# GENDER IDENTIFICATION FROM VOICE USING MACHINE LEARNING

Report submitted to the SASTRA Deemed to be University
in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology
in
Electronics and Communication Engineering

A MINI PROJECT-1 REPORT

Submitted by

**GUDI VAMSIKRISHNA**
**(121004095-ECE)**
**SHAIK MUHAMMAD ARSHAD ALI**
**(121004234-ECE)**
**DALAVAI GAGANNATH**
**(121004288-ECE)**

**May 2020**


**School of Electrical and Electronics Engineering**

**THANJAVUR, TAMIL NADU, INDIA-613401**


**School of Electrical and Electronics Engineering**

# BONAFIDE CERTIFICATE

This is to certify that the report titled "**GENDER IDENTIFICATION FROM VOICE USING MACHINE LEARNING"** submitted as a requirement for the course, **BECCEC608: MINI PROJECT** for B.Tech Electronics and Communication engineering programme, is a bonafide record of the work done by **1.Gudi VamsiKrishna (121004095-ECE-3rd year)** , **2. Shaik Muhammad Arshad Ali (121004234-ECE-3rd year), 3.Dalavai Gagannath (121004288-ECE-3rd year)** during the academic year 2019-2020, in the school of Electrical & Electronics , under my supervison.

**Signature of Project Supervisor :**

**Name with Affiliation :**

**Date :**

**Examiner 1** **Examiner 2**


## School of Electrical and Electronics

## Declaration

We declare that the report titled '**GENDER IDENTIFICATION FROM VOICE USING MACHINE LEARNING'** submitted by us is an original work done by us under the guidance of **Dr.RAMKUMAR K, Professor,School of Electrical and Electronics,SASTRA Deemed to be University** during the final semester of academic year 2019-2020, in the **School of Electrical and Electronics.**The work is original and wherever We have used materials from other sources,We have given due credit and cited them in the text of report.This report has not formed the basis for the award of any degree,diploma,associate-ship,fellowship or other similar title to any candidate of any university

# ACKNOWLEDGEMENT

First of all, we express our gratitude to **Dr. S Vaidhyasubramaniam**, Vice-Chancellor, SASTRA Deemed to be University who provided all facilities and necessary encouragement during the course of our study. We extend our sincere thanks to **Dr. R Chandramouli**, Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this mini project.

We dedicate our wholehearted thanks to **Dr. K. Thenmozhi,** Dean (SEEE) and **Dr. John Bosco Balaguru**, Dean(Research), **Dr. K Sridhar**, Associate Dean(ECE), who motivated us during the mini project-I work.

I would like to express my deepest appreciation to all those who provided me the possibility to complete this mini-project and whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report. Further, I would like to acknowledge with much appreciation the crucial role of my teammates who helped me and gave suggestions about the task "**GENDER IDENTIFICATION FROM VOICE USING MACHINE LEARNING**". Last but not least, many thanks goes to the mentor of our mini-project, **Prof. K RAMKUMAR** who has invested his full efforts in guiding the team and in achieving the goal. I have to appreciate the guidance given by another supervisor as well as the panels especially in our mini project presentation that has improved our presentation skills, thanks to their comments and advice.

# ABSTRACT

Gender Identification is one of the important aspects of speech analysis today. The gender of the person has become very crucial in economic markets in the form of AdSense and also has many applications. Identification of gender using the features of acoustic data like mean frequency, median, Q25, IQR, etc. using Machine learning can give good results.

We used a Kaggle dataset that consists of 3168 voice samples that were recorded from both (Male & female) gender speakers which are processed with speech analysis. Out of these features, we found Mean fundamental frequency, Q25, and IQR play an important role. To build a model, we are using different algorithms: Logistic Regression, Random Forest, Support Vector machine, Decision tree.

We are using parameters like Precision, F1score, Recall, and Accuracy to evaluate our algorithms. We found the Support Vector machine algorithm performing best in prediction. Finally, We took raw voice samples and processed it using the acoustic analysis to extract features in Python and predicted the gender.

# TABLE OF CONTENTS

# List of Figures



# List of tables

# CHAPTER 1

# INTRODUCTION

## 1.1 Brief Introduction on Machine Learning

Machine Learning definition: "A program is said to learn as for some set of assignments T and performance parameter K, if its performance at assignments in T, as calculated by P, improves with experience."

Machine Learning comes under Artificial Intelligence. This innovation gives computer the capacity to naturally take in and improve as a matter of fact without being programming explicitly.

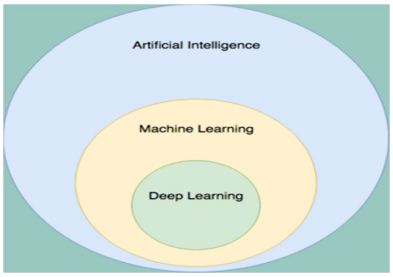



Figure 1.1 (AI vs ML vs DL)

ML has three important branches/type:

- Supervised Learning
- Unsupervised machine Learning
- Reinforcement machine Learning

**Supervised Learning** In this case we will have a dataset in which the outcome is already known. They mostly used to handle 'Regression' and 'Classification' problems.

Eg: (a)Regression – Predicting the foot size from its height.
(b)Classification – Predicting gender from the facial features.

**Unsupervised Learning** We can learn by clustering/grouping the data based on patterns among the features in the data. Here we need to understand the structure of data and analyse find the pattern.This can be done by Partitioning based clustering, Gaussian mixture models etc.

Eg: Recommender systems

**Reinforcement Learning** learns by associating with its situation. The operator gets compensations by performing effectively and punishments for performing mistakenly. The operator that learns without human mediation by boosting its prize and limiting its punishment. Eg: Chess game with computer

**1.2 Gender Identification and its applications:**

Gender Identification is one of the important aspects today. The gender of the person has become very important in economic markets as it has wide Applications like

1. Better Ads based on gender by Google Adsense
2. Promotional Telephone calls by gender categorization
3. Better Video Suggestions from Youtube.
4. Improved Human – Machine Interaction.
5. And many more..

Reason why we preferred Voice to identify gender:

- Gender of a person can be Identified using DNA, Facial and body features, Fingerprint patterns, Iris , ECG & EEG , Voice/Speech.
- From the above attributes, Most of them are difficult to collect from a user.
- We can easily get the voice/speech of user compared to all other attributes. From the voice, Speech analysis should be done to extract spectral,statistical audio features.
- Python/R language provides us extensive libraries to analyse and extract the featuresfrom the voice samples.
- In Order to predict or identify gender from voice, Machine Learning comes into the picture. Machine Learning has powerful and capable algorithms to handle this problem.
- So we need a dataset that has audio features with label. This comes under the branch of Supervised Machine Learning Classification problem.

### 1.3 Motivation and Objectives:

In this cutting edge world, Artificial Intelligence(AI) is changing lives. It has become most demanded skill. We also felt interested after knowing the capabilities of data, AI and Machine Learning. Watching the modern world technologies like voice assistant, Chat box etc., We decided to do a mini-project that might help improving current technologies.

**Objectives:**

1. To find a suitable dataset to Identify the gender.
2. To learn how to apply libraries to perform Exploratory data Analysis.
3. To Understand and Apply suitable machine Learning Algorithms.
4. To evaluate the suitable algorithm for predicting the gender.
5. Try to do the feature extract features from raw voice samples.

# CHAPTER 2

# DATA SET AND FEATURES

## 2.1. INFORMATION ON DATA SET:

The dataset we used was downloaded from 'Kaggle'.This data set totally contains 3,168 observations with 21 different features. These 21 variables are the audio features of male and female voice samples. All the samples in dataset are already preprocessed.

Out of 21 columns, 20 columns are used for each feature and final label column is identify whether it is male or female. With all 21 features it will be easy to identify small changes and modulations that are present in voices. These variations can be pitch levels like low and high, periodic waves in air, consonants unvoiced sounds etc..

Out of 3168 samples, there are 1584 male voice samples and 1584 female voice samples.

| | meanfreq | sd | median | Q25 | Q75 | IQR | skew | kurt | sp.ent | sfm | mode | centroid | meanfun | minfun | maxfun | meandom | mindom | maxdom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| label | | | | | | | | | | | | | | | | | | |
| female | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 |
| male | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 | 1584 |

Figure 2.1 Features and its count in dataset

## 2.2. VARIOUS ACOUSTIC FEATURES:

These are the audio spectral and statistical features that are extracted and used in identifying gender from voice:

**1. meanfreq**:(mean frequency in kHz).It is obtained by calculating mean of frequencies of the sample .The peak frequency present in voice sample is taken as the frequency mean.

**2. sd**: (standard deviation) It is the amount of variation or dispersion of the frequencies in the voice.

**3 .median**: (median frequency in kHz).It is obtained by adding the intensities of the signals and then cumulative intensity is selected.

**4.Q25**: (first quantile in kHz).It is frequency at which the signal is split into 2 frequency intervals i.e. 25% and 75% energy .

**Q75**: (third quantile in kHz).It is the frequency at which the signal is split into 2 frequency intervals i.e.75% and25% energy .

**IQR**(Interquartile frequency range in kHz). It extends somewhere in the range of third quantile and first quantile. **IQR**=$Q_3$-$Q_1$

**skew**:(skewness) It is the amount of deviation/distortion in the samples. It can be less than zero, greater than zero, 0 or undefined.

**kurt**:(kurtosis) It is obtained by measuring the tails of a frequency that are compared to gaussian distribution

**sp.ent**:(spectral entropy) It is distribution of energies of frequencies and its spectrum

**sfm**: (spectral flatness) It is obtained by measuring noisiness in the voice samples.

**mode**: ( mode frequency) It is mode of the frequencies of voice samples.

**Centroid:** It is central frequency. It is obtained by calculating mean of frequencies in signal with its magnitudes.

**meanfun**:(mean fundamental frequency) It is obtained by calculating the average of fundamental frequencies of signal.

**minfun**:(minimum fundamental frequency) It is obtained by calculating minimum fundamental frequencies of signal.

**maxfun**:(maximum fundamental frequency) It is maximum fundamental frequency of signal.

**meandom**:(mean of dominant frequency) It is obtained by calculating average of superior frequencies in signal.

**mindom**:(minimum of dominant frequency )It is minimum frequency measured at superior part of signal.

**maxdom**:(maximum of dominant frequency) It is maximum frequency measured at superior part of signal.

**drange**:(range of dominant frequency) It is range of dominant frequencies measured in the signal.

**modindx**:(modulation index) It is obtained by calculating difference between major frequencies and adjacent measurements. If difference is negative mod is taken.

# CHAPTER 3

# METHODOLOGY

## 3.1. BLOCK DIAGRAM

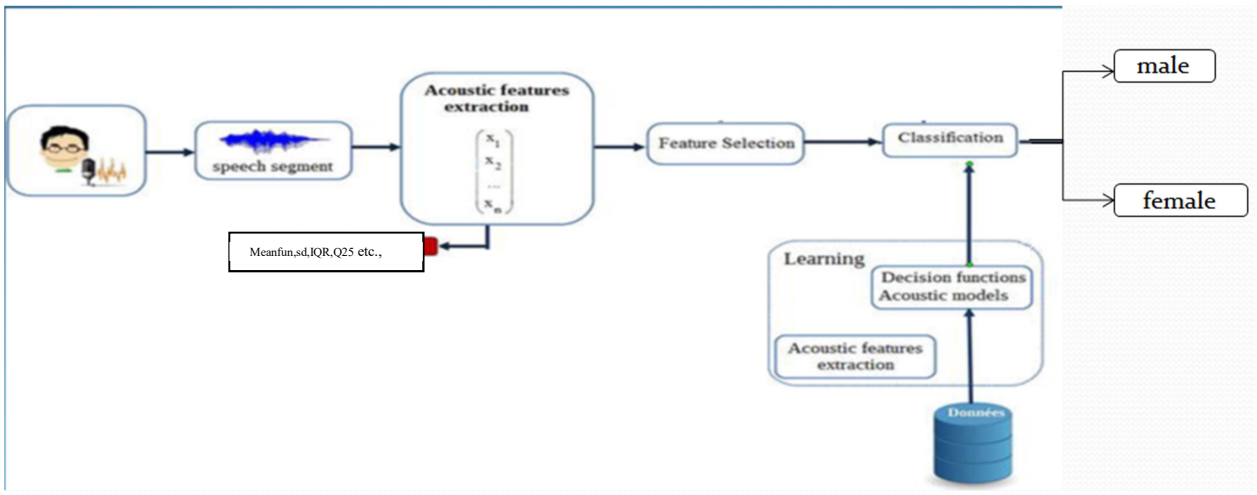



Figure 3.1: Block diagram

## 3.2. SOFTWARE AND LIBRARIES

The language we used is 'python', and IDE we used for python is 'Jupyter Notebook'. Python has various inbuilt libraries. Here we have used various libraries which have their own importance. Various libraries used are

1. Pandas: Pandas is a high level data manipulation language. It provides good functions, simple structures and data analytics tools. It is speed and flexible. It has dataframe & series data-structures with inbuilt functions that help us to do data analysis in python.

2. Numpy: Numpy means numeric python. It is used for performing all the mathematics in python. It is used exclusively to handle arrays. It has various functions related to linear algebra, fourier transforms and matrix operations.

3. Matplot: It is a plotting library for python and also it is numerical maths extension for numpy. Using matplot we can make different graphical representations like histograms, barplots etc..

4. Sklearn: Sklearn means scikit-learn .It's a library in python language for machine learning algorihms & its evauation . It includes many useful classification, regression and clustering algorithms.

5.Seaborn : *Seaborn* is a Python library based on matplotlib for data visualization. It provides a intuitive interface for making graphs, tables attractive and also to draw informative statistical graphics.

### 3.3. DESIGN AND IMPLEMENTATION

- Voice samples from various males and females are collected by kaggle.
- From the collected samples different acoustic features are extracted. These features were already preproccesed and extracted using R languages packages.
- First we imported all the imported all the required libraries.
- We have load the dataset ad a data frame
- Then we checked whether all features have same no of samples or if any missing values are present in the dataset. We make use of pandas function to determine the missing values.
- Then we Checked the statistical properties of all features.

```
#Checking for missing values in the dataset
data.isnull().sum()
```

```
meanfreq    0
sd          0
median      0
Q25         0
Q75         0
IQR         0
skew        0
kurt        0
sp.ent      0
sfm         0
mode        0
centroid    0
meanfun     0
minfun      0
maxfun      0
meandom     0
mindom      0
maxdom      0
dfrange     0
modindx     0
label       0
dtype: int64
```

```
data.describe()
```

| | meanfreq | sd | median | Q25 | Q75 | IQR | skew | kurt | sp.ent | sfm |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 |
| mean | 0.180907 | 0.057126 | 0.185621 | 0.140456 | 0.224765 | 0.084309 | 3.140168 | 36.568461 | 0.895127 | 0.408216 |
| std | 0.029918 | 0.016652 | 0.036360 | 0.048680 | 0.023639 | 0.042783 | 4.240529 | 134.928661 | 0.044980 | 0.177521 |
| min | 0.039363 | 0.018363 | 0.010975 | 0.000229 | 0.042946 | 0.014558 | 0.141735 | 2.068455 | 0.738651 | 0.036876 |
| 25% | 0.163662 | 0.041954 | 0.169593 | 0.111087 | 0.208747 | 0.042560 | 1.649569 | 5.669547 | 0.861811 | 0.258041 |
| 50% | 0.184838 | 0.059155 | 0.190032 | 0.140286 | 0.225684 | 0.094280 | 2.197101 | 8.318463 | 0.901767 | 0.396335 |
| 75% | 0.199146 | 0.067020 | 0.210618 | 0.175939 | 0.243660 | 0.114175 | 2.931694 | 13.648905 | 0.928713 | 0.533676 |
| max | 0.251124 | 0.115273 | 0.261224 | 0.247347 | 0.273469 | 0.252225 | 34.725453 | 1309.612887 | 0.981997 | 0.842936 |

Figure 3.2.Dataset information

- From the extracted acoustic features, features that are best suitable for differentiating male or female are selected by plotting their distribution graphs. From graphs,we can understand the correlation with the label of features. For plotting those graphs, We make use of Matplot library and seaborn library.
  :
- Label Encoding is to be done as the machine cannot understand by names.So we used that and done male: ‘1’ and female: ‘0’.

- For the purpose of Cross-validation, we then split the voice.csv dataset into training and testing datasets.

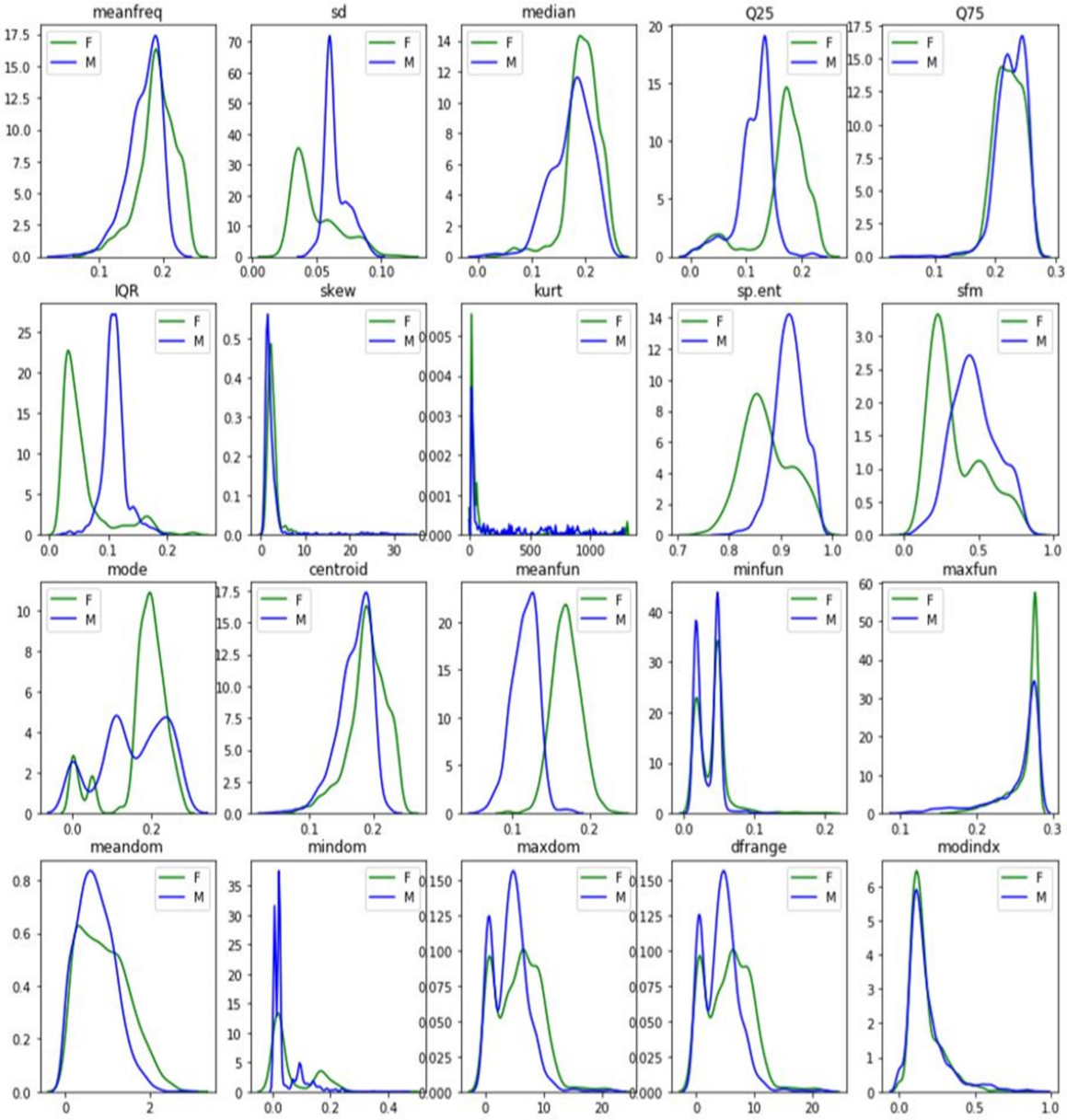



Figure 3.3 KDE Plots

- For training we give 70% of dataset and for testing we give 30% of dataset.
- All columns of dataset except label column are taken as ‘X’ an independent variable(input)

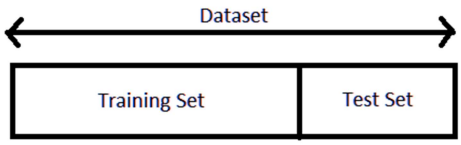



Figure 3.4: Cross-Validation

- Label column is taken as ‘Y’ a dependent variable(output).
- ‘X’ has Xtrain , Xtest ,’Y’ has Ytrain,Ytest.
- Standardization is done to avoid machine errors that are influenced by outliers and values of two independent features. We have choosen standardisation as algorithms follow Gaussian distribution.
- Next step is to train the machine with collected features from the dataset to make machine capable to classify genders of voice.
- Here machines are trained with 4 machine learning algorithms like LogisticRegression, Decision trees, RandomForest and Support vector machines.

### 3.4 Algorithms:

**a. Logistic Regression:** Logistic Regression analysis is supervised machine learning algorithm .It helps to estimates the relationship between a dependent features and an independent feature based on maximum likelihood. It gives results in binary form which is used to estimate the outcome of a categorical dependent feature. So the outcome will be in discrete form.

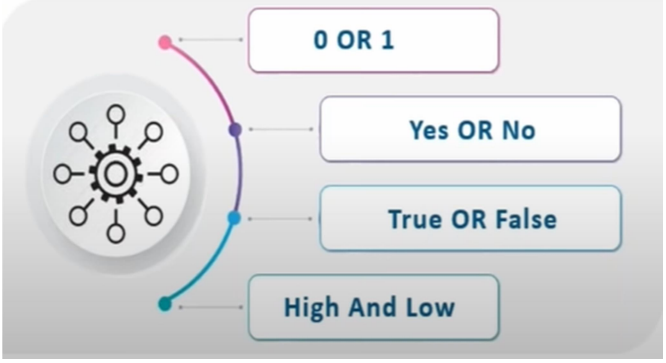



Fig 3.5 Logistic Regression

Logistic regression uses sigmoid curve.This S curve is drawn by using sigmoid function which converts any infinite value to discrete values which a logistic regression wants which are 0 and 1.

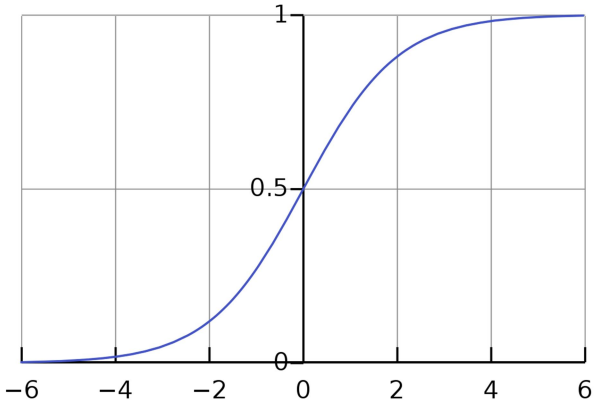




Figure 3.6. S-Curve(sigmoid)

It can be used to predict whether its cloudy or not ,sunny or not. In our project we have to tell whether the given voice is masculine or feminine .By using logistic function the voice attributes are corverted into binary format.It has only two outcomes .so we can use logistic regression for gender identification.

Estimation is done through maximum likelihood.The dependent variable in logistic regression follows Bernoulli Distribution.

Advantages:

Easy Implementation, Doesn't require high computational power and efficient.

Dis-advantages:

Could not able to solve the non-linear problems and also vulnerable to overfitting.

b. **Random Forest:**

Random forest which is used to solve classification problem. It is a supervised machine learning model as the classifier already has a set of classified examples and from these examples, the classifier learns to assign inseen new examples. It is an ensemble method classifier made using many decision tree models. Ensemble model combine the results from different models .More accurate and stable predictons are made by this algorithm

It is trained with Bagging method

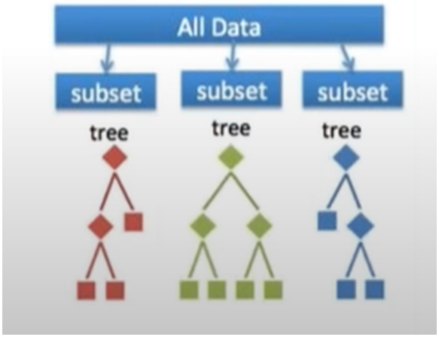



Fig 3.7 Random Forest

In our project we are having 21 parameters so using random forest is very feasible as it develops trees for each parameter and analyses from all those outputs, it will go by the call of the majority voting and gives final result.

1. Select some N data points from the training dataset randomly. They can also be repetitive. Consider this as a subset of given dataset.
2. Then build decision trees associated with the selected data points.
3. Choose the number K for decision trees that you want to build.
4. Repeat Step 1&2
5. Find the predictions of each decision tree, decide the result which gets maximum votes.

It takes less training time as compared to other algorithms.

It predicts output with high accuracy, even for the large dataset it works effectively.

It can also maintain accuracy when large proportion of data is missing. But it does not perform as good with regression problems compared to classification.

**c.Support vector machine(SVM):**

SVM is a discriminative classifier that is formally designed by separative hyperplane. It is a representation of examples as points in space that are aseperated by a gaps as wide as possible. An svm is supervised machine learning algorithm that has high efficient results. It is capable of classification regression and outlayer detection as well.it can also perform non linear classification.

Its main objective is to segregate the given data in best possible way. when the segregation is done the distance between the nearest points is margin and the approach is to select a hyperplain with maximum possible margin between the support vectors in the given data set. now to select the maximum hyperplain in the given sets the svm follows.

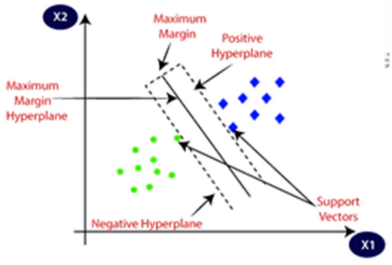



Fig 3.8 Support Vector Machines

1)it develops a hyperplain which segregates the classes in best possible way

2)Then it selects a hyperplain with maximum segregation from the nearest points

Advantages:

In high dimension spaces,it works effective.

It is memory efficient as it uses a subset of training points in the decision function.

Disadvantages:

If the no of the features are much larger that the no of samples we have to avoid over fitting in choosing kernel functions

Svm do not directly provide probability estimates these are calculated using five fold cross validation

d. **Decision Trees**:

It is Graphical presentation of all possible solutions to a decision based on certain condition. These decision trees are easy to read and understand. It starts with a root and keeps on growing with conditions and decisions. In this project we have used decision tree so that we can easy understand why computer has given this outputs as is very easily understandable based the conditions

Steps used for construction of a decision tree :

1: Start the tree from the root node, example R, that has full data.

2: Using Criterion like ‘gini’ or ‘entropy’ measure, find the best attribute .

3: Now R can be divides into subsets with best attributes

4: Then we have generated decision tree node which has best attributes.

5: By using the subsets created from the dataset recursively construct decision tree.

Repeat the procedure till a stage is reached where you will not be able to further classify the nodes and called the last/end node as a leaf node.

# CHAPTER 4

# RESULTS

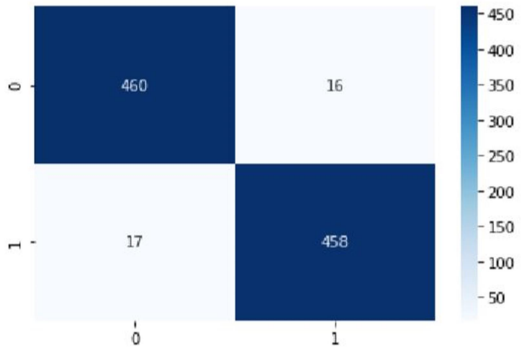
## 4.1. Accuracy

| Algorithm used | Training set | Test set |
|---|---|---|
| 1. Logistic Regression | 0.9779 | 0.9653 |
| 2. Random Forests | 0.9995 | 0.9653 |
| 3. Support Vector Machines | 0.9892 | 0.9737 |
| 4. Decision Trees | 0.9829 | 0.9611 |

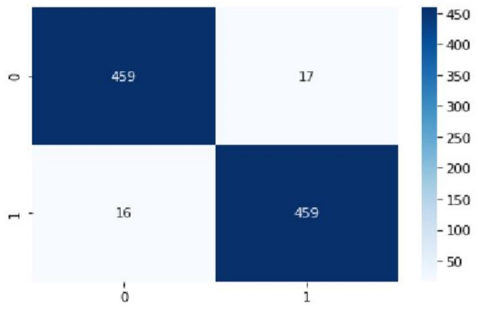
Table 4.1 Accuracy of algorithms

## 4.2. Confusion matrix of algorithms

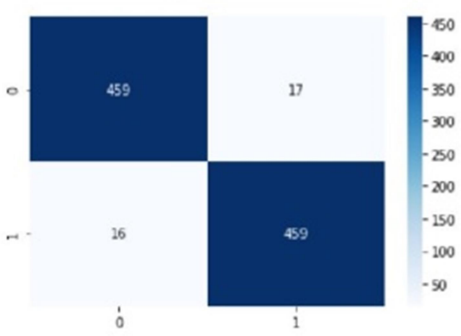
a.LogisticRegression:




b. Random Forests:




c.Support Vector Machines




d. Decison Tree

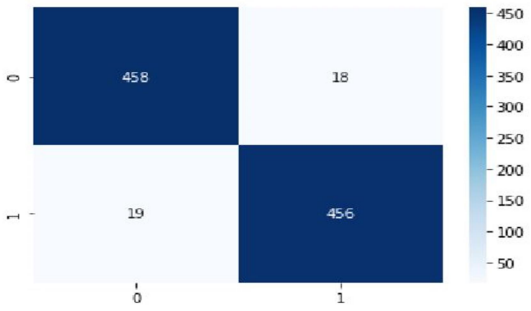



Figure 4.1 Confusion matrix of algorithms

| Algorithm | Precision | Recall |
|---|---|---|
| Logistic Regression | 0.9652 | 0.9652 |
| Random Forest | 0.9652 | 0.9652 |
| SVM | 0.9652 | 0.9652 |
| Decision trees | 0.9610 | 0.9610 |

Table 4.2 Precision and Recall

## 4.3.Decision tree formed with this dataset and Challenges:

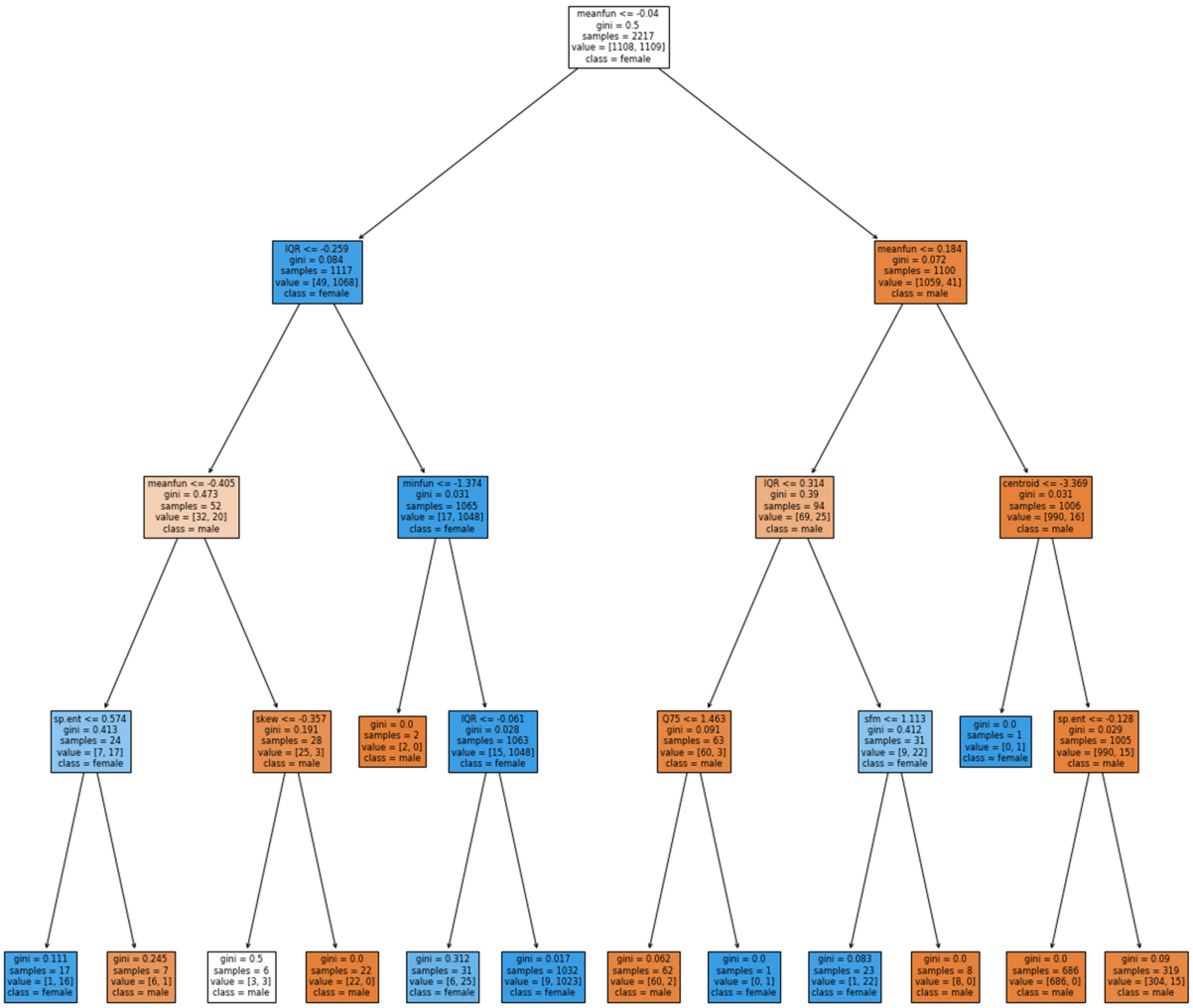



Fig.4.2 Decision tree formed with this dataset

Challenges:

After evaluating the model results, We decided to try to extract the features from raw voice sample to use them as test samples for our model. As the datset we have choosen was preprocessed using R.

However, python exclusively does not provide a library to get those features from voice samples. We wrote code using the formulae available and could able to derive 9 audio spectral statistical features. But it is not as good enough to move further to predict gender.

## 4.4 Feature Importance

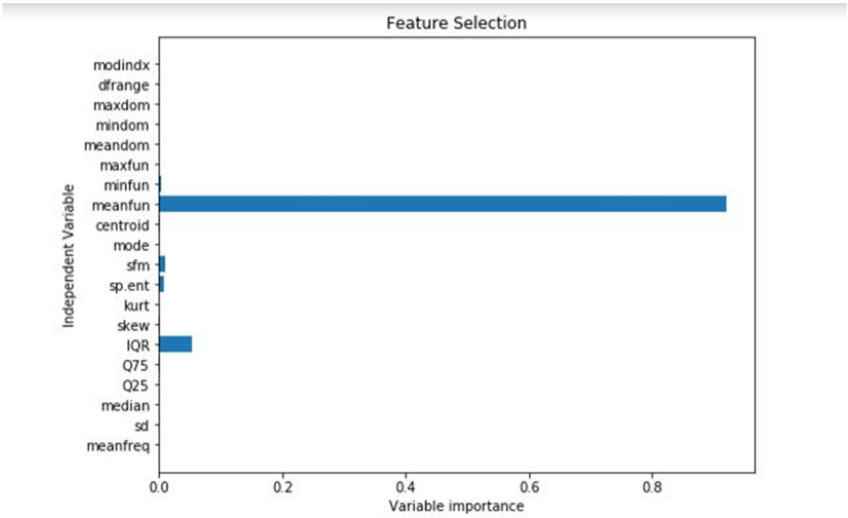



Figure 4.3 Feature Importance in decision trees

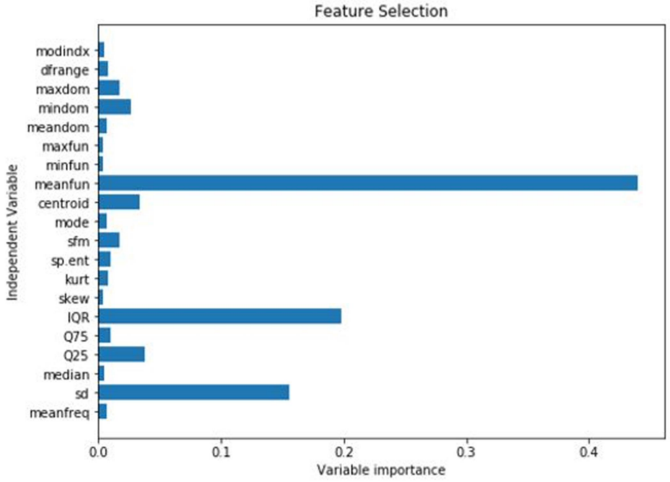



Figure 4.4 Feature Importance in random forest

# CHAPTER 5

# CONCLUSION AND FURTHER WORK

## 5.1. CONCLUSION

In this work we utilised supervised learning classification algorithms.The process of selecting best features among all features plays a major role in designing a gender identification system. From all the features mean function, spectral entropy and IQR results in best classification rate. From the results obtained SVM algorithm performs best classification with less error rate and with accuracy of 97.37%. The results obtained are only for this dataset and varies for other dataset.

## 5.2.FURTHER WORK

Our further work is focused on improving prediction(test)accuracy by applying several other algorithms. We further like to expand our project by collecting raw data voice samples preprocessing them, and then performing feature extraction. Also further we would like to implement our project with hardware.

# REFERENCES


[1] R Praveen Kumar, P Sree Varsha , L Sandhya Rani, G Bharadwaj., "Gender Prediction by Voice using Logistic Regression". IJRASET Volume 7,Issue IV. Apr 2019
Available at : http://ijraset.com/fileserve.php?FID=21756

[2] Whiteside S P., "Temporal-Based Acoustic-phonetic Patterns in read speech". International Phonetic Association 26 23-40 1996.

[3] Zeng Y M,Wn Z Y,Falk T and Chan W Y., "Robust GMM based gender classification using pitch and RASTA-PLP Parameters of speech".Proceeding of the International Conference on Machine Learning and Cybernetics 3376-3379 2006.

[4] Udry J. R., "The nature of gender," Demography, vol. 31, pp. 561-573, 1994.

[5] Ting, H, Yingchun, Zhaohui, W. "Combining MFCC and Pitch to Enhance the Performance of the Gender Recognition", IEEE 2006.