Capstone Project

Speech Emotion Recognition

Team Member

Mohd. Navaid Ansari

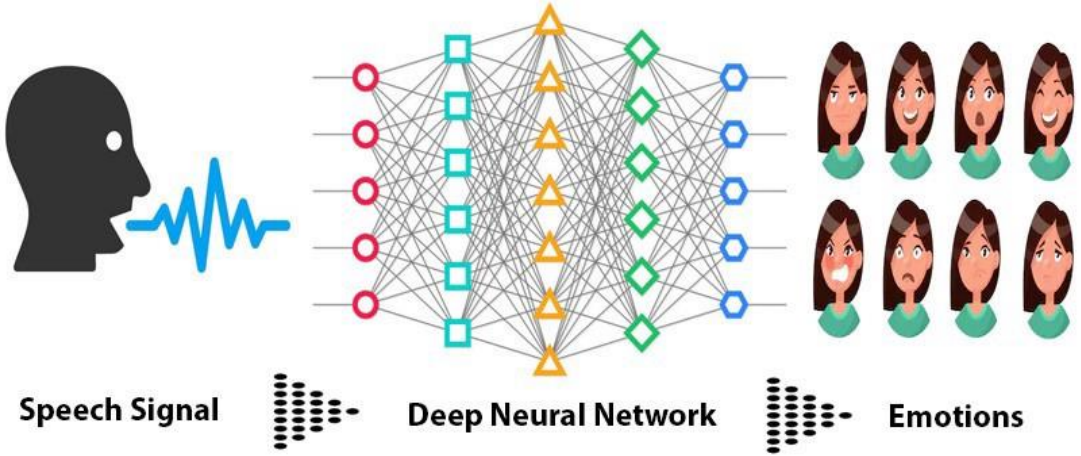
Mohammad Anas Ansari

Content

Speech emotion recognition

Introduction:

- Speech is the fast and best normal way of communicating amongst human.
- This reality motivate many researchers to consider speech signal as a quick and effective process to interact between computer and human.
- Although, there is a significant improvement in speech recognition but still researcher are away from natural interplay between computer and human, since computer is not capable of understanding human emotional state.



- The recognition of emotional speech aims to recognize the emotional condition of individual utterer by applying his/her voice automatically.
- Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis.

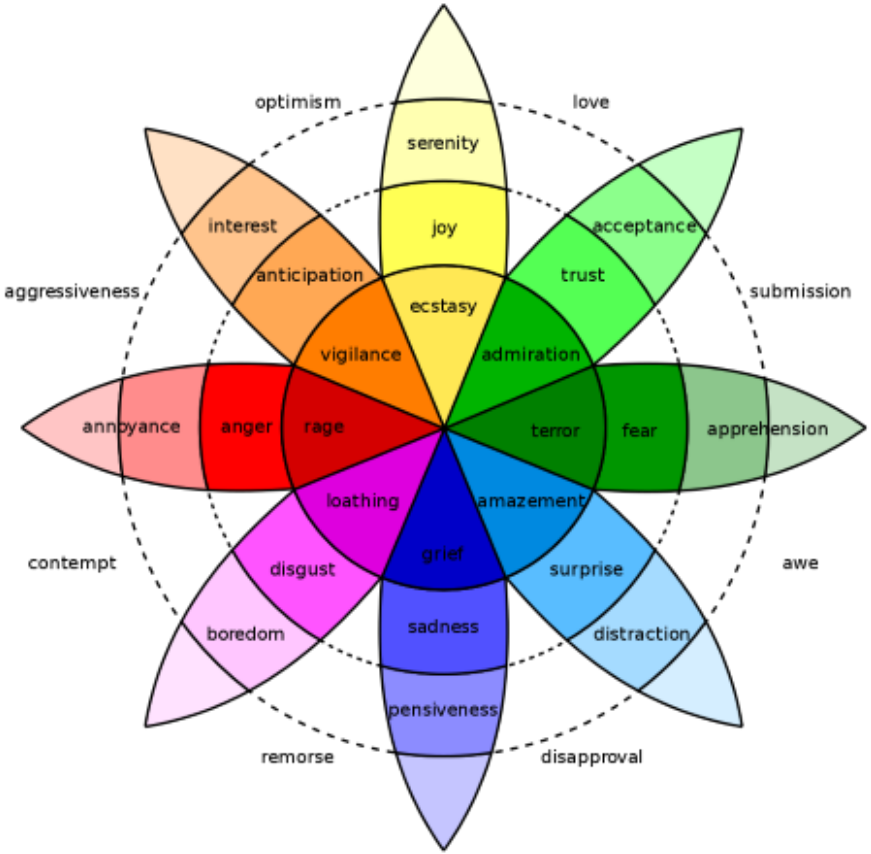
Introduction

- Speech emotional methods is selecting the best features, which is powerful enough to distinguish between different emotions.

- The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features include pitch, energy.

- Furthermore, it is possible to have a more than one specific emotion at the same in the same speech signal, each emotion correlate with a different part of speech signals. Therefore, defines the boundaries between parts of emotion in very challenging task.

WHAT KIND OF EMOTIONS CAN BE DETECTED AND RECOGNIZED?



Data Preparation:

The data used in this project was combined from five different data sources as mentioned below:
1.TESS (Toronto Emotional Speech Set): 2 female speakers (young and old), 2800 audio files, random words were spoken in 7 different emotions.

2.SAVEE (Surrey Audio-Visual Expressed Emotion): 4 male speakers, 480 audio files, same sentences were spoken in 7 different emotions.

3.RAVDESS: 2452 audio files, with 12 male speakers and 12 Female speakers, the lexical features (vocabulary) of the utterances are kept constant by speaking only 2 statements of equal lengths in 8 different emotions by all speakers.

4.CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset): 7442 audio files, 91 different speakers (48 male and 43 female between the ages of 20 and 74) of different races and ethnicities, different statements are spoken in 6 different emotions and 4 emotional levels (low, mid, high and unspecified).

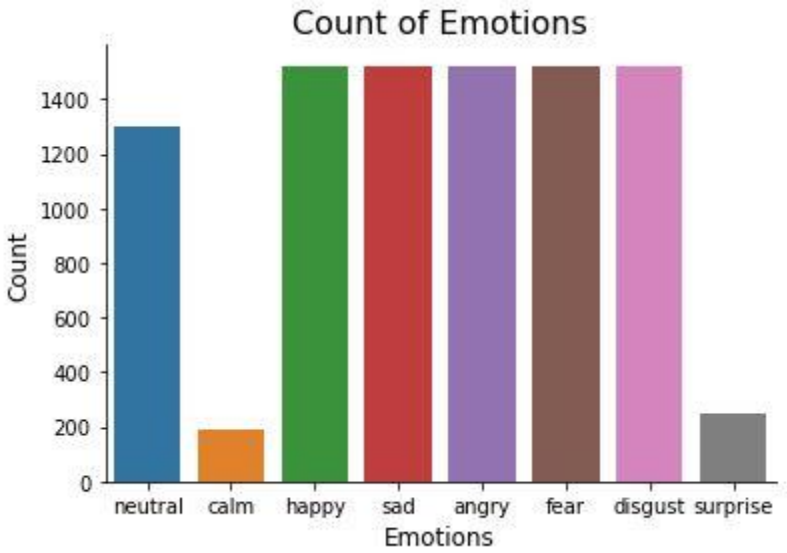
Exploratory Data Analysis:

The combined data set from the original 4 sources is thoroughly analyzed with respect to the following aspects

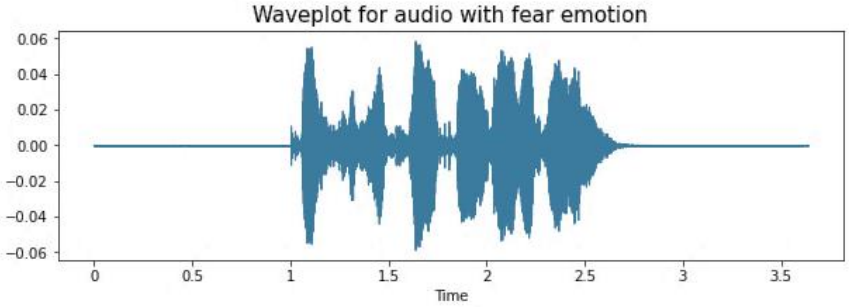
- Emotion distribution by gender
- Variation in energy across emotions
- Variation of relative pace and power across emotions

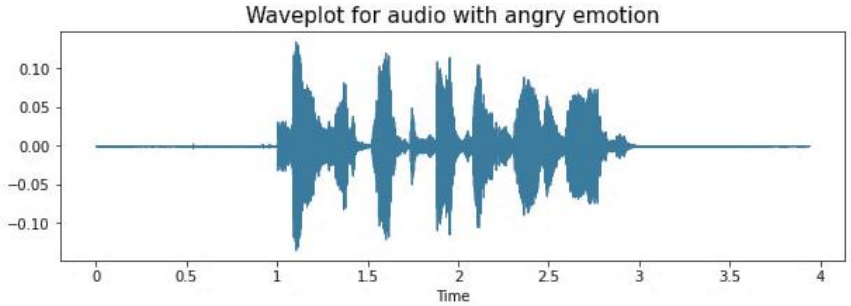
We checked the distribution of labels with respect to emotions and gender and found that while the data is balanced for six emotions viz. ***neutral, happy, sad, angry, fear*** and ***disgust***, the number of labels was slightly less for ***surprise*** and negligible for ***calm***. While the slightly fewer instances of surprise can be overlooked on account of it being a rarer emotion, the imbalance against calm was rectified later by clubbing sadness and calm together due to them being similar acoustically. It's also worth noting that calm could have been combined with neutral emotion but since both ***sadness*** and ***calm*** are negative emotions, it made more sense to combine them.

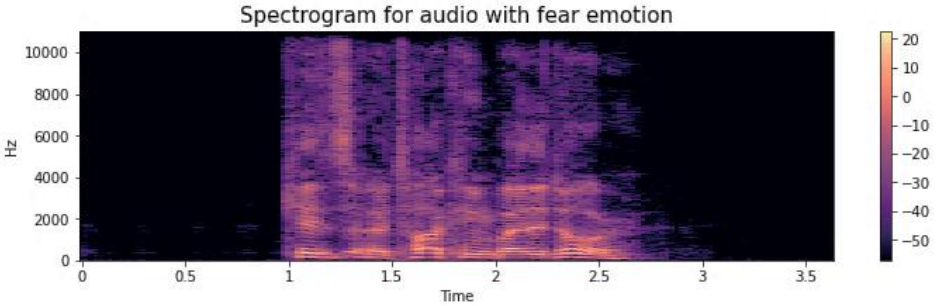
```
plt.title('Count of Emotions', size=16)
sns.countplot(data_path.Emotions)
plt.ylabel('Count', size=12)
plt.xlabel('Emotions', size=12)
sns.despine(top=True, right=True, left=False, bottom=False)
plt.show()
```

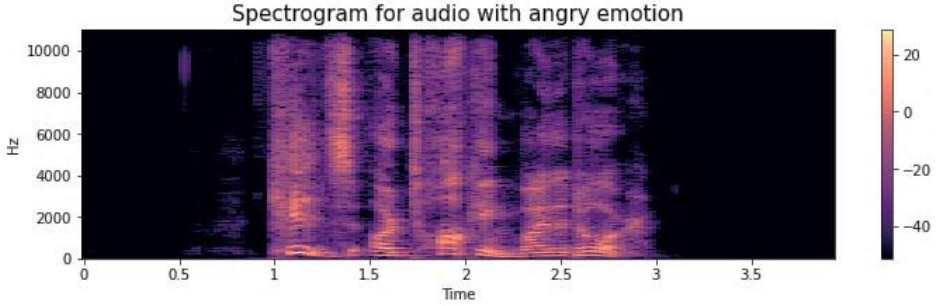


Wave plots and Spectrograms





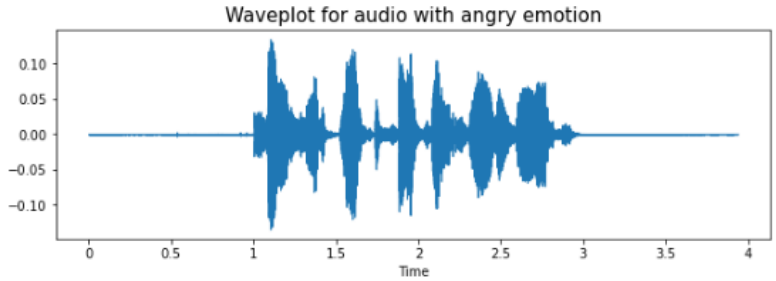


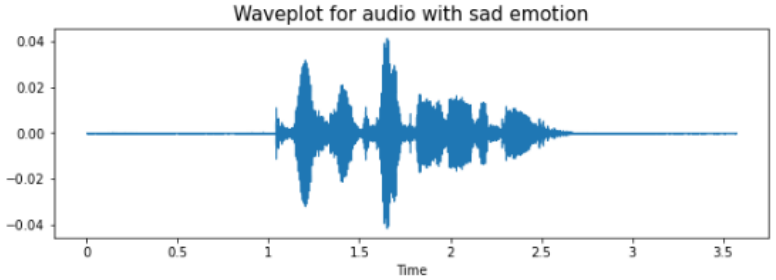


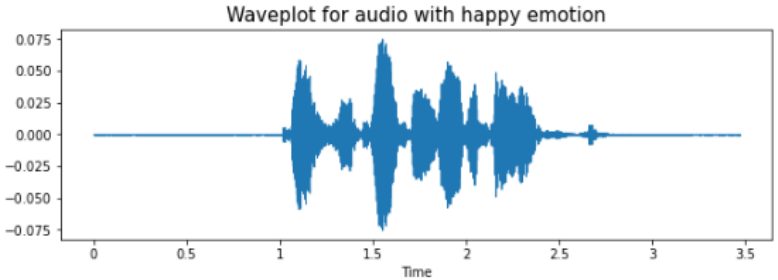
Fear audio

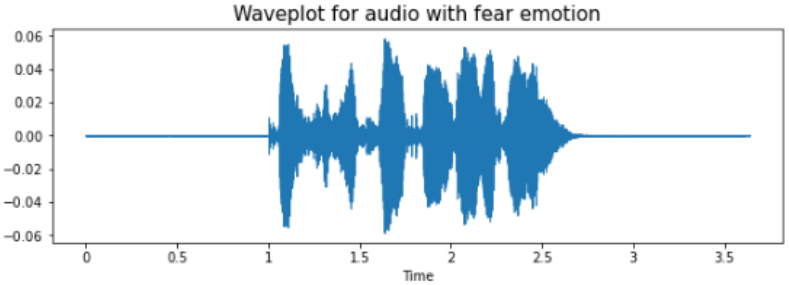
Angry audio

Wave Plots for Different Emotions







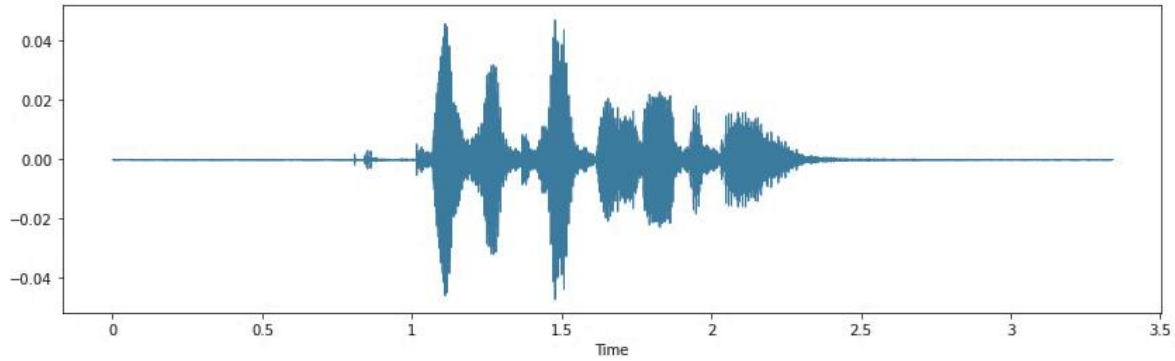


Data Augmentation

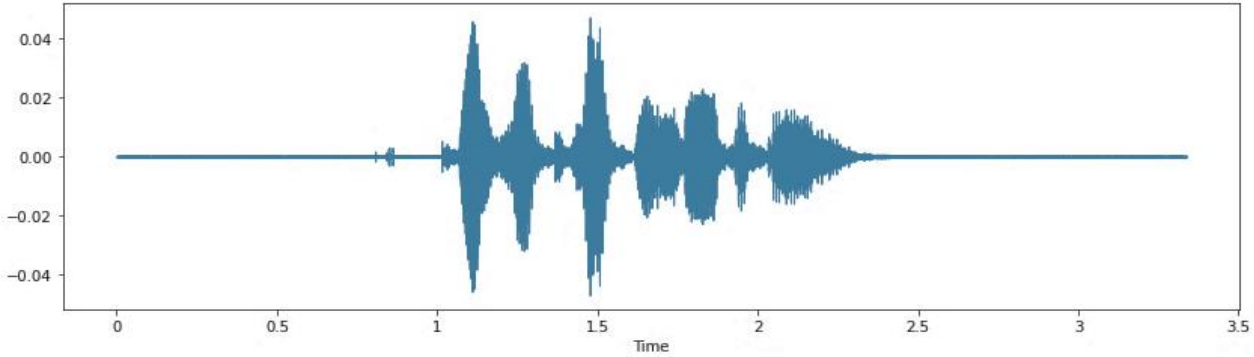
- Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set.
- To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed.
- The objective is to make our model invariant to those perturbations and enhance its ability to generalize.
- In order to this to work adding the perturbations must conserve the same label as the original training sample.

Simple audio

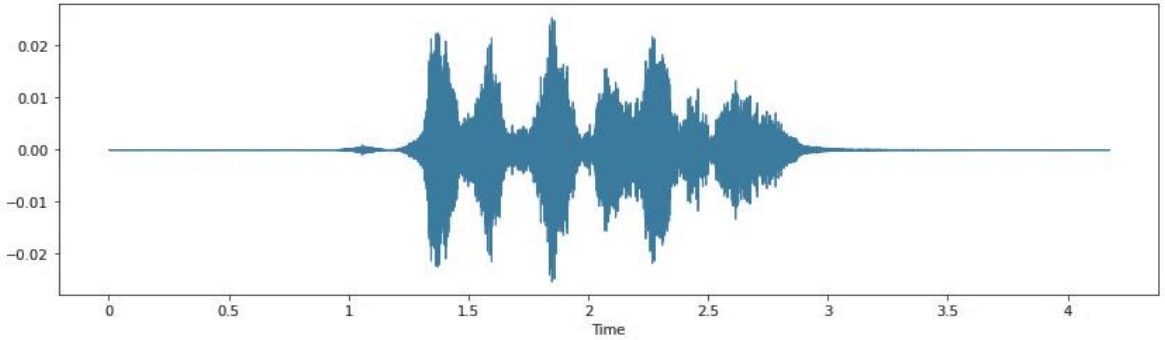
Out[52]:

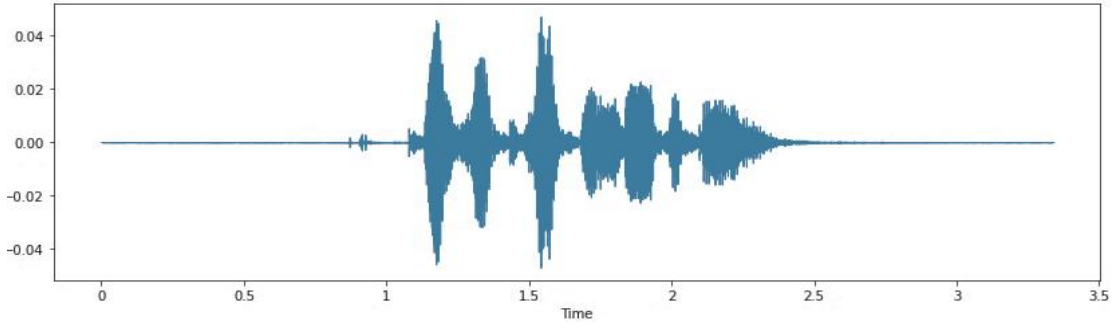


Noise injection



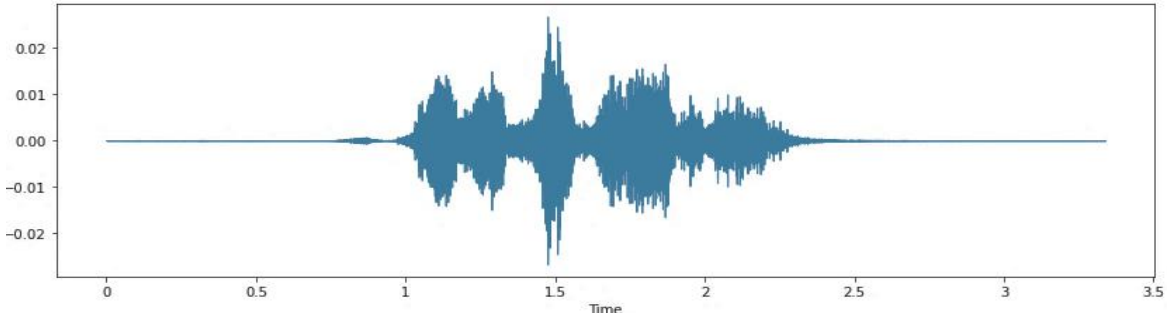
Stretching





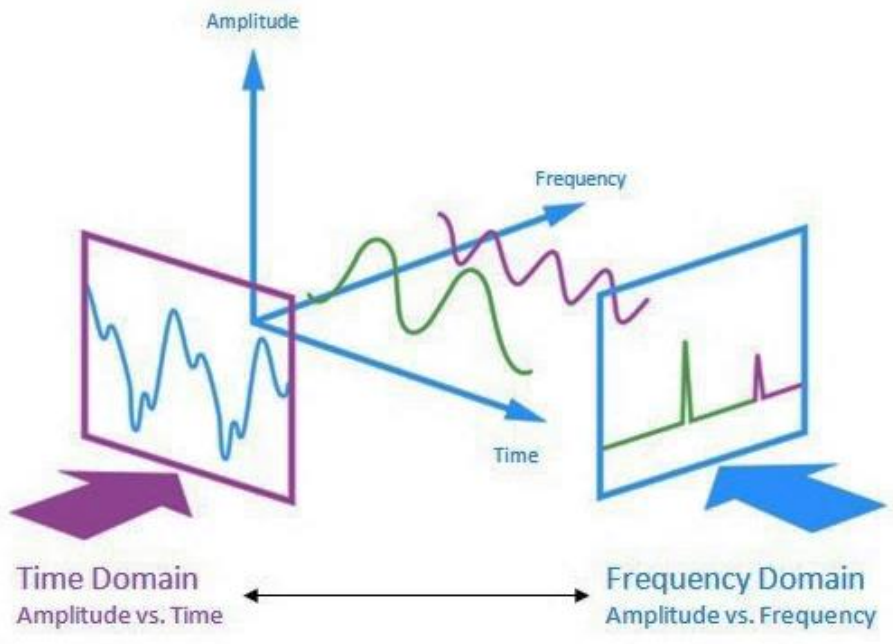
Shifting

Pitch



Feature Extraction

- Extraction of features is a very important part in analyzing and finding relations between different things.
- As we already know that the data provided of audio cannot be understood by the models directly.
- so we need to convert them into an understandable format for which feature extraction is used.



- As stated there with the help of the sample rate and the sample data, one can perform several transformations on it to extract valuable features out of it.
- In this project we are not going deep in feature selection process to check which features are good for our dataset rather we are only extracting 5 features:
    - Zero Crossing Rate
    - Chroma_stft
    - MFCC
    - RMS(root mean square) value
    - MelSpectogram to train our model.

List of feature:

See Table 1 for the features that were extracted for each frame of the audio signal, along with their definitions

Table.1 List of features present in an audio signal

Feature ID	Feature Name	Description
1	Zero Crossing Rate	*"The rate at which the signal changes its sign."*
2	Energy	*"The sum of the signal values squared and normalized using frame length."*
3	Entropy of Energy	*"The value of the change in energy."*
4	Spectral Centroid	*"The value at the center of the spectrum."*
5	Spectral Spread	*"The value of the bandwidth in the spectrum."*
6	Spectral Entropy	*"The value of the change in the spectral energy."*
7	Spectral Flux	*"The square of the difference between the spectral energies of consecutive frames."*
8	Spectral Rolloff	*"The value of the frequency under which 90% of the spectral distribution occurs."*
9-21	MFCCs	*"Mel Frequency Cepstral Coefficient values of the frequency bands distributed in the Mel-scale."*
22-33	Chroma Vector	*"The 12 values representing the energy belonging to each pitch class."*
34	Chroma Deviation	*"The value of the standard deviation of the Chroma vectors."*

Methodology:

Feed forward artificial neural network which is well known as ANN is super class of MLP which stands for multi-level perceptron. It has minimum of three layers which are one input layer, an output layer and other hidden layers.

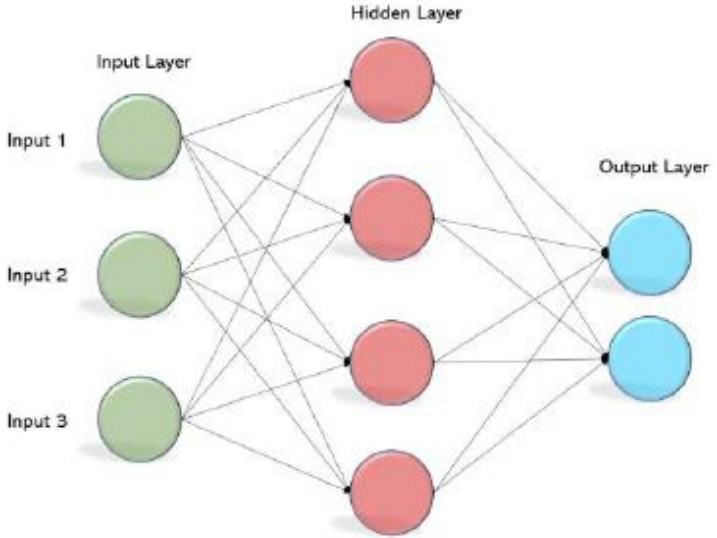
$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(w^T x + b)$$

Where, x = input vector

W = Weights vector

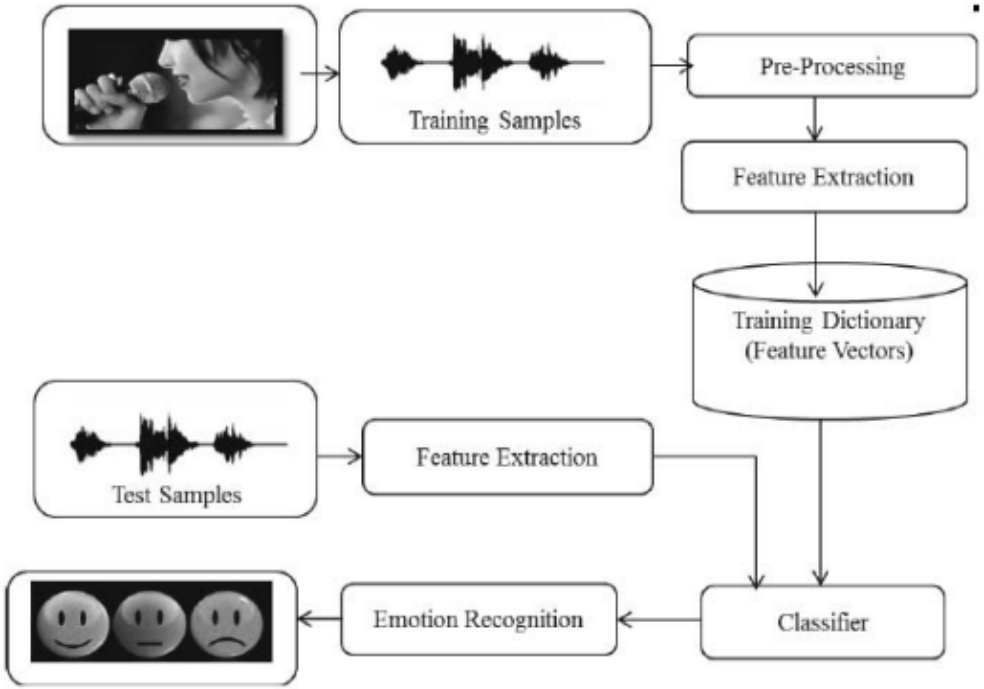
B = Bias

φ = Activation Function



Artificial Neural Network

- Convolution
- ReLu Activation Function
- Pooling
- Classification

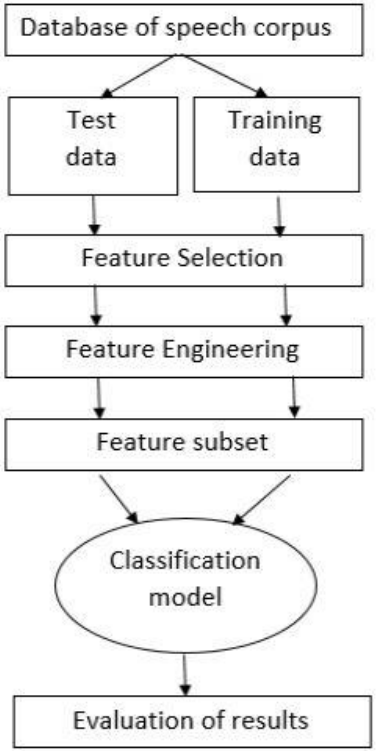


Model Structure

Flow Chart

The flowchart represents a pictorial overview of the process.

- The first step is data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data.
- The second step, called feature engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures address the several data representation and data quality issues.
- The third step is often considered the core of an ML project where an algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to respond to any new data it is exposed to.
- The final step is to evaluate the functioning of the built model. Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms. Comparison results help to choose the appropriate ML algorithm most relevant to the problem.

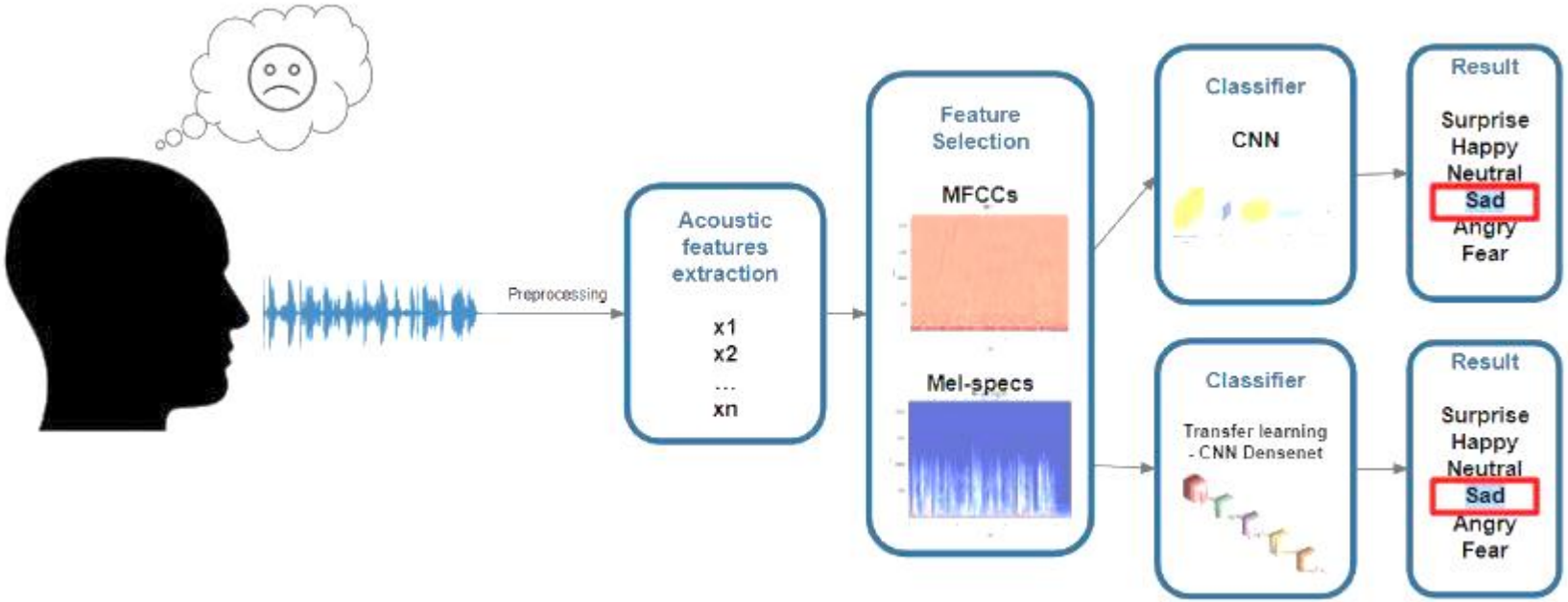


IMPLEMENTATION USING:

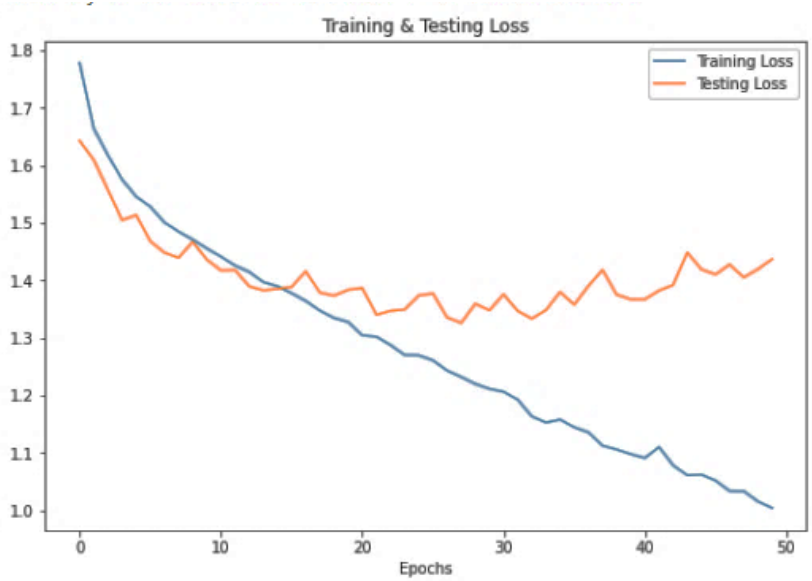
In [2]:

```
import pandas as pd
import numpy as np
import os
import sys

# Librosa is a Python library for analyzing audio and music.
# It can be used to extract the data from the audio files we will see it later.
import librosa
import librosa.display
import seaborn as sns
import matplotlib.pyplot as plt
# For data visualization matplotlib
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# To play the audio files
from IPython.display import Audio
# For deep learning models
import keras
from keras.callbacks import ReduceLROnPlateau
from keras.models import Sequential
from keras.layers import Dense, Conv1D, MaxPooling1D, Flatten, Dropout, BatchNormalization
from tensorflow.keras.utils import to_categorical
from keras.utils import np_utils
from keras.callbacks import ModelCheckpoint

import warnings
if not sys.warnoptions:
    warnings.simplefilter("ignore")
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

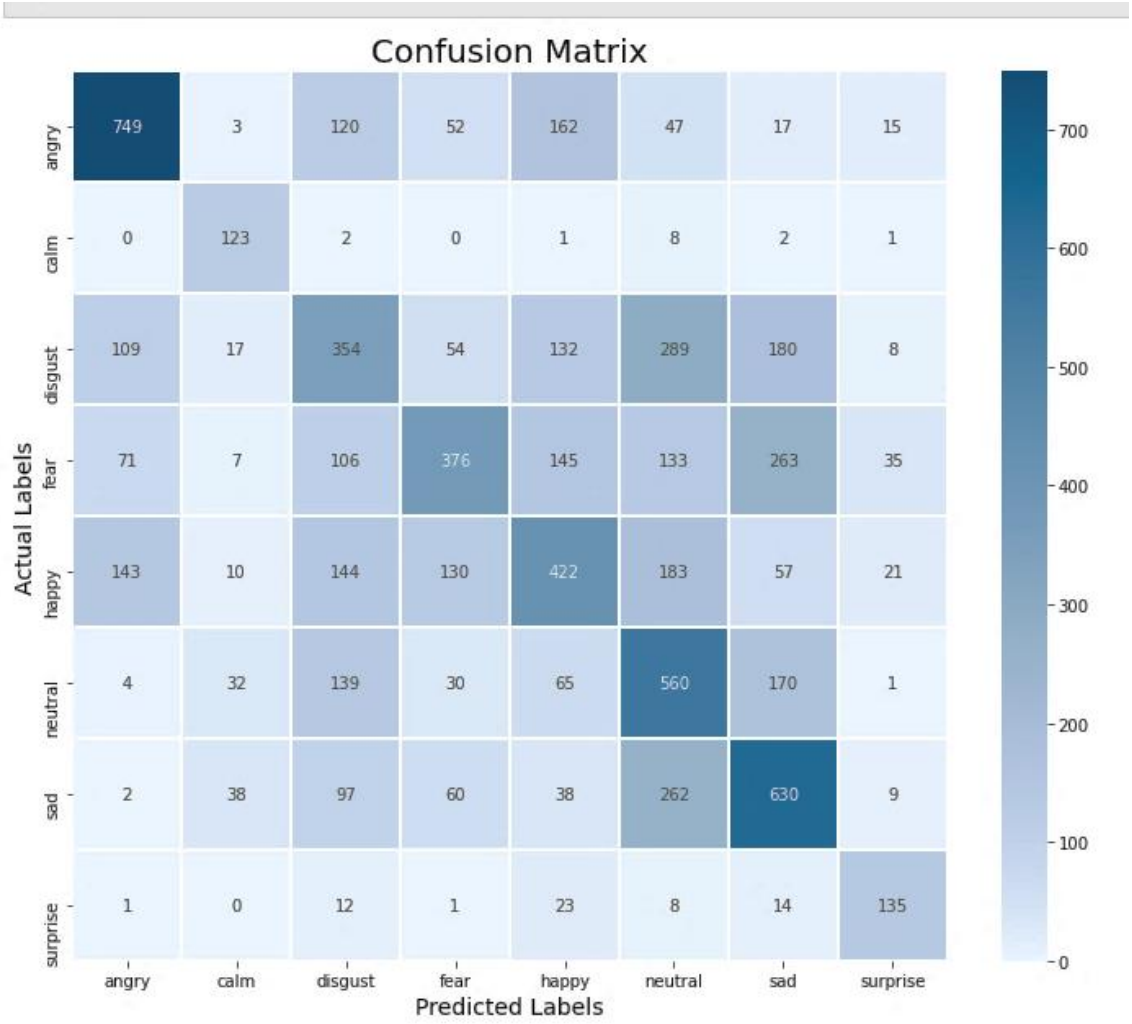
How speech emotion work :



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 162, 256)          1536

 max_pooling1d (MaxPooling1D  (None, 81, 256)          0
 )

 conv1d_1 (Conv1D)           (None, 81, 256)           327936

 max_pooling1d_1 (MaxPooling  (None, 41, 256)          0
 1D)

 conv1d_2 (Conv1D)           (None, 41, 128)           163968

 max_pooling1d_2 (MaxPooling  (None, 21, 128)          0
 1D)

 dropout (Dropout)           (None, 21, 128)           0

 conv1d_3 (Conv1D)           (None, 21, 64)            41024

 max_pooling1d_3 (MaxPooling  (None, 11, 64)           0
 1D)

 flatten (Flatten)           (None, 704)               0

 dense (Dense)               (None, 32)                22560

 dropout_1 (Dropout)         (None, 32)                0

 dense_1 (Dense)             (None, 8)                 264

=================================================================
Total params: 557,288
Trainable params: 557,288
Non-trainable params: 0
```



Model accuracy and loss on training and testing set of data


Confusion Matrix
Actual Labels
Predicted Labels
angry
calm
disgust
fear
happy
neutral
sad
surprise
749 3 120 52 162 47 17 15
0 123 2 0 1 8 2 1
109 17 354 54 132 289 180 8
71 7 106 376 145 133 263 35
143 10 144 130 422 183 57 21
4 32 139 30 65 560 170 1
2 38 97 60 38 262 630 9
1 0 12 1 23 8 14 135
700
600
500
400
300
200
100
0

Evaluation Matrix

	precision	recall	f1-score	support
angry	0.69	0.64	0.67	1165
calm	0.53	0.90	0.67	137
disgust	0.36	0.31	0.33	1143
fear	0.53	0.33	0.41	1136
happy	0.43	0.38	0.40	1110
neutral	0.38	0.56	0.45	1001
sad	0.47	0.55	0.51	1136
surprise	0.60	0.70	0.64	194
accuracy			0.48	7022
macro avg	0.50	0.55	0.51	7022
weighted avg	0.49	0.48	0.47	7022

For each emotion in the data set the corresponding values for precision, recall and f1-score is represented.

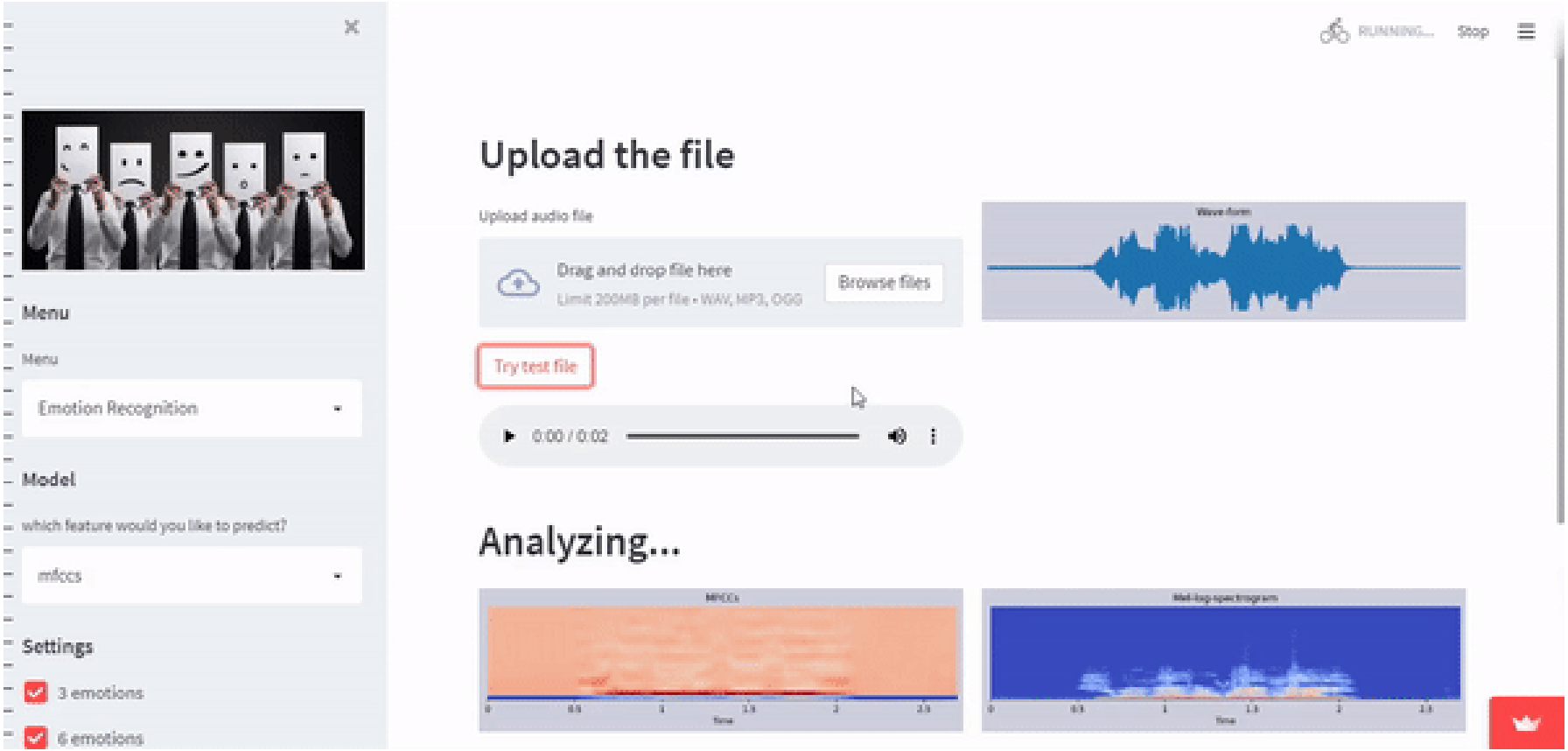
The overall model accuracy shown on the test data set which is 48%.

Model Deployment on Streamlit

A web app has been created on streamlit.io, where you can use the model for checking the emotion of any audio/speech of file type WAV, OGG, and MP3. With file size not more than 200 MB. The link for the app is:

https://share.streamlit.io/navaidansari786/speech_emotion_recog/main/app.py

App on streamlit



Fields of emotion recognition

Emotion recognition in text: Text data is a favorable research object for emotion recognition when it is free and available everywhere in human life. Compare to other types of data, the storage of text data is lighter and easy to compress to the best performance due to the frequent repetition of words and characters in languages. Emotions can be extracted from two essential text forms: written texts and conversation (dialogues). For written texts, many scholars focus on working with sentence level to extract "words/phrases" representing emotions.

Emotion recognition in audio:

Different from emotion recognition in text, vocal signals are used for the recognition to extract emotion from audio.

Emotion recognition in video:

Video data is a combination of audio data, image data and sometimes texts (in case of subtitle).

Emotion recognition in conversation:

Emotion recognition in conversation(ERC) extracts opinions between participants from massive conversational data in social platform, such as Facebook, Twitter, YouTube, and others. ERC can take input data like text, audio, video or a combination form to detect several emotions such as fear, lust, pain, and pleasure.

WHAT IS SPEECH EMOTION RECOGNITION USED FOR TODAY?

Nowadays, emotion recognition is used for various purposes that some people do not even notice on a daily basis. Here are some of the areas that would show that emotion recognition is beneficial:

- Education
- Customer Services
- Healthcare Applications
- Disability Assistance

Results :

	Predicted Labels	Actual Labels
0	Angry	Angry
1	Sad	Sad
2	Neutral	Neutral
3	Fear	Happy
4	Sad	Neutral
5	Fear	Fear
6	Sad	Sad
7	Fear	Disgust
8	Neutral	Neutral
9	Happy	Happy

Conclusion:

- We can see our model is more accurate in predicting angry, surprise emotions and it makes sense also because audio files of these emotions differ to other audio files in a lot of ways like pitch, speed etc.
- We overall achieved 47% accuracy on our test data and its not decent but we can improve it more by applying more augmentation techniques and using other feature extraction methods.

Reference:

- 1) https://www.almabetter.com/
- 2) https://www.wikipedia.org
- 3) https://www.kaggle.com/
- 4) https://github.com/