy to determine the class of test data by checking the side of the hyperplane it belongs [20]. I used SVM because there are 2 classes 1 and 0 i.e. male and female. Two parameters involved in SVM are regularization parameter or C parameter that tells about the SVM optimization and the gamma parameter. To implement it, SVM from sklearn library is used. SVC class along with radial basis function default kernal , value of C as 10 and value of gamma as 0.1 is used to fit the model from train data. The fitted model is predicted on the test data and the output is obtained. This output is compared with actual output and the accuracy, recall, precison and F1 score are calculated.

5) Support Vector Machine Classifier using poly kernel

In Support vector Machine the default kernel used is radial basis function. But there are many other kernels that can also be used which include poly and sigmoid. This technique is used to find the better kernel than radial basis function and to implement it. At first, I performed cross validation technique to find the scores associated with each kernel. The kernel with highest score is better and chosen. For creating a better model, C and gamma values associated with kernel should also be calculated. For calculating these values, again cross validation techniques are performed and accurate value of C and gamma is chosen. Using the kernel with high score and its associated C and gamma values, the classifier is created. The model is trained using this classifier along with the X and Y values of train data. Then this model is tested on X values of test data and corresponding Y values are obtained. These Y values are

compared to actual Y of test data and accuracy, recall, precision and F1 score are calculated.

6) Gradient Boosting Classifier

Gradient boosting is a machine learning technique which produces a classifier in the form of an ensemble of weak prediction models. It builds an additive model in a forward stage-wise fashion and is generally used when individual classifiers do not provide accuracy. It is implemented using Gradient Boosting in ensemble-based methods from sklearn library. To fit the model, X and Y columns of training dataset is required. After fitting the model, it is predicted on the testing data set. The output is obtained, and this output is compared with actual output and the accuracy, recall, precision and F1 score are calculated.

IV. IMPLEMENTATION

A. *Mean Frequency v/s label*

Mean frequency is the key component to distinguish the voice between male and female. The boxplot of mean frequency v/s label i.e. either male or female is given in Fig. 2 as follows.

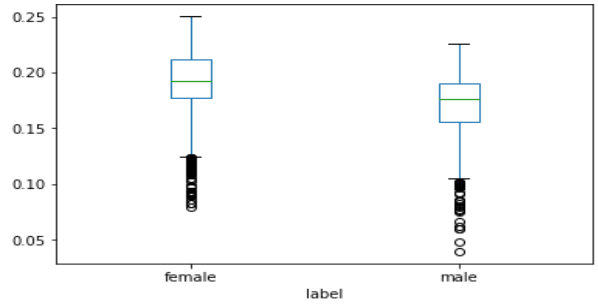

Fig. 2. Boxplot of mean frequency v/s label

It can be seen from the boxplot that there is difference in mean frequency of male and female voices i.e. the frequency of female voice is higher than a male voice. This can also be seen from Fig.3 and Fig.4 which shows individual mean frequency plot of male and female. On the other hand, boxplot also shows that male and female voices cannot be determined only based on mean frequency, but other features are also required.

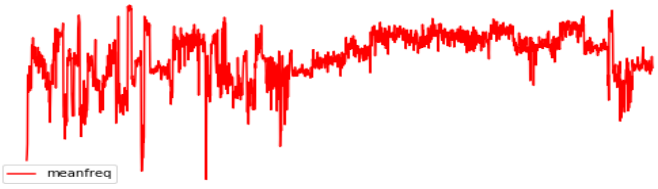

Fig.3. Mean frequency of male

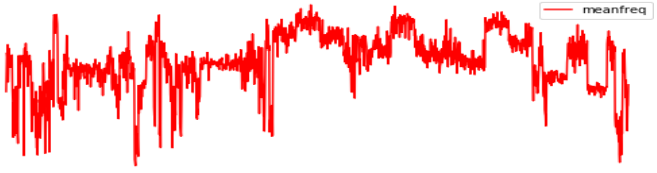

Fig.4. Mean frequency of female

B. *Application of Principal Component Analysis*

Principal Component Analysis (PCA) is a dimensionality reduction method as it helps in removing correlated variables in the dataset. But it is not always the case that PCA will give promising results. In this project, I used PCA and compared the results with that of without PCA and I can see that PCA does not improve the performance of the classifier. The graph of results after performing PCA is given in Fig. 5 as follows.

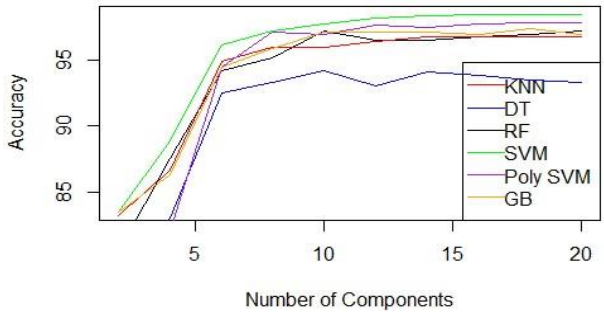

Fig.5. Number of components in PCA v/s accuracy

From Fig.5, it can be deduced that as number of components increases accuracy also increases. PCA is performed to increase the accuracy. For almost all the classifiers, accuracy remains constant after number of components reaches 16. But there is no use of performing PCA and reducing the dimensions to 16 from 20 as it will not make much difference. The reason that PCA does not perform better is that for each different classifier, different number of features are showing importance. Performance depends on the feature which the classifier takes while building the model. This can be seen in Fig.6 which shows feature extraction of Gradient Boosting classifier and Fig.7 which shows feature extraction of Random Forest.

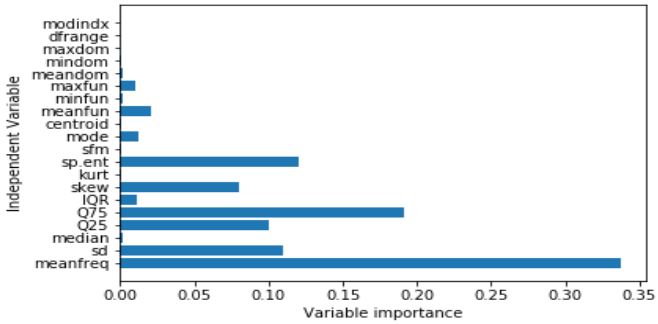

Fig.6. Feature extraction of Gradient Boosting

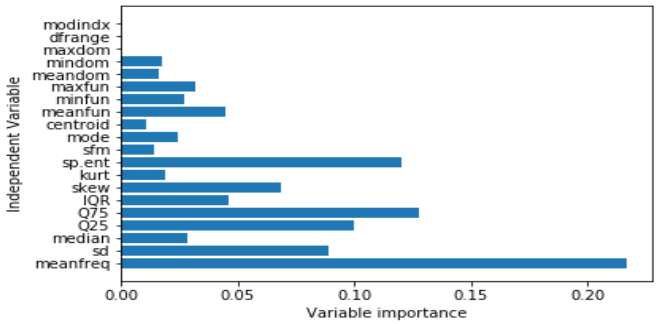

Fig.7. Feature extraction of Random Forest

When PCA can use all the features, it can be seen from Fig.6 and Fig.7 that features chosen by Gradient Boosting differ from those of Random Forest classifier. Random forest has used all the features except maxdom, dfrange and modindx. But Gradient Boosting excludes maxdom, dfrange and modindx along with many other features like mindom, and meandom. Moreover, Gradient Boosting is only using 10 out of 20 independent variables to create the model.

C. Confusion Matrix of Classifiers

A confusion matrix is a table that is used to describe the performance of a classifier on a set of test data for which the output is to be predicted but actual output is known. For creating confusion matrix, actual output and predicted value is required and then True Positive, False Positive, True Negative and False Negative can be evaluated. With the help of confusion matrix accuracy, recall, precision and F score can be easily calculated. Crosstab function in pandas library is used to implement the confusion matrices. Table II, Table III, Table IV, Table V, Table VI, Table VII shows the confusion Matrix of K Nearest Neighbors classifier, Decision Tree, Random Forest, Support Vector Machine, Support Vector Machine using Poly kernel and Gradient Boosting classifier respectively.

1) K Nearest Neighbors Classifier

TABLE II. CONFUSION MATRIX K NEAREST NEIGHBORS

Predicted/Actual	0	1
0	384	16
1	10	382

2) Decision Tree Classifier

TABLE III. CONFUSION MATRIX OF DECISION TREE

Predicted/Actual	0	1
0	372	28
1	25	367

3) Randon Forest Classifier

TABLE IV. CONFUSION MATRIX OF RANDOM FOREST

Predicted/Actual	0	1
0	391	9
1	9	383

4) Support Vector Machine Classifier

TABLE V. CONFUSION MATRIX SUPPORT VECTOR MACHINE

Predicted/Actual	0	1
0	395	5
1	7	385

5) Support Vector Machine Classifier using poly kernel

TABLE VI. CONFUSION MATRIX SVM USING POLY KERNEL

Predicted/Actual	0	1
0	393	7
1	11	381

6) Gradient Boosting Classifier

TABLE VII. CONFUSION MATRIX OF GRADIENT BOOSTING

Predicted/Actual	0	1
0	385	15
1	9	383

V. RESULTS AND DISCUSSIONS

Classifiers which include K Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, Support Vector Machine using poly kernel and Gradient Boosting are compared with each other based on Accuracy, Precision, Recall and F score. Accuracy is the fraction of predictions that are predicted correctly. Precision is the fraction of relevant instances among the retrieved instances and is also called positive predictive value. It is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances and is also called sensitivity [22]. It is the ratio of correctly predicted positive observations to the all observations in actual class - yes. F1 Score is the weighted average of Precision and Recall [23]. To calculate accuracy, precision, recall and F1 score, I am using metrics function in sklearn library.

TABLE VIII. ACCURACY, PRECISION, RECALL AND F SCORE OF CLASSIFIERS IN PERCENTAGE

Classifier	Accuracy	Precision	Recall	F1 score
K Nearest Neighbors	96.71	97.44	95.98	96.71
Decision Tree	93.33	93.62	92.91	93.26
Random Forest	97.72	97.70	97.70	97.70
Support Vector Machine	98.48	98.71	98.21	98.46
SVM using Poly kernel	97.72	97.19	98.19	97.19
Gradient Boosting	96.96	97.70	96.23	96.96

From the table, it can be clearly seen than that accuracy, recall, precision and F1 score associated with the Support Vector machine is higher as compared to any other classifier. So, according to my results Support Vector Machine is the best classifier.

The performance of Random forest and Support Vector Machine using Poly kernel is almost same. Gradient Boosting is expected to perform better than Support Vector Machine as it is an ensemble classifier. On the contrary, this is not the

case. The reason may be Gradient Boosting is the combination of weak classifiers. But in this case Support Vector Machine is acting as a strong classifier. So, there is not much need for Gradient Boosting.

VI. Conclusion

For the dataset I used, Support Vector Machine can be considered as best classifier for recognizing gender using voice as it provides best result. Support Vector Machine gives accuracy of 98.48%, precision of 98.71%, recall of 98.21% and F1 Score of 98.46%. In this case, I applied Principal Component analysis, but the results were not satisfactory. While creating the model, all the classifiers use different features. Therefore, while creating the best fit model, feature selection plays an important role. No classifier can always act as best classifier. It depends on the problem and the solution to it.

References

[1] A Raahul, R Sapthagiri, K Pankaj and V Vijayarajan, “Voice based gender classification using machine learning”, IOP Conf. Series of Materials Science and Engineering, 2017.

[2] K. Becker, “Identifying the Gender of a Voice using Machine Learning”, 2016, *unpublished*.

[3] J. M. Hilbe, “Logistic Regression Models”, CRC Press, 2009.

[4] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, “Classification and Regression Trees”, CRC Press, 1984.

[5] L. Breiman, “Random forests”, Machine Learning, Springer US, 45:5–32, 2001.

[6] J.H. Friedman, Stochastic Gradient Boosting, 1999.

[7] C. Cortes, V. Vapnik, “Support-vector networks”, Machine Learning, 20 (3): 273–297, 1995.

[8] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, 1999.

[9] L. Breiman, “Stacked regressions”, Machine Learning, Springer US, 45:5–32, 2001.

[10] Ali, Md Sadek, Md Shariful Islam, and Md Alamgir Hossain, "Gender recognition system using speech signal." International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2, No. 1, 2012, pp. 1-9.

[11] Alías, Francesc, Joan Claudi Socoró, and Xavier Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds." Applied Sciences 6, No. 5, 2016, pp. 143.

[12] Subramanian, Hariharan, P. Rao, and S. D. Roy, "Audio signal classification." EE Dept, IIT Bombay, 2004, pp. 1-5.

[13] Bach, Jörg-Hendrik, Jörn Anemüller, and Birger Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features." Speech Communication, Vol. 53, No. 5, 2011, pp. 690-706.

[14] Richard, Gaël, Shiva Sundaram, and Shrikanth Narayanan, "An overview on perceptually motivated audio indexing and classification." Proceedings of the IEEE, Vol. 101, No. 9, 2013, pp.1939-1954.

[15] Moghaddam, Baback, and Ming-Hsuan Yang, "Gender classification with support vector machines." In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, 2000, pp. 306-311, IEEE.

[16] Harb, Hadi, and Liming Chen. "Gender identification using a general audio classifier." In *Multimedia and Expo, ICME'03. Proceedings. International Conference on*, Vol. 2, 2003, pp. II-733, IEEE.

[17] Meinedo, Hugo, and Isabel Trancoso. "Age and gender classification using fusion of acoustic and prosodic features." In Eleventh Annual Conference of the International Speech Communication Association. 2010.

[18] https://toolbox.google.com/datasetsearch/search?query=Gender%20Recognition%20by%20Voice&docid=3HuasGDpBXlmWrU0AAAAAA%3D%3D

[19] https://scikit-learn.org/stable/modules/preprocessing.html.

[20] Saptarshi Sengupta, Ghazaala Yasmin and Arijit Ghosal, “Classification of Male and Female Speech Using Perceptual Features,” in International Conference on Computing, Communication and Networking Technologies, 2017.

[21] https://scikit-learn.org/stable/modules/tree.html

[22] https://en.wikipedia.org/wiki/Precision_and_recall

[23] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score interpretation-of-performance-measures/