A Voice Signal Based Gender Prediction Model Using Random Forest Classifier

Group members:

Saif Ahmed - 12201021

Sajjad Hossain - 13101252

Gazala Chowdhury - 13201036

Maliha Mehnaz - 13301015

Supervised by: Dr. Jia Uddin

Assistant Professor

Department of Computer Science and Engineering,

BRAC University

Date: 25 March, 2018

DECLARATION

We, cordially declare that every result, research included in this thesis is found by the experiments and work done by ourselves. We took some references from other researchers which are mentioned in the reference section. This Thesis, neither in whole or in part, has been previously submitted for any degree.

Signature of Supervisor

Dr. Jia Uddin

Signature of Author

Saif Ahmed

Sajjad Hossain

Gazala Chowdhury

Maliha Mehnaz

ACKNOWLEDGMENTS

In the beginning we would like to thank the Almighty ALLAH the creator and protector of this world. Our gratefulness to him is beyond limit as He provided us with wellbeing and abilities, for which we could complete our thesis successfully. May He be glorified and exalted for the blessings that He has bestowed upon us in both spiritual and worldly term.

Even though thank you cannot express the unconditional support and guidance Dr. Jia Uddin, our thesis supervisor, provided us with throughout our project, we would like to thank Sir from the bottom of our heart. Without Dr. Jia Uddin Sir, our thesis project wouldn't have seen the light of success.

We can never forget the help we received from the BRAC University Faculty Staffs of the Computer Engineering & Engineering Department. They have been with us throughout the whole journey and were right beside us for guidance any time we needed them, particularly in developing our basic knowledge and enhancing our capabilities. Thank you for always being there.

At the end, we would be proud to express our thankfulness and appreciation to our respectable parents, brothers and sisters for their motivation and support. We are grateful to all of our family like friends who helped to complete this project.

CONTENTS

LIST OF FIGURES

Abstract

In the proposed model, Classification and Regression Tree (CART) was used as a classifier to classify gender using four different algorithms which were tested with changing dataset frames, layer sizes and samples to get best options for our model. We had to tune our dataset with Principal Component Analyzer(PCA) which improved the accuracy rate a bit and also worked along with the algorithms. The intelligible idea of voiceprints and human-computer interaction gave us the motivation to predict gender by using different proposed classifiers that we are using in our model .Besides the overall efficiency and outcome of human-computer interaction gave us the inspiration to select this model for our thesis paper. In this existing system there are quite a lot of problem that arose while dealing with our proposed model those are over fitting of the dataset, having different layer sizes, number of decision tree and most importantly solving the hidden layer sizes. We did successfully solved most of the problems by running five different algorithms on our model which are Decision Tree Classifier, Logistic Regression, Support Vector Machine (SVM) , Multi-Layer Perceptron Classifier (MLP) and Random Forest (RF) Classifier. To use the total dataset on this algorithm we used 75% training and 25% testing of the total dataset. Due to different layers we had different accuracy result for each of the algorithms. The worst accuracy result was given by Multi-Layer Perceptron (MLP) which was 75% in two implementations and the best accuracy result was given by Random Forest Classifier which was 97.34 % from our proposed model.

Keywords: *Classification and Regression Tree (CART), Logistic Regression, SVM (Support Vector Machine), Random Forest, MLP (Multi-Layer Perceptron, Principal Component Analysis (PCA)*

CHAPTER-1

Introduction

1.1 Motivations

Among the means of communication voice is the main media of communication. When a sound generates from a sound source it vibrates the air molecules. This vibration with the help of some media reaches human ear and depending of the sound frequency human beings can distinguish the voice from different speaker [2]. A voice contains linguistic information of a speaker and with this human beings can verify an even an unknown speaker (gender, age, origin). Therefore, voice features are considered as voiceprints [13]. In this proposed work we extract features from a voice signal and these features are then used to determine the gender of the speaker. This system of gender recognition can be leveraged upon in some very useful applications, for instance it can provide more targeted services based on gender interoperability. Moreover in HCI (Human-computer interaction), this system could get the user interface scope and develop the experience in most Internet of Things applications. These information can be used to for gender customizations in such Internet of Things apps and extend the security in these applications [1].Even though we don't realize but it is important to identify the speaker's gender automatically with the help of voice signals. First of all, the interactive information system can be developed with user gender identification as it can automatically select the proper interaction service for different genders. In human-machine interaction, the system service is unlike for both male and female, for example,

interface styles and color. For this reason, the differentiation of gender using such voice authentication would improve the overall efficiency and outcome of human-computer interaction. Second of all, in social networking system, we can control access through the gender determination of user. This can also reduce the advertising cost by narrowing down and targeting the exact market [13].

1.2 Contribution Summary

For our project feature selection is one of the most important factors in designing a gender recognition system. From the study of different previous research works it was observed that among the different features the power spectrum results in best classification rate. PCA was used to select the most important and dominant features. PCA is a statistical and mathematical technique. Identification of patterns in data and expressing the data in such a way that highlights the difference and similarities between the data is done by the principal components analysis. Since it is a very difficult task to find the patterns in data, the main objective of the principal components analysis is to reduce the dimensionality of the given data and it can also use for the data compression. The largest dimension of the data space has reduced into the smaller natural dimension of the feature vector is the main purpose of the principal components analysis.

1.3 Thesis orientation

- Chapter II includes the necessary background information regarding the proposed approach of the Gender recognition system.

- Chapter III includes the methodology of our new proposed model for gender detection using audio with the help of PCA and Classifier.

- Chapter IV includes the experimental analysis, results of the new efficient model which shows the performance.
- Chapter V discusses about the future work and concludes this thesis in short.

CHAPTER-2

Background Information

2.1 Dimensionality Reduction

The basic working procedure of our desired goal is to reduce the dimensions of a d-dimensional data-set by culturing it into a k-dimensional subspace (where k<d).This method is used to ensure the efficiency to retain most of the information. While working with the data-sets we will ignore the class labels and we will be using multiple matrix data-sets in this model referred as d-dimensional data-sets. We need find the mean of each and every dimension of the whole data-set. We need to work on the mean values in order to get a good subspace and this good space will be helping us to get a good. Next we need to apply the covariance matrix for better distribution of data-set and it's better to avoid the scatter matrix. Another good point of using covariance matrix is it has its very own built-in function numpy.cov().

2.1.1 Principal Component Analysis (PCA)

Principle Component Analysis is a mathematical procedure to analyze a set of observation of correlated variable converged into a data file to convert it into uncorrelated variables, which brings out the strongest patterns from those set of correlated variables [8]. In one sentence we can say that it is a reduction tool used to reduce large set of variable designed in a dataset file into small set of variable. Normally it reduces the linear dimensionality into singular value which

is the decomposition, in order to lower down the spaces between the variables of the required dataset file. Basically it can be described as the variance convergent structure used for a set of variables of correlated variables. We can claim that it retains most of the information while converting maximum variance in high-dimensional data onto a smaller dimensional of that required data-set.

2.2 Regression

Random Forest Classifier is an estimator that can control over fitting and improve the predictive accuracy rate by averaging and sub-samples of dataset. The procedure of analyzing a specific outcome for various continuous-valued attribute associated with various variables in data-set compiled in a CSV, it is a type of spreadsheet used in relational database application that be put in a table oriented application so that we can present it as an input. File is the main objective of a regression. This methodology works with the regression problems that are faced while working with classifiers of analyzing the various variables in our algorithm [10]. For solving the regression problem in our model we are using Classification and Regression Tree (CART).

2.2.1 Classification and Regression Tree

The most commonly used Decision Tree Algorithm is the Classification and Regression Tree. In CART (Classification and Regression Tree) it splits each value or node into two child nodes which makes it as a Binary Decision Tree. Each root node represents a single input variable (x) and a split point on that variable. So, when this algorithm needs to make any decision, it splits

the value or parent according to the level or its child [11]. The Figure-1 illustrates the analyzation of trees by using CART

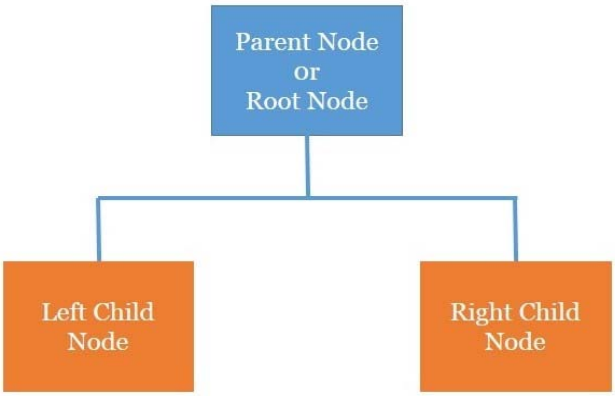


Fig-1: CART Analyzing Tree

There are following two types of decision trees such as:

- Classification Trees: If the variable is categorical and when the tree is used to define the classes then Classification Trees are used. It is used for the discrete values.
- Regression Trees: In Regression Tree, it is commonly used for the continuous variable and only to identify the approximate value.

In CART the main abstracts are-

- It needs when to split a data in a node according to a value of one variable.
- When there is no more data in a branch then it stops
- Finally it figures out the desire variable in each terminal.

CART Algorithm Steps: Decision Tree building algorithm involves a few steps are described below-

- Take Labeled Input data: It works with a target variable and the list of independent input variable. Here we are using more than 3000 samples of audio files both of the male and the female sample. We are using the data sample as a 'CSV' format.
- Best Split: For each of the independent variables Find Best Split
- Best variable: Select the Best Variable for the split
- Split the input data: Into Left and Right Nodes
- Decision Tree Pruning: Steps to prune Decision Tree built

2.2.2 Classifier

With the help of various classifiers we were able to train our data set and have observed various co-related identification among the multiple features which are organized in our data-set. The main purpose of using different type of classifier algorithms in our model is to check which algorithm would give us the best accuracy result by training our data-set with each of the classifier algorithms. The next is to compare those results against our data-sets which are compiled in a CSV file and get the best outcome for our model. In this chapter I would like describe in brief about the different classifier we are using for our model.

2.2.2 Logistic Regression

This classifier is very explicit which always represents two outcomes either it's either a yes or a no. This type of statistical classifiers always tends to deal with the probability of any model by vigorously analyzing the required data-set. We can also call this a binomial logistic regression which deals with the

dependent variable in the data-set and predicts a result of having a failed outcome if we use the independent variables instead of the dependent variable while working with the data-sets.

2.2.4 Support Vector Machine

Now coming over to Support Vector Machine commonly known as SVM is more of a non-probabilistic linear classifier. This is a much supervised linear learning algorithm which works with high-dimensional variable feature in the required data-set [3]. Basically it maps a space as wide as possible between the different categories in the data-set and creates a new space by creating a new category among the variable in the data-set and compares between the two spaces to convert it into a high-dimensional linear space [9]. It is not a supervised learning system and it mainly works in a nonlinear classification way. The Figure-2 shows us the two-dimensional spaces between two non-linear classifications in SVM

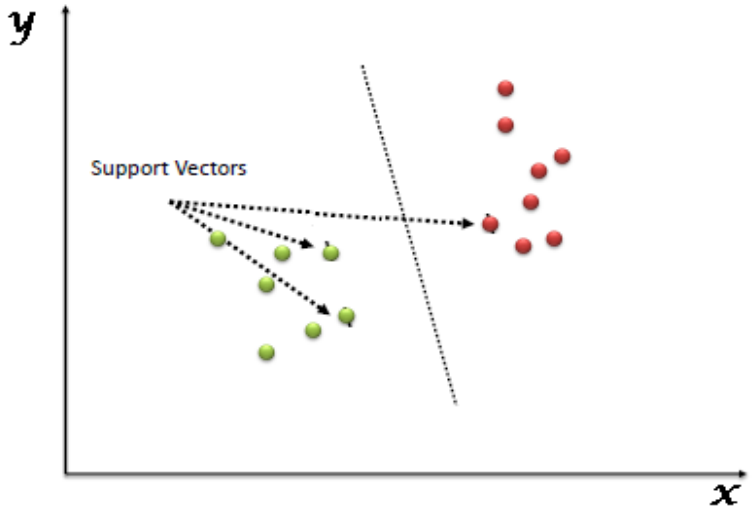


Fig-2: Non-Linear Classification In Two-Dimensional Spaces.

The neural networks in a SVM classifier works with the mode process and finds non-linear signals and associates them with an appropriate class after analyzing the classifier in a proper neural network

approach [4]. Neural networks are tending to be a quadratic programming and are not that much of over fitting as that of the result. SVM are tend to be having fewer parameters but Neural Networks are always been well known for being great in problem solving for single models and its approach towards any data-set is very accentuate [14,15].

2.2.5 Random Forest

Random Forest Classifier is a estimator that can control over fitting and improve the predictive accuracy rate by averaging and sub-samples of dataset. Among all the classifier that we will be using, this is the most powerful machine learning classifier relatively known for working with more conventional pattern recognizer. Basically when this classifier works its finds out the various number of decision tree classifier by estimating with a very different meta estimator and at last it predicts an result by averaging them and it also look overs the over-fitting of the data-set. With the help of bootstrap which is always true by default it maintains the size of the sub-sample and the original one of the required data-set. So in short in our working procedure we will be seeing that, the classifier is mainly working with leaf node, weight fraction of the leaf node , impurity of the split, decrease in the impurity of the split and the random state generator.

2.2.6 Multi-Layer Perceptron

MLP classifier commonly known as Multi-Layer Perceptron Classifier deals with the logarithmic function of the data-set and it can be found out with the exclusive help of incremental gradient descent. Basically the classifier finds out the accuracy of our model by penalizing the false classification. Incremental gradient descent deals with the negative gradient of the various variables found in the data-set and it also has a very strong use in the field of neural networks because of its fast convergence .In each loss of function the MPL classifier trains the data repeatedly with help of the modal parameters in order to

give us a better result of accuracy [18,19]. This classifier prevents us from over-fitting by regularizing the loss function of the required data-set by shortening down the modal parameters. While analyzing the data-set we need to convert the variable in the data-set into a sparse scipy array of floating point values in order to properly implement the MPL classifier [20]. The Figure-3 showcases the one hidden layer in the classifier

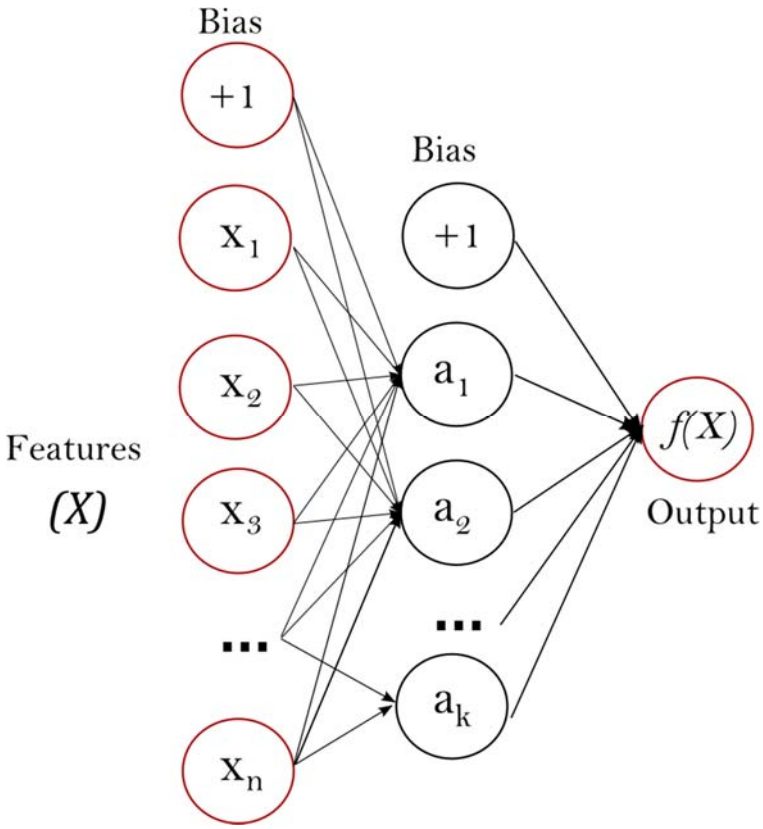


Fig-3: One Hidden Layer MLP

This is the least output we have found among all the classifier we have worked up to but this is the easiest way to solve our model and also the last classifier we have worked with in our model.

CHAPTER-3

Proposed Method Methodology

3.1 Introduction

In our proposed model we have worked with the Classification and Regression Tree which is basically a Binary Decision Tree. Feature selection is just deciding which variable to include in our model. In case of CART feature selection is done by the model itself. CART does the feature selection, so the more variables/features can be discarded before running the CART algorithm, the higher the chances that we discard an important variable. This leads to a loss in performance. For CART it does not matter if we use many number of features, it searches for the best of those features. We are using a CVS file of more than 3000 audio files of male and female sample having 16 properties. By Principle Component Analysis (PCA) we sort out the most important properties which are more distinguishable from each other. We are working with some different types of model which are Logistic Regression model, Support Vector Machine, Random Forest Classifier and Multi-Layer Perceptron Classifier [24]. Among those four models Random Forest Classifier give us the best accuracy. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes. Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression. Here the neural network model is Multi-Layer Perceptron Classifier that maps sets of input data onto a set of appropriate outputs. While

working on it we had use 25% data for testing and 75% data for training and the Figure-4 showcases the flowchart of our model's methodology :-

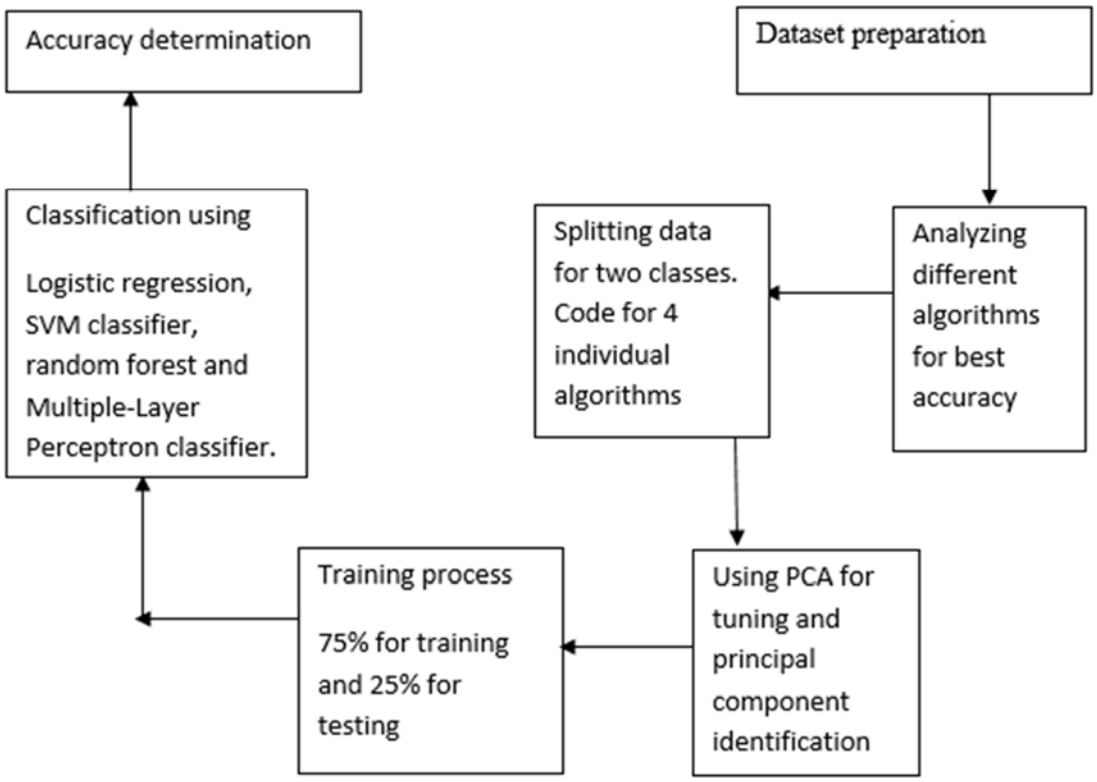


Fig-4: Flowchart of Methodology

After reviewing the flowchart it might seem confusing to understand the flow of work, so we would like to give a brief review on our work flow along with the description of each sections working procedure. The very first step of our model was the data preparation and working on procedure of analyzing the best possible way to put our data in such an efficient way that it may be workable so that it can be easier to implement it properly .Then we will be moving towards the analysis of different algorithms to find out which one would be suitable to work with or which algorithm would give us the best accuracy rate. Next we will be splitting the data into two classes which are known as male and female and we will be

applying four individual algorithms to get the best accuracy rate. After that we will be using principal component identification for tuning the data-set to get the best possible pattern from analyzing the required data-set .Now basing on the data-set we will be processing the data-set in such a way that 75% of the algorithm will be learned by the data-set and then 25% will be pre-processed so that we can test the data-set with our required algorithm. Finally we need to observe the data-set and analyze the data-set and find the best training set or classification method. At last we will be checking all the accuracy rate against all the classifier and determine the best accuracy rate.

3.2 Processing of Data-Set

The data-sets are collected by measuring various acoustic properties of various male and female voice files collected from various speeches and voice clips .These voice clips are pre-processed by analysis of acoustic properties of these voice clips. The noise have been reduced from these voice clips and pre-processed .The data-sets are consist of nearly 3246 recorded voice samples .The different properties that we found in each voice files are illustrated in the Table below of data which were collected in the CSV file that are showcased in the figure-5

	meanfreq		sd		median		Q25	
	Male	Female	Male	Female	Male	Female	Male	Female
3	0.059781	0.158108	0.064241	0.082782	0.032027	0.191191	0.015071	0.06235
4	0.066009	0.182855	0.06731	0.067789	0.040229	0.200639	0.019414	0.175489
5	0.077316	0.199807	0.083829	0.061974	0.036718	0.211358	0.008701	0.184422
6	0.151228	0.19528	0.072111	0.072087	0.158011	0.204656	0.096582	0.180611
7	0.13512	0.208504	0.079146	0.05755	0.124656	0.220229	0.07872	0.190343
8	0.132786	0.168426	0.079557	0.071344	0.11909	0.189051	0.067958	0.158394
9	0.150762	0.179982	0.074463	0.077964	0.160106	0.204277	0.092899	0.127086
10	0.160514	0.204518	0.076767	0.068569	0.144337	0.216026	0.110532	0.19223

	Q75		IQR		skew		kurt	
	Male	Female	Male	Female	Male	Female	Male	Female
3	0.090193	0.224552	0.075122	0.162202	12.86346	2.801344	274.4029	19.92962
4	0.092666	0.226068	0.073252	0.050579	22.42329	3.00189	634.6139	19.86548
5	0.131908	0.235687	0.123207	0.051265	30.75715	2.543841	1024.928	14.92196
6	0.207955	0.255954	0.111374	0.075344	1.232831	2.392326	4.177296	10.06149
7	0.206045	0.249759	0.127325	0.059416	1.101174	1.707786	4.333713	5.670912
8	0.209592	0.216642	0.141634	0.058248	1.932562	3.020346	8.308895	17.79122
9	0.205718	0.240377	0.112819	0.113291	1.530643	2.76467	5.987498	12.82216
10	0.231962	0.255276	0.12143	0.063046	1.397156	1.918543	4.766611	7.537126

	meanfun		minfun		maxfun		meandom	
	Male	Female	Male	Female	Male	Female	Male	Female
3	0.084279	0.185042	0.015702	0.023022	0.275862	0.275862	0.007813	0.272964
4	0.107937	0.15959	0.015826	0.018713	0.25	0.266667	0.009014	0.25897
5	0.098706	0.156465	0.015656	0.016194	0.271186	0.266667	0.00799	0.250446
6	0.088965	0.182629	0.017798	0.024922	0.25	0.275862	0.201497	0.269531
7	0.106398	0.162043	0.016931	0.016807	0.266667	0.262295	0.712813	0.260789
8	0.110132	0.168343	0.017112	0.015936	0.253968	0.271186	0.298222	0.253125
9	0.105945	0.162259	0.02623	0.016842	0.266667	0.275862	0.47962	0.207332
10	0.093052	0.163077	0.017758	0.016754	0.144144	0.275862	0.301339	0.205

sp.ent		sfm		mode		centroid	
Male	Female	Male	Female	Male	Female	Male	Female
0.893369	0.952161	0.491918	0.679223	0	0.049926	0.059781	0.158108
0.892193	0.910458	0.513724	0.506099	0	0.05002	0.066009	0.182855
0.846389	0.904432	0.478905	0.425289	0	0.049961	0.077316	0.199807
0.963322	0.907115	0.727232	0.524209	0.083878	0.193435	0.151228	0.19528
0.971955	0.879674	0.783568	0.343548	0.104261	0.050165	0.13512	0.208504
0.963181	0.901022	0.738307	0.491709	0.112555	0.050073	0.132786	0.168426
0.967573	0.927424	0.762638	0.580274	0.086197	0.241551	0.150762	0.179982
0.959255	0.896919	0.719858	0.409583	0.128324	0.050066	0.160514	0.204518
0.970723	0.902083	0.770992	0.449973	0.219103	0.050062	0.142239	0.198292

	mindom		maxdom		dfrange		modindx	
	Male	Female	Male	Female	Male	Female	Male	Female
3	0.007813	0.046875	0.007813	0.742188	0	0.695313	0	0.339888
4	0.007813	0.054688	0.054688	0.804688	0.046875	0.75	0.052632	0.269231
5	0.007813	0.054688	0.015625	0.898438	0.007813	0.84375	0.046512	0.329521
6	0.007813	0.054688	0.5625	0.703125	0.554688	0.648438	0.247119	0.294717
7	0.007813	0.054688	5.484375	0.8125	5.476563	0.757813	0.208274	0.251546
8	0.007813	0.054688	2.726563	0.914063	2.71875	0.859375	0.12516	0.407792
9	0.007813	0.054688	5.3125	0.546875	5.304688	0.492188	0.123992	0.309841
10	0.007813	0.046875	0.539063	0.757813	0.53125	0.710938	0.283937	0.235348

Fig-05 : CSV files of Data-Set

All these properties are firstly written in excel worksheet than they were converted into a CSV file which will help the classifiers that we are using to implement it properly.

3.3 Methodology

Our first approach was to use the dataset in most effective way to determine best accuracy or prediction results for testing data. This could be done using linear regression algorithms and co-relation of mean fundamental frequency but, there was possibility of biased output and low accuracy.

The very first classifier that we have used for our training process is the Logistic Regression and we did had a satisfactory output for this classifier but the process had to be done by few of the following process which are stated below

- By analyzing the sample we had to predict a pretty good confidence score.
- By reading the training data again and again we had to fit the model as per the training data.
- For the estimator we need the best possible parameters.
- In the samples of X we had to predict class labels.
- We need to calculate the log of probability.
- Finally on the given test data and label we need to calculate the mean accuracy.

These are the steps we used for process the Logistic Regression classifier.

For the second classification we have used Support Vector Machine (SVM) which mainly works with the effectiveness of the high dimensional spaces [13]. It also uses a subset of training points

to finds its decision function which is memory efficient which the benefits of using this classifier.

This classifier is of no use if we face over-fitting of our data-set that is if the number of features are greater than the number of samples. But such problems didn't occur in our data-set. Our data-set is a multi-class classification so we can use both dense and convertible classes as per our requirement. Another problem of our model is solved by SVM as it is quadratic programming problem. The following equation has been used to check the complexity of the data-set as per the application of the SVM. :

O (n_ {features} \times n_{samples}^3)……(1)

The support vectors which are the subsets of the training data are the functions which decide the decision function [16]. At last the decision function gives a perceptive score for each of the sample and then the probability function decides the accuracy rate by the application of this algorithm.

The next classifier is very work efficient we didn't have to work on it much comparatively to the other classifier which is MLP [7] [22]. It only purpose is to learn a built-in function as per the data-set and work on it .Basically it is a supervised learning algorithm. The function that we will be using for learning the data-set is given below:

$$f(\cdot) : R^m \rightarrow R^o \quad \text{……..…(2)}$$

Here "m" is the number of dimensions in the input and "o" is the number of dimensions we will get as an output. MLP has to decide either classification or regression by targeting a set of features as it also works as a non-linear function too. This classifier has many hidden layers in between its output and input layer.

The best accuracy result that we have found is from Random Forest Classifier and its efficiency result is on the best output we would have found till now [17].This classifier really works on many of the decision tree and it can also work on many sub-sample sizes and it can also control the over-fitting so that it can predict the best accuracy rate [20]. First it will control the values of the parameters and the can be put in the parameter values and split it by training the data-set. It also checks the behavior and then controls the fitting of the data-set. This classifier also has given us the best possible outcome for our accuracy rate.

Algorithms were compiled and executed in Anaconda along with Py-charm libraries. Each of the codes for individual algorithms were tested with changing dataset frames, layer sizes and input samples and best options were chosen for experiment and analysis. Then the dataset was tuned with principal component analyzer what slightly improved the accuracy rate and worked along with the algorithms.

At first we need to Import the CSV file and frame the data. We use the compiler anaconda for our entire classification. In the next coding part regression happens. For that using the test size whose value is 0.25 so that we can split the data frame. We need to label the X axis. Similarly for Y axis, we need to label the axis. Moving on, we need to create a decision tree classifier and fit the X & Y axis. Next we use method "predict" and get the initial prediction. Then using score and other methods we get and then print the accuracy score of decision tree classifier. After observing all the classifier we need to found out which of the classifier has given us the best output and the best output is given by Random Forest Classifier.

CHAPTER-4

Experimental Analysis

4.1 Introduction

In order to fulfill the purpose of creating the best model as a classifier to identify a voice sample as of Male or Female from our dataset, we tried 5 different algorithms on this. These are: Decision tree classifier, Linear support vector machine, Multi-layer preceptor, (MLP), Logistic regression and finally Random forest classifier. All these classifiers gave us fairly good accuracy results except of MLP. Among these Random forest classifier was chosen as the best model for purpose depending on accuracy and efficiency.

4.2 Analysis of the dataset

This dataset was created to predict a voice as male or female, based upon acoustic properties of the voice. The dataset contains almost 3000 recorded voice samples, collected from male and female speakers, in different environment. The voice samples are pre-processed by acoustic analysis in R, which can also be done with python using the seewave and tuneR packages, (externally imported library in Anaconda) with an analyzed frequency range of 0hz-280hz. Few of these attribute columns were numerically tuned for the ease of implementation of the proposed algorithms. There are 21 attributes in the dataset with an identified label for each row. The machine learning algorithm was trained with all of those but principal component identification was required for efficiency of the algorithm.

A simple MATLAB graph plotting was done to visually show the importance and variances of the attributes. Two blue and red represent the plot of male and female respectively. The different properties that we found in each voice files are

1. Mean frequency

2. Standard deviation of the found frequency

3. Median frequency

4. First quartile of the data-set

4. Third quartile of the data-set

5. Interquartile range between the first and third quartile

6. Skewness which is the statistical distribution of the data-set

7. Kurtosis is the sharpness of the peak of the frequency

8. Spectral entropy is the density of the signal

9. Spectral flatness is the dealing with noise or tone in the data-set

10. Mode frequency of the voice samples

11. Frequency centroid of the voice samples

12. Peak frequency of the voice samples

13. Average of fundamental frequency which is measured by every signal found in the voice samples

14. Minimum fundamental frequency which is measured by every signal found in the voice samples

15. Maximum fundamental frequency which is measured by every signal found in the voice samples

16. Average of dominant frequency which is measured by every signal found in the voice samples

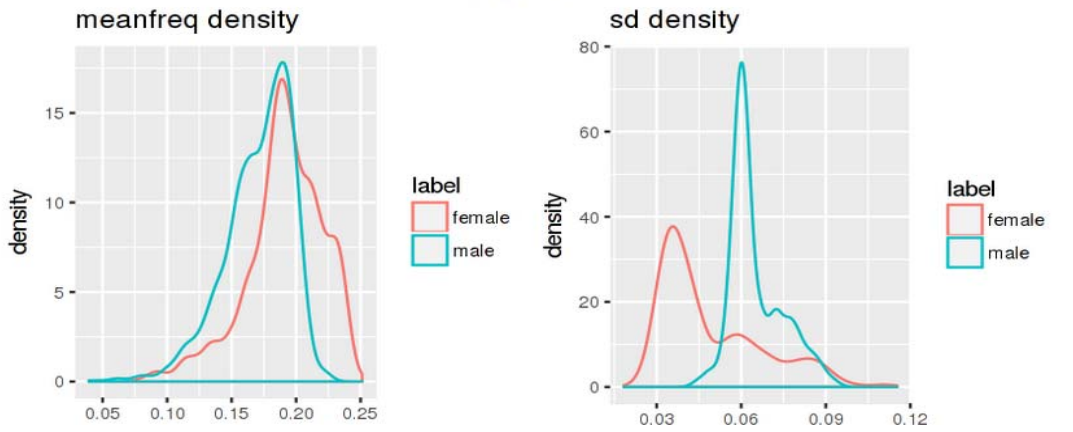
17. Minimum of dominant frequency which is measured by every signal found in the voice samples

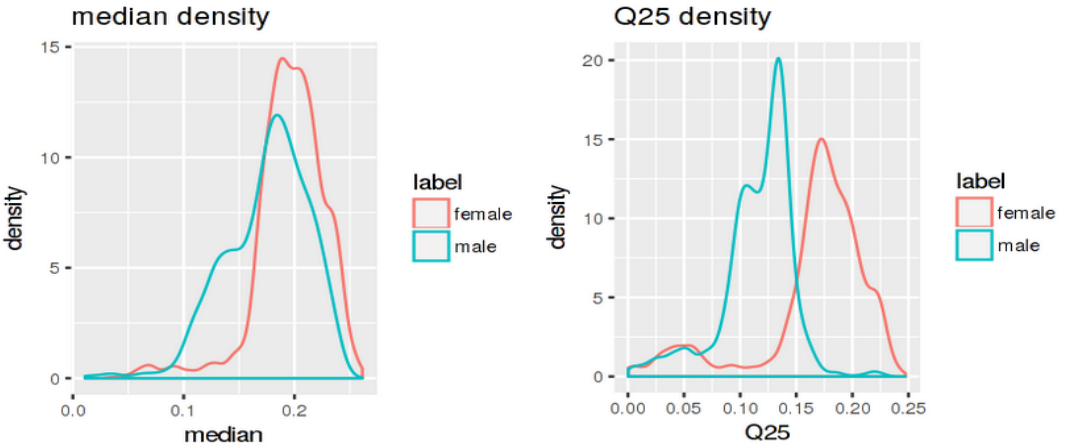
18. Maximum of dominant frequency which is measured by every signal found in the voice samples

19. Range of dominant frequency which is measured by every signal found in the voice samples

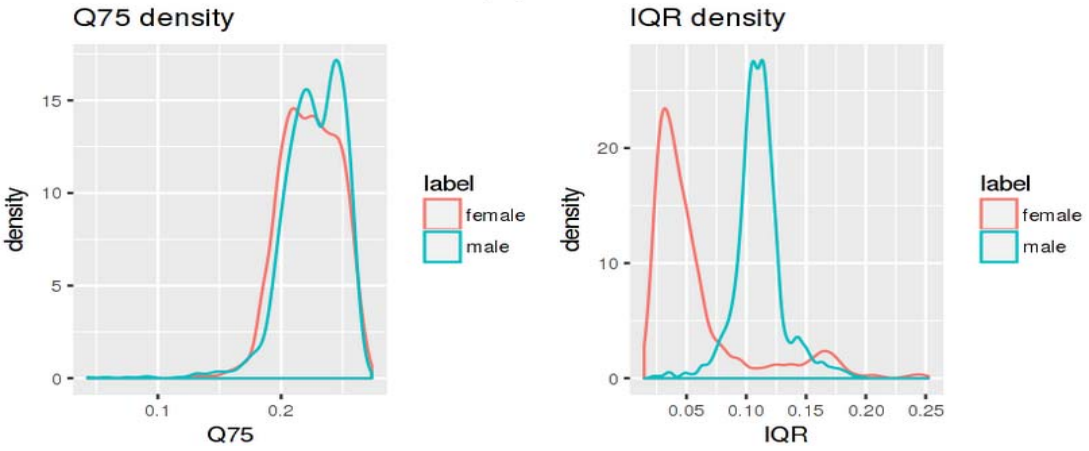
20. Modulation index is the calculation between the differences between fundamental
21.frequencies divided by the frequency range

22. Label to predict whether it is male or female


meanfreq density
density
label
female
male
sd density
density
label
female
male


median density
density
median
label
female
male
Q25 density
density
Q25
label
female
male

page 1 of 1


Q75 density
density
Q75
label
female
male
IQR density
density
IQR
label
female
male

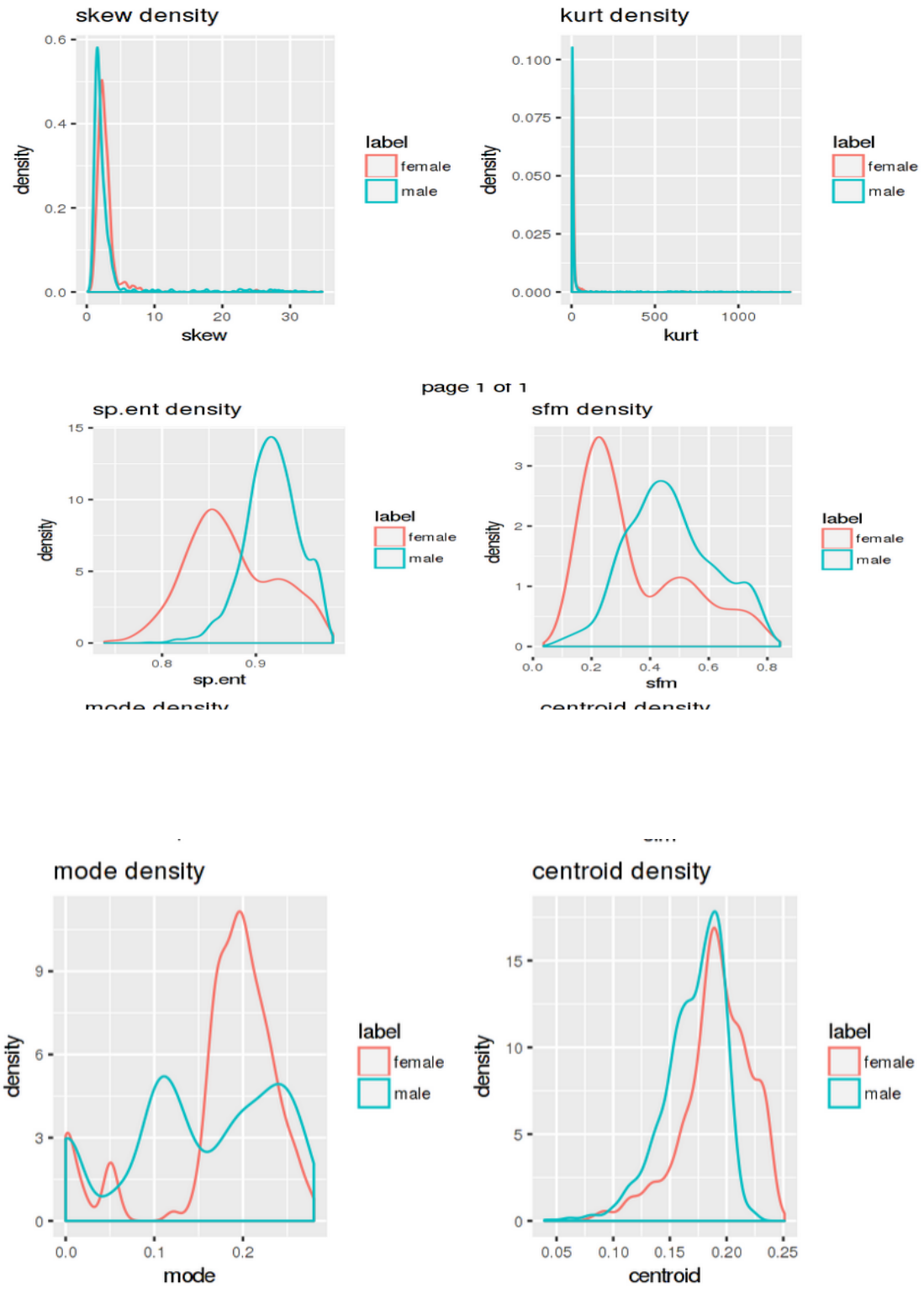


Fig-6: Graph plot of acoustic properties (visualizing importance)

4.3 Advantages of the Random forest classifier

The advantage was mainly from the branching of the trees. If the properties that contribute most in distinguishing the samples, having larger variance of the mean of the values of male and female , are in the front columns , then the identification will be easier and faster and the classifier will learn efficiently.

It puts the sample in the expected class depending on the most important property and later changes the decision if necessary depending on less important property vary with the learned sample largely.

Random forests overcome several problems with decision trees and as a result with the overall accuracy regarding the training part. Firstly this algorithm Reduces over fitting: by averaging several trees, solving correlation of mean variance. Thus, there is a significantly lower risk of over fitting in the process. Another specific advantage is less variance. By using multiple trees in tree building process, it reduces the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data.

Random forest is an ensemble method in which a classifier is constructed by combining several different *Independent* base classifiers. The independence is theoretically enforced by training each base classifier on a training set sampled with replacement from the original training set. This technique is known as bagging, or bootstrap aggregation. In Random Forest, further randomness is introduced by identifying the best split feature from a random subset of available features.

The ensemble classifier then aggregates the individual predictions to combine into a final prediction, based on a majority voting on the individual predictions.

It can be shown that an ensemble of independent classifiers, each with an error rate e, when combined significantly reduces the error rate.

For example if we consider 10 impendent classifiers, each with error rate of 0..2

$$\epsilon = 0..2$$

In this setting, the error rate of the ensemble can be computed as below , considering majority of this is the predictions.(An ensemble makes a wrong prediction only when more than half of the base classifiers are wrong)

$$\epsilon ensemble = \sum i = 10, i = 6, (10 i)\ \epsilon i (1-\epsilon) 10 - i \approx 0.04$$

It can be seen that with the theoretical guarantees stated above an ensemble model, in this case random forest performs significantly well.

More importantly, an important disadvantage of using a single decision tree was that decision trees tend to be prone to over fitting the training data. A random forest creates lots of individual decision trees on a training set. It's expected that each of the individual trees in a random forest should do well at predicting the target values in the training set but should also be constructed to be different in some way from the other trees in the forest (randomness). This difference is calculated by introducing random variation into the process of building each decision tree.

Also, the computational cost of training a random forest is quite low.

Disadvantages

1. AS THIS Algorithms deterministic, it leaves less option for interpretability
2. In the process, over fitting might occur with wrong requirements
3. Number of trees has to be chosen manually.

But in the proposed model of ours, Random forest is used as it fits the requirements, the number of decision class is less with large number of samples and attributes.

4.4 Experimental analysis

In the beginning, all of the stated classifier models were trained and tested for 75% and 25% of the total dataset respectively. Then in different layers of our experiment it was changed to 70% with 30% and 80% and 20%. Individual processes are described briefly below.

Decision tree:

In the first implementation of decision tree algorithm, it was trained with 75% of the data and tested with 25%. This came up with an accuracy result of 96.21%, which was fairly satisfying but not enough. According to our analysis, the result was expected to rise with another algorithm also with increased efficiency in case of dealing with larger dataset.

SVM:

Support vector machine algorithm was tested for the same amount of training and testing data.It came up with a accuracy score of 92.04% in the first test and 95.20% in the second test. The change in result was brought randomly and then changed drastically when the amount of test set was changed, it began to fall.

MLP:

MLP or Multilayer perceptron was used as a neural network model which did not give good score in testing for the same portion of data with (5,2) hidden layer size. It was also not suitable

for this type of two class prediction model according to our analysis. However it gave an accuracy result of around 75% in two implementation. And drastically dropped as the layer size was increased. The Figure-7 illustrates the screenshot of the code we run and the accuracy result after that

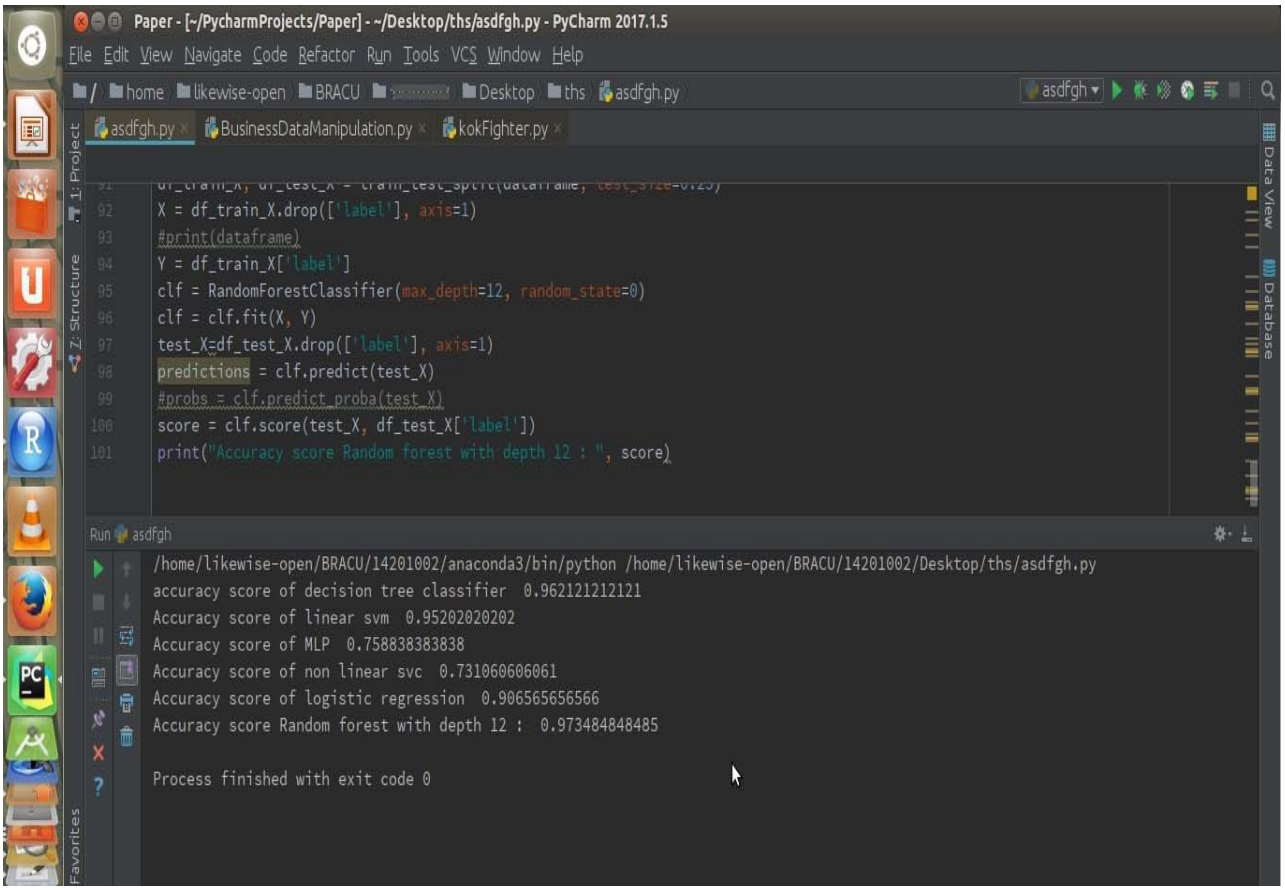


Fig-7: Screenshot of the Code run and Accuracy result

LR:

Logistic regression is a largely used model to deal with two class identifier .We trained and tested this model for a increase in accuracy but it had only 90.06% and 90.65% accuracy. The implementation is simple and efficient but the score was not satisfactory.

After these simulations, A tuning of the dataset was required, but it did not help with the accuracy. Then Principal Component Analysis was performed on both classes of the dataset. Also it did not had any remarkable effect in the accuracy.

RF:

Random forest was the model with best accuracy from same amount of training and testing data (which was changed in proceeding rounds) among these all and also , it could be more efficient for big data set. First round was done with maximum depth of 5 but the result was relatively lower . after several implementations, we found the best accuracy score of 97.34% with maximum depth of 12. After given forest depth of 25, the score got down in big scale. The Figure-8 illustrates the screenshot of the code we run and the accuracy result after that

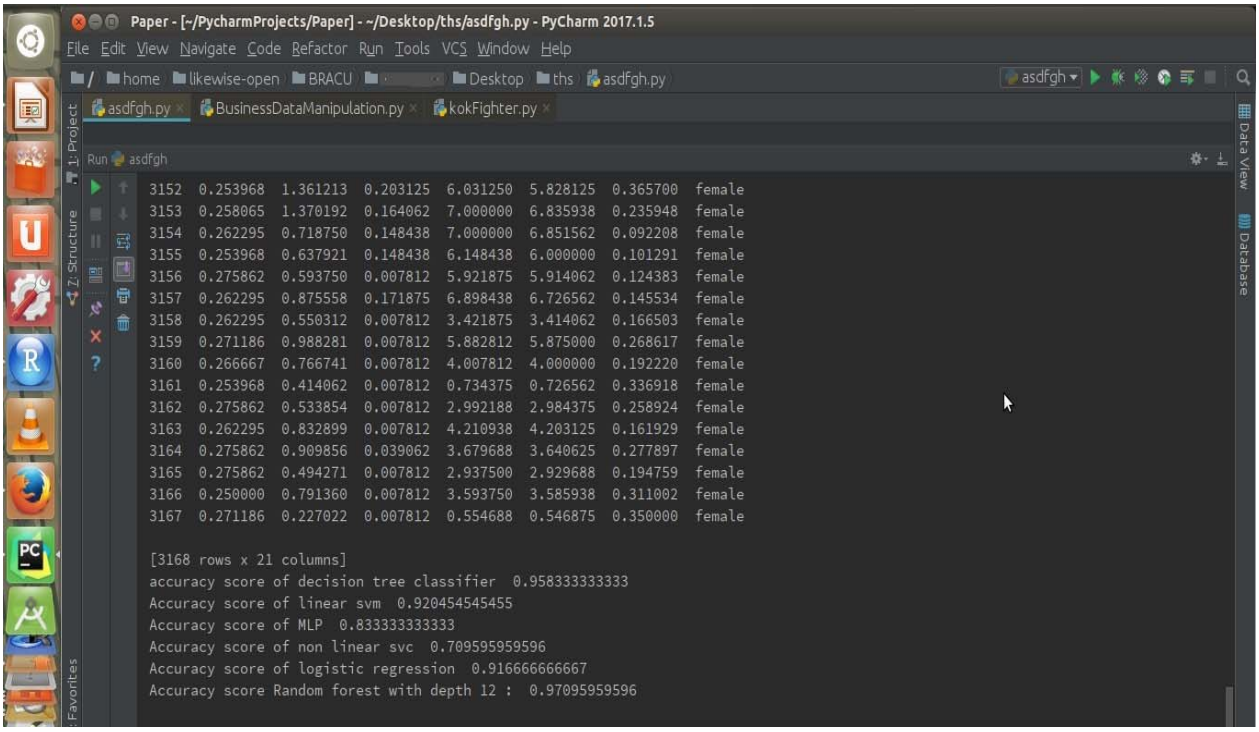


Fig-8: screenshot of the Code run and Accuracy result

CHAPTER-5

Conclusions And Future Work

The objective of this thesis is to build a speaker gender recognition system. The experiment, we tuned our dataset with PCA (Principal Component Analyzer) for a more accurate result. For using the dataset properly we used 25% testing and rest with training. Then we used Classification and Regression Tree as a classifier and five different Algorithms, which are Decision Tree Classifier, Logistic Regression, Support Vector Machine (SVM), Multi-Layer Perceptron and Random Forest Classifier. As we had different layers and different layer sizes we got different accuracy rates using different algorithms but the same data set. Yet we tried all the five algorithms to compare them and get the best result. According to the test results, we got the highest accuracy using Random Forest Classifier and that was 97.34 %. We had limited time, otherwise the accuracy would have increased. In future the main goal is to achieve 100% accuracy. If we can use this method in the sectors we want to, this thesis work can serve a lot of important purposes and can help us technically develop

For future work, we would like to increase the accuracy of gender recognition system. The performance of a few more different classifications would be compared to obtain a conclusion and the classification method with the best performance will be used in the speaker recognition system. Moreover, we would like to develop a perfect speaker recognition system with many

functions. The speaker gender is helpful in many applications like advertisement, targeting to particular people, to provide facility and service to customer in a call center, in some cases voice of a user can be used as the biometric security because human beings have unique voice patterns and unique features [6]. Then the research work will be focused on emotion recognition and region recognition. We wanted to design a voice recognition system which might be applied in Internet of things, such as user authentication by voice, where voice input will be used. The final goal of this thesis is to design a system that can predict the age of the speaker alongside the region from the voice and for that we need an experimental speech database with participants both male and female, extended age range and expanded nationalities [5][23]. The future research will focus more on these functionalities.

References

[1] Burnett M., & Kulesza T. (2015) End-User Development in Internet of Things: We the People. In International Reports on Socio-Informatics (IRSI), Vol.12, Iss.2, pp. 81-86.

[2] Faundez, M., Roure, J., Espinosa-Duró V., Ortega J.A.: An efficient face verification method in a transformed domain. Pattern Recognition Letters 28, 854–858 (2007)

[3] Brudzewsk K., Osowskib S., Markiewicz T., Ulaczykd J. (17 January 2006), Classification of gasoline with supplement of bio-products by means of an electronic nose and SVM neural network,Volume67(),Pages93-104.Available: https://www.sciencedirect.com/science/article/pii/S0924271611001304

[4] Hwanjo Y. and Sungchul K., SVM Tutorial: Classification, Regression, and Ranking, .Available: https://pdfs.semanticscholar.org/cbc3/d8b04d37b2d4155f081cd423380220a91f13.pdf

[5] Brown C., Marlin. 'Human-Computer Interface Design Guidelines' Intellect Book, 1998.

[6] Rachel M., Patriya T., and Darrel D. M., The impact of gender on ad processing: A social identity perspective, 2003.

[7] Quinlan R., Decision Tree Discovery, Blue Martini Software 2600 Campus Dr. Suite 175, San Mateo, CA & Samuels Building, G08 University of New South Wales, Sydney 2052 Australia.

[8] Roth G., Levine M.D, 'Geometric primitive extraction using a genetic algorithm'. IEEE Transactions on Pattern Analysis and Machine Intelligence, 901–905 (1994), ISSN 0162-8828 9.

[9] Marc F., Cheung-Chi L., Claude B. and Jean-Luc G., 2010 "Comparison of Speaker Adaptation Methods as Feature Extraction for SVM Based Speaker Recognition" ,IEEE Transaction 1558-7916.

[10] Breiman L., Friedman J, Olshen R., Stone C., Classification and Regression Trees. Belmont, CA: Wadsworth International Group; 1984

[11] Harrell, Frank E. (2001). Regression Modeling Strategies (2nd ed.). Springer-Verlag. .

[12] Rollins D. K., Zhai D., Joe A.L., Guidarelli J.W., Murarka A., Gonzalez R.: A novel data mining

method to identify assay-specific signatures in functional genomic studies, BMC Bioinformatics 2006, 7:377

[13.] Karhunen J., Joutsensalo J.: Generalizations of principal component analysis, optimization problems, and neural networks. Neural Networks, 1995, 7: 113–127

[14] Bishop C.M.,: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)

[15] Michel P. and Kaliouby R. E., *Real time facial expression recognition in video using support vector machines*, In Proceedings of ICMI'03 258-264, (2003).

[16] Cristianini N., Shawe-Taylor J., An Introduction to Support Vector Machines, Cambridge Univ. Press, 2000.

[17] Mukherjee S., Rifkin R., Denison D. D., Hansen M. H. and Poggio T. (2002). Regression and classification with regularization: Nonlinear Estimation and Classification, 107–124. New York:

[18] Yoonkyung L., Lin Y. & Wahba G. (2001). " Multicategory Support Vector Machines" . Computing Science and Statistics.

[19] Yoonkyung L., Lin Y., Wahba G. (2004). "Multicategory Support Vector Machines". Journal of the American Statistical Association. 99 (465): 67.

[20] Ho T. (1998). "The Random Subspace Method for Constructing Decision Forests" IEEE Transactions on Pattern Analysis and Machine Intelligence

[21] Pal S.K., Mitra S.," Multilayer perceptron, fuzzy sets, and classification", Volume: 3 (Issue: 5). Available at: http://ieeexplore.ieee.org/document/159058/?anchor=relatedarticles

[22] Ahad A.,Fayyaz A., Mehmood T. (2002) "Speech recognition using multilayer perceptron" http://ieeexplore.ieee.org/document/1215948/citations

[23] Hassam U. S., "Who is speaking ? Male or female?" Ph.D dissertation, dept. of engineering and sciences,University of Manchester.

[24] Speybroeck N. Classification and regression trees. International Journal of Public Health. 2012;57(1):243–246