MAIS 202 - PROJECT DELIVERABLE 1

Noa Kemp (260898203)
Aymen Boustani (260916311)
Sanjeev Lakhwani (260954760)

1. Choose your dataset:
We are considering the two following data sets:

Kaggle Data set: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

- We choose this data set because it has more than 175k songs already updated to 2021 releases. There are also more than 10 features per song such as the tempo, the speecheness and danceability.

Free Music Archive; A Dataset for Music Analysis: https://github.com/mdeff/fma

- We choose this dataset because of its broadness -- metadata for each track such as ID, genres and play counts -- and its compatibility with Librosa, an audio and music processing library
- Most of files are CSV (comma separated values) files, which are highly compatible with Pandas, the library we'll use to process data

Million Song: Audio features and metadata for a million contemporary popular music track
http://millionsongdataset.com

- We think of choosing this data set because it is very large and englobing additional datasets which could be very relevant for the project.
- However it seems like it has not been updated since 2012; that's why we are firstly considering the other two datasets.

If we do use both; or all three of them, we will merge them into matching csv files; for instance, if we have features csv files from both/all the datasets, we'll merge them into one feature csv file.

2. Methodology:

a.

- We think that the datasets we chose are feasible because they are not too large in scale and mostly contain important features relevant to our predictive model.
- To preprocess the data, we will use Pandas' methods to extract from the CSV files the features needed for our machine learning model.

- The most useful information from the kaggle data set are the following features per song: tempo, danceability, valence -- feature representing the mood of the song --, accousticness, energy, duration, instrumentalness, tempo, liveliness, loudness (goes from -60 to 0; we could divide it into intervals) and speecheness. Both in the kaggle data set and free music archive data we have access to the artist, the name of the song, the key of the song and the release date. A great addition from free music archive data set is the amount of play counts per song.

- The million song data set could be very useful as it contains a lot of data; such as the full listening history of 1M users. It also contains a cluster of complementary data sets, which propose even more user data; as well as song tags and a lyrics dataset.

b.

- By studying and inputting the relevant features into our machine learning model, we would like to predict and propose high compatibility songs according to the user's music taste, based on what the user already listens to.
- For now, the main machine learning models we've studied are linear regression and KNN algorithms. Therefore, these are the models we'll use. We'll consider more complex models further in the course.

- Besides the material studied in class, we're trying to consider other machine learning models and see what models would be more relevant to the features we'll use for our predictions.
- KNN algorithms are useful to classify data. We don't know however if the linear regression model is completely applicable in our case.

c. Evaluation Metric:

Since it is the only evaluation metric we've covered so far, we'll settle on MSE for the moment. Again, as we'll evolve in the course and study other models, we'll consider other evaluation metrics, as well as other ML models.

d. Final conceptualization: For demo purposes, we want you to be able to showcase your project!

- Application
We want you to integrate your model in a simple landing-page webapp. For those of you who have experience, you are welcome to integrate your model in more sophisticated technologies (eg. mobile, hardware, webapps). Over the semester, we will host an Intro to Flask Workshop, make sure to come to this if you aren't familiar with web applications. Work with something that you are comfortable with :). Give the general idea of your application, and the technologies you plan on using. Here is an example of a simple webapp that uses Computer Vision to estimate people's age.
There are no limitations to the final product. You are free to build anything you want! :)
Furthermore, you should be able explain the specific problem's accepted metrics. Keep track of the average baseline results which you hope to beat (eg. predict x with y% accuracy, < x mean squared error, etc).