Mid-term Project

The mid-term project is **due at the start of class** on **Wednesday, March 27, 2019**, without exceptions unless permission was obtained from the instructor **in advance**.
You may collaborate in **groups of up to 3** and submit a joint report. However, you must submit **test set predictions individually** on eLearning. That is, if you worked in a group of 3 and created a file of test set predictions together, all three students must submit that file individually on eLearning.

Problem. What does it take to make a **Billboard Top 100 Hit**? **Hit song science** aims to predict if a song will become a hit prior to its release and is one of the applications within **music information retrieval**[1].

Task: We will use the machine-learning approaches we've studied to to train and validate a model for this problem. The training data consists of around $1,900$ songs from the 2000s taken from the Million Songs Dataset[2] created by Columbia University cross-referenced with the list of Billboard Top 100 Hits[3]. Each song is characterized by audio features (such as danceability, speechiness and tempo; more information below) extracted from the Spotify API. **There is no missing data**. You will:

- Compare any **three machine-learning algorithms** to predict if a song will become a Billboard Top 100 Hit using the **training data**. You may choose among: Support Vector Machines, Decision Trees, Naïve Bayes, Logistic Regression and k-Nearest Neighbors. You may pre-process the data in any appropriate manner (make sure you check out the documentation for `sklearn.preprocessing`). You should also aim to design an appropriate cross-validation mechanism via Grid Search; it may be convenient to design composite estimators that pre-process, learn and cross validate the model using `sklearn.pipeline`[4].
- Select the best model and make predictions on the **test set**. Note that the true labels of the test set and identifying information such as the names of artists, songs and song years have been **removed**. **Your test set predictions must be uploaded to eLearning by the due date for grading**.
- Write a brief report (see instructions below) describing your steps and justifications. **A hard-copy of your report must be submitted in class on the due date for grading**.

Data. The training data consists of **meta-data**: Artist, Track, Year, and **features**: PreviousHit, Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence and Tempo. The feature PreviousHit is 1 if the artist previously had a Billboard Top 100 Hit between 1986 and 2010, and 0 otherwise. Details of all other features can be found in the Spotify API documentation[5].

Report. **Using the report templates** (.tex/.docx) provided, write a brief report detailing: (a) pre-processing, (b) machine-learning methods compared and your reasons for choosing them, (c) bar plots comparing your chosen machine-learning methods, (d) your final choice and your reasons for choosing it and (e) thoughts for future work on how your model might be improved. Your report must not exceed 3 pages. **Make sure you detail each team member's contribution to the project, otherwise the team will receive no grade**.

Submission Instructions. Once you have your final model, you can make predictions on the test set provided. Generate a file of test set predictions named `predictions.txt`, **with either 1 or 0 per line** indicating your model's prediction. This file must be uploaded to eLearning individually.

Acknowledgement. This data set was originally created by Elena Georgieva, Marcella Suta, and Nicholas Burton at Stanford University.

[1] https://en.wikipedia.org/wiki/Music_information_retrieval
[2] https://labrosa.ee.columbia.edu/millionsong/
[3] https://www.billboard.com/charts/hot-100
[4] https://scikit-learn.org/stable/modules/compose.html#pipeline
[5] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/