Peer-graded Assignment

Supervised Machine Learning: Classification

Michał Bożyk

1. Dataset summary

The aim of the project was to compare a performance of different classification techniques including **logistic regression, support vector machine, decision tree, XGBoost and simple neural network** to assess which one is the most suitable one for the task of music genre classification. The original dataset included 18 features e.g. artist_name, track_name, acousticness, duration_ms, energy and contained data on 50000 different songs. It was compiled based on Spotify data.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   instance_id       50000 non-null  float64
 1   artist_name       50000 non-null  object
 2   track_name        50000 non-null  object
 3   popularity        50000 non-null  float64
 4   acousticness      50000 non-null  float64
 5   danceability      50000 non-null  float64
 6   duration_ms       50000 non-null  float64
 7   energy            50000 non-null  float64
 8   instrumentalness  50000 non-null  float64
 9   key               50000 non-null  object
 10  liveness          50000 non-null  float64
 11  loudness          50000 non-null  float64
 12  mode              50000 non-null  object
 13  speechiness       50000 non-null  float64
 14  tempo             50000 non-null  object
 15  obtained_date     50000 non-null  object
 16  valence           50000 non-null  float64
 17  music_genre       50000 non-null  object
dtypes: float64(11), object(7)
memory usage: 6.9+ MB
```

2. Exploratory Data Analysis

The dataset contained both numerical and object data types. It also contained 5 NaN value in each feature, therefore whole rows that contain NaN values were dropped. Next, object data was encoded using LabelEncoder(). It was also spotted that duration feature contains negative values and corresponding rows were dropped. After applying changes the dataset consisted of 45061 columns. After that obtained_date column was removed as it provides no information since there are only 5 unique values. Dataset was later split into features and target subsets. The balance of classes were checked. Target subset was balanced.
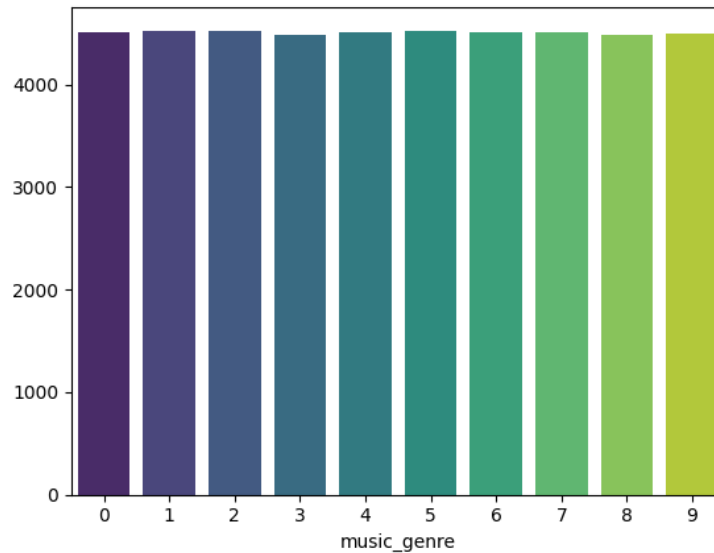

3. Methods

The following methods were applied and compared: logistic regression, supported vector machine, decision tree, XGBoost and simple neural network. All hyperparameters, except for neural in which case they were chosen arbitrarily, were selected based on GridSearch.

3.1 Logistic regression

Performed was multiclass logistic regression. The following hyperparameters were used:

- Penalty: elasticnet
- Solver: saga

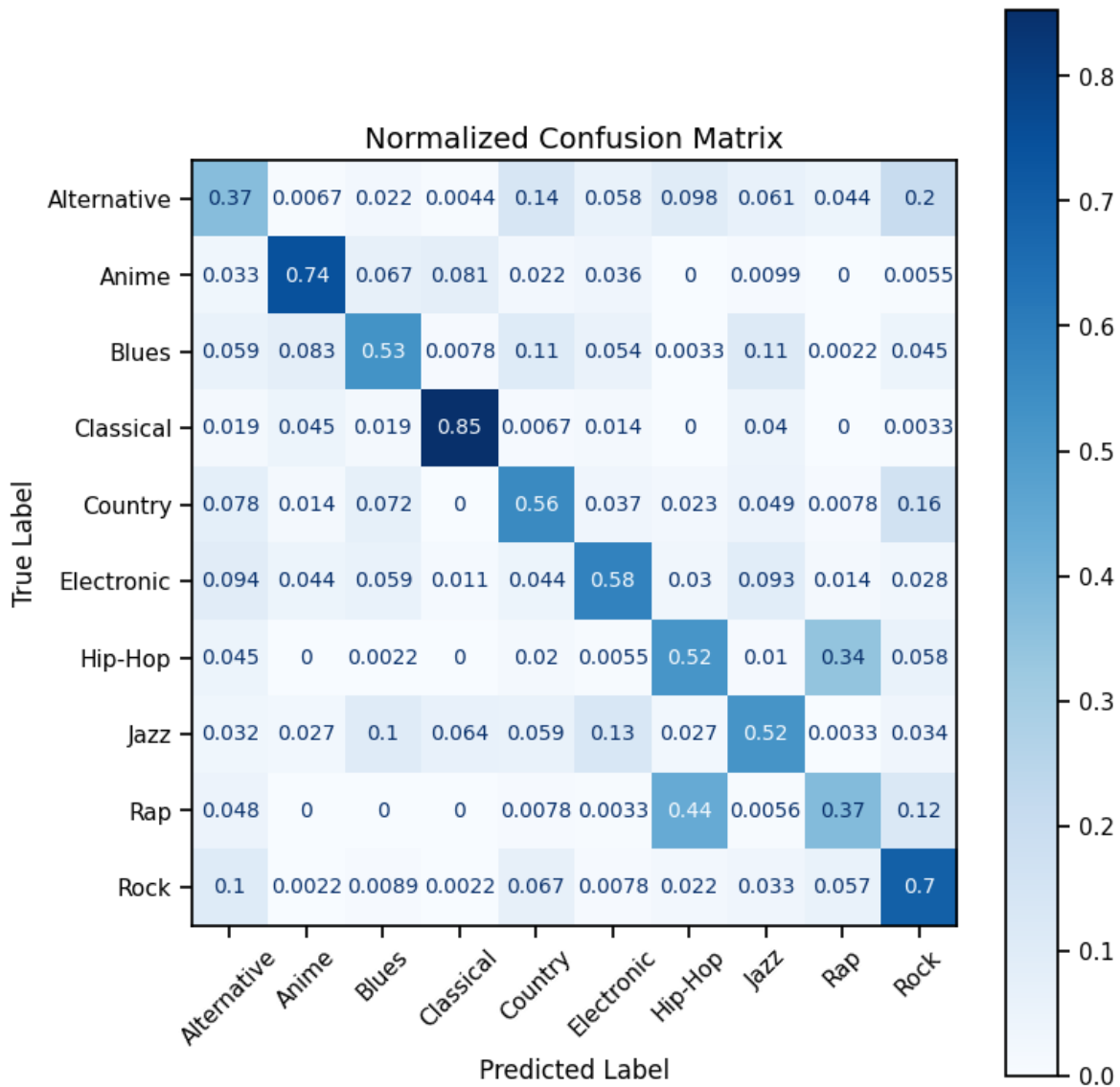The achieved accuracy was c. **0.53**. Obtained was confusion matrix:

Normalized Confusion Matrix
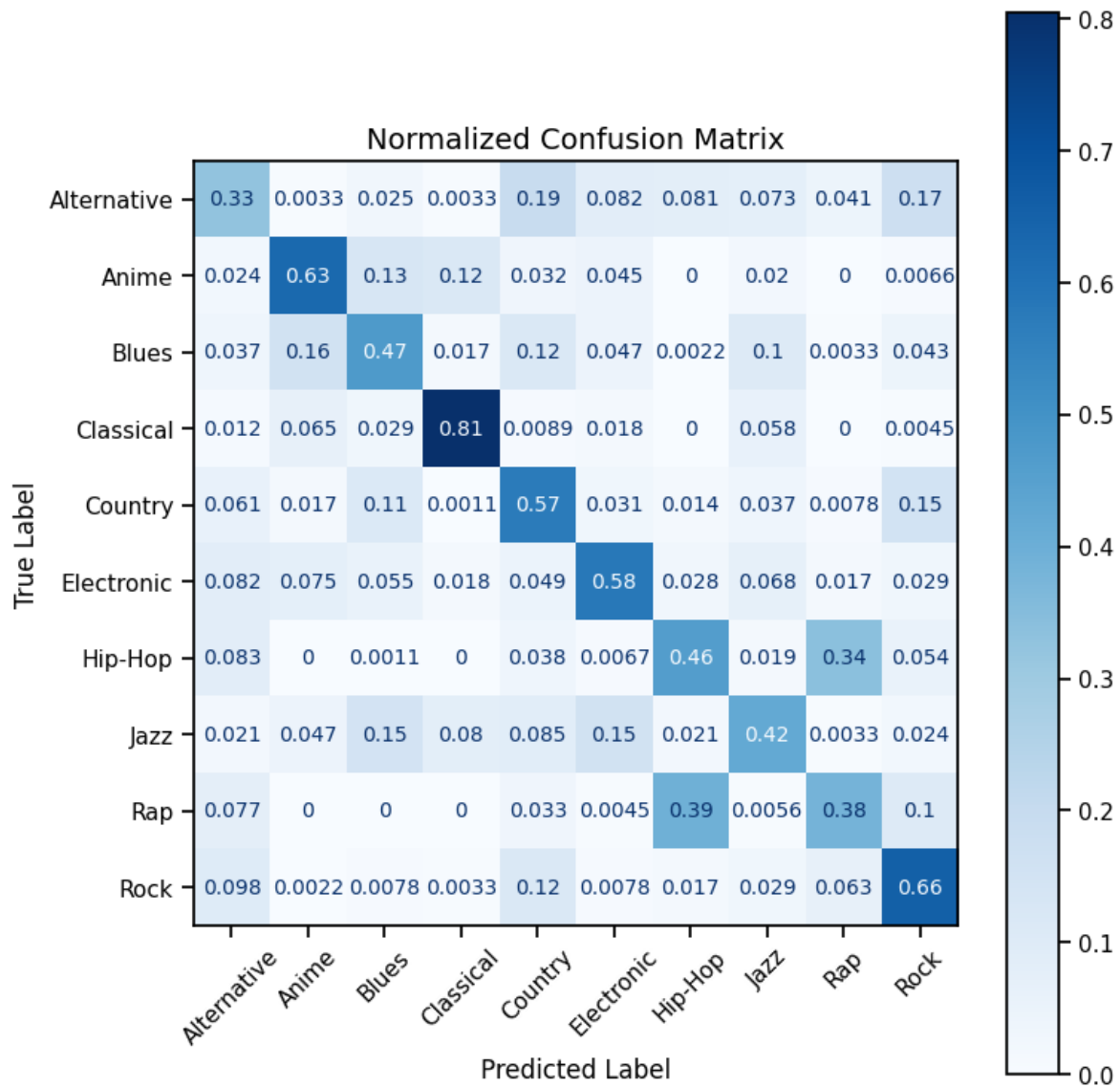
Commonly the algorithm confused Rap with Hip-Hop.

3.2 Support vector machine

Performed was support vector machine. The following hyperparameters were applied:

- C: 1
- Kernel: rbf

The achieved accuracy was c. **0.57**.

Normalized Confusion Matrix

The model shows improvement compared to logistic regression however, it is rather slight.

3.3   Decision tree

Next, decision tree was utilized for classification task. The following hyperparameters were applied:

- Criterion: gini
- Max_depth: 10
- Min_samples_leaf: 5

The accuracy was c. **0.53**. Confusion matrix was generated:

Normalized Confusion Matrix
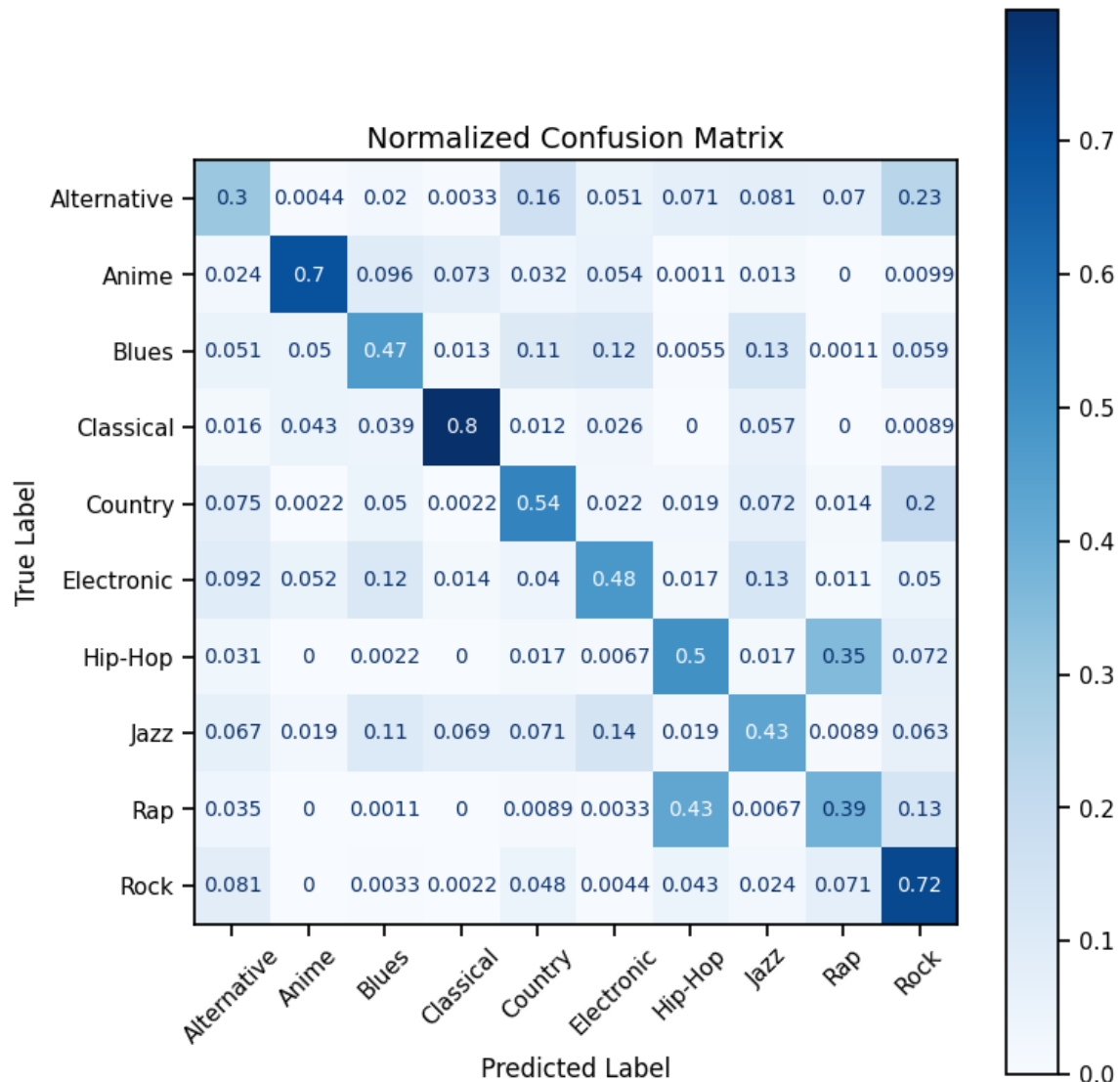
3.4 XGBoost

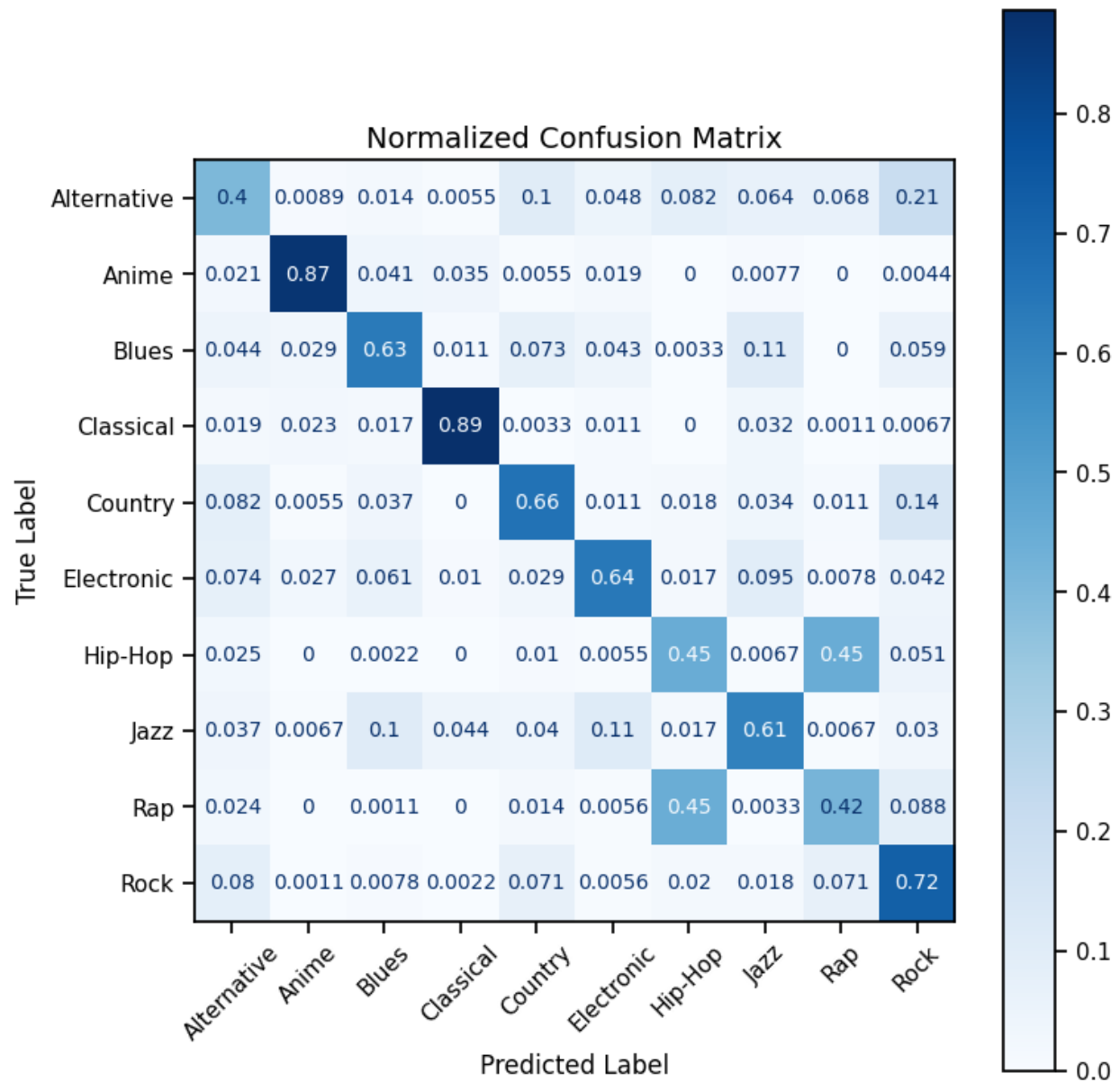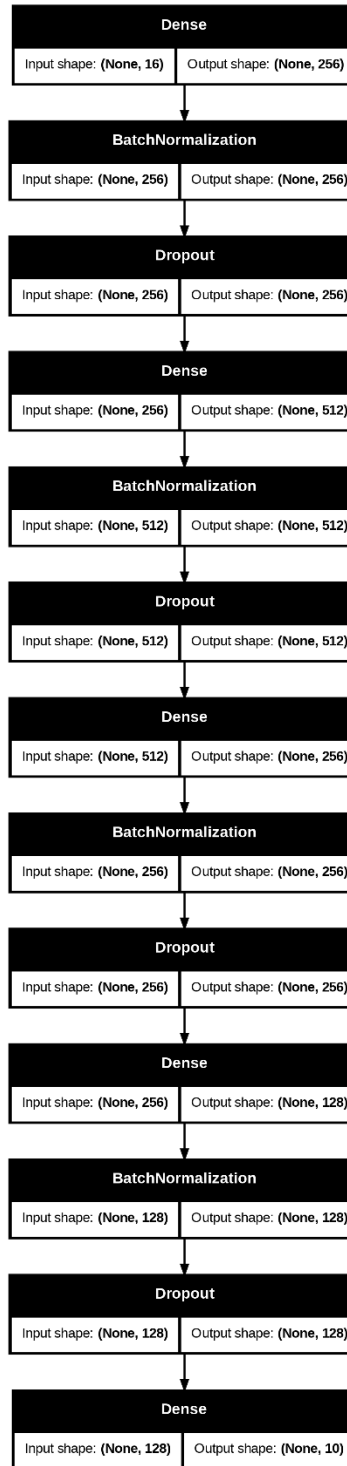Next, XGBoostClassifier was applied. Following hyperparameters were chosen:

- colsample_bytree: 0.8
- learning_rate: 0.1
- max_depth: 6
- n_estimators: 150
- subsample: 0.8

The accuracy was c**. 0.63**. The improvement is visible. Confusion matrix was generated:

Normalized Confusion Matrix

3.5 Neural Network

Lastly, fully connected feed forward neural network was created. The network has a following architecture:

Dense	
Input shape: **(None, 16)**	Output shape: **(None, 256)**

BatchNormalization	
Input shape: **(None, 256)**	Output shape: **(None, 256)**

Dropout	
Input shape: **(None, 256)**	Output shape: **(None, 256)**

Dense	
Input shape: **(None, 256)**	Output shape: **(None, 512)**

BatchNormalization	
Input shape: **(None, 512)**	Output shape: **(None, 512)**

Dropout	
Input shape: **(None, 512)**	Output shape: **(None, 512)**

Dense	
Input shape: **(None, 512)**	Output shape: **(None, 256)**

BatchNormalization	
Input shape: **(None, 256)**	Output shape: **(None, 256)**

Dropout	
Input shape: **(None, 256)**	Output shape: **(None, 256)**

Dense	
Input shape: **(None, 256)**	Output shape: **(None, 128)**

BatchNormalization	
Input shape: **(None, 128)**	Output shape: **(None, 128)**

Dropout	
Input shape: **(None, 128)**	Output shape: **(None, 128)**

Dense	
Input shape: **(None, 128)**	Output shape: **(None, 10)**

```
Model: "sequential_4"
```

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 256)	4,352
batch_norm_1 (BatchNormalization)	(None, 256)	1,024
dropout_13 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 512)	131,584
batch_norm_2 (BatchNormalization)	(None, 512)	2,048
dropout_14 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 256)	131,328
batch_norm_3 (BatchNormalization)	(None, 256)	1,024
dropout_15 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 128)	32,896
batch_norm_4 (BatchNormalization)	(None, 128)	512
dropout_16 (Dropout)	(None, 128)	0
ouput_layer (Dense)	(None, 10)	1,290

```
Total params: 306,058 (1.17 MB)
Trainable params: 303,754 (1.16 MB)
Non-trainable params: 2,304 (9.00 KB)
```

The neural network has a total of 306,058 parameters. In addition to neural network, early stopping mechanism was applied. Adam optimizer was used and sparse categorical cross entropy was chosen to be loss function.
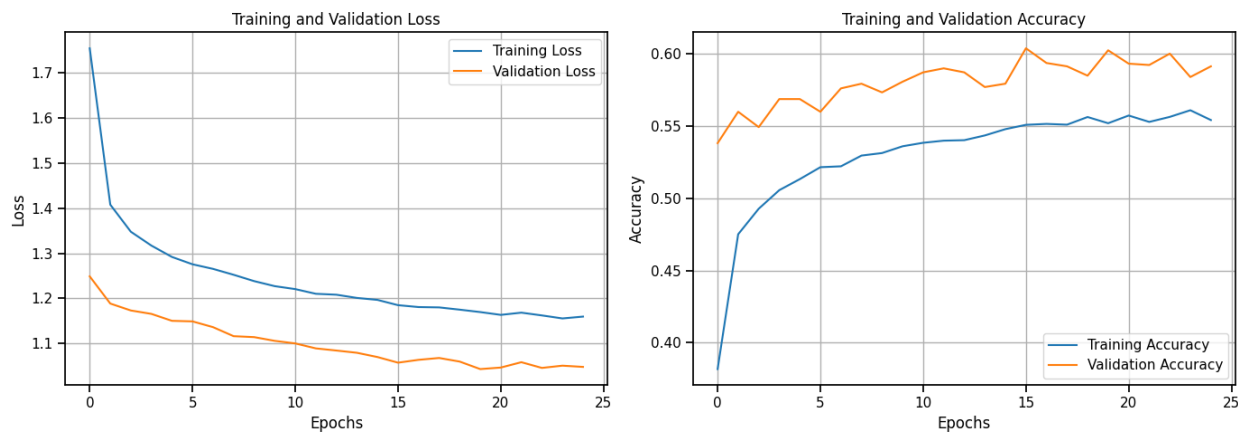
Both validation and training losses are steadily decreasing what indicates that learning process is successful. On the other hand, both accuracy values are increasing which is also a good sign.

However, the learning process stopped on 25 epoch with early stopping set on validation loss and epoch patience equal to 5. The accuracy on test set is c. 0.6.

Confusion matrix was created:
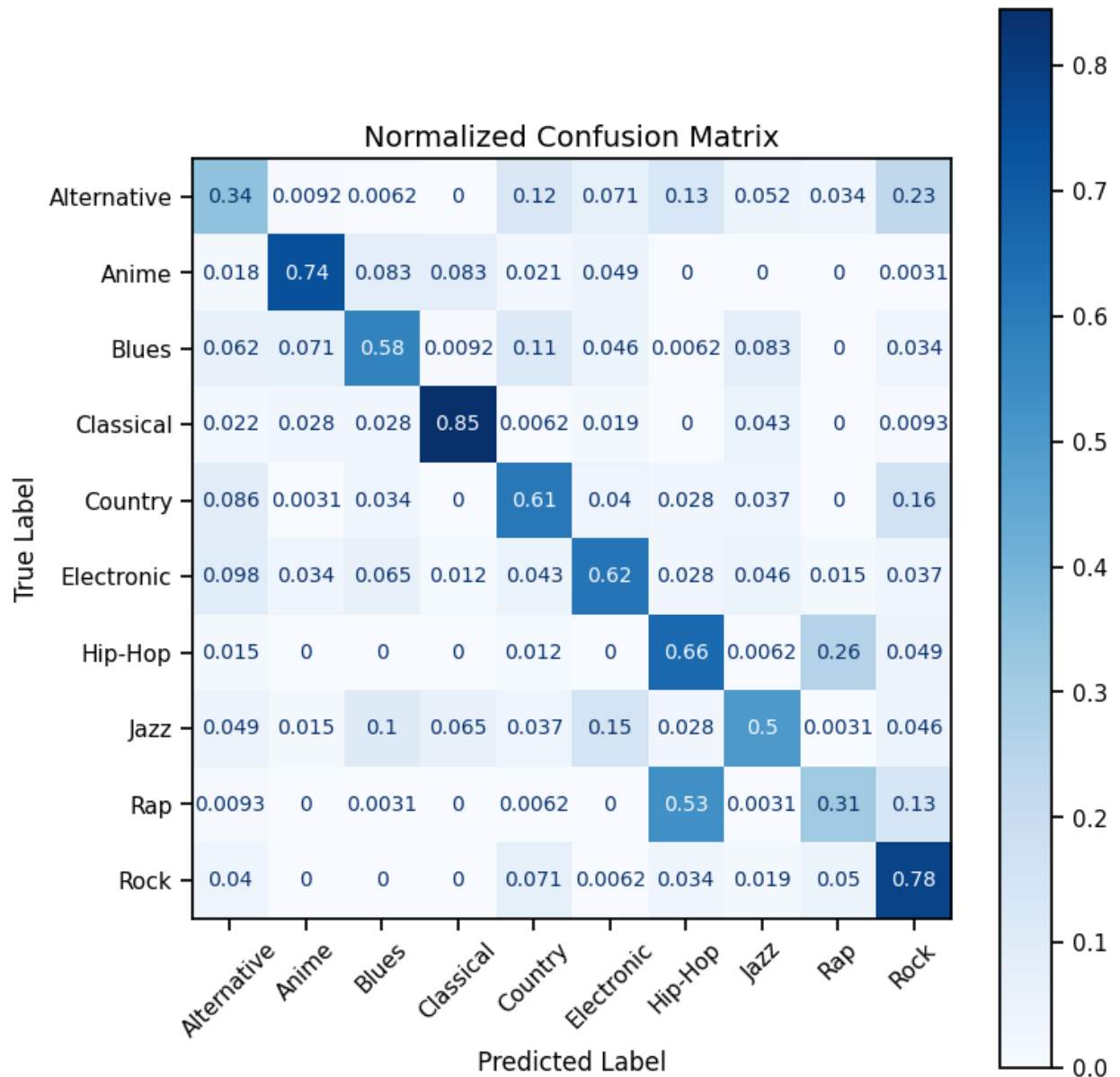
4. Conclusions

The performance of all 5 approaches seems to be comparable. More complex methods such as SGBoost or neural network have slightly higher accuracy than SVM, logistic regression or decision tree.

Model Accuracy Comparison
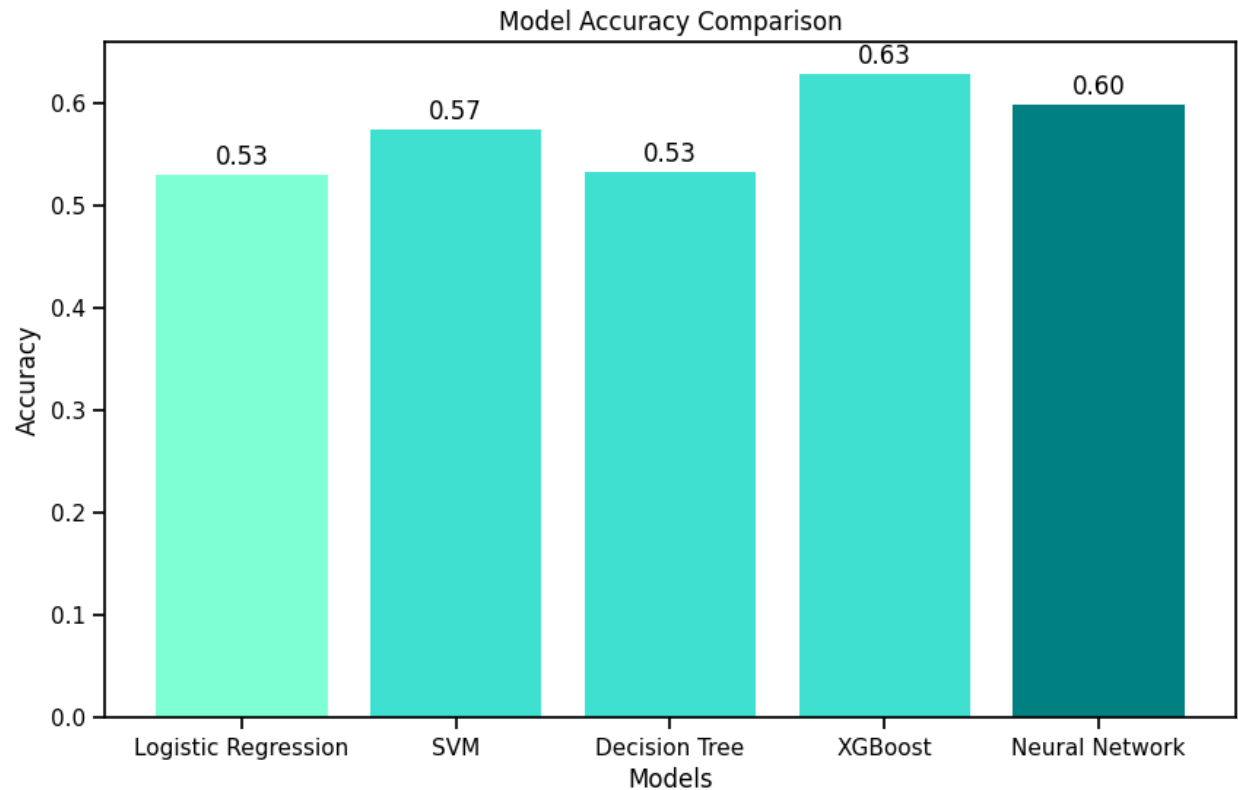
As per analysis of confusion matrices hip-hop and rap are confused the most. However, most unique music genres are rock, classical and anime. As it comes to further improving the predictions and the efficiency of training irrelevant features can be dropped leaving only those that are correlated with target value. The dataset is large and complex therefore further models will be most likely more complex neural architectures.