Jagjeet Singh, V00970743
Jeet Ajmani, V00942302
Jackson Walde, V00899888
Ella Kuypers, V00910928
Kenil Shah, V00903842

The Problem

Music, specifically music streaming, is a multi-billion dollar industry, and as such, machine learning is present in various capacities, whether for recommendations, classification, or even production. The basis of our project is to use data from existing songs to predict metrics such as genre, chart placement, and popularity, with the hopes to dive deeper and uncover the relation between them.

The importance of the problem lies in the fact that the metrics we plan on predicting are often volatile or difficult to categorize and will require some care as to how we approach the task. Predicting labels of this nature is a multi-class classification problem, and therefore the accuracy of the algorithms utilized in tackling this problem is a significant focus.

The results we will produce with this project will be interesting as the popularity of a song is not an easily predicted thing since several factors affect a song's reception by the public. Can we apply some of the data mining methods learned in class to determine what factors play the most significant role in a song's popularity? Or what genre the songs will fit into? Or how danceable an individual song is? It will also be interesting to look at the connections between the many data points we study.

Due to the extensive and well-documented history of music streaming, there is a plethora of data available that can provide a solid baseline for our project. The dataset we will use for the project comes from Kaggle, a community-based machine learning website. The dataset contains metrics for every song from Spotify's daily top 200 charts worldwide from December 2016 to December 2020. There are tens of thousands of songs included in the dataset and along with them are metrics relating to their popularity, genre, danceability, tempo, loudness, duration, and other information about the song.

Suppose the dataset doesn't meet our expectations or proves to be too difficult to work with. In that case, we have a backup plan in the form of "Million Song Dataset," a freely available dataset that contains audio information and metadata for one million songs.

Goals

The end goal of this project is to compare and contrast the effectiveness of machine learning methods in predicting/analyzing various features of the dataset. During the project timeline each team member will have an individual goal of determining which learning models and feature selection have the most successes in predicting the label of their choosing. Towards the end of the project timeline, and at points before, team members will compare and discuss their results to reflect on how effective their modeling and feature selection works to predict their assigned label. In general, a model and feature selection combination will be rated using training/test accuracy. If a team member decides to use more specific forms of evaluation for their models, they can. Each team member should explore at least three different regression or classification models for their label. Additionally, each team member should explore feature selection and pre-processing techniques. Again, the baseline measure of success for feature and model selection will be training/test accuracy.

Plan
To achieve our goals, we will be using Python as our programming language to manage datasets and run machine learning algorithms. The Python scikit-learn package contains tools for constructing machine learning algorithms and provides robust customization for hyper-parameters.

We will be performing experimentation on the dataset by comparing and contrasting decision trees, random forests, and neural networks. Using Python, we will be able to generate visually pleasing plots that will make comparisons easy and simple.

As the datasets we plan to use are significantly large, we will use a subset of the data to verify the success of each machine learning technique. Some of the parameters have definite and clear values that will provide well defined and clear success rates. Meanwhile, other metrics, such as the propensity for a song to reach the top charts, can be arbitrary. We plan to verify the success of algorithms measuring abstract success by using test cases that are known to be successful, top-charting songs, to ensure the validity of the algorithm.

To ensure that we develop a successful algorithm for reaching our set of goals, we will be adjusting the hyper-parameters for each machine learning algorithm carefully. Once again, the Python plots will be necessary in determining the ideal values for the hyper-parameters. Comparisons can be made easily and adjustments can even be done dynamically.

Furthermore, there are numerous resources available online for music classification techniques that will aid our group in achieving our goals. If we require more aid, our group will consider consulting with University of Victoria professor George Tzanetakis, who is an expert in machine learning algorithms directed towards music-based datasets.

Our project timeline will be as follows:

Task	**Deadline**
Formal proposal	February 14th
Set up framework for machine learning algorithms	February 21st
Preliminary experimentation with dataset	February 28th
Progress report	March 14th
Optimization of machine learning algorithms	March 21st
Further testing and success verification	March 30th
Presentation	April 1st, 5th, and 7th
Final Report	TBD

Task Breakdown
While the tasks of individual team members may grow and evolve as we explore with this project, this table shows what each member will work on. We will each take a different label within the dataset and form predictions individually.

Name	**Label**
Jeet	Popularity
Kenil	Genre
Jag	Country
Ella	Danceability
Jackson	Valence