A PROJECT REPORT

ON

**SPEECH EMOTION RECOGNITION**

**using**

**MLP CLASSIFIER**

Submitted in partial fulfillment for the requirement of the award of

Training In

*Data Analytics, Machine Learning and AI using Python*



Submitted By

**Saloni Singh**

(B.Tech(CSE)-1st year student at Indian Institute of Technology, Dharwad)

Under the guidance of

**Mr. Bipul Shahi**

**ABSTRACT**

Emotions play a very important role in all our conversations. Unlike humans, machines cannot perceive or show emotions. This cavity in machine-human interaction can be repleted to some extent by automating the emotion recognition process by techniques such as machine learning. The primary objective it serves is that it improves machine-human interface. Speech features like MFCC, mel and chroma are extracted from the dataset, which are used to classify emotions into four main categories which are the emotions of being neutral, angry, happy and calm. Various classification-based models like KNeighbours classifier, Logistic Regression, Bagging Classifier, Decision tree classifier and Neural network-based model like MLP Classifier are used to develop this recognition project. Inference about the performance of speech emotion recognition system based on each of these models is discussed.

**INTRODUCTION**

Emotions play an extremely important role in human life. It is a medium of one’s perspective or one’s mental state to others. The emotional detection is natural for humans but it is very difficult task for machine. Therefore, the purpose of emotion recognition system is to use emotion related knowledge in such a way that human machine communication will be improved. This is mainly based on the fact that speech often reflects the emotion within it through tone, pitch, amplitude and other features of audio signal. This is also the phenomenon that animals like dogs and horses use to be able to understand human emotion.

It has a potentially wide applications, such as an interface with robots, banking, car board systems, computer games etc. It can be used at places such as call centres and recommendation systems, as it will adjust recommendations based on a person’s mental state at that instant, as emotions are a reflection of a person’s mental state. This helps in improving the service and relevance. For class-room coordination or E-learning, information about the emotional state of students can provide focus on enhancement of teaching quality.

This process of speech emotion recognition mainly involves basic speech processing and machine learning techniques. Speech processing helps in extracting the features such as mfcc, mel, chroma ,stft which contain information about the emotion.The machine learning techniques then help in the prediction process.

In this project we will be able to predict whether the emotion of the person’s speech is neutral, happy, angry or calm.

**DATASET**

In this project, the RAVDESS dataset is used. This Ryerson Audio-Visual Database of Emotional Speech and Song dataset is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. It has 7356 files rated by 247 individuals 10 times on emotional validity, intensity, and genuineness. However, the sample rate has been lowered on all the files.

**TOOLS, TECHNOLOGY AND LIBRARIES**

Machine Learning

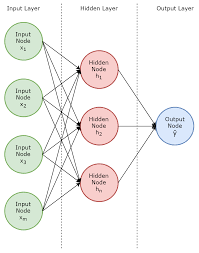
Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence). Machine learning algorithms build a [mathematical model](https://en.wikipedia.org/wiki/Mathematical_model) based on sample data, known as "[training data](https://en.wikipedia.org/wiki/Training_data)", in order to make predictions or decisions without being explicitly programmed to do so. It is basically categorised as supervised learning and unsupervised learning. In this project we use supervised learning, i.e. we have labels corresponding to features. The project is based on classification of data into different emotions, so we use the following classification based models, KNeighbors classifier, Logistic Regression, Decision Tree classifier, Bagging classifier. These algorithms are used for prediction of emotion. The algorithm with most accurate prediction is used for recognition.

Neural Networks

Neural networks are a set of algorithms, modelled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. The patterns they recognize are numerical, contained in vectors.

In this project, the neural network algorithm used is MLP Classifier. A multilayer perceptron (MLP) is a class of [feedforward](https://en.wikipedia.org/wiki/Feedforward_neural_network) [artificial neural network](https://en.wikipedia.org/wiki/Artificial_neural_network) (ANN). The term MLP is used to refer to networks composed of multiple layers of [perceptrons](https://en.wikipedia.org/wiki/Perceptron" \o "Perceptron) (with threshold activation). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear [activation function](https://en.wikipedia.org/wiki/Activation_function). MLP utilizes a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) technique called [backpropagation](https://en.wikipedia.org/wiki/Backpropagation) for training.

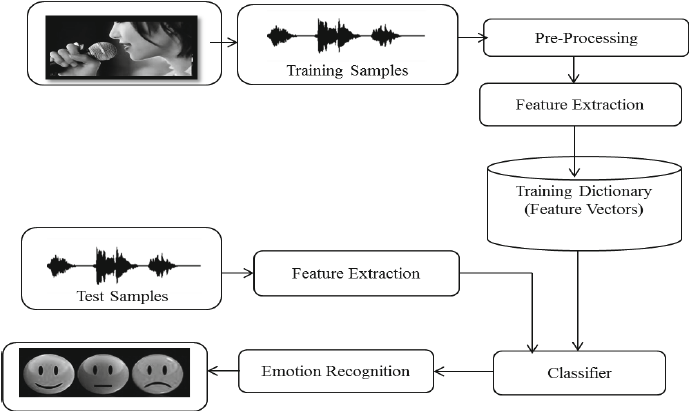
In this project the activation is set to default i.e. relu, hence activation function is f(x)=max(0,x).



Librosa

Librosa is a [Python library](https://data-flair.training/blogs/python-libraries/) for analyzing audio and music. It has a flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code. This is the library which helps in speech processing and feature extraction.

**Methodology**

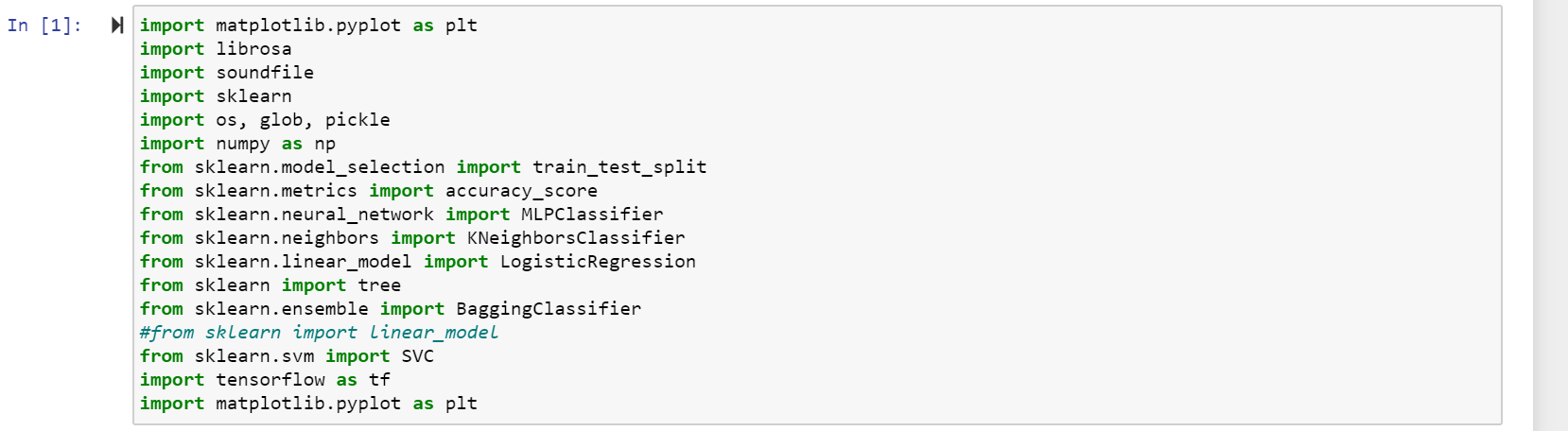


**Source:** <https://www.researchgate.net/figure/Architecture-of-Speech-Emotion-Recognition-System_fig1_270899631>

The above process flow chart shows the architecture of the speech emotion recognition project.

**Importing Libraries**

We import all the necessary libraries. Librosa and soundfile for speech processing; numpy, matplotlib,os,etc for data analysis and sklearn for machine learning models.

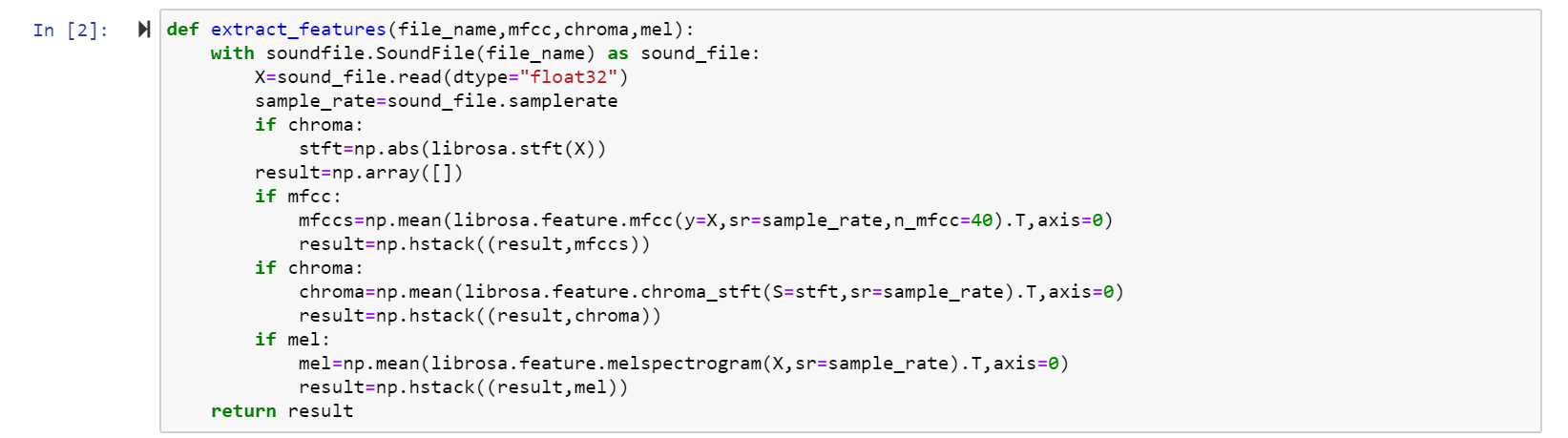


**Feature Extraction**

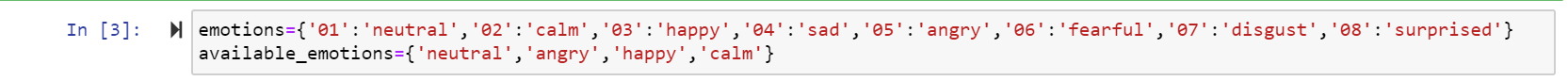
In this process we define extract\_features function to extract the features using librosa from the audio signals.It takes filename as parameter as well as three Boolean parameters which are

* mfcc- Mel Frequency Cepstral Coefficient, represents the short-term power spectrum of a sound.
* mel- Mel Spectrogram Frequency
* chroma- Pertains to the 12 different pitch classes

Based on the parameters, the features mfcc, mel and stft(short term fourier transform) are extracted and returned.

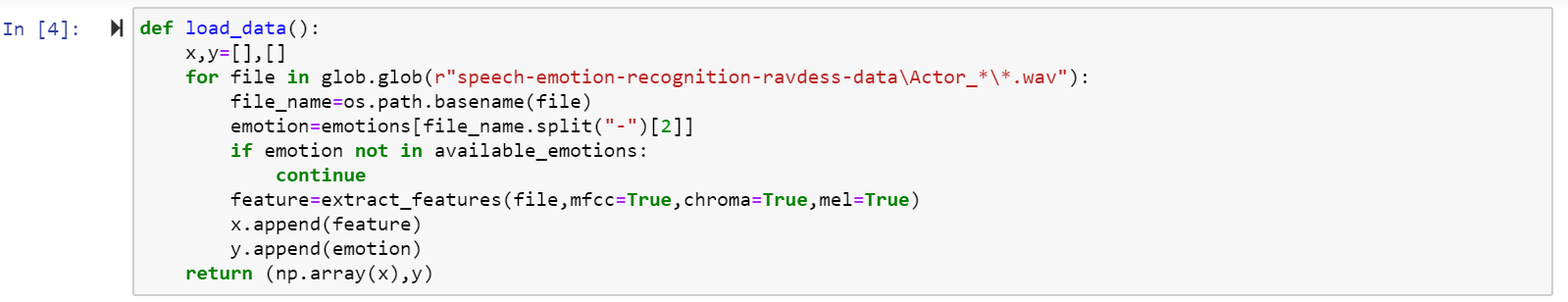


**Defining emotions**

There are eight emotions present in the dataset namely, neutral, calm, happy, sad, angry, fearful, disgust, surprised. However, we will be classifying emotions only as neutral, angry, happy and calm. 

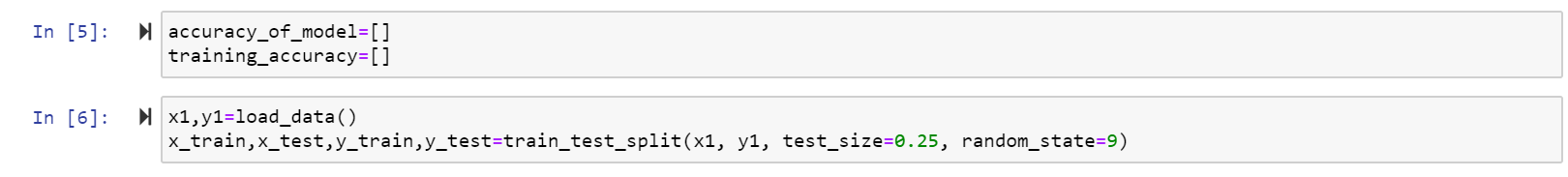
**Loading data**

Now the data is uploaded with the load\_data() function. The glob() function from the glob module is used to get all the pathnames for the sound files in our dataset. For each such path, get the basename of the file, the emotion by splitting the name around ‘-’ and extracting the third value. This value is then mapped to the emotion using the dictionary defined in the previous step. Thus, the function returns an array of features and another array of emotions.

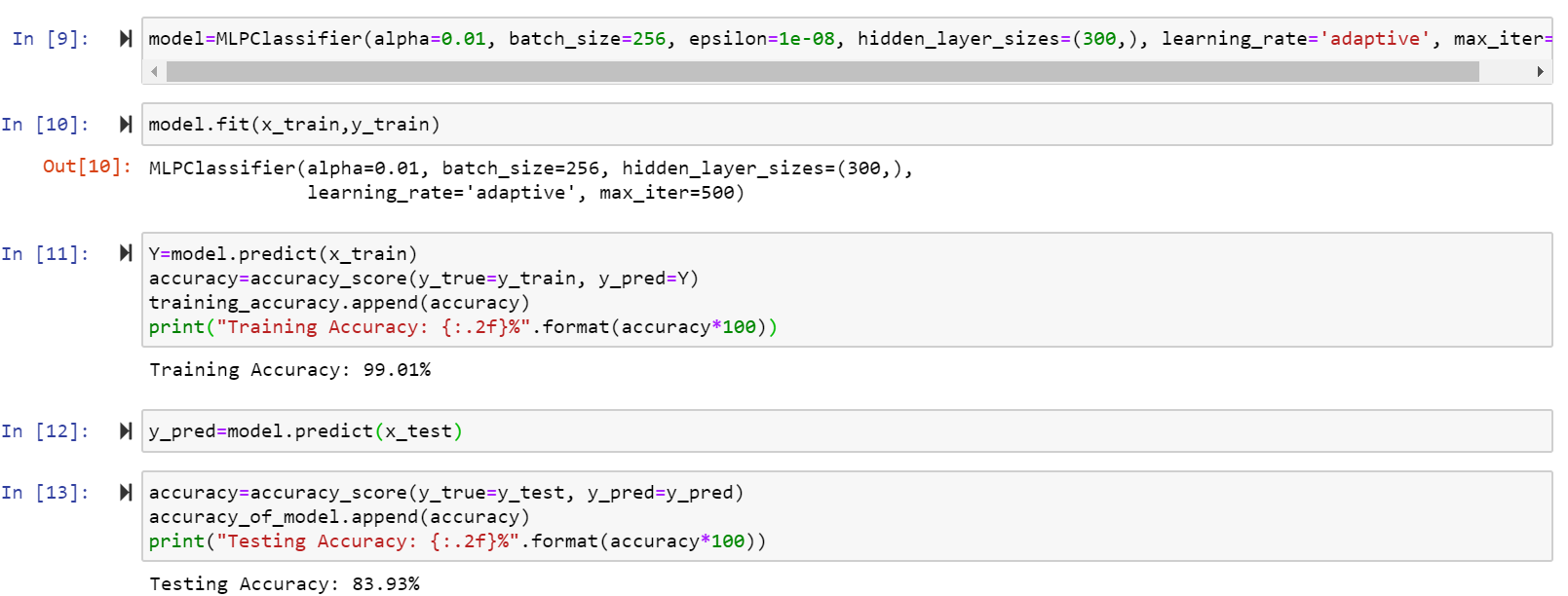


**Training the model**

Now the load\_data() function is called and the data is split into training and testing data and trained on various models and their training and testing accuracies are stored for analysis.



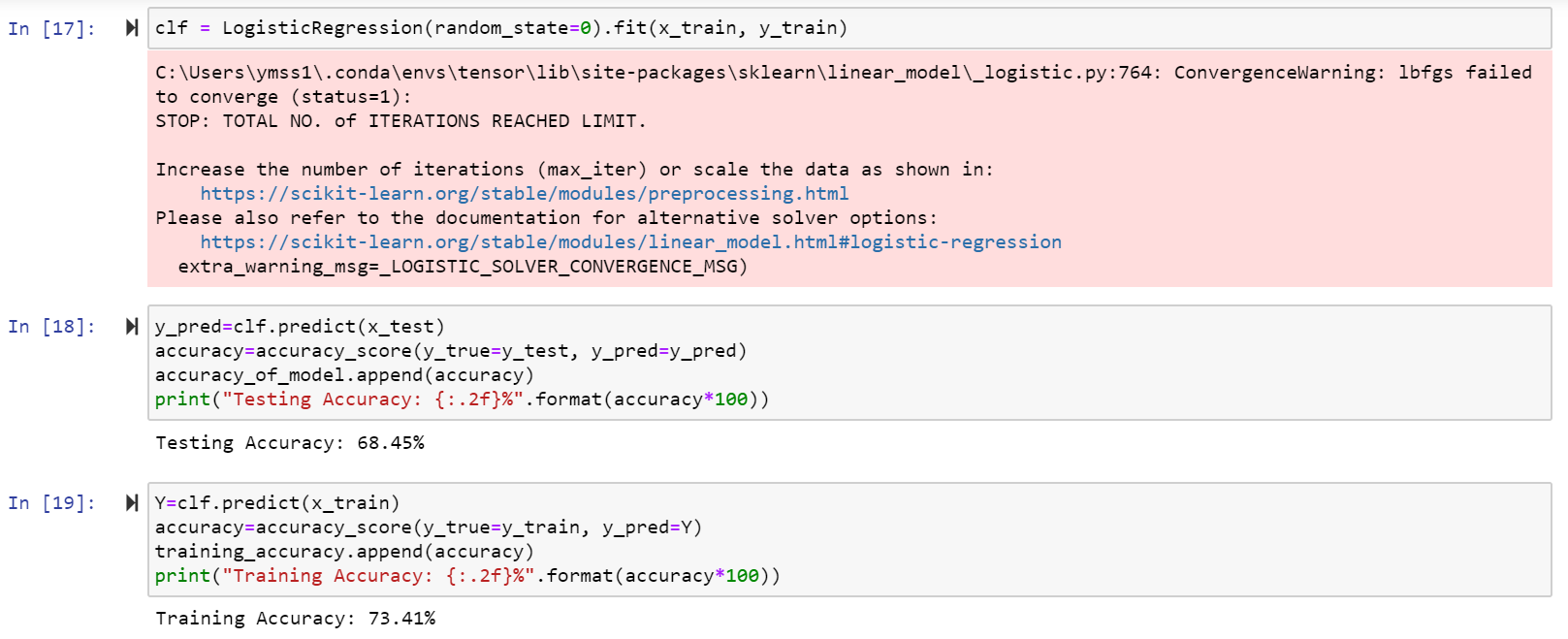
**MLPClassifier**



**KNeighbours Classifier**



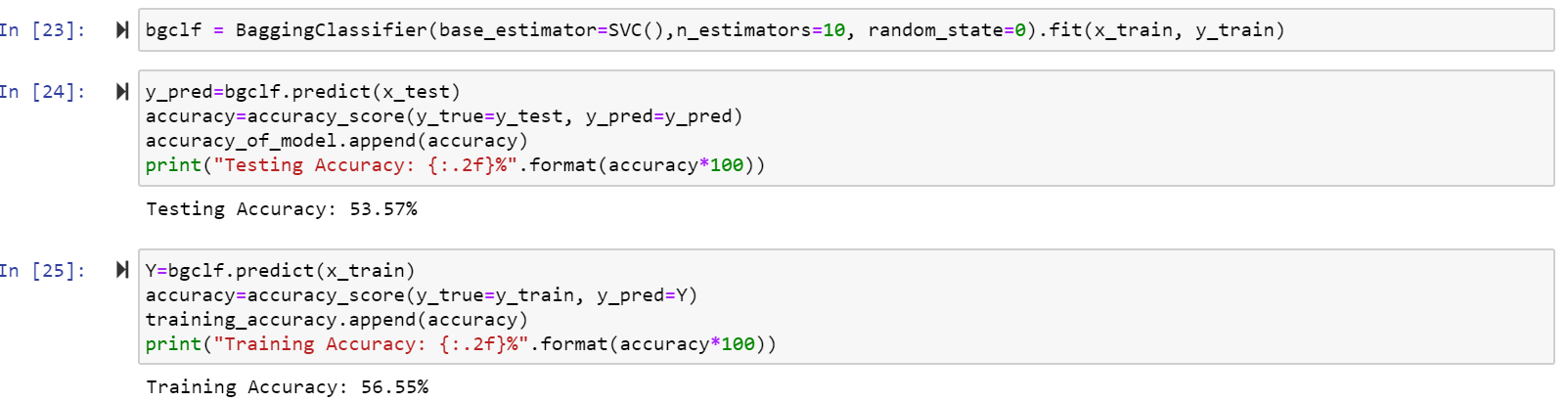
**Logistic Regression**



**Decision Tree Classifier**

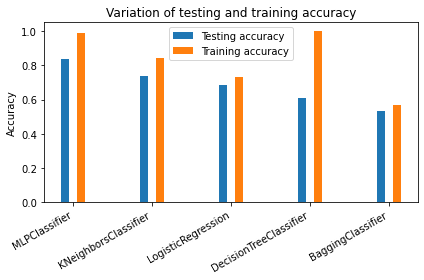


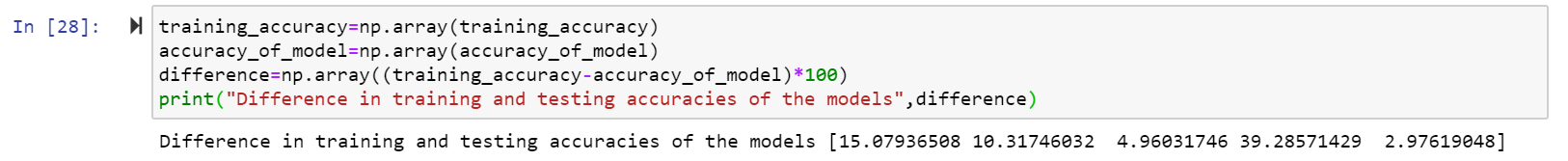
**Bagging Classifier**

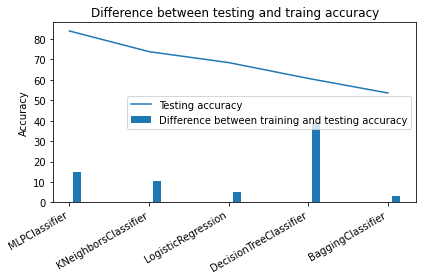


**Analysis**

In the project using the mentioned dataset, the number of features extracted if 180.The size of training and testing data respectively is 504 and 168.







Bagging Classifier gives a very poor training and testing accuracy of around 50%.Decision Tree Classifier presents an overfit model with training accuracy of 100% and testing accuracy of around 60% , showing a very large difference in both these accuracies. Logistic Regression does provide quite acceptable difference between training and testing accuracies although they are quite low. KNeighbours Classifier also gives reasonable output, with less difference between training and testing accuracies but the accuracies are quite low. MLPClassifier gives the most acceptable result, with quite high accuracies.

**Conclusion**

In this project, we tried and tested different machine learning models to recognise the emotion of the speech. The best results were obtained with Neural Network based MLPClassifier, although KNeighbours Classifier also gives good results. Thus, the emotions from the speech were successfully recognised with an accuracy of 83.93%. These results were obtained when only four emotions (i.e. neutral, anger, happy, calm) were recognised. Further upgradations in the model can help in detection of more emotions.

**Bibliography**

* <https://librosa.github.io/librosa/>
* <https://pathmind.com/wiki/neural-network>
* <https://data-flair.training/blogs/>
* <https://www.researchgate.net/publication/322873355_Speech_Emotion_Recognition_Methods_and_Cases_Study>
* <https://www.researchgate.net/publication/299185942_Human_speech_emotion_recognition>
* <https://www.thepythoncode.com/>