Lisa Siefker
Music Selection – Final Project
Data Mining Module – CPDA
SP2024

PROBLEM STATEMENT

I am a data miner assisting the new manager of an exercise studio. The manager would like me to select the best combination of ten songs for each exercise class offered at the studio.

The class schedule is as follows:

	SUN	**MON**	**TUES**	**WED**	**THURS**	**FRI**	**SAT**
5:00 AM	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength
6:00 AM	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic
12:00 PM	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength	aerobic + strength
6:00 PM	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic
7:00 PM	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic
8:00 PM	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic	aerobic

PROBLEM ANALYSIS

First, I met with the studio manager to better understand the business problem. I gathered the following information at that meeting:

- All music must be selected from a provided list of 2,000 songs that the studio has purchased.
- Songs with explicit lyrics can be included in the playlists.
- Three different instructors (Molly, Piper, and Teddy) teach classes at the studio, and the instructors have a set schedule. Any songs selected must meet the preferences of each instructor.

Next, I met with the instructors to further my understanding of the business problem:

- The instructors have the following preferences for the music played in the classes they teach:
    - Molly prefers music released in the 2000's
    - Piper prefers pop music
    - Teddy prefers hip hop music
- Each class lasts one hour and includes a 10-minute warm up and a 10-minute cool down
- The aerobic + strength classes include 20 minutes of aerobics and 20 minutes of strength training
- The aerobic classes include 40 minutes of aerobics

- Research[1] indicates the following tempos are ideal for the following types of exercise:
  - dance (aerobics): 130 to 170 bpm
  - weightlifting (strength): 130 to 150 bpm
  - warm up: 100 – 140 bpm
  - cool down: 60 – 90 bpm

	SUN	MON	TUES	WED	THURS	FRI	SAT	DUR	TEMPO (bpm)
5:00 AM	aerobic + strength PIPER (pop)	aerobic + strength TEDDY (hip hop)	aerobic + strength TEDDY (hip hop)	aerobic + strength TEDDY (hip hop)	aerobic + strength PIPER (pop)	aerobic + strength PIPER (pop)	aerobic + strength TEDDY (hip hop)	10 m 20 m 20 m 10 m	100 -140 130 -170 130-150 60-90
6:00 AM	aerobic PIPER (pop)	aerobic TEDDY (hip hop)	aerobic TEDDY (hip hop)	aerobic TEDDY (hip hop)	aerobic PIPER (pop)	aerobic PIPER (pop)	aerobic TEDDY (hip hop)	10 m 40 m 10 m	100 -140 130 – 170 60-90
12:00 PM	aerobic + strength TEDDY (hip hop)	aerobic + strength TEDDY (hip hop)	aerobic + strength TEDDY (hip hop)	aerobic + strength PIPER (pop)	aerobic + strength TEDDY (hip hop)	aerobic + strength TEDDY (hip hop)	aerobic + strength PIPER (pop)	10 m 20 m 20 m 10 m	100 -140 130-170 130-150 60-90
6:00 PM	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic PIPER (pop)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	10 m 40 m 10 m	100 -140 130-170 60-90
7:00 PM	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic PIPER (pop)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	10 m 40 m 10 m	100 -140 130-170 60-90
8:00 PM	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic PIPER (pop)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	aerobic MOLLY (2000's)	10 m 40 m 10 m	100 -140 130-170 60-90

Lastly, I read online reviews for the exercise studio to see if I could gain additional understanding of the business problem. Four reviews were posted in the last year. The reviews focused on the cleanliness of the studio and availability of parking but did not provide any customer insight into the song selection problem.

DATA USED

[1] https://www.cnet.com/health/fitness/can-a-playlist-boost-your-performance-yes-with-the-right-songs/

I considered the songs_normalize dataset provided by the studio manager. I reviewed the dataset on Kaggle and learned that the song list comprised was comprised of the top hits on spotify from 2000 – 2019. I also reviewed the data categories and input[2].

DATA ANALYSIS & INSIGHTS

An analysis of the data shows that there are no nulls in the dataset. I identified 59 duplicate rows and removed them.

Popularity analysis: All songs on the list are the top hits on spotify; however, almost 175 songs have a popularity rating of '0'. The data input does not explain how popularity is calculated, but states 'the higher the value the more popular the song is'. Since this is a list of top hits, I will make the assumption that all of the songs are popular for purposes of my analysis and drop the 'popularity' column.

Instrumentalness analysis: Kaggle input says that instrumentalness "predicts whether a track contains no vocals." Values range from 0.0 to 0.99; however, the mean is very close to 0. It's not clear how this value is measured and most values are 0 or very near 0, so I will drop the 'instrumentalness' column from the dataset.

Genre analysis: Many songs in the data set are categorized into multiple genres. I made the assumption that it is possible for a song to be correctly classified into multiple genres and that all genre classifications in the dataset are accurate.

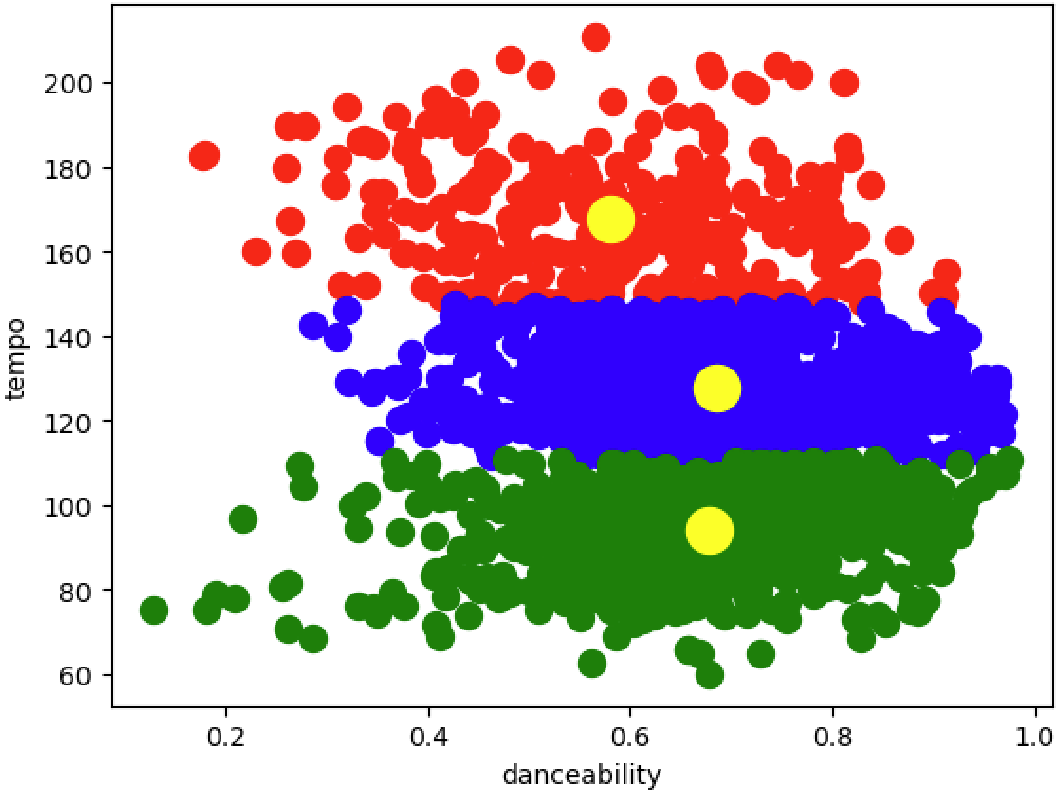
Duration_ms analysis: The duration of each song is measured in milliseconds. There are 60,000 milliseconds in one minute, and I will need to choose songs from particular genres and tempo ranges for each class. I make an assumption that a song longer than 5 minutes (300,000 ms) is too long for an exercise class. I decide to drop songs longer than 300,000 ms.

Danceability analysis: I made the assumption that all songs chosen for the exercise classes should have higher than average (> 0.67) danceability to keep energy up and customers excited and happy with the selections.

K-means

I implemented the K-means algorithm in an attempt to identify clusters based on danceability and tempo data. I used a copy of the original data to cluster since I was interested in the full range of danceability. Using the elbow method, I chose to segment the data into four clusters. This technique did create three distinct tempo clusters; however each cluster was fairly evenly distributed in the range of 0.4 – 0.9 danceability. The clustering data was ultimately not very useful in analyzing the business problem.

[2] https://www.kaggle.com/code/soldatovda/spotify/input

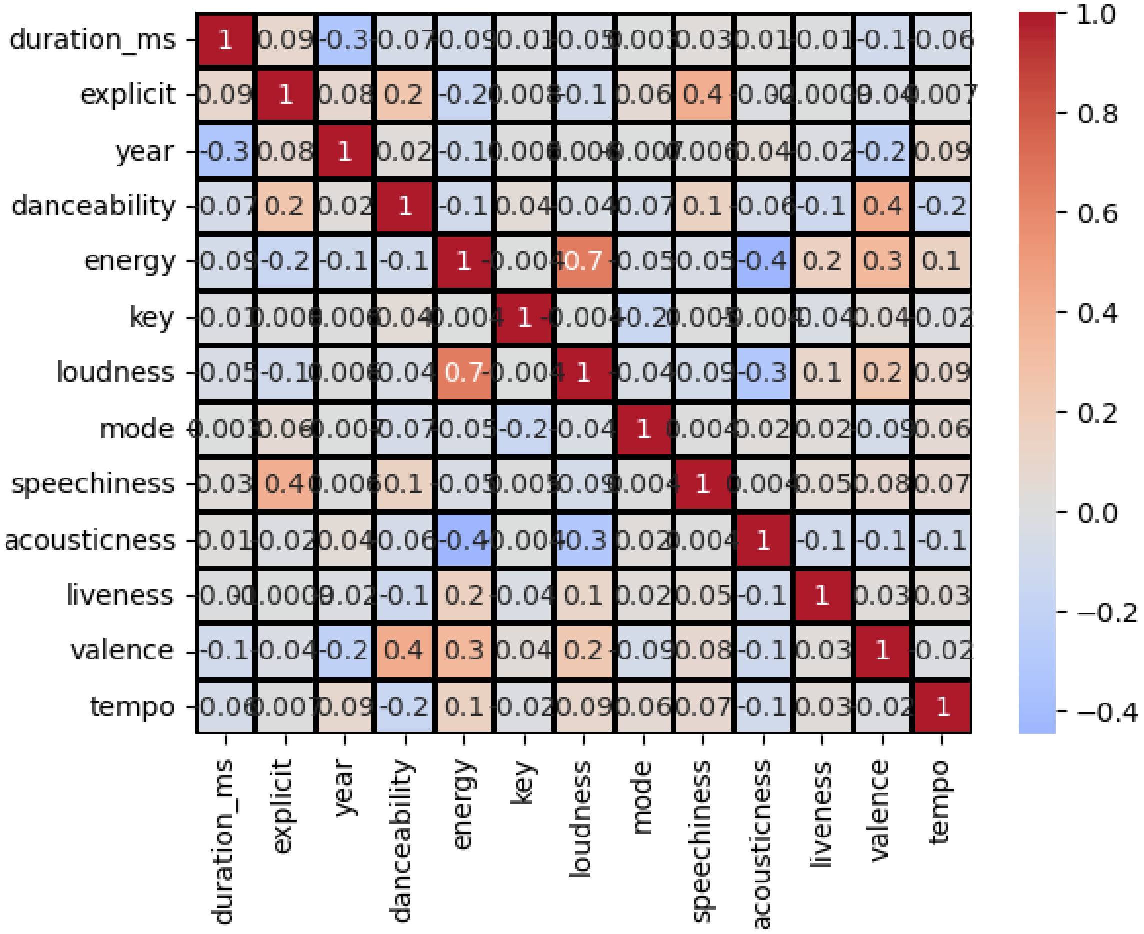


Association Rule Mining

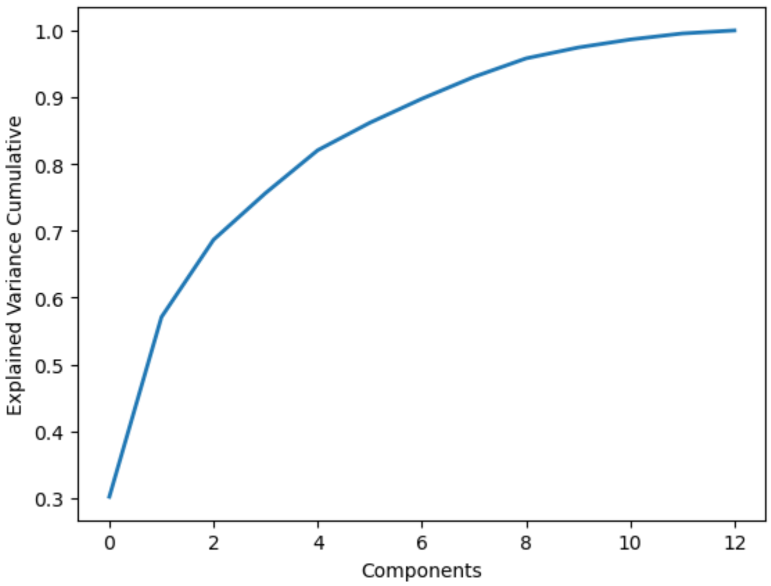
My data does not lend itself to association rule mining. The songs_normalize dataset provides information for individual songs but does not include playlists and does not indicate which songs are frequently played together. That information could have potentially been beneficial to an associated rule mining analysis. If songs are frequently played together, I could consider also playing them together in the exercise class playlist.

PCA, Decision Tree and Random Forest

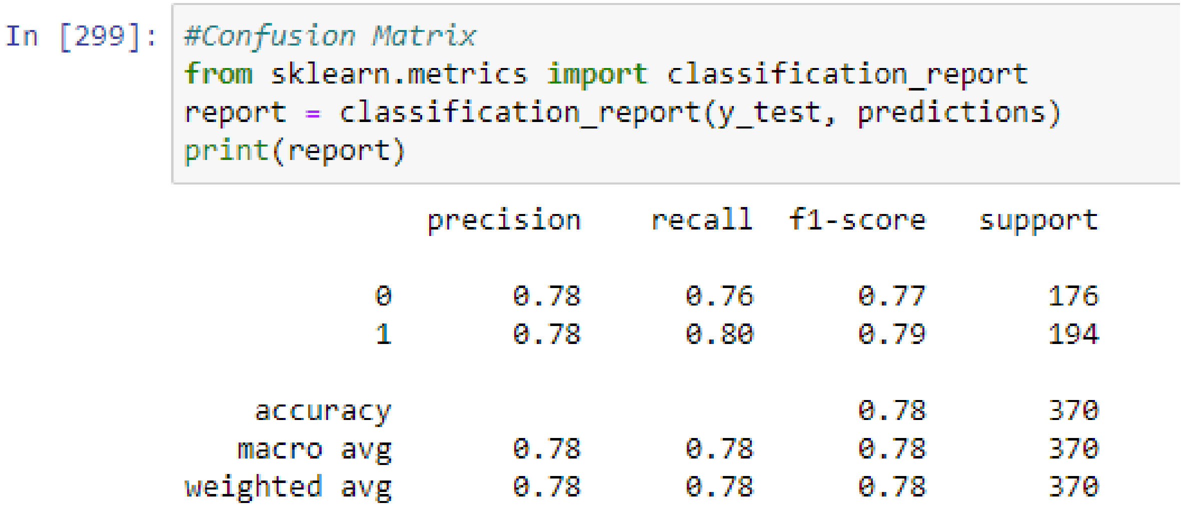
Each song in the data set was measured by 14 variables. A correlation matrix shows that energy and loudness were the most strongly correlated (0.6). Energy and valence had some correlation (0.4) and loudness and valence were also somewhat correlated (0.3).



Next, I applied principle component analysis (PCA) to my data frame and decided to use 4 components, which would explain around 75% of variance.



I also wanted to determine whether I could predict “danceability” using a random forest classifier. I created a new ‘danceable’ variable and classified a song as ‘danceable’ (1) if danceability for that song was more than the mean danceability (0.67). A song is classified as ‘not danceable’ (0) if the danceability was less than the mean danceability. The model was able to predict danceability with 78% accuracy.



```
In [299]: #Confusion Matrix
          from sklearn.metrics import classification_report
          report = classification_report(y_test, predictions)
          print(report)
```

```
              precision    recall  f1-score   support

           0       0.78      0.76      0.77       176
           1       0.78      0.80      0.79       194

    accuracy                           0.78       370
   macro avg       0.78      0.78      0.78       370
weighted avg       0.78      0.78      0.78       370
```

Survival Analysis

The songs_normalize dataset does not have an event or time associated with it, so I was not able to use survival analysis.

REFLECTION & NEXT STEPS

I learned a lot about my data as I used techniques to explore, analyze, and understand the data. I was able to clean the data and think critically about whether each variable was useful to the business problem. PCA, decision tree and random forest classification were the most interesting techniques for analyzing this business problem.

Next steps should include building a model to automatically generate playlists for each class using appropriate genre and tempo classifications for the class segments.