SAPIENZA
UNIVERSITÀ DI ROMA

DEPARTMENT OF COMPUTER SCIENCE

Title

BIG DATA COMPUTING

Professors:
Gabriele Tolomei

Students:
Andrea Bernini - 2021867,
Donato Francesco Pio Stanco - 2027523

Academic Year 2021/2022

1 Project Proposal

1.1 Problem

The problem we want to address is to classify the songs according to their popularity. According to Spotify's Web API the popularity of a track is ranked with a number between 1 and 100, so we will discretize this problem by using 10 classes. We also analyze which aspects 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). Another possible idea is to classify the genre of a track (such as pop, rock, etc.), by using the aspect of the song such as the "loudness", "energy", "bpm", etc.

1.2 Dataset

For the project we chose the following 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). Furthermore we will use a Python script that uses Spotify's Web API to add additional features to the dataset, related to the track and the artist, such as the `popularity` and `genres` features, which are not present in the starting dataset. For the python's script we will use `spotipy` library which allows you to interface with the Spotify Web-App.

1.3 Methods

The method we are going to experiment will be the Random Forest, which belongs to the ensemble method category. Performance comparison with other classification models (such as SVM) is also included during development.

1.4 Evaluation framework

As an evaluation framework we will use the confusion matrix (which also provides precision and recall) and accuracy.