Spotify Genre Classification

1. Team Members: Míriam Méndez Serrano & Eliya Tiram

2. Dataset choice

Since we both have a passion for music. We were interested in exploring the relationship between music and data, and how machine learning algorithms could help identify different genres of music.

Therefore we decided to choose the Spotify music dataset which contains a lot of tracks from a wide range of genres, including pop, rock, hip-hop, and classical music. Excited by the prospect of working with such a large and diverse dataset, we decided to use it for the machine learning project.

Moreover, this dataset fulfils the requirements and we have found that it is based on real-world data from the Spotify platform, which means that it reflects practical relevance.

The dataset contains detailed information about each track, including its title, artist, duration, tempo, and key, as well as its popularity, acousticness, danceability, energy, and other characteristics.

3. References of previous work

As the data is in kaggle we can find several people using this dataset and publishing their work. In fact, there are 40 notebooks published there. Most of them are only exploring the data. The most voted uses this dataset for a recommender system.

4. Information on the data.

Raw dataset contains 22 columns and 42,306 rows. However there is a column called Unnamed, which we remove since contains 21525 null values and it behaves as the index of the title feature.

Features	Type + Description
danceability	Double Values [0,1] - how suitable a track is for dancing
energy	Double Values [0,1] - intensity and activity of the track
key	Integer - key of the track 0=C, 1=C#, 2=D etc.
loudness	Double - how load in DB (decibels)
mode	Integer - 0 or 1 indicates major or minor scale, 0 for minor and 1 for major
speechiness	Double [0,1] - spoken words in a track
acousticness	Double [0,1] - A confidence measure whether the track is acoustic, 1.0 represents high confidence
instrumentalness	Double [0,1] - no vocals measure, the closer to 1 the less vocals on the track
liveness	Double [0,1] - probability that the track was performed live
valence	Double [0,1] - high valence sound is more positive
tempo	Double - BPM (beats per minute)
duration_ms	Integer
time_signature	Integer - Beats on each bar
genre	String - Categorical variable contain 15 different values of music genres
song_name	String
title	String - whenever song_name is empty title has a value