Machine Learning - Assignment 2

Publication date: 15/05/2025

Due date: 12/06/2025

Data

Data source: Spotify and Youtube Song Data
This dataset contains a comprehensive analysis of over 20,000 songs and their properties, collected from their Spotify and YouTube pages. The data encompasses the song's name, artist name, album name, links for the song on Spotify and YouTube, popularity data, as well as musical properties such as key, loudness etc.

Your task is to predict whether a song is published as part of an album, or as a single. For that purpose, the feature 'Album_type' serves as our label. The values are 'single', 'album' and 'compilation'. For our purposes, 'compilation' counts as 'album'.

Note: The label counts are imbalanced. You are expected to handle that as we learned in class, or if you choose to not handle that, provide an explanation as to why.

A detailed explanation of each column can be found in the Kaggle page of the dataset.

Requirements

Section A - Data Exploration & Visualization (10 pts)

Explore the data using tables, visualizations, and other relevant methods.

- Plots should have an informative main title, axis labels and a legend (if needed).
- For each plot or table, provide a short description of **key observations**. Make sure to only include content which would be **meaningful/informative**.
- The visualizations should be detailed and cover all relevant aspects of the data.

- The visualizations should highlight any interesting patterns or trends that can be observed in the data, as well as key statistics such as the mean, mode etc. of each feature.
- Perform **at least 5 visualizations**, in **at least 3** plot types.

The goal of this section is to get insights on the data which may or may not be relevant for the following sections.

Section B - Data Pre-processing (30 pts)

Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections. The results of this section have a direct impact on your models' performances in the next sections. Be sure to read and understand them before you start working on this section.

- Perform feature engineering on the data, including the specific features provided below and at least six additional features of your own choosing.

  Explain why you chose these features and how they may improve model performance.

  Create the following features - add them to the current data:

  - Album Song Count: The total number of songs in the current song's album.
  - Average Artist Song Views: The average number of YouTube views a song by the current song's artist gets (total number of song views by current artist divided by the total number of songs by the current artist).

    Note: Similar features (such as average number of likes, streams, comments etc.) would only count as one of the six features you need to engineer.
  - Song Name Length: The number of words in the name of the current song (track).
  - Total Album Length: The total length of the songs in the current song's album.
  - Fitness for Clubs: How fit the song is to be played at clubs. An average of the 'Danceability', 'Energy', 'Loudness', and 'Valence' features (**Note:** Loudness is not on the same scale as the other features).
  - At least six other features of your own choosing.

- Apply at least one type of imputation (if needed), one transformation, and one exclusion (i.e., feature selection).
  Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it’s needed.
  Note: Not all features should be used for our task.

IMPORTANT: In the following sections, you do **not** have to implement the models/algorithms yourselves. You may use existing models from scikit-learn. If you use another library, explain the usage and algorithm.

Section C - Single/Album Song (25 pts)

Use at least **three** different machine learning models we have studied in class to predict whether the song is a single or not, according to the ‘Album_type’ feature (**Note** the clarification for its values in the ‘Data’ section).

- The implementation must include parameter tuning.
- **IMPORTANT:** The data is not currently split into train/test/validation sets. You need to split the data into three sets with a 80/10/10 % split (train, test and validation respectively), and train your model with the train set only. Hyperparameter tuning should only be done on the **validation** set, and once you find the best parameters, evaluate your model on the **test** set.
- Report a suitable measure to evaluate the performance of each model.
- Present the models’ results in a plot.
- Compare the results of the different models, discuss them.

Bonus - Feature Importance (10 pts)

Not all features affect model performance in the same way. Perform feature importance analysis in any way you choose:

- Calculate correlation between a feature and a label.

- Drop features and check performance with/without them.
- Check correlation between features to see if they don’t provide additional useful information.
- Any other method of your choice.

You may use existing methods and libraries (if you use them, explain their usage and the method). Be sure to annotate properly, report your findings and discuss the results.

Section D - Clustering (25 pts)

Apply at least **two** clustering algorithms we studied in class on the data to cluster the songs. You may use all of the features you used for Section C, drop some of them or create additional features if you think they’re appropriate.

- Use parameter tuning based on the algorithms you selected.
- Identify the most important features that contribute to the differences between the clusters. Discuss your findings and find a way to demonstrate **visually** what similarities the clusters may have.
- Use a method (of your own choice) to estimate the quality of the clusters you created with each clustering algorithm. Visualize the results according to the method you selected. If you use an evaluation method we haven’t learned in class, explain the method, its results and what they mean.
- Try to explain the meaning of the different clusters you found.

Presentation (10 pts)

Create a short presentation (no more than 6 slides) that includes interesting findings of your choice. A few presentations will be chosen to be presented in front of the class. The goal is to learn from other students’ work.

This section is mandatory, not a bonus!

Section E - Exploring Artists- Bonus (15 pts)

In this section, you will explore the different artists and their characteristics.

- Using the existing dataset, create a new dataset where each row represents an artist and the features represent their properties and song characteristics (e.g. Average/Total views, Total number of singles, Average Song Fitness for Clubs etc.)
- Formulate a question that can be asked about this set, and suggest a machine learning algorithm that can answer it (this can be a classification/regression task or a clustering task). If you decide to implement a method that was not discussed in the course so far, include references to where you studied it.
- Apply the suggested machine learning algorithm.
- Discuss the results and reflect on your question and choice of solution.

Guidelines

Please read the following section carefully before submitting the assignment.

Coding Guidelines

- Use familiar packages with explicit explanations.
- If you have installed any libraries beyond those presented in the exercises, please specify this in the report.
- The code should run without warnings or errors.
- Good documentation is **critical**.
- Indicate the exercise sections in the code as well.
- Use meaningful variable names.
- Do not use reserved words.
- Use constants where possible.

Submission Guidelines

- The assignment should be submitted in pairs (only one submission).

- You are required to submit two files including all the sections. One in **.ipynb** format and one in **.html** format. **Both files should also include the program's outputs**. In addition, you are required to upload a **PDF** file of the presentation you prepared.
- The files' names should be of the form: **ML_HW2_ID1_ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to one week. Later submissions will not be accepted.

Grading

You can get more than 100 points for the exercise. The exercise will be graded according to correctness, clarity, efficiency of implementation, elegance of implementation.

Self-learning

As we mentioned at the beginning of the course, self-learning is an important part of the course. Treat all sources of information carefully and critically.

Usage of LLMs is allowed, but reference it where it was used.

You can and should consult with other students in the course, but each pair must write their own work.

It is reasonable to assume that not all results and algorithms will be identical.

Questions and Reception hours

- Please post your questions on the exercise forum in Moodle, after you have read the previous posts. Professional questions sent by email will not be answered.
- If you want to schedule a reception hour with one of the instructors - please send your questions by email in advance.
- In any other case (personal questions, request for an extension with a justified reason, etc.) please email the instructor.

Good luck