ML Project:
Challa Venkata Raghava Reddy
AM.EN.U4CSE17506

Problem Statement:

To design an efficient machine learning algorithm that will be able to identify and classify the Music into its genre.
The model is said to learn and differentiate the type of songs based on their genre.

- Task(T): Classify the songs into their genre
- Experience(E): A data set of songs with its genre
- Performance(P): Classification accuracy, the number of songs predicted correctly out of all songs.

Dataset:

This dataset has various genres like pop,rock, blues etc.

Data Preprocessing:

We first converted all the data into wav format using librosa package and then convert them into data that can be used in CSV format.

librosa is a **python** package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems

I had loaded the data which is in wav format into librosa. load which returns us the time series and the default rate is 22050, then this time series is used to extract multiple features like

chroma short-time Fourier transform: It is a variant of Fourier transform which splits the audio signal into frames and then takes the Fourier transform of each frame.

root-mean-square (RMS) energy for each frame, using the time series or audio samples generated above.

Spectral - centroid: Each frame of a magnitude spectrogram is normalized and treated as a distribution over frequency bins, from which the mean (centroid) is extracted per frame or indicates at which frequency the energy of a spectrum is centered upon.

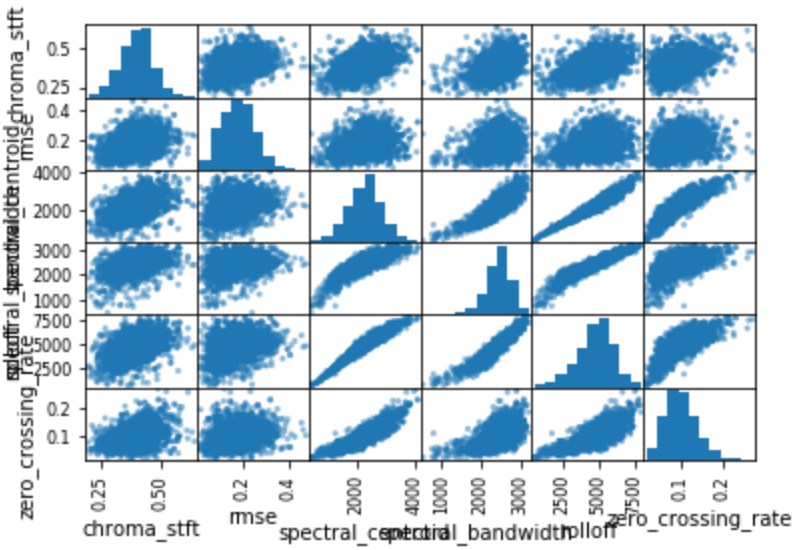
Spectral roll-off is the frequency below which a specified percentage of the total spectral energy, e.g. 85%, lies.

We also used the zero-crossing rate and mfcc in our feature set.

Data Visualization and summarization:

In this, we visualize the data and find the relation between the attributes present in the data.

We use scatter plot and scatter matrix to find the correlation between necessary attributes and then we can remove some attributes which are similarly related.



Data Interpretation:

We then converted the wav format to csv format dataset and check if there are any null values and used standardization to make our dataset uniform.

Python Packages:

Numpy: *NumPy* is a *python* library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. This is used for numpy arrays and functions.

Pandas: **Pandas** is a high-level data manipulation tool. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables. This is used for storing data in the data frame and manipulating data for data visualization and preprocessing.

Matplotlib: **Matplotlib** is a plotting library for the **Python** programming language and its numerical mathematics extension NumPy. This is used for plotting graphs.

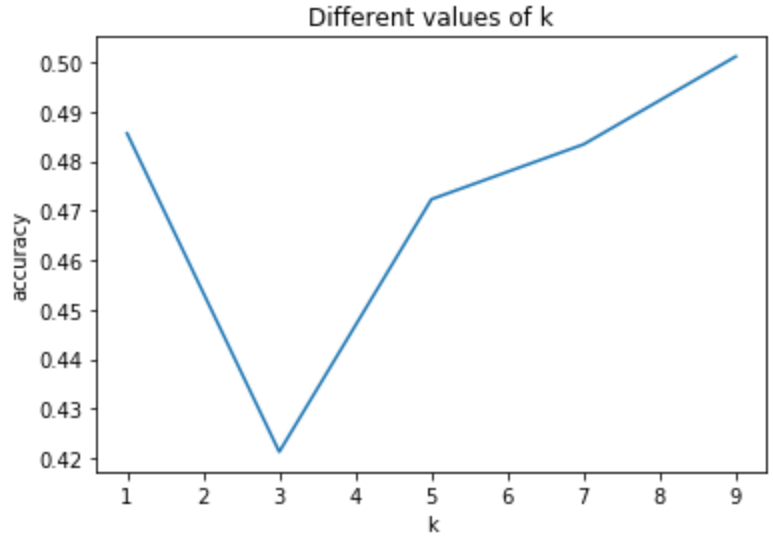
Librosa: L**ibrosa** is a **python** package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. This is used for getting information from audio files.

Sklearn: **Scikit-learn** is a free machine learning library for **Python**. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports **Python** numerical and scientific libraries like NumPy and SciPy. This is used to implement machine learning algorithms.

Supervised Learning Algorithms:

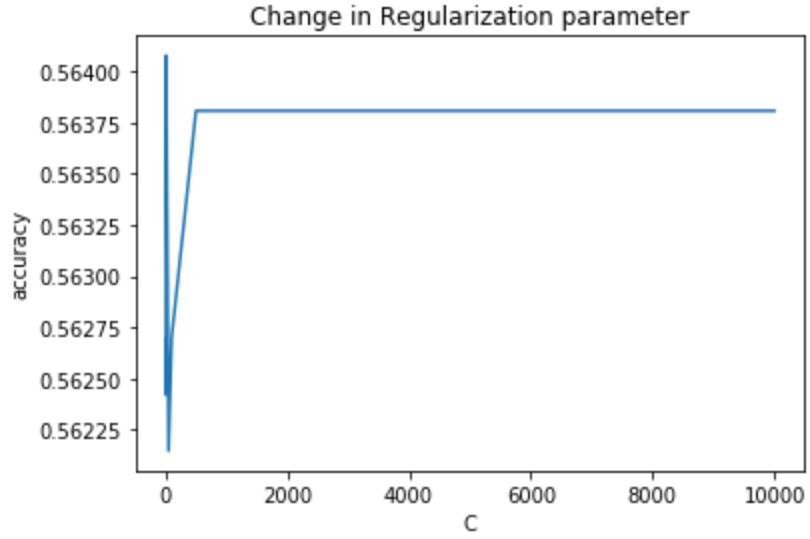
KNN: K-Nearest Neighbors, or *KNN* for short, is one of the simplest machine learning algorithms and is used in a wide array of institutions. *KNN* is a non-parametric, lazy learning algorithm.

The change in accuracy when varying K will be shown in the graph below



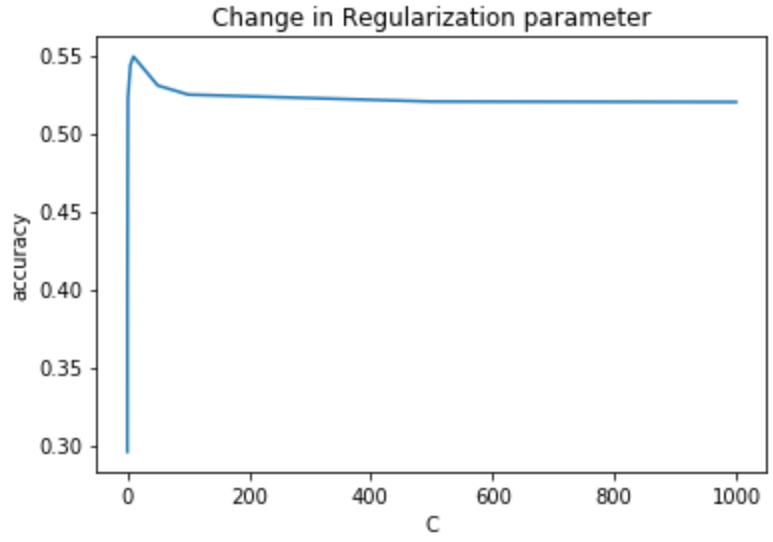
Logistic Regression: This is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables

There is a change in accuracy while varying the regularization parameter.



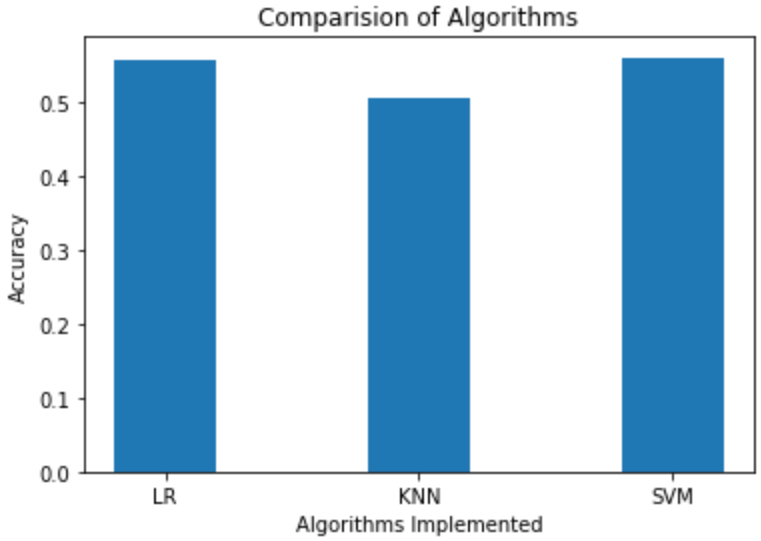
SVM: A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for n group classification problems.

The change in accuracy while varying the parameter C is given below



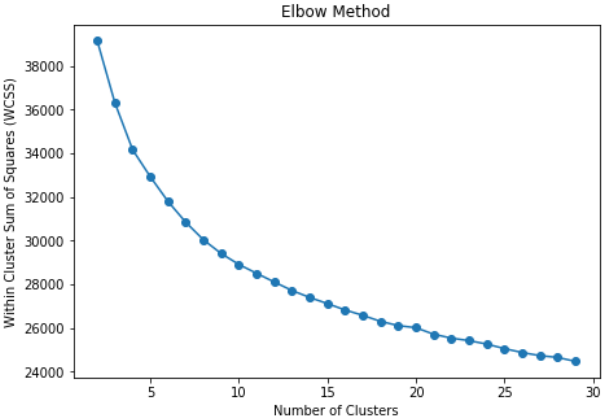
Comparison between algorithms:

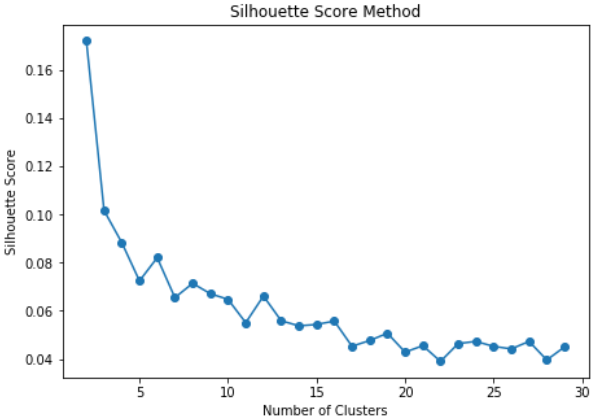
On comparing the algorithms the conclusion is that performance of SVM is good on both validation and testing data.



Unsupervised Learning algorithm

For Unsupervised Learning algorithm I used KMedoids and KMeans and found that the wcss Values are very high and thus unsupervised learning algorithms doesn't work well for this dataset and the score is found to be 0.06.


Elbow Method
Within Cluster Sum of Squares (WCSS)
38000
36000
34000
32000
30000
28000
26000
24000
5
10
15
20
25
30
Number of Clusters


Silhouette Score Method
Silhouette Score
0.16
0.14
0.12
0.10
0.08
0.06
0.04
5
10
15
20
25
30
Number of Clusters