MAIS 202 - PROJECT DELIVERABLE 1

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

Due: TBD
Over the course of MAIS202, you will be completing a machine learning based project of your choice for the final project. You will also demo your project by integrating it into a webapp (or something more advanced). To conclude, you will be writing a blog post to share your project with others.

Submission
This is an individual deliverable. All deliverables should be electronically submitted on Github and completed with the same academic integrity and standards expected at McGill University. Include appropriate citations.
To submit, create a repository under your own Github, you can name it whatever you want, and push your report there. List your repository link in this spreadsheet . All of the code and report for this project should be found in this repository. Make sure to maintain it with properly documented README and structured code.
Submit this deliverable as “Data Selection Proposal.pdf”
Max length: 1 page
Deliverable Description
The first step of the project is to choose the dataset that you want to work with and propose your project idea. Your project idea can either come from a list of pre-approved deliverables available here, or it can be an idea you create yourself. Note that the ideas on the pre-approved project list are first come, first served (no duplicates). Once you have an idea, discuss it with your assigned TPM. The two of you will work on ironing out the details of your idea, so that it is doable in the next 10-weeks. Only once your TPM has approved the project should you begin writing up your deliverable.

1. Choose your dataset: You can choose any public dataset of your choice (don’t forget to cite them!). There are also a couple of useful databases that are available: Google Dataset Search and Kaggle. Explain the reasons why you choose this dataset. Furthermore, you can also look into creating your own custom dataset by scraping websites. If so, explain what kind of data you will be scraping.

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

If we do use both, we will merge them into matching csv files; for instance, if we have features csv files from both datasets, we’ll merge them into one feature csv file.

2. Methodology: Describe how you plan on approaching the project. This should be a high level overview of your plans, and this will allow us to judge the feasibility of your project. Be as thorough as you can, so we can give you critical feedback.

a. Data Preprocessing: Is the dataset you chose feasible? What information provided is/are the most useful? How are you planning on preprocessing the dataset to

extract this information? You can take a look at these F2019 slides on data preprocessing.

- We think that the dataset we choose is 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 github we have access to the artist, the name of the song, the key of the song and the release date. A great addition from the github data set is the amount of play counts per song.

b. Machine learning model: What do you want to predict/estimate from this dataset? Propose a machine learning model/algorithm for it, and explain your reasoning. Have you considered other alternative models? What are the pros and cons?

- 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: Analysis requirements differ in every field, but some things to consider reporting include but should not be limited to:
i. Confusion matrix and accuracy/precision-recall/logistic loss (classification problems).
ii. Mean squared error (regression problems)
iii. Rand index (unsupervised models)
iv. BLEU score with brevity penalty (text generation)
v. Variance of the dimension reduced set vs variance of the initial dataset (dimensionality reduction/PCA)
If you are not sure, ask your Buddy.

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).