Rishika Anil Gupta
Satria Bagus Wicaksono

Machine Learning
Project Proposal

Title: Genre Classification Based on Spotify Dataset

Problem Statement:
Given a dataset containing song attributes (loudness, tempo, duration, etc), predict the genre of the song.

Motivation for the problem:
Spotify is a popular music streaming service that allows users to access a vast library of songs from various artists and genres. With over 70 million tracks and counting, Spotify's catalog is one of the largest in the world. It offers a wide range of genres, including pop, hip-hop, electronic, classical, jazz, country, and many more.

The motivation for this project is due to the benefits that it can provide to multiple significant parties as described below:-

- Users: To *better music recommendations* by identifying the different genres that a song belongs to, streaming services can recommend songs that are more likely to appeal to the user's musical tastes or *more efficient music organization* by categorizing songs into multiple genres, it becomes easier to find specific types of music when searching through a music library or playlist.
- Researchers: To *better understand the evolution and influence* of different music genres.
- Producers: To *improve the production of music* by analyzing the characteristics of successful songs across different genres, producers can identify the elements that make a song appealing to a wide audience and incorporate those elements into their own work.
- Spotify: To *automate genre tagging* for the new songs added without human intervention. It also helps to *create automated playlists* easily for the users, as Spotify adds 60.000 new tracks every day.

Data Description:
The dataset is taken from https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify, we will use **genre_v2.csv**

Rows	**Cols**	**Target Variable (1)**	**Categorical Variable (3)**	**Numerical Variable (10)**	**Object Variables (5)**
42,305	20	Genre	- Key - Mode - Time_signature	Attributes - danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms	Attributes - type, id, uri, track_href, analysis_url

In the preliminary analysis there are several columns that might not be useful as a variable such as: id, url, track_href, analysis_url. These columns are unique identifiers of the song so the value will be different for each song. Additionally, there is one column without a name and description so we will need to do deep analysis into the data for determining its usage.