Dataset: Spotify Dataset | https://www.kaggle.com/datasets/vatsalmavani/spotify-dataset/data

1. What is the project about?

The core objective of this project is to develop a Hit Song Prediction Model using a large-scale Spotify dataset of approximately 244,930 tracks. We are framing this as a supervised binary classification task.

Specifically, the model will predict whether a song will become a "Hit"—defined as achieving a popularity score of 80 or higher. This project is non-trivial because it addresses:

- **Class Imbalance:** Only a small percentage of songs reach the "Hit" threshold, requiring specialized evaluation metrics beyond simple accuracy.
- **Dataset Shift:** Music trends evolve; therefore, we will implement a Temporal Split (training on older tracks and testing on more recent ones) to evaluate how well the model generalizes to shifting listener preferences.

2. What data features might be used?

We will focus on the perceptual audio features provided by Spotify to ensure the model learns musical patterns rather than just memorizing famous artists.

- **Target Variable (y):** A binary indicator where 1 (Hit) represents popularity >=80, and 0 (Non-Hit) represents popularity < 80.
- **Input Features (X):** Non-identifying numerical features including danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.
- **Excluded Features:** We will drop unique identifiers such as track_name, id, and artist_name to prevent data leakage and ensure the model's objectivity.

3. What would be your first step?

The first phase involves Exploratory Data Analysis (EDA) and Data Integrity Verification. Our priority is to investigate the distribution of the target variable to confirm the severity of the class imbalance. Furthermore, we will analyze the relationship between audio features and the release year. This will help us determine the most effective "cutoff year" for our temporal train-test split, ensuring that the training set provides a historically diverse foundation for predicting modern hits.