Lab 2 - Music Mood Classification - Report

2024-03-20 Project Report - Stefano Scola, Robin Doerfler

for ATSMC Course 2023/2024 by Rafael Ramirez Melendez

Feature Selection

Methodology

The pipeline presented in this section utilizes scikit-learn for data preprocessing, feature selection, and the training phase. Before the analysis started, the data underwent the usual preprocessing steps *(see data_loader.py)*. After a quick descriptive inspection, the columns "lm" and "ld" were dropped as they contained mainly missing values. Consequently, missing values were imputed using the mean with the SimpleImputer. Regarding the scaling of the numerical features, the current implementation allows easy selection of either StandardScaler or MinMaxScaler. Labels were encoded using LabelEncoder. It's important to note that both the scaler and encoder were fitted only on the training data to avoid biasing the test set.

Once the data had been preprocessed, feature selection was conducted in two steps *(see model_factory.py).* Firstly, from the initial 62 features, 50% of them were selected using SelectPercentile by examining the statistical correlation between the labels and the features. The pipeline allows for the selection of the percentile to retrieve a specific portion from the initial set of features. Secondly, the statistically selected set of features was further reduced by employing SequentialFeatureSelection, both in a forward and backward fashion. Contrary to the previously mentioned statistical feature selection, this type of feature selection makes use of a sklearn estimator, for our purpose we have used random forest, gradient boosting and k-nearest neighbors.

The final set of selected features was initially set to 5. However, in order to gain more insight into the features, experimentation was conducted with values of 10 and 20 as well (as shown in the graph below).

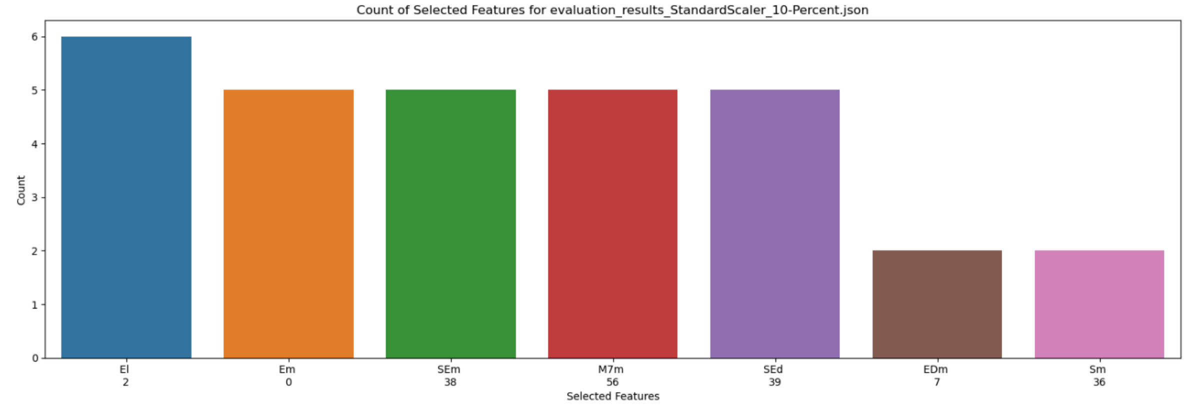
Results

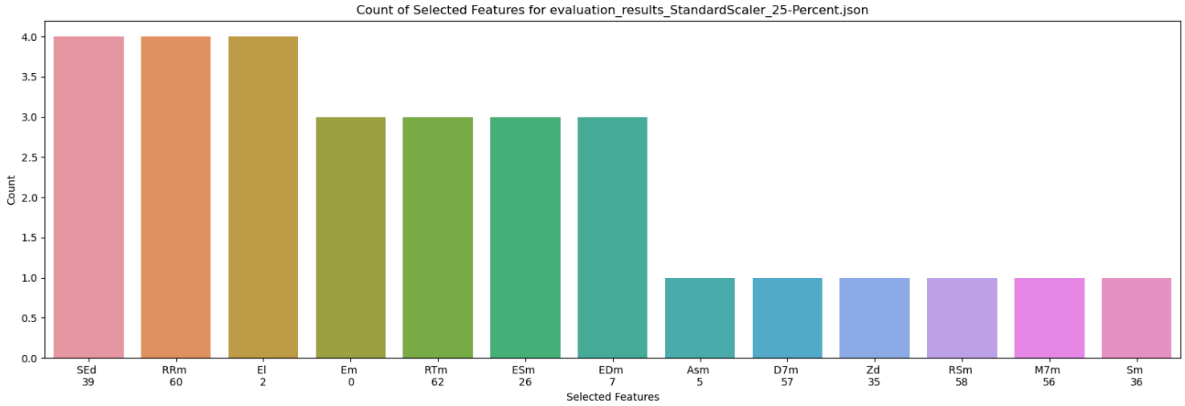
Based on the methodology described above, we evaluated the most informative features and investigated their categories.

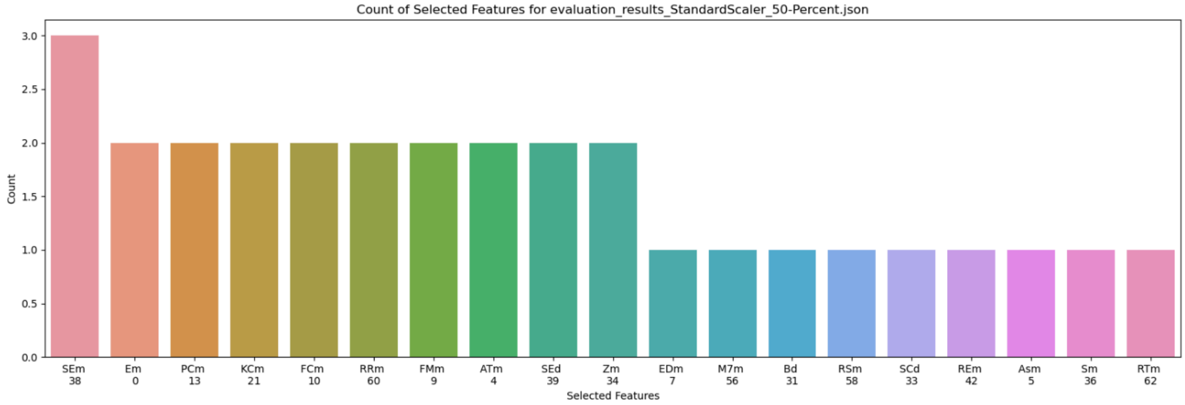
Most informative features

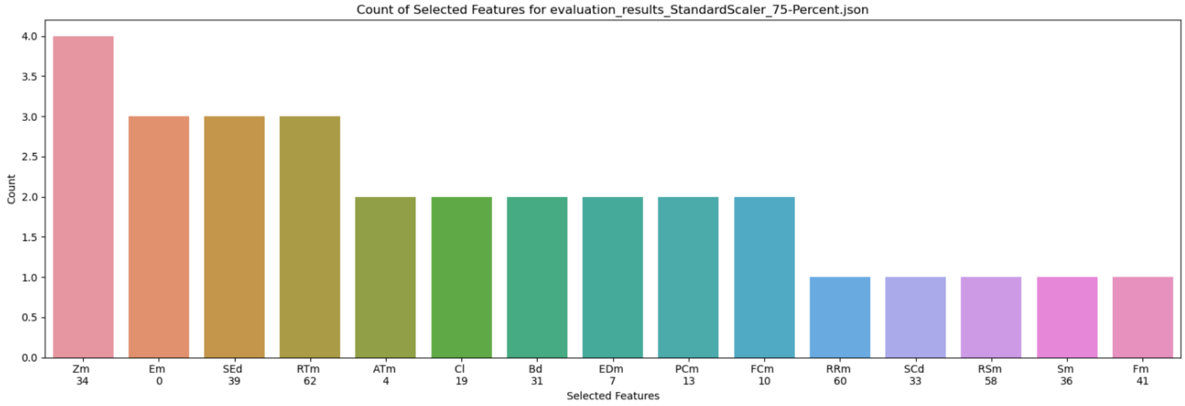
The first three graphs below show the results of the feature selection process. Each graph represents the aggregation of the results of the three different algorithms mentioned above. As indicated by their titles, each figure represents the results of a different configuration. The graphs show the selection of k=5 best features, when provided with a pre-selected subset of 10 %, 25%, 50% and 75% of the features, normalized using the StandardScaler. Both

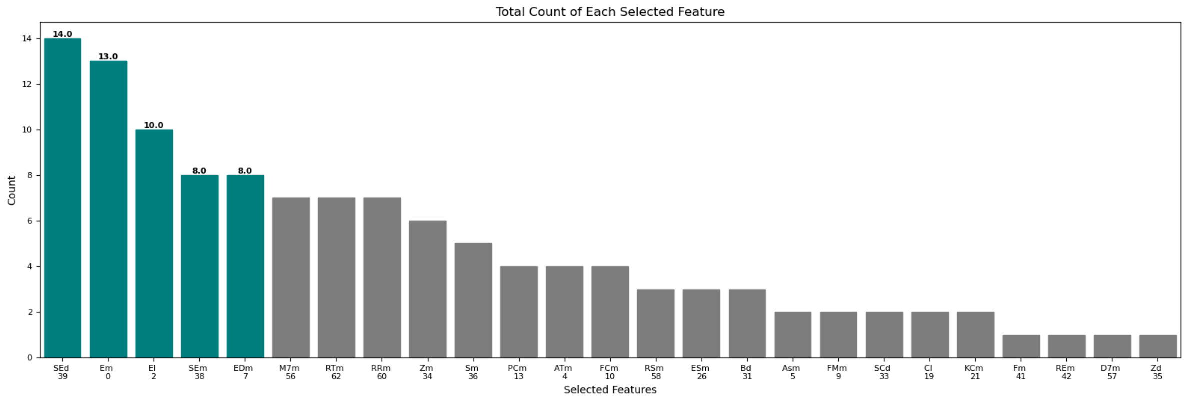
StandardScaler and MinMaxScaler were compared for the task, but did not change the results of the selection drastically.











We can see that all the feature reduction and selection configurations differ to some extent in their results. The 5 most informative features among the sum of all feature selection configurations were *SEd, Em, El, SEm* and *EDm.*

Em & El: RMS Energy [Mean and Entropy]

Average energy level of an audio signal within a specified time frame.

- average energy level of a signal over a certain time frame.
- Root Mean Square (RMS): Square root of the mean of the squared values of the samples within the specified window.
- provides insight into the overall intensity or loudness of an audio signal.

SEd & SEm: Spectral Entropy [STD and Mean]

Diversity or randomness in the distribution of frequency components within an audio signal.

- measure of the disorder or unpredictability of a signal's spectral content.
- quantifies the distribution of energy across different frequency bands within a segment of audio. Higher spectral entropy values indicate a more diverse or spread-out distribution of spectral energy, whereas lower values indicate a more concentrated or organized distribution.

EDm: Event Density [Mean]

The rate or concentration of distinct audio events within a specified time interval.

- measure of the frequency or concentration of distinct audio events within a given time frame or segment.
- quantifies how busy or active a particular section of audio is in terms of events such as notes, sounds, or other auditory phenomena.

M7m: 7th MFCC [Mean]

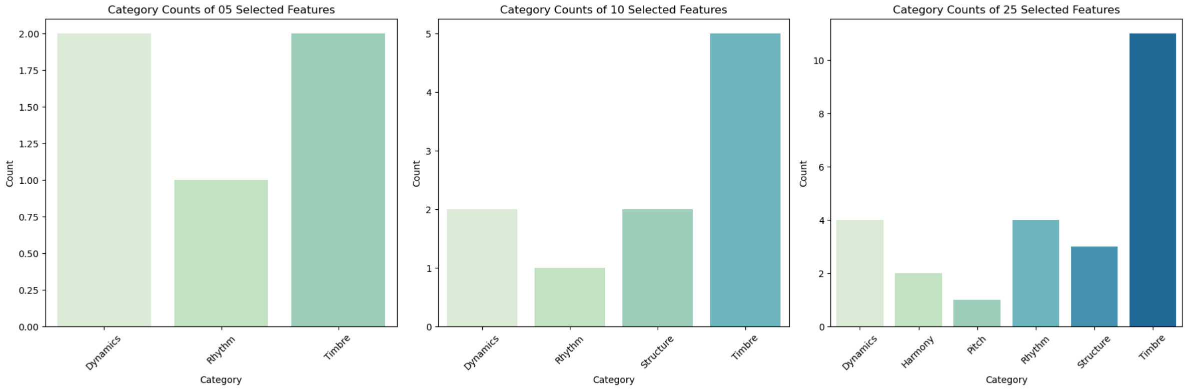
The seventh Mel-frequency cepstral coefficient, a logarithmic, mel scaled filterbank capturing the ‘Spectral Shape’ of a signal as perceived by the human hearing system.

- measures the logarithmically compressed magnitudes of a mel-scaled filterbank applied to the signals spectrum.
- captures important spectral characteristics of the audio signal, providing information about its timbral qualities.

Most relevant feature category

Out of the five most informative features, two are found in the category of ‘Dynamics’, another two in ‘Timbre’’ and one in the ‘Rhythm’ category. In sum, the most important category informing the decision of the classification was the Dynamics, more precisely the average energy of the signal and how much this energy varies across the song. Almost equally relevant has been the category of ‘Timbre’, capturing how the ‘chaotic’ or ‘organized' spectrum is in average and deviation. Even though the ‘Event Density’ is part of the Rhythm category, it is not unrelated to both the other measures. It therefore seems that the most informative features are rather describing how things are happening, how frequently things are changing, how organized these changes are, rather than what exactly is happening, such as features related to the Pitch or Harmony categories.

When looking at the most 10 and most 25 informative Features, we can see that more and more Timbre information is taking part in the decision, making it the most contributing category from a macro perspective. However, it is also the category with by far the most features, therefore we can not conclude that it necessarily contains the strongest information.



Model Improvement

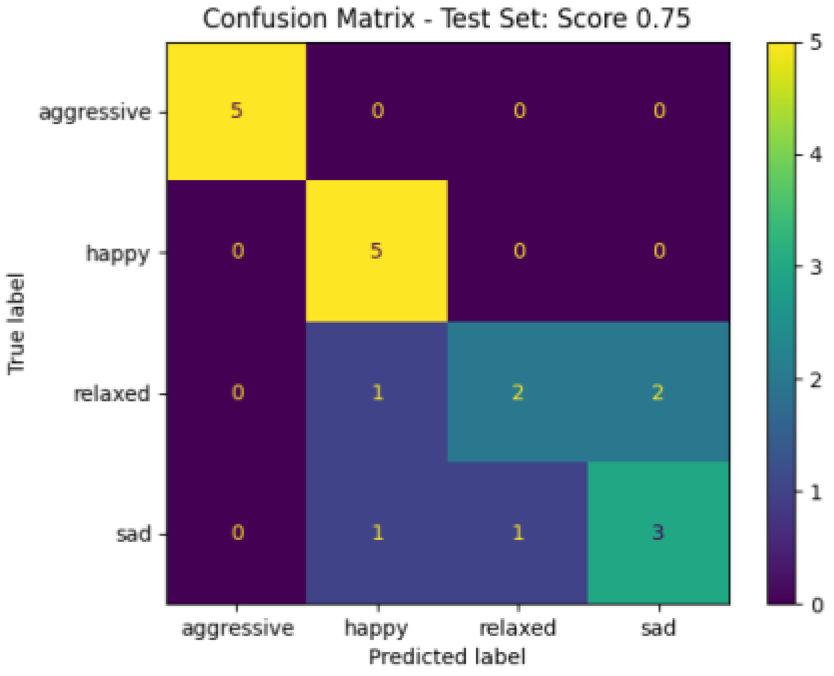
In order to improve the performance of the model, hyper parameter tuning has been implemented for random forest and k-nearest neighbors *(see models_gym*.py and *train_best.py).* The current implementation uses random search and allows for a more extensive search by increasing the number of iterations, or more refined search by specifying the score metrics to optimize. Finally, KNN yielded an accuracy of 0.75 on the test set.

Error Analysis and Misclassified Songs

To gain insights into the errors made by the two aforementioned algorithms, we computed a confusion matrix, which is presented below. Additionally, we saved the misclassification results in a JSON file (accessible in the "results" folder), enabling us to listen to the songs that were incorrectly classified. Let us now examine one instance of misclassification found in the results of the best performing model.

```
{
    "file_name": "02-padraic my prince.mp3",
    "original_label": "sad",
    "misclassified_label": "happy"
},
```

Classifying this song as "happy" is clearly a mistake, as it can be heard. This specific data point might be considered an outlier, which biases the model predictions. Nonetheless, upon observing the confusion matrix below, we noticed that the model exhibits the worst performance regarding the class "relaxed." Indeed, relaxed tracks are misclassified as sad twice, and once as happy. Moreover, one song, originally labeled as sad, gets classified as relaxed. Overall, these results indicate the difficulty of disambiguating and assigning a single emotion label to a track. In fact, we believe that a song can evoke both sadness and relaxation simultaneously. The ambiguity of this problem will be briefly discussed in the next section.



KNN Confusion Matrix

Reflection on the Problem

Data Ambiguity

Even though most of the examples in the dataset are unambiguous, it can be argued that a musical experience is always more complex and nuanced than a single label can appropriately capture. There are songs which are exactly built on the contrast between a ‘happy’ chord sequence and sad lyrics. Furthermore there is no mutual exclusiveness between songs which are ‘relaxed’ and other emotional qualities. For example ‘01-homesick’ which is labeled as ‘relaxed’ could equally be perceived as sad. Especially among the ‘relaxed’ tracks this ambiguity can be observed more frequently to some extent. If we would like to expand on the task of mood classification, breaking the high-level emotional labels into sub-categories and more precise tags, or adding labels which explicitly imply a certain ambiguity, like ‘melancholic’ might be helpful.

Posing of the Problem

The problem of music mood classification can be considered ill-posed due to its inherent complexity and subjectivity. Assigning a single emotional label to a piece of music oversimplifies the multifaceted and dynamic nature of human emotional responses to music. Music is capable of evoking a wide range of emotions simultaneously or sequentially, often eliciting different reactions from different listeners. Moreover, emotional responses to music are influenced by various factors such as cultural background, personal experiences, and even physiological states. These factors contribute to the subjective and fluid nature of emotional perception in music, making it challenging to establish a universally agreed-upon classification system. Therefore, the task of music mood classification can be seen as ill-posed, as it involves reducing a complex and nuanced phenomenon into a single categorical label, which fails to capture the richness and variability of emotional experiences in music.

Accuracy of a human performing the task

As described in the above, the problem of music mood classification with a single label is not perfectly well posed to begin with. Besides the complexity of moods in music, the verbal language, with its limitations in capturing the intricate nuances of emotional experiences, is another obstacle for this task. Additionally, human perception itself is highly subjective and multifactorial, influenced by cultural backgrounds, individual biases, and varying sensitivities to emotional nuances. As a result, the accuracy of such classification would be expected to be inconsistent among individuals. Attempting to assign a single label to the complex emotional landscape of a song would inevitably lead to frustration, as it oversimplifies the richness of the musical experience. Precisely and accurately describing feelings with words is a formidable task in any context, further complicating the process of emotional classification in music.

Discussion

Through the course of this assignment, we have come to realize that the attempt to create a generalizable model for the task of music mood classification with such limited data is somewhat impractical.

One major limitation is the small number of mood categories used for classification. With only four categories, the complexity of musical emotions is oversimplified, leading to ambiguity in labeling the dataset. Subtle variations and nuances in musical expression are restricted and not reflected in the labels. Therefore, expanding the range of emotion labels could lead to more accurate predictions and provide deeper insights into the relationship between music and emotion.

The size of our training dataset was another constraint. Having only 60 examples limits the capacity of the model to learn from diverse inputs and make reliable predictions across different genres and styles. To address this issue, significantly increasing the volume of the dataset should yield better results. Moreover, incorporating additional sources of information, such as song names or lyrics, may help identify meaningful patterns within the data and enhance prediction accuracy.

Lastly, reconsidering the targets themselves might prove beneficial. Labels like 'happy,' 'sad,' 'aggressive,' and 'relaxed' overlap and do not form mutually exclusive categories. As a result, these broad terms fail to account for complex combinations of emotions present in many pieces of music. By refining target definitions and allowing for multiple simultaneous labels, the model might achieve greater precision and reflect real-world scenarios more closely.

To summarize, while our project has shed light on several challenges faced when attempting to predict musical emotions using audio features alone, further investigation remains necessary. For a more generalizable model, it may be a good idea to address the issues related to the breadth of emotion categories, expand the dataset, consider alternative input types, and to reassess target specifications for improved performance and understanding.