MUSIC GENRE CLASSIFICATION

Data Science and Artificial Intelligence- FCSB (group 8)

By- Dahiya Adit, Dahiya Advik, Kakkar Dhairya

MOTIVATIONS

It’s frustrating when you can’t easily discover the songs you want to hear. Our group thought about this situation and realised that it is a common and big problem most of us face.

Worldwide listeners struggle to discover new songs within their preferred genres due to the vast volume of music uploads, lack of time to explore each track and even a lack of mood of doing so.

PROBLEM DEFINITION

Predicting the music genres after analyzing the characteristics of the songs to improve user experience on music platforms.

EXPLORATORY DATA ANALYSIS

```
instance_id          float64
artist_name           object
track_name            object
popularity           float64
acousticness         float64
danceability         float64
duration_ms          float64
energy               float64
instrumentalness     float64
key                   object
liveness             float64
loudness             float64
mode                  object
speechiness          float64
tempo                 object
obtained_date         object
valence              float64
music_genre           object
dtype: object
```

```
print("Data type : ", type(music_df))
print("Data dims : ", music_df.shape)
```

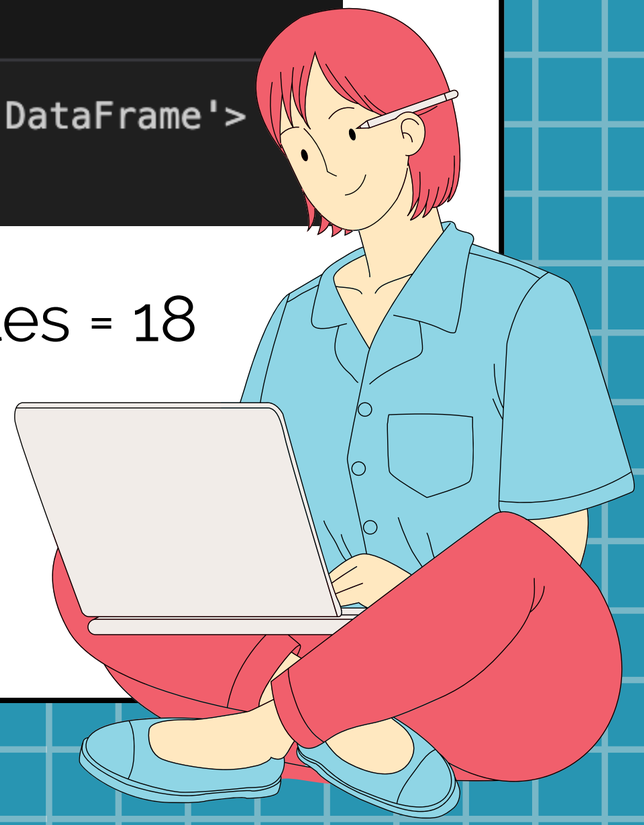
```
Data type :  <class 'pandas.core.frame.DataFrame'>
Data dims :  (50005, 18)
```

Number of columns/variables = 18

Number of rows = 50,005

Float variables = 11

Object variables = 7



EXPLORATORY DATA ANALYSIS

Possible Genres:

'Electronic', 'Anime', 'Jazz', 'Alternative', 'Country', 'Rap', 'Blues', 'Rock', 'Classical', 'Hip-Hop'.

Selected Variables:

```
music.head()
```

	popularity	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence
0	27.0	0.00468	0.652	0.941	0.79200	0.115	-5.201	0.0748	100.889	0.759
1	31.0	0.01270	0.622	0.890	0.95000	0.124	-7.043	0.0300	115.00200000000001	0.531
2	28.0	0.00306	0.620	0.755	0.01180	0.534	-4.617	0.0345	127.994	0.333
3	34.0	0.02540	0.774	0.700	0.00253	0.157	-4.498	0.2390	128.014	0.270
4	32.0	0.00465	0.638	0.587	0.90900	0.157	-6.266	0.0413	145.036	0.323

CLEANING THE DATASET

	mode	speechiness	tempo	obtained_date	valence
00000000000	Minor	0.0748	100.889	4-Apr	
00000000000	Minor	0.03	115.00200000000000	4-Apr	
-4.617	Major	0.0345	127.994	4-Apr	0.33300
-4.498	Major	0.239	128.014	4-Apr	
-6.266	Major	0.0413	145.036	4-Apr	0.32300
-10.517	Minor	0.0412	?	4-Apr	
-4.294	Major	0.351	149.995	4-Apr	
-9.339	Minor	0.0484	120.008	4-Apr	0.76100
-3.175	Minor	0.268	149.94800000000000	4-Apr	
-7.091	Minor	0.173	139.933	4-Apr	
-13.787	Minor	0.0345	57.528	4-Apr	
-5.439	Minor	0.0609	178.543	3-Apr	
-3.464	Major	0.0645	128.043	4-Apr	
-10.536	Minor	0.0424	154.745	4-Apr	
00000000000	Major	0.185	139.911	4-Apr	
-2.51	Major	0.0904	100.024	4-Apr	

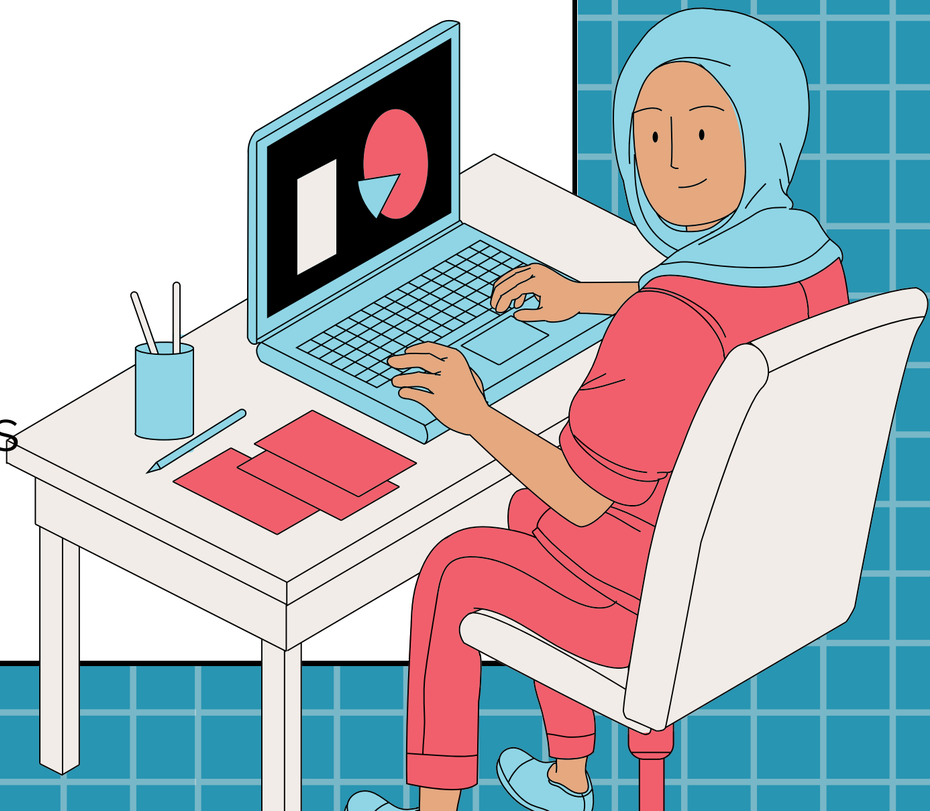
Some values of “tempo” contained a **“?”** value making it an object type variable. Therefore the data **had to be cleaned**

CLEANING THE DATASET

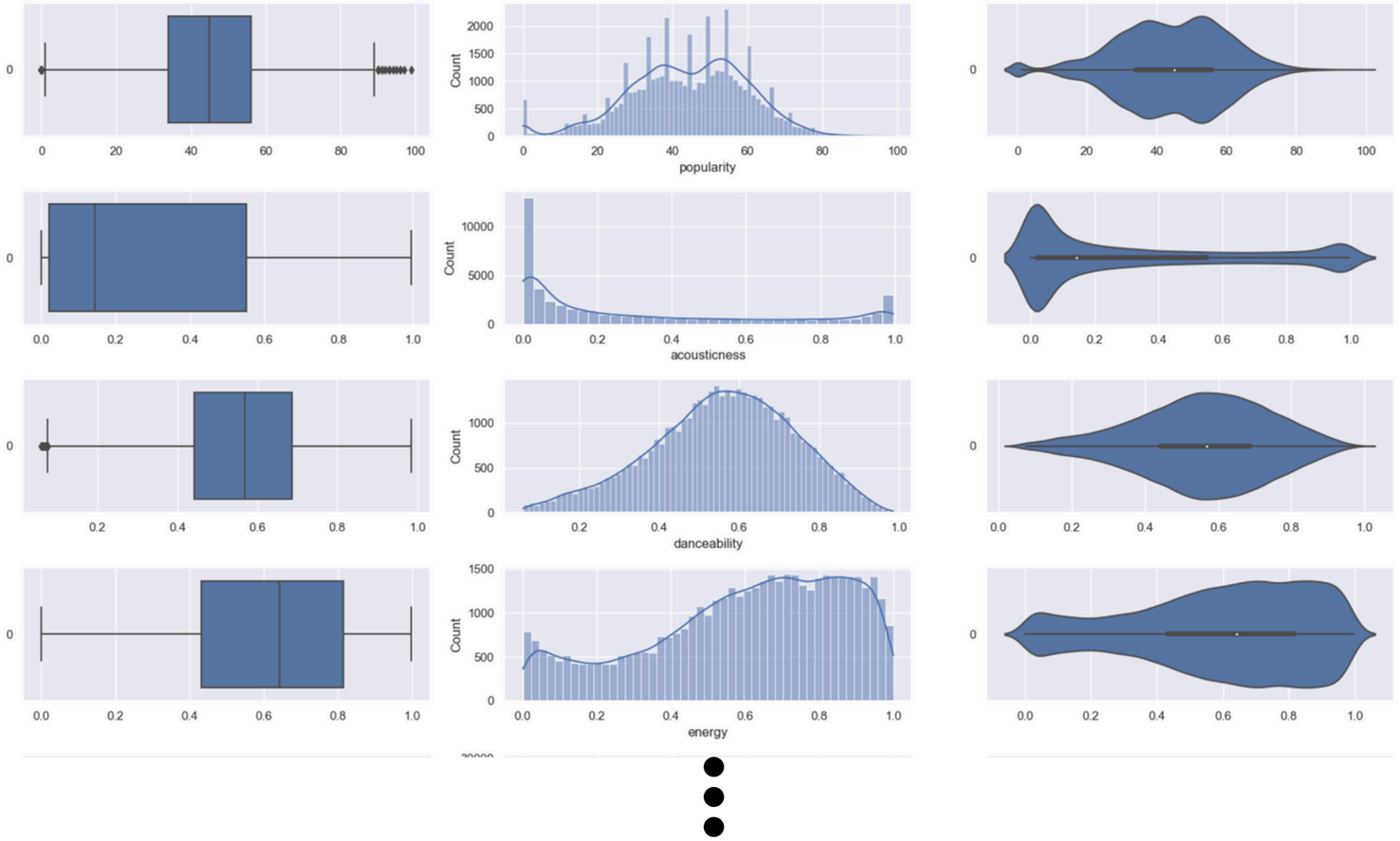
```
[9]  music['tempo'] = pd.to_numeric(music['tempo'], errors = 'coerce')

[10] music_df['tempo'] = pd.to_numeric(music_df['tempo'], errors = 'coerce')
     music_df = music_df.dropna()
     print("Data dims : ", music_df.shape)

     Data dims :  (45020, 18)

[11] music = music.dropna()
     print("Data dims : ", music.shape)
```

Using “**.dropna()**”

Tempo (object) → Tempo (float64) → “?” changed to NaN → NaN rows deleted



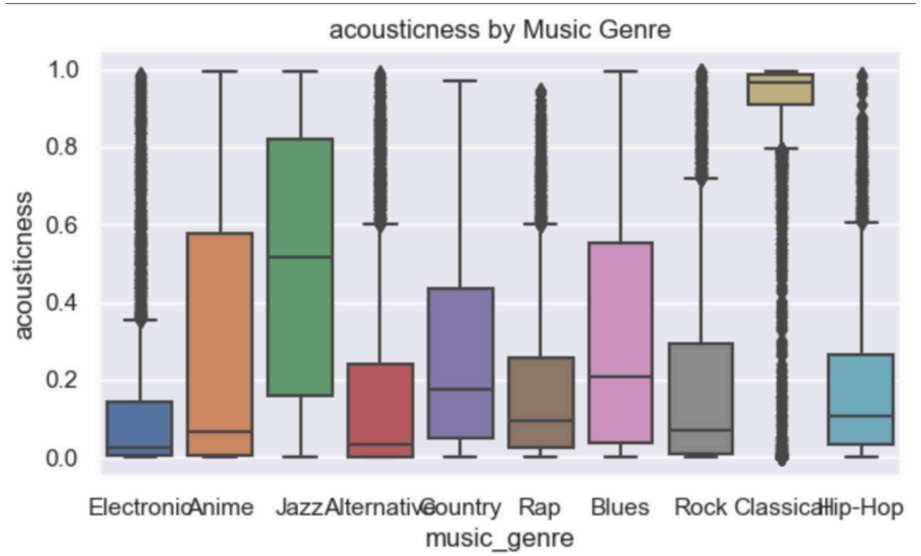
VISUAL ANALYSIS



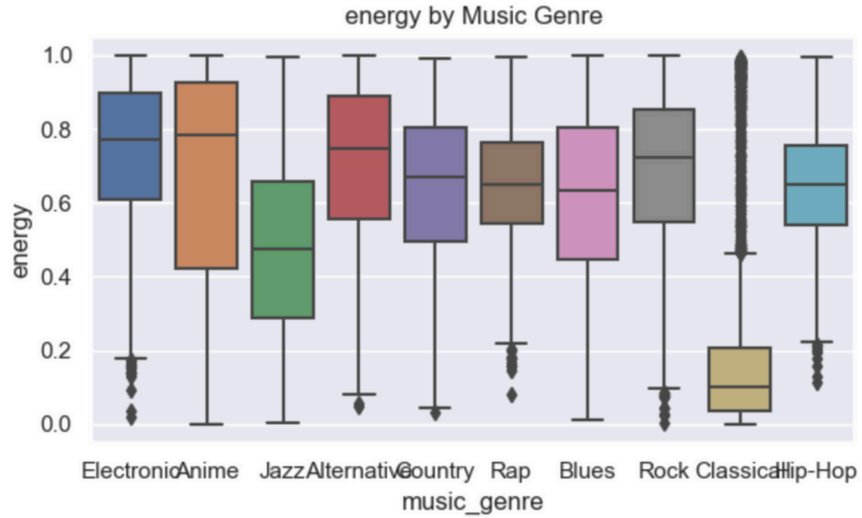
Making Uni-Variate statistics of the selected variables

- Bar plot
- Histogram
- Violin plot

VISUAL ANALYSIS



More than 50% of the classical music genres have acousticness 0.9 and 1.0



More than 50% of the classical music genres have energy 0.1 and 0.4

MODELS IMPLEMENTED

Machine Learning Technique used:

Classification

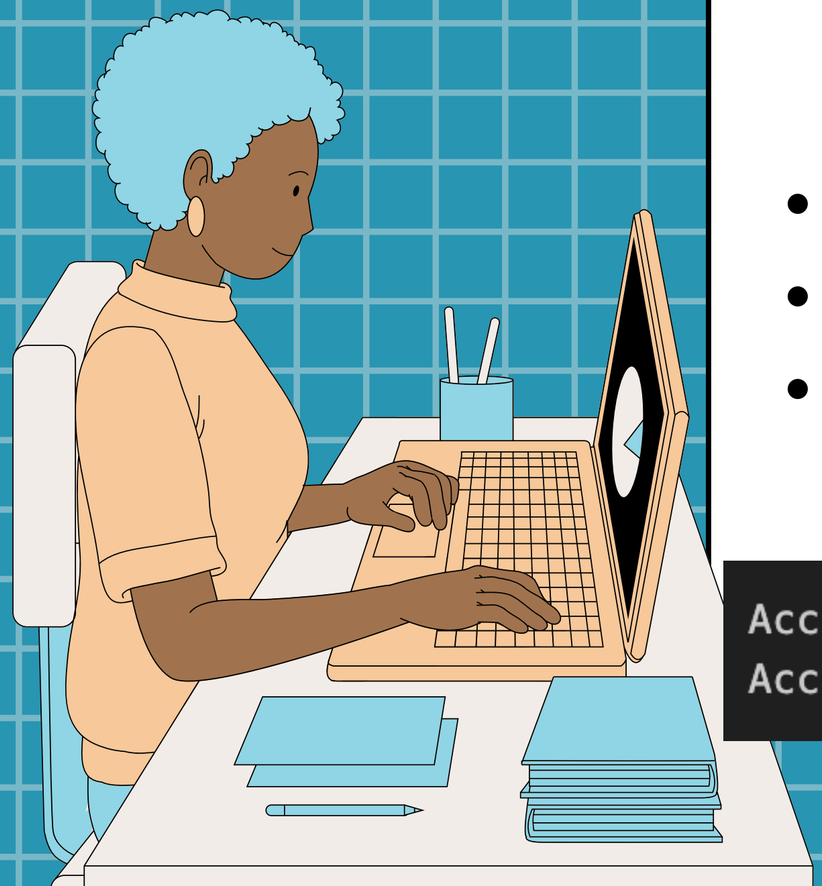
Classification Methods Used:

1) Random Forest Classifier
2) Decision Tree Classifier
3) Gradient Boosting Classifier
4) Gaussian Naive Bayes Method

RANDOM FOREST CLASSIFIER

- Constructs multiple decision trees
- Selects random feature subsets
- Combines each tree to give result

```
Accuracy of Random Forest Classifier(Train set): 96.7919020118042 8%
Accuracy of Random Forest Classifier(Test set): 54.575744113727 23%
```



DECISION TREE CLASSIFIER

- Tree-like structure
- Recursively partitions the data based on features.
- Selects the feature that best splits the data forming a decision path

```
Accuracy of Decision Tree Classifier(Train set): 96.79190201180428%
Accuracy of Decision Tree Classifier(Test set): 43.995261365319 12%
```

GRADIENT BOOSTING CLASSIFIER:

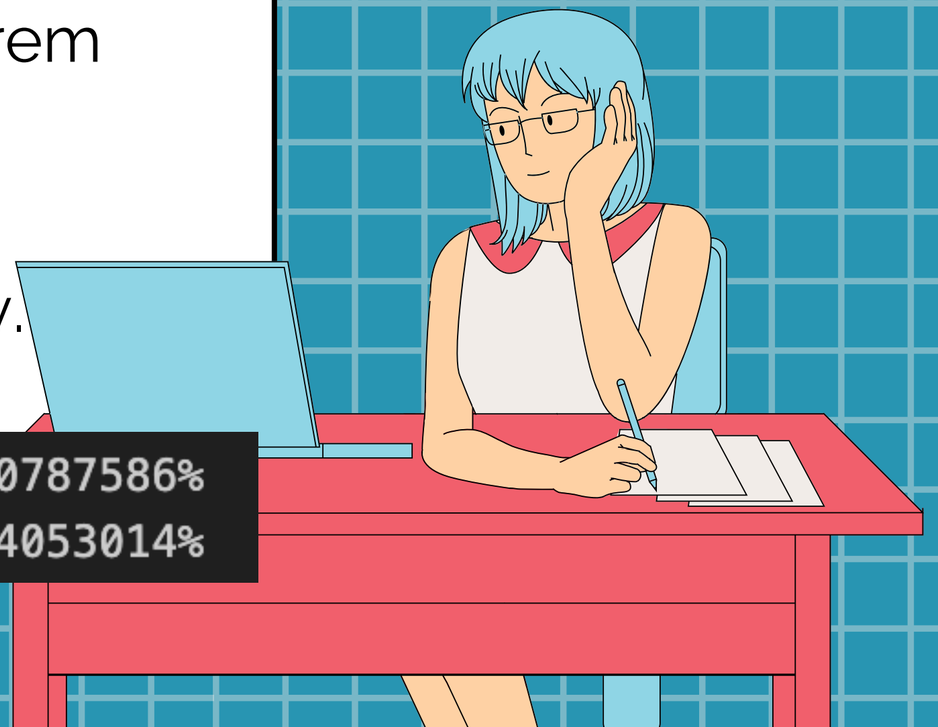
- Boosts algorithm by adding onto weak models through iteration.
- Corrects previous errors
- Focuses on data points with large residuals

```
Accuracy of Gradient Boosting Classifier(Train set): 63.0291299105159 55%
Accuracy of Gradient Boosting Classifier(Test set): 56.626684436546 72%
```

GAUSSIAN NAIVE BAYES METHOD

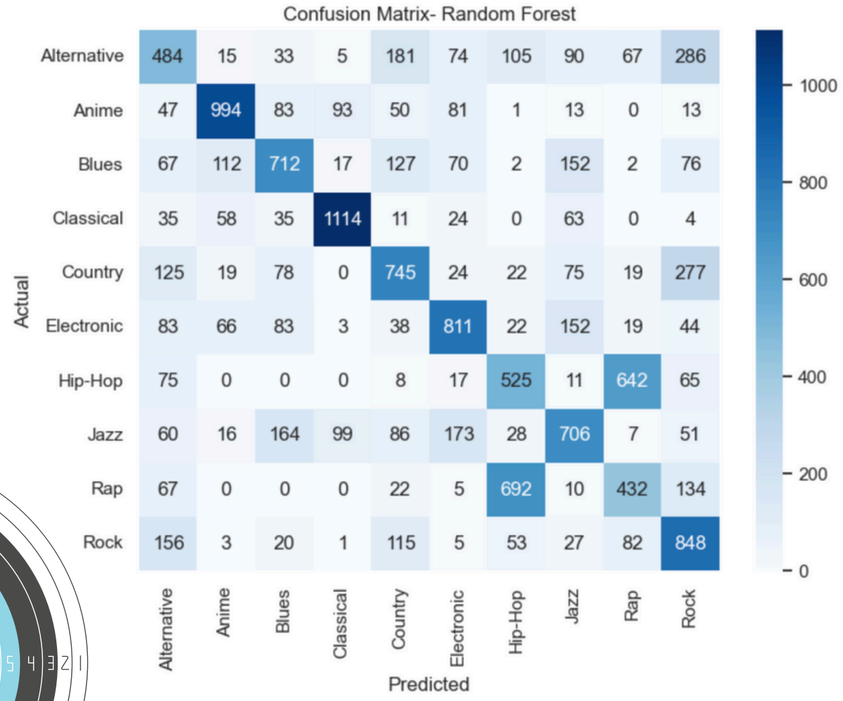
- Probabilistic classifier based on Bayes' theorem
- Follows a Gaussian distribution
- Calculates the probability of each class
- Selects the class with the highest probability.

```
Accuracy of Gaussian Naive Bayes(Train set): 43.101478707875 86%
Accuracy of Gaussian Naive Bayes(Test set): 43.5732267140530 14%
```

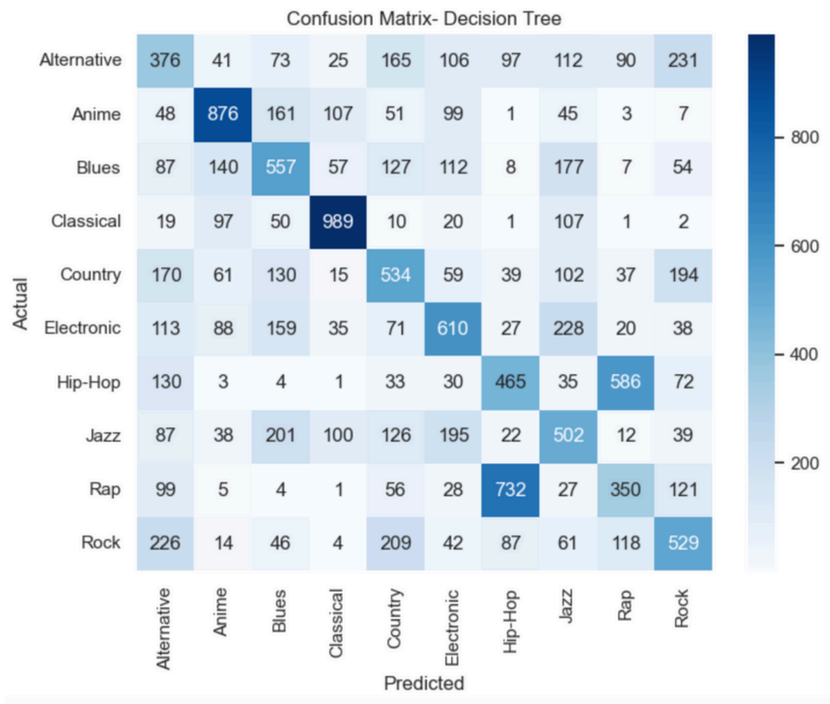


CONFUSION MATRIX

Random Forest

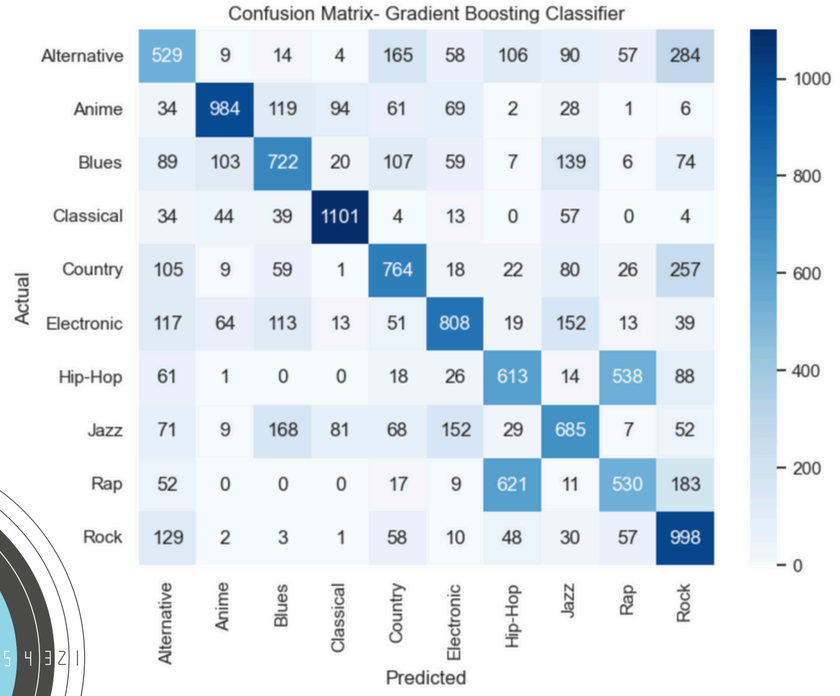


Decision Tree

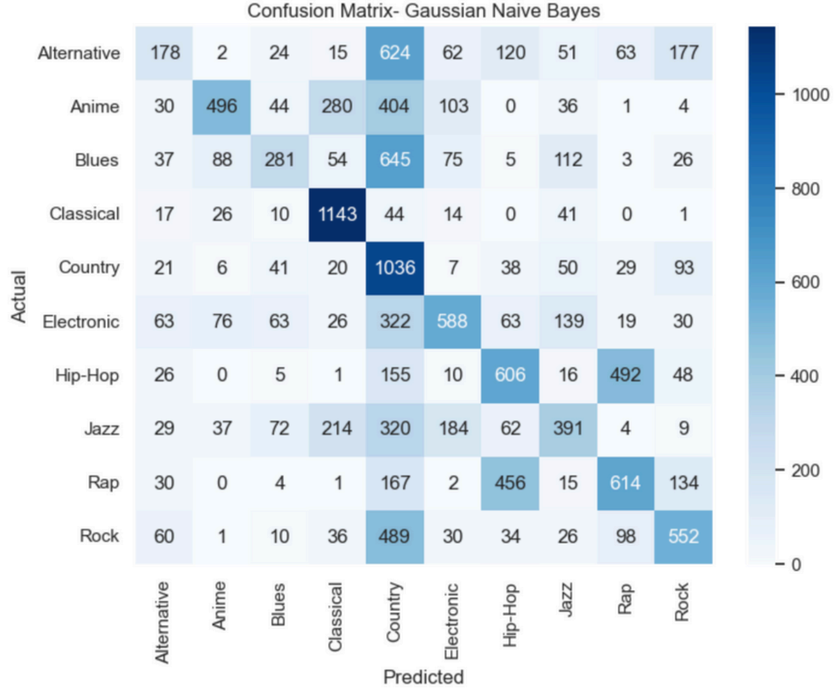


CONFUSION MATRIX

Gradient boosting



Gaussian Naive Bayes



OUTCOMES

Gradient Boosting classifier achieved an accuracy of 56%

Future advancements can come from refining algorithms

Interesting fact - Basic characteristics like tempo, energy etc. can moderately predict music genres

THANK YOU