# Finding the Genre - Multi-Class Classification

Date: 12.11.24  
Made by: Ortal Lasry & Or Cohen Raviv

## Business Question:

How can we accurately classify music genres using audio features to improve recommendations and content organization on streaming platforms?

Background: We utilized a Spotify dataset from Kaggle, which contains 114 track genres and 14 main numeric features (e.g., duration\_ms, danceability). Additionally, it includes a boolean feature indicating explicit content. The dataset originally had 20 columns and 114,000 rows, but after removing duplicates (keeping only the first occurrence), we retained 87,867 rows. The dataset also presents significant class imbalance in the target variable, track\_genre.

## Analysis Summary:

Upon preliminary exploration, we observed that certain genres had more track IDs than others, contributing to data imbalance and complicating pattern interpretation due to overlapping characteristics within broader genres. For example, subgenres like 'rock', 'rock-n-roll', and 'alt-rock' fall under the broader 'rock' category. To address this, we grouped them into 10 classes. In addition, we used class imbalance when we fitted the ml models.

## Feature Engineering and Descriptive Analysis:

Descriptive statistics revealed right-skewed distributions in variables such as duration\_ms, speechiness, acousticness, and instrumentalness, while loudness was left-skewed. These patterns guided our feature engineering approach. We applied log transformations to right-skewed variables to capture feature characteristics better and boost model accuracy. We also created polynomial features for tempo and danceability due to their non-linear distributions. We investigated the interaction effects between 'tempo' and 'danceability' (since songs with a higher tempo are more likely to be more danceable) and between 'duration\_ms' and 'energy'.

The correlation matrix between all quantitative features revealed a high correlation between energy and loudness (r=0.75) and between energy and acousticness (r=-0.72).

Bar Chart for the Categorical Feature: 'Explicit' (Presence of Explicit Language):

As expected, some genres have higher frequencies of explicit content than others. For example, Hip Hop and Death Metal show the highest frequencies of explicit content, while Classical music has none.

## Modeling Approach and Evaluation:

In the first step, we split the dataset into training and testing sets (0.3%/0.7%) and used 5 features that exhibited high variance across music genres. We predicted only 5 music genres based on ML models, including Logistic Regression, Voting Classifier, and Random Forest.  We assessed feature importance and found that ‘explicit,’ ‘danceability,’ and ‘speechiness’ made the greatest contributions to reducing Gini impurity. The Random Forest model achieved the highest accuracy (0.8) with a Test ROC AUC score of 0.95.

Following this, we expanded our approach to predict all genres (114 music genres grouped into 10 classes) and included eight features in total. We tested the accuracy score for each model with the original features vs. engineered features. The results showed minor accuracy improvements, leading us to proceed with XGBoost using only the engineered features, which achieved a highest accuracy of 0.5 and a Test ROC AUC of 0.84.

## Project Insights and Future Research:

The main results of our project show:

* Differences in accuracy scores when using small vs. large datasets for music genre classification.
* A stronger contribution of certain features to reducing Gini impurity in music genre classification.
* Some machine learning models outperform others in terms of accuracy (e.g., Random Forest vs. Logistic Regression).
* The importance of using boosting methods to reduce classification bias.
* The need to address data imbalance for better model performance.
* Modest improvements in model accuracy through feature engineering.

**Future Research:**

* The strong correlations found between energy, loudness, and acousticness indicate the potential value of performing dimensionality reduction in future studies.
* Exploring additional text-based features or contextual audio data could further enhance model performance.