Divya Thomas

Class: CS 677

Date: 4/26/2023

Code Instructions

1. Initial_analysis.py

   This file contains mean, standard deviation calculations for song popularity data and feature data. It also holds all analytical and preprocessing functions required throughout this project, such as train-test-split, accuracy measurements and counts. Outputs of these data is summarized in the supplemental summary document. Therefore this code is not required to be re-run.

2. Knn.py

   This file contains the algorithm for k-Nearest Neighbors. When run it will perform a series of runs with different k values ([3, 5, 7, 9, 11]). The accuracies of these values will then be plotted and saved in the datasets directory under the knn_plot.png filename. This file also contains a function to calculate and output count measurements. These are used later in the project.

3. L_kernel_svm.py

   This file contains the algorithm for linear-kernel SVM. It also contains a function to calculate and output count measurements. These are used later in the project. This file has no output.

4. Logistic_regression.py

   This file contains the algorithm for the logistic regression model. It also contains a function to calculate and output count measurements. These are used later in the project. This file has no output.

5. Naive_bayesian.py

   This file contains the algorithm for Naïve Bayes model. It also contains a function to calculate and output count measurements. These are used later in the project. This file has no output.

6. Consolidated.py

   This file contains the algorithm for consolidated ensemble model. It also contains a function to calculate and output count measurements. These are used later in the project. This file has no output.

7. All_model_accuracy.py

   This file will run all the models on the same training and testing split and output their comparable counts and accuracy measurements.

8. Feature_removal.py

   This file, when run, will implement the consolidated model on training and testing data, as the previous file does. However, each run of this model will be missing a feature. After each feature has been removed at least once, the accuracies resulted from each run are plotted and saved in the datasets directory under the filename missing_feature.png

9. Predictor.py

   This file will display a simple interactive UI. The user will be asked prompts to input feature values. These values will then be run against the consolidated classifier after it trains on the full dataset in order to output a prediction of whether or not the classifier deems the inputted song to be successful and popular.