SAPIENZA
UNIVERSITÀ DI ROMA

COMPUTER SCIENCE DEPARTMENT

Project Proposal

BIG DATA COMPUTING

Professors:
Gabriele Tolomei

Students:
Niccolò Erasmi
1851448

Academic Year 2022/2023

1 Project Proposal

1.1 Problem

The problem I want to address is to classify the songs according to their popularity. Based on Spotify's Web API, the popularity of a track is ranked with a number between 0 and 100, so I will discretize this problem by using 10 classes.
I want to analyze which factors of the song influence this aspect the most, such as the popularity of the artist, the year of publication, the catchiness, the positivity and the feeling of the song (sad, happy) and the aspect of the song such as the "loudness", "energy", "bpm", etc.
I also want to work on this project because, having studied music and harmony for several years, I would like to observe how different factors such as key (major or minor), the chords used and other factors influence the popularity of a song.

1.2 Dataset

For the project I chose the dataset from Kaggle Spotify 1.2M+ Songs, which contains information about the songs of the Spotify music platform. This dataset has 1.2M records and 24 columns (features).
I also want to create my own dataset from scratch using the Spotify Web API that provides queries to get:

- the list of all genres classified by Spotify;
- a list of up to 100 recommended songs given a specific genre;
- the audio characteristics of a track given its Spotify ID;

1.3 Methods

I will be conducting the experiment using the Random Forest method, which falls into the ensemble method category. Furthermore, I intend to employ additional classification models, including SVM, during the development phase.

1.4 Evaluation Framework

As a evaluation framework I will use the confusion matrix, which offers metrics such as precision and recall, along with accuracy.