Carlos Gonzalez

**Title:**

Using Clustering Algorithms to differentiate genres and predict how popular a song will be using the features from the Spotify Dataset from 1921-2021

**Introduction:**

Throughout time many people have aspired to become popular musicians to no avail, this project would be able to help classify what makes a popular song. On the other hand, differentiating the genres of songs using their features allows for the development of the Spotify app because it allows the app to choose what songs belong to what genre with minimal interference from a person. This is all-important because it allows for the advancement of music apps in general such as Apple Music, Pandora, and Spotify. Each song in the data set has different features such as acousticness, artists, and what key the song is played in. This dataset there includes various files such as data.csv which gives the data described and will be used the most. There also exists 4 other files as data\_by\_artist.csv, data\_by\_genres.csv, data\_by\_year.csv, and data\_w\_genres.csv. Some of these files require a lot of preprocessing data because of incorrect dates or mismatching information which can be fixed with some research about certain songs. Throughout the Kaggle tasks, 5 tasks are open for submission. None of them have confirmed solutions that would allow me to create a solution that will hopefully be good enough to be confirmed as a solution to the problems listed above. The overall problem that makes me want to do this project is the fact that there exist people out there who don’t know how to get popular This will not only cover the popular genre that people like it but it will look at the features of this huge dataset of music. It will incorporate all the years to find a song that would not only be popular during the present time but also from the past to create a song that has a higher chance of becoming popular. Of course, this isn’t meant to make an automatic popular song but it helps people create something popular. For the other part, I want to help find a way that would allow songs to be differentiated based on genre because this is something that can be automated to be extremely accurate and not need someone to do it. Of course, it won’t be completely accurate but it will allow for a lot of time-saving.

**Proposed Work:**

This dataset does require a bit of preprocessing. After preprocessing the data, I will do exploratory data analysis on the data to see if there are any imbalances in the data or to see any trends before applying a machine learning algorithm. For the first part, I will be using various types of clustering algorithms such as k-means clustering and other clustering algorithms that were not covered in the class. I will be using various machine algorithms in this case because I want to figure out which algorithm is best suited for this type of problem and to learn more about the various types of clustering problems that exist. For the second portion, I plan on using various types of classifier machine learning algorithms to see what algorithms work best. For this portion, I do plan on taking a more business approach and plan on finding which algorithm is more accurate and efficient. The types of classifier machine learning programs that I will use in this are logistical regression and different decision trees. This is not only to get exposure to different algorithms but also prepare to be able to explain these types of machine learning algorithms.

**Timeline:**

**Timeline

Description automatically generated**

**References:**

Kaggle Dataset:

<https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks/tasks>  
Algorithms:

<https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks/tasks?taskId=2173>

<https://scikit-learn.org/stable/modules/tree.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>