**Project Title:** Music Genre Classification

**Team Members:** Eduardo Cruz, Ethan Ngo, Justin Vuong, Kevin Tang

**Description:**

The project takes various column details such as artist name, energy, loudness, etc. and places them into different “classes”. The “classes” consists of many genres such as Rock, Alt, Pop, and more. By analyzing the data, we will categorize them as accurately as possible. (The actual categorization will actually only be ‘Rock’ or ‘Not Rock’ for this project. This is explained below)

**Goals:**

The goal for this project is to make accurate music genre classifications given a training set consisting of music features and its corresponding genre labels. The initial goal will be to classify the songs into a binary classifier such as “rock and not rock” in order to simplify the process for the purposes of this course. If time allows, we will also try to add more categories to classify the songs into.

**Details about the Dataset:**

As is already outlined in the description above, there are various features that will be used to classify music genres. There are 16 in total. The data varies somewhat in its format with some being things like name of the artist or song while others are values from 0.0 to 1.0 such as with the speechiness or acousticness. Of note is that there are also some features such as danceability, energy, valence, etc. that seem to be based off of some other form of analysis. The project description on Kaggle does not reveal how this data was determined though. It does not seem like the ambiguity on this matter would affect the integrity of the project in any way, however.

Moreover, the dataset has feature columns containing NaN (Not a Number) values. These values created issues during the process of splitting the dataset into training and testing sets, so we replaced the NaN values with the mean of all values for that specific column. The feature columns that consisted of NaN values are ‘Popularity’, ‘key’, and ‘instrumentalness’.

An important note is that after discussing the project proposal with the TA, he suggested that in order to keep the project more feasible, we should have a binary classification for the genres. This means that instead of having multiple genre classes like Rock, Pop, Metal, etc. it would instead just be categorized into ‘Rock’ or ‘Not Rock’. In the code, this means that songs that fall under the ‘Rock’ category will have a class value of 1 and all others will have a class value of 0.

**Developed Algorithms:**

* **Random Forest Classifier**
* **Decision Tree Classifier**
* **KNN Classifier**

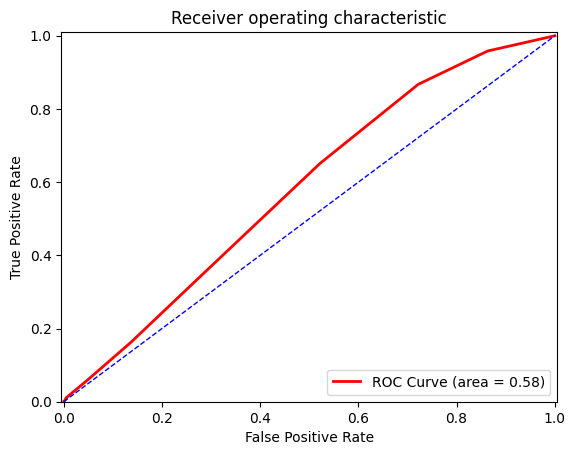
**Performance of Developed Algorithms:**

We developed accuracy scores, the AUC, and ROC Graphs of the three developed algorithms to compare them to each other.

* **KNN Classifier (Containing NaN Values)**

Accuracy score: 69.4%

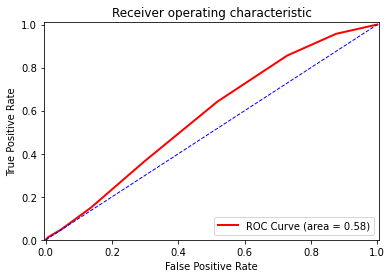
AUC: 0.58

ROC Graph: 

* **KNN Classifier (Not including NaN Values)**

Accuracy score: 69.2%

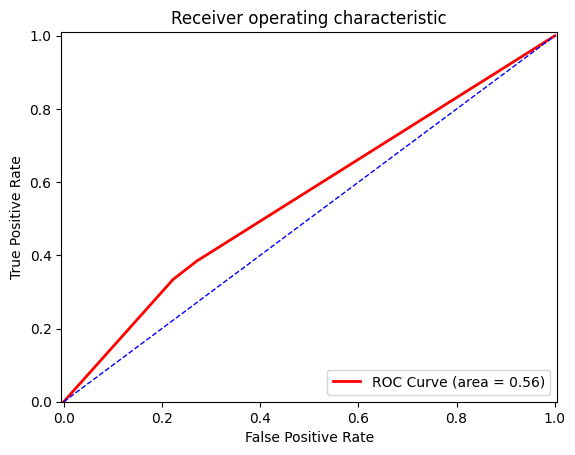
AUC: 0.58

ROC Graph: 

* **Decision Tree Classifier (Containing NaN Values)**

Accuracy score: 64.9%

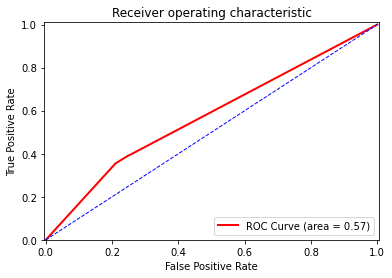
AUC: 0.56

ROC Graph: 

* **Decision Tree Classifier (Not Including NaN Values)**

Accuracy score: 66.4%

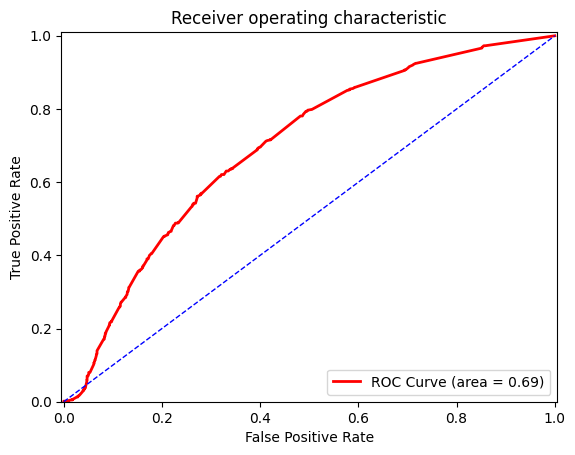
AUC: 0.57

ROC Graph: 

* **Random Forest Classifier (Containing NaN Values)**

Accuracy score: 70.6%

AUC: 0.69

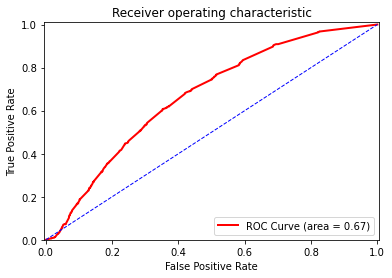
ROC Graph: 

* **Random Forest Classifier (Not Including NaN Values)**

Accuracy score: 68.6%

AUC: 0.67

ROC Graph:



**Team Member Responsibilities:**

Kevin Tang, Justin Vuong:

Responsibilities include evaluating the data, selecting the best features for the creation of our prediction model, and working closely with the other team members in improving the prediction model when performing test runs via testing data.

Eduardo Cruz, Ethan Ngo:

Responsibilities include creating an early build of the prediction model with feedback from team members that evaluated the appropriate data, and working with them to continue improving the prediction model’s accuracy.