# MUSIC GENRE CLASSIFICATION

*KR Abishekraswanth* CB.EN.U4CCE20001
*Akshara R* CB.EN.U4CCE20003
*R Rangashree Dhanvanth* CB.EN.U4CCE20048
*Swathi P* CB.EN.U4CCE20062

*Abstract - Music genre prediction is one of the interesting topics that audio signal processing is interested in. Digital signal processing methods were used in this study to extract the acoustic features of the music, and machine learning techniques were used to classify the music's genres and generate music recommendations. The GTZAN database was used in the study, and the SVM algorithm had the highest accuracy overall.*

## 1. INTRODUCTION

Significant changes have occurred in the music industry as well as other fields due to the growing use of the Internet. Examples of these advances include the widespread usage of online music listening and purchasing platforms, copyright management for musical works, genre classification, and music recommendations. People may listen to music anywhere, at any time, this is due to the development of music broadcast platforms. Utilising the features extracted from a raw music wav file using audio processing techniques, the study's goal is to classify different types of music based on their genres. The features that are to be extracted are certain time and frequency domain characteristics of the wave of the music signal.

To compare all the performances, classification analysis is done on the GTZAN dataset and the accuracy is compared. With this the conclusion is made that the genre predicted or classified by the most accurate classifier is the genre of the music.

## 2. METHODS

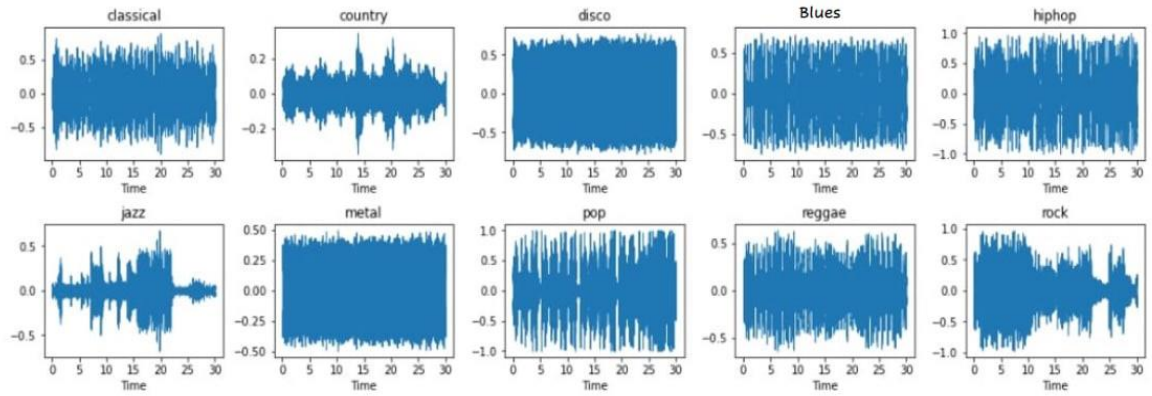
### 2.1 DATASET:

One of the most widely used datasets for music signal processing is GTZAN, which was first proposed by G. Tzanetakis. It has 1,000 tracks with a sampling frequency of 22050 Hz, 30 seconds, and 16 bits. The GTZAN has 100 songs in the blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock genres. The dataset consists of music files along with 2 csv files, containing the features of the audio music in the duration of 3 and 30 secs. The features include time and frequency domain, which is used to train the model for classification. The following table gives us the number of music files in each genre present in the dataset.

| SNo. | GENRE | NUMBER |
|---|---|---|
| 0 | blues | 100 |
| 1 | classical | 100 |
| 2 | country | 100 |
| 3 | disco | 100 |
| 4 | hip-hop | 100 |
| 5 | jazz | 100 |
| 6 | metal | 100 |
| 7 | pop | 100 |
| 8 | reggae | 100 |
| 9 | rock | 100N |

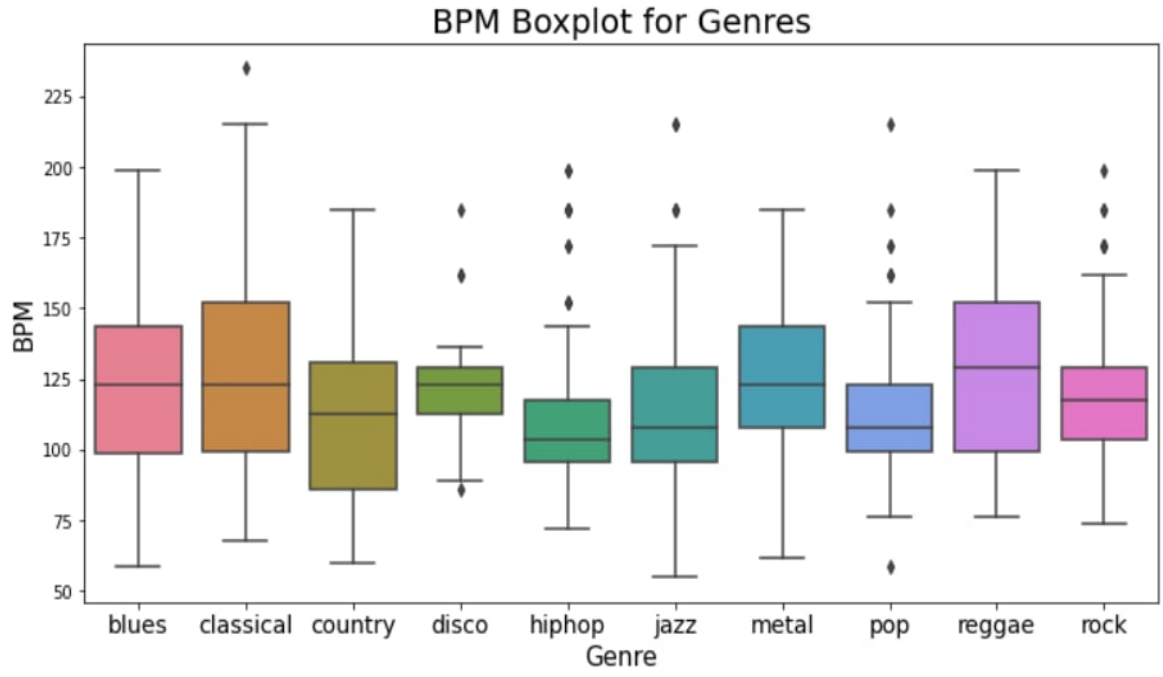
## 2.2 DATA VISUALISATION:

The data of the music files are visualised in a wave format using the python libraries such as librosa. For each genre the wave is generated, the following picture gives us an idea about how each genre looks when varying with respect to time. The difference of each wave can be seen.



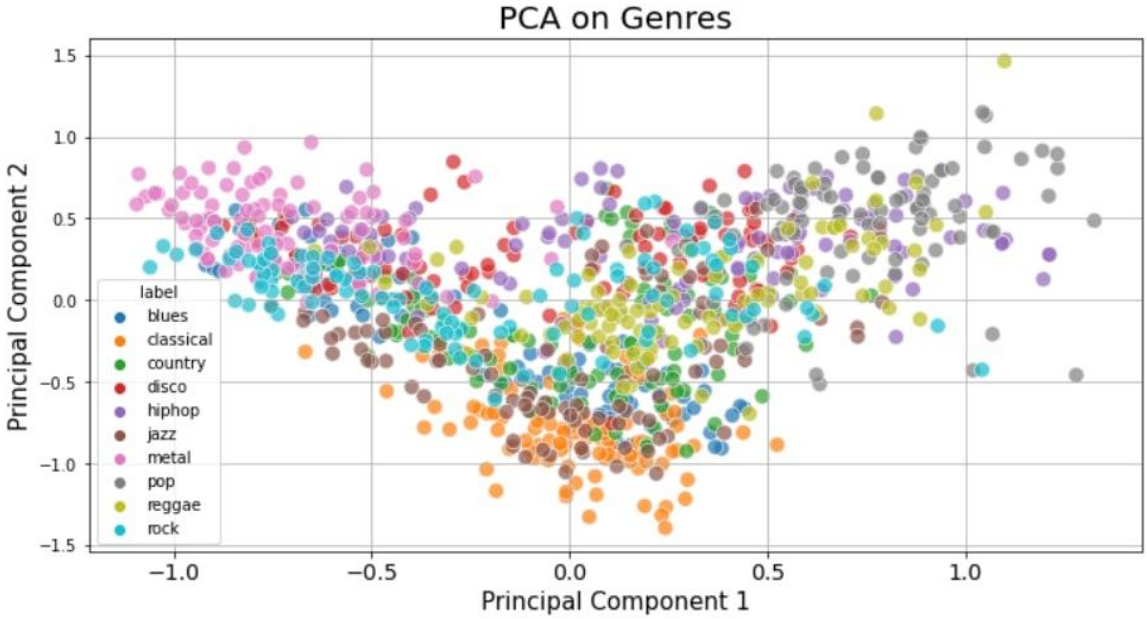

- BOX PLOTS:

Box plots are a method to understand and visualise data easily as they give the five-member summary of each of the 10 genres present. With the help of this five-member summary, it is easy to understand the minimum, maximum and average variations of the data. This also gives the idea of the outliers present and how it varies with respect to the beats per minute of the music.



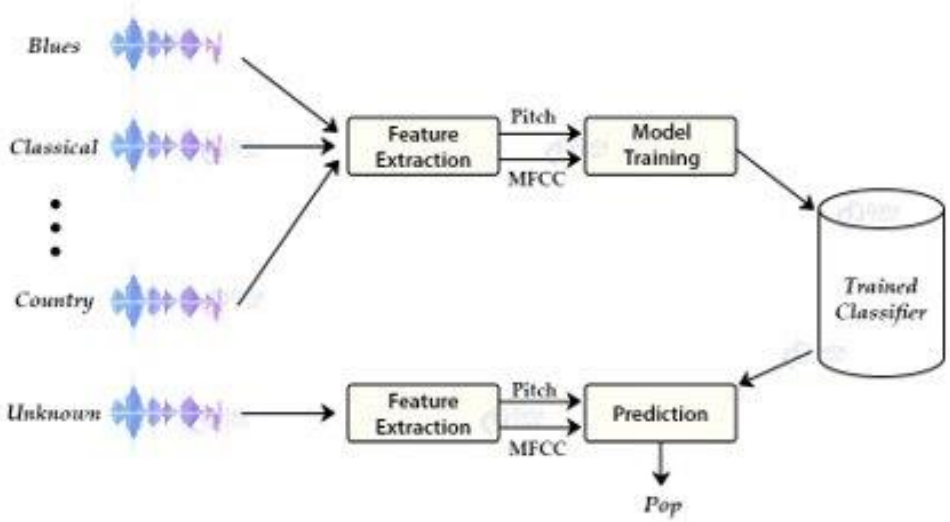
- PRINCIPAL COMPONENT ANALYSIS:

Principal Component Analysis (PCA) is an unsupervised statistical technique used to examine the interrelations among a set of variables. This has been implemented to visualise the data with the help of 2 principal components, where there is a reduction in dimensionality. The interrelations among the genres are visualised in the form of a graphical representation.




## 2.3 FEATURE EXTRACTION:

In audio processing, feature extraction is another important process that should be done. These features are fed to the model for training and testing in the model building process, hence predicting the genre of the music.



The features of the audio signal are extracted and statistical measures such as mean, standard deviation and variance are calculated, and are used to train the model that is built. There are multiple features of an audio signal; the few features that are selected are represented in the table below. Considering the 57 extracted features, the classifier is trained.

| FEATURE | Statistical Functions | No of Features |
|---|---|---|
| Chroma stft | Mean<br>Variance | 2 |
| rms | | 2 |
| Spectral centroid | | 2 |
| Spectral bandwidth | | 2 |
| Rolloff | | 2 |
| Zero crossing rate | | 2 |
| Harmony | | 2 |
| Percussive rate | | 2 |
| Tempo | | 1 |
| MFCC (20 coefficients) | | 40 |

## 2.3.1 TIME DOMAIN FEATURES:

Features that are extracted from raw audio signals are said to be the time domain features of the wave.

- ZERO CROSSING RATE:

ZCR gives a general picture of the frequency content. In a certain time span, it shows the rate of signal change. When a signal switches from being negative to positive, or vice versa, it is considered a signal change.
ZCR is computed at time T as:

$$ZCR(T) = \frac{\sum_{n=1}^{L}[signal_n * signal_{n+1} < 0]}{L}$$

where L is the total number of signals in time slot T. If $n^{th}$ signal, $signal_n$, times its next signal, signal+1 is negative, a change will be counted. ZCR is the sum of changes divided by the total number of signals L.

- TEMPO:

The speed at which the music is played and is measured in beats per minute. It is expressed in terms of Beats Per Minute (BPM). As the tempo varies with time, the mean and variance of it is considered.

- ROOT MEAN SQUARE (RMS):

The energy in a music signal is calculated using the following formula;

$$\sum_{n=1}^{N} |x(n)|^2$$

Hence, the root mean square value can be computed as:

$$\sqrt{\frac{1}{N}\sum_{n=1}^{N} |x(n)|^2}$$

This is calculated across all frames; mean and variance are computed and considered.

## 2.3.2 FREQUENCY DOMAIN FEATURES:

Audio signals that are transformed into their frequency domains using Fourier transform are called frequency domain features.

- MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC):

The Mel Frequency Cepstral Coefficients (MFCCs), a reduced set of characteristics, are used to define the general features of a spectral envelope. It might be regarded as a timbre feature. The MFCC is used to modify the cepstral coefficients for the benefit of the human hearing system. Cepstral coefficients are linear scales.

MFCC is one of the commonly used features in speech and speaker recognition systems. Steps of MFCC are Frame Blocking, Windowing, Fast Fourier Transform, Mel Frequency Wrapping and Spectrum, respectively.

Short-time Fourier transform is applied on a continuous signal by first breaking them into shorter signals like frames. The window functions typically overlap one another to ensure continuity. After the Fourier transform, adding some window functions, such as the Hanning or Hamming windows, tends to reduce noise.

- SPECTRAL CENTROID:

The brightness of a sound can be measured using a spectral centroid. It is the average frequency component of any audio signal's frequency spectrum. The average frequency divided by the sum of the amplitudes generates the individual centroid of a spectral frame, or:

$$Spectral \quad Centroid \quad = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]}$$

F [k] is the amplitude corresponding to bin k in the DFT spectrum.

- SPECTRAL BANDWIDTH:

The weighted average amplitude difference between frequency magnitude and brightness is provided by spectral bandwidth. It provides information about the frame's frequency range.

- SPECTRAL ROLLOFF:

The normalised frequency at which the sum of the sound's low frequency power values reaches a particular rate in the total power spectrum is known as the spectral rolloff frequency. It can be summed up as the frequency value corresponding to a particular ratio of the spectrum's distribution.

- CHROMA STFT FEATURES:

In the world of music information retrieval, the chroma feature is a common audio melody feature. It was created using a twelve-tone equal temperament. Knowing the distribution of chroma without knowing the absolute frequency (i.e., the original octave) can still provide useful musical information about the audio and may even reveal perceived musical similarity that is not visible in the original spectra because in music, notes exactly one octave apart are perceived as being particularly similar.

A typical representation of a chroma feature is a 12-dimensional vector, v=[V(1), V(2), V(3),... V(12)], where each element of the vector corresponds to a different component of the set of letters {C, C#, D, D#, E, F, F#, G, G#, A, A#,B} which reflects the local energy distribution of the audio signal.

# 3. MUSIC GENRE CLASSIFICATION USING MACHINE LEARNING

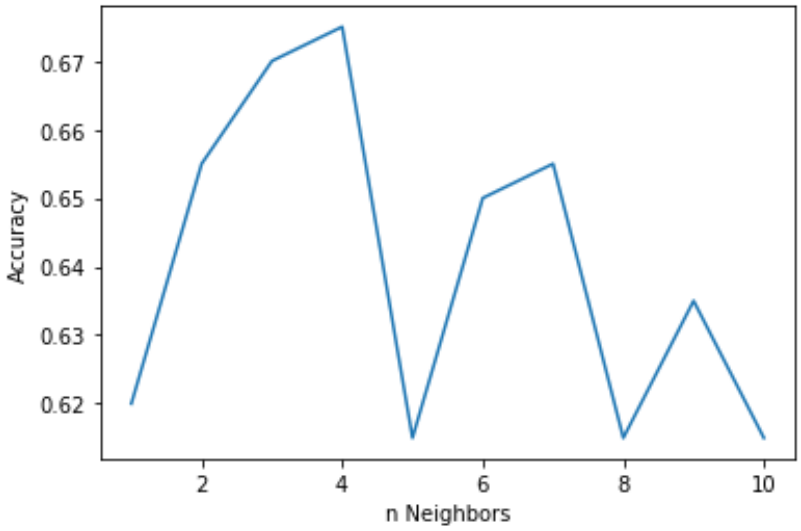
Music genre classification is done by training and testing different classifiers with the features of the music signals. These classifiers include K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree (DT) and Naive Bayes (NB).

## 3.1 CLASSIFIERS:

- K- NEAREST NEIGHBOURS-KNN:

KNN is one of the distance-based supervised learning algorithms. This approach does not include the building of a model; instead, the test operation is carried out on the data set's labelled samples. The distance from the examples in the dataset will be used to construct a new

instance of the class label. By voting on the class labels of the closest k, the class tag is estimated using these determined distances. The Euclidean and Manhattan distance formulas are commonly implemented when calculating distance. The following graph shows the accuracy of KNN considering different number of nearest neighbours.




- SUPPORT VECTOR MACHINE-SVM:

One of the model-based supervised learning algorithms is the SVM. SVM is built on the idea of training for a decision surface that will allow the two classes to be distinguished. The boundary regions of the two classes are optimised to produce this decision surface. Other than two-class data sets, SVM can be applied to multi-class data sets using OneVsRest and OneVsOne aspects.

- DECISION TREE-DT:

Decision trees are supervised and model-based learning techniques. It aims to determine the most distinguishing feature in the data set as the tree's root node. When the most differentiating feature is identified, an entropy calculation is performed. There are also several measures in the literature that provide distinguishing characteristics.

- NAIVE BAYES-NB:

The Naive Bayes algorithm is a probabilistic supervised learning technique that develops a classification model by calculating initial probabilities from the data in the data set and then classifies new data using this model. It is an algorithm that can be applied to a variety of issues because it is compatible with all types of data and only needs basic statistical computations.

These four models built are trained using the given dataset and by testing them, the accuracy and classification report can be considered to find out the best model for music genre classification.

## 3.2 METRICS:

In order to evaluate the performance of the models described above, the following metrics will be used.

1. ACCURACY: Refers to the percentage of correctly classified test samples

$$accuracy = \frac{true\ positive + true\ negative}{true\ positive + false\ positive + true\ negative + false\ negative}$$

2. PRECISION: Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

3. RECALL: The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples.

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

4. AUC: The area under the receiver operator characteristics (ROC) curve is a standard criterion for judging the performance of a multi-class classification system. The true positive rate and the false positive rate are plotted on a graph called the ROC. Because the system being created is intended to have a greater AUC than 0.5, it is assumed that the baseline model, which randomly predicts each class label with equal probability, will have an AUC of 0.5.

| AUC Interpretation | |
|---|---|
| **AUC Value** | **Interpretation** |
| **>= 0.9** | Excellent Model |
| **0.8 to 0.9** | Good Model |
| **0.7 to 0.8** | Fair Model |
| **0.6 to 0.7** | Poor Model |
| **<0.6** | Very Poor Model |

# 4. EXPERIMENT RESULTS:

After training the models, the model characteristics are obtained by testing the model with the test set which is 20% of the dataset without the labels. The following are the classification reports of the models with their respective performance metrics and accuracy.

## 4.1. CLASSIFICATION REPORT:

### A. KNN

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| blues | 0.73 | 0.61 | 0.67 | 18 |
| classical | 0.73 | 0.83 | 0.78 | 23 |
| country | 0.59 | 0.76 | 0.67 | 21 |
| disco | 0.42 | 0.62 | 0.50 | 16 |
| hiphop | 0.81 | 0.59 | 0.68 | 22 |
| jazz | 0.64 | 0.47 | 0.55 | 19 |
| metal | 0.94 | 0.71 | 0.81 | 24 |
| pop | 0.71 | 0.80 | 0.75 | 15 |
| reggae | 0.68 | 0.72 | 0.70 | 18 |
| rock | 0.58 | 0.58 | 0.58 | 24 |
| accuracy | | | 0.67 | 200 |
| macro avg | 0.68 | 0.67 | 0.67 | 200 |
| weighted avg | 0.69 | 0.67 | 0.67 | 200 |

### B. SVM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| blues | 0.69 | 0.61 | 0.65 | 18 |
| classical | 0.87 | 0.87 | 0.87 | 23 |
| country | 0.73 | 0.76 | 0.74 | 21 |
| disco | 0.39 | 0.69 | 0.50 | 16 |
| hiphop | 0.78 | 0.64 | 0.70 | 22 |
| jazz | 0.73 | 0.84 | 0.78 | 19 |
| metal | 0.77 | 0.71 | 0.74 | 24 |
| pop | 0.75 | 0.80 | 0.77 | 15 |
| reggae | 0.68 | 0.83 | 0.75 | 18 |
| rock | 0.73 | 0.33 | 0.46 | 24 |
| accuracy | | | 0.70 | 200 |
| macro avg | 0.71 | 0.71 | 0.70 | 200 |
| weighted avg | 0.72 | 0.70 | 0.70 | 200 |

## C. DECISION TREE

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| blues | 0.42 | 0.44 | 0.43 | 18 |
| classical | 0.82 | 0.78 | 0.80 | 23 |
| country | 0.50 | 0.43 | 0.46 | 21 |
| disco | 0.18 | 0.19 | 0.18 | 16 |
| hiphop | 0.44 | 0.36 | 0.40 | 22 |
| jazz | 0.52 | 0.68 | 0.59 | 19 |
| metal | 0.75 | 0.62 | 0.68 | 24 |
| pop | 0.50 | 0.73 | 0.59 | 15 |
| reggae | 0.32 | 0.39 | 0.35 | 18 |
| rock | 0.35 | 0.25 | 0.29 | 24 |
| accuracy | | | 0.49 | 200 |
| macro avg | 0.48 | 0.49 | 0.48 | 200 |
| weighted avg | 0.50 | 0.49 | 0.49 | 200 |

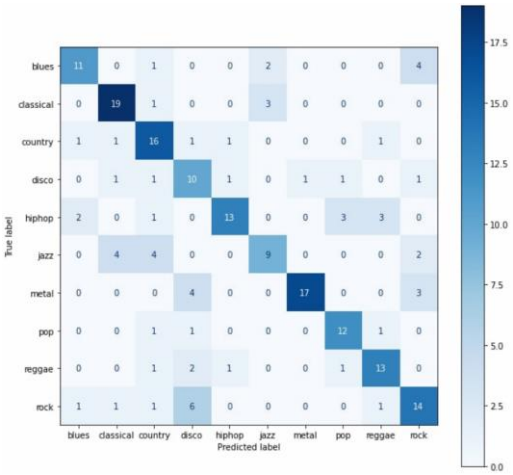
## D. NAÏVE BAYES

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| blues | 0.40 | 0.33 | 0.36 | 18 |
| classical | 0.87 | 0.87 | 0.87 | 23 |
| country | 0.52 | 0.67 | 0.58 | 21 |
| disco | 0.25 | 0.38 | 0.30 | 16 |
| hiphop | 0.56 | 0.23 | 0.32 | 22 |
| jazz | 0.67 | 0.63 | 0.65 | 19 |
| metal | 0.52 | 0.71 | 0.60 | 24 |
| pop | 0.54 | 0.87 | 0.67 | 15 |
| reggae | 0.71 | 0.67 | 0.69 | 18 |
| rock | 0.30 | 0.12 | 0.18 | 24 |
| accuracy | | | 0.54 | 200 |
| macro avg | 0.53 | 0.55 | 0.52 | 200 |
| weighted avg | 0.54 | 0.54 | 0.52 | 200 |

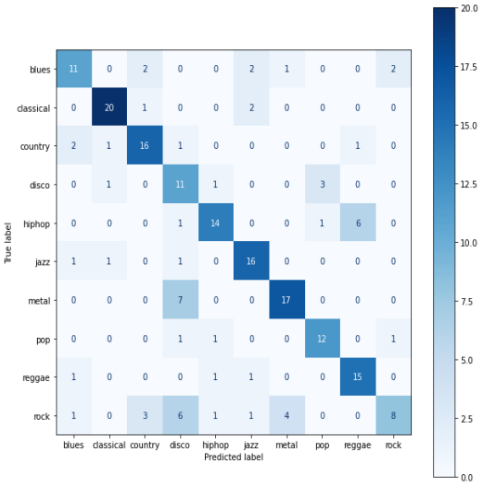
# 4.2. CONFUSION MATRIX:

The confusion matrix is a tabular representation that allows us to better understand the model's strengths and flaws. The number of test examples of class i that the model predicted as class j is represented by matrix element $a_{ij}$. The diagonal elements ($a_{ii}$) match the accurate predictions. The confusion matrix for all four models is shown, to understand the true and false predicted labels.

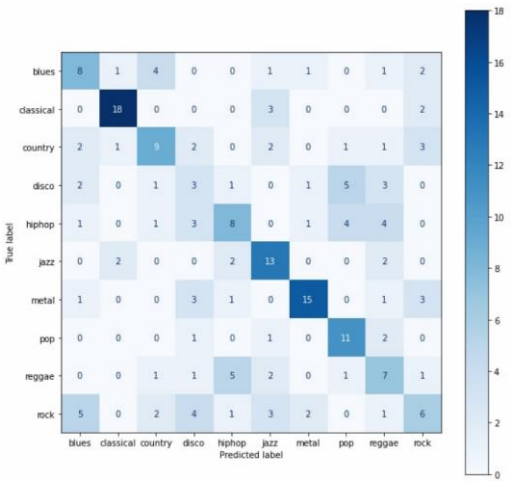
A. KNN CLASSIFIER



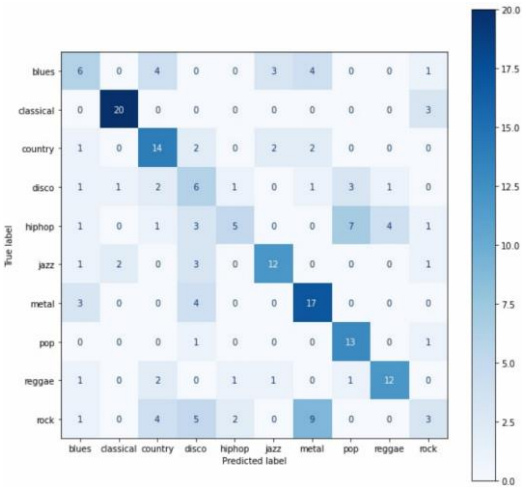

B. SVM CLASSIFIER



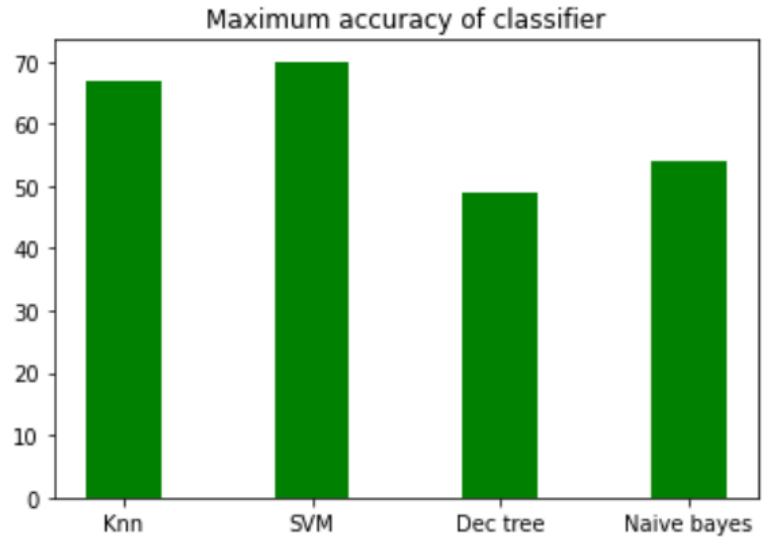

C. DECISION TREE CLASSIFIER




D. NAÏVE BAYES CLASSIFIER



The accuracy of the models thus obtained are 67% for KNN, 70% for SVM, 49% for decision trees and 54% for naive bayes classifiers as shown below.




We can see that SVM model gives a better result with 70% accuracy and is considered as the best classifier to predict the music genres.

## 4.3. AUC-ROC SCORE:

AUC is an abbreviation for area under the curve. It is used in classification analysis in order to determine which of the used models predicts the classes best. An example of its application are ROC curves. Here, the true positive rates are plotted against false positive rates. It is seen that SVM has an AUC score of 0.96 which is near to the ideal score. Hence the two curves true negative and true positive has nearly less overlap which means the model has ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.

| | classifier | Auc Roc score |
|---|---|---|
| 0 | Knn | 0.922004 |
| 1 | SVM | 0.962217 |
| 2 | Dec tree | 0.723916 |
| 3 | Naive bayes | 0.899126 |

# 5. CONCLUSION:

The purpose of this study is to categorise and recommend songs based on auditory data gathered utilising digital signal processing technologies. The GTZAN dataset is used to understand the classification process, and find out the best classifier to predict the genre of the music. The study was done in two stages: figuring out how to get the features of the music and creating a service that classifies music genres. Digital signal processing methods were used to extract features. The findings listed in the previous tables show that SVM outperformed other methods in terms of classification outcomes.

# REFRENCES:


[1] Musical Genre Classification of Audio Signal',George Tzanetakis and Perry Cook,IEEE Transactions on Speech and Audio Processing, 10(5), July 2002.

[2] https://docs.google.com/presentation/d/1Y024bJxO-XtvH163Te59dOHUYZXpMY3nDPI2lb48RKo/htmlpresent

[3] https://www.analyticsvidhya.com/blog/2022/03/music-genre-classification-project-using-machine-learning-techniques/

[4] https://www.researchgate.net/publication/324218667_Music_Genre_Classification_using_Machine_Learning_Techniques

[5] https://www.researchgate.net/publication/329396097_Music_Genre_Classification_and