MSCI 446: Project Proposal
Yasmin Abu Helal - 20953398
Aashrita Pillutla - 20937920
Parto Aflatounian - 20957200
Adeena Syed - 20938935

The aim of this project is to develop a music recommendation system using Spotify's API, focusing on specific features like danceability, energy, and loudness. Advanced machine learning techniques will be used to match these song features with user preferences, aiming to provide more personalized music recommendations. The main benefit of this system is the enhancement of the user experience on music streaming platforms. It will offer recommendations that are closely aligned with individual user tastes, making it easier for users to find music they enjoy. This approach not only caters to specific preferences but also encourages users to explore a wider variety of music. Accurate and tailored suggestions are key to improving user satisfaction with the streaming service, as users are more likely to find music that suits their taste. This project aims to improve how users interact with music streaming services, making the experience more personalized and enjoyable.

Creating a content-based music recommendation model using Spotify data involves several steps. First, we will need to use Spotify's Web API to fetch detailed information about tracks on Spotify and the kind of music the user listens to. This information includes audio features of tracks (danceability, energy, etc.), metadata of tracks (artist, album, genre, etc.), user playlists and listening history. After obtaining this data, we would need to clean it by filling in missing values/removing rows with missing values and removing duplicate rows. We also need to ensure quantitative data is within the same scale by normalizing it and categorical data is encoded using LabelEncoder or one-hot encoding. Because we are creating a model that recommends music based on unlabelled data (user activity) we will be creating an unsupervised machine learning model. Clustering methods can be used to group similar songs together and make recommendations.

To compare methods and validate results in our unsupervised learning music recommendation project, we will implement several strategies. We'll split our data into training and test sets, using internal measures to assess cluster consistency. Our adaptation of k-fold cross-validation will focus on the consistency of data clustering across different folds. User feedback will serve as an invaluable indirect measure of our model's effectiveness, helping us gauge the success of our recommendations. We'll ensure our model offers diverse and novel music selections, crucial for an engaging user experience. Finally, sensitivity analysis will be conducted to evaluate how input data changes affect our model, ensuring its robustness and adaptability.

With our choice to complete an empirical evaluation, in which we will collect data and create an ML algorithm to recommend music based on user listening patterns, we have identified a few key risks we may encounter in completing this project. Potential risks identified include:

- **Model Scalability**: We may face challenges in ensuring the model is scalable and can handle Spotify datasets of varying sizes to provide accurate recommendations. The system must be able to analyze data for several users, and it also needs to be

efficient enough to process the Spotify data and deliver recommendations quickly enough to satisfy the end user.

- **Breadth of Data**: As the model's intended purpose is to provide personalized recommendations based on the individual user's Spotify data and listening history, there may be a risk of having a limited data source to test the model with, and we may require external Spotify listening data. Similarly, newer users may not have enough data to make recommendations with a high level of accuracy, as many users may only interact with a small subset of tracks, leading to sparse data.
- **Model Bias**: There could also be a risk of introducing bias into the system, where the model may favour and recommend certain types of music over others instead of being impartial. This could happen if certain types of data occurred more frequently, or due to inherent biases in the algorithm which could lead to increased promotion of artists who are already very popular on the platform, skewing recommendations away from lesser-known artists.

Task breakdown via Gantt Chart: