Yaara Kaplan
CSCI 1420: Capstone
May 15, 2025

Spotify classification project

Abstract

Modern social media platforms often have algorithms that recommend new content that a user might like, based on previous data. In this project, we apply machine learning algorithms to music data to predict whether or not a song will be “liked”. This is an interesting task because it uses data that is somewhat subjective and attempts to make predictions based on “energy” or “vibes” – categories that are inherently human and not binary. We use machine learning models including K-means clustering, Logistic Regression, and K-Nearest Neighbors to learn the data. Due to technical challenges the results of this project are not as strong as we hoped for, but do show a correlation between certain features of the data and the quality of being “liked”.

Yaara Kaplan
CSCI 1420: Capstone
May 15, 2025

Objective

The initial goal for this project was to build a playlist recommendation application that would output a list of suggested songs given a playlist. Due to technical challenges with accessing data from the Spotify API (as will be addressed later), I chose to simplify the project to a classification task that outputs the prediction of whether or not a user will like a song, based on data describing song attributes and whether or not it is liked by the user.

Data

The data used for this project comes from Spotify's song data, and includes the following attributes[1]:

- *Acousticness*: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- *Danceability*: describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- *duration_ms*: The duration of the track in milliseconds.
- *Energy*: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- *Instrumentalness*: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- *Key*: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- *Liveness*: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- *Loudness*: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.

[1] Definitions are taken directly from Spotify's webpage which is cited below

- *Mode*: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- *Speechiness*: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- *Tempo*: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- *time_signature*: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".
- *Valence*: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

An additional feature added to the dataset is the “target” attribute, referring to whether the song is liked by the user of the original data. Other features that exist in Spotify’s data that are not numerical were not considered in the scope of this project.

Challenges

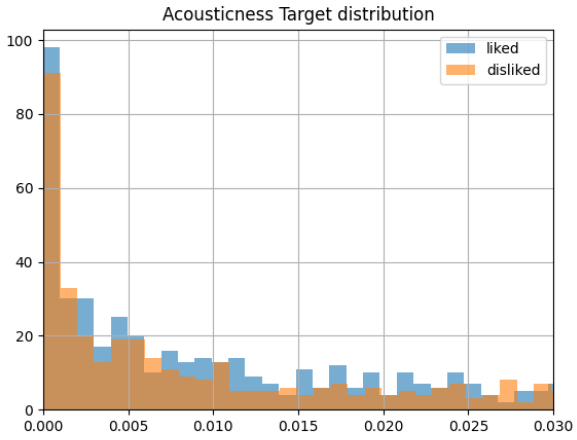
As mentioned in the introduction, the biggest challenge I faced in this project was data acquisition. My original intention was to scrape data from Spotify’s featured playlists and then access a user’s personal liked songs to recommend new songs from the given playlists. After working for some time and learning how to properly use the Spotify API, I discovered that in the past 6 months Spotify had removed the functionality of the endpoints to get featured playlists, so I had to find a new way to access the data. I instead found a dataset online, similar to the one I wanted to create, that was scraped before Spotify removed the endpoint. Due to the limited amount of data I had, which was less than I had hoped to scrape from the API, I chose to change my project to a binary prediction task.

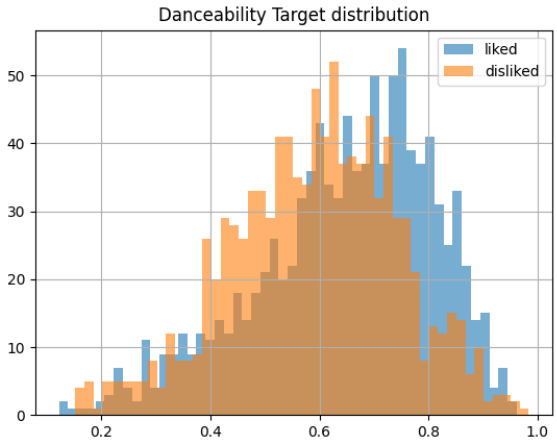
Yaara Kaplan
CSCI 1420: Capstone

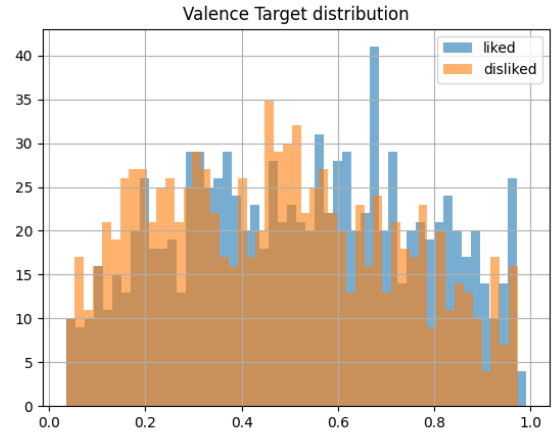
May 15, 2025

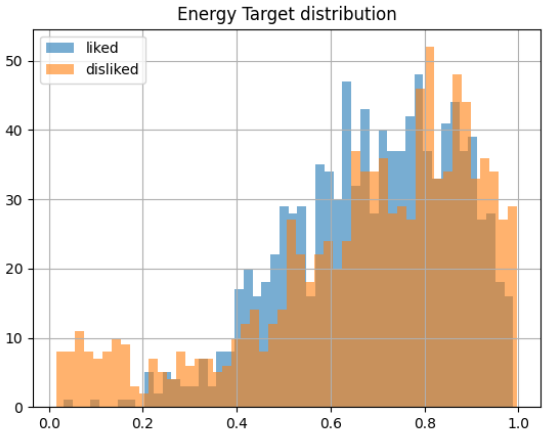
Exploratory Data Analysis (EDA)

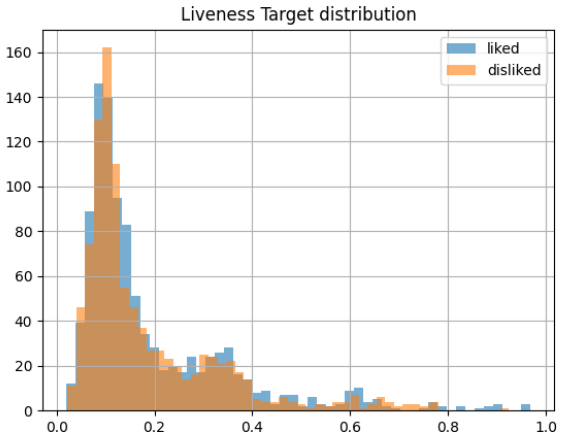
Before beginning the machine learning portion of the project, I wanted to get a sense of the data, so I performed some EDA to analyze which attributes of the data seemed most relevant to the task of predicting likeability. Some attributes were more helpful than others, as illustrated in the figures below:





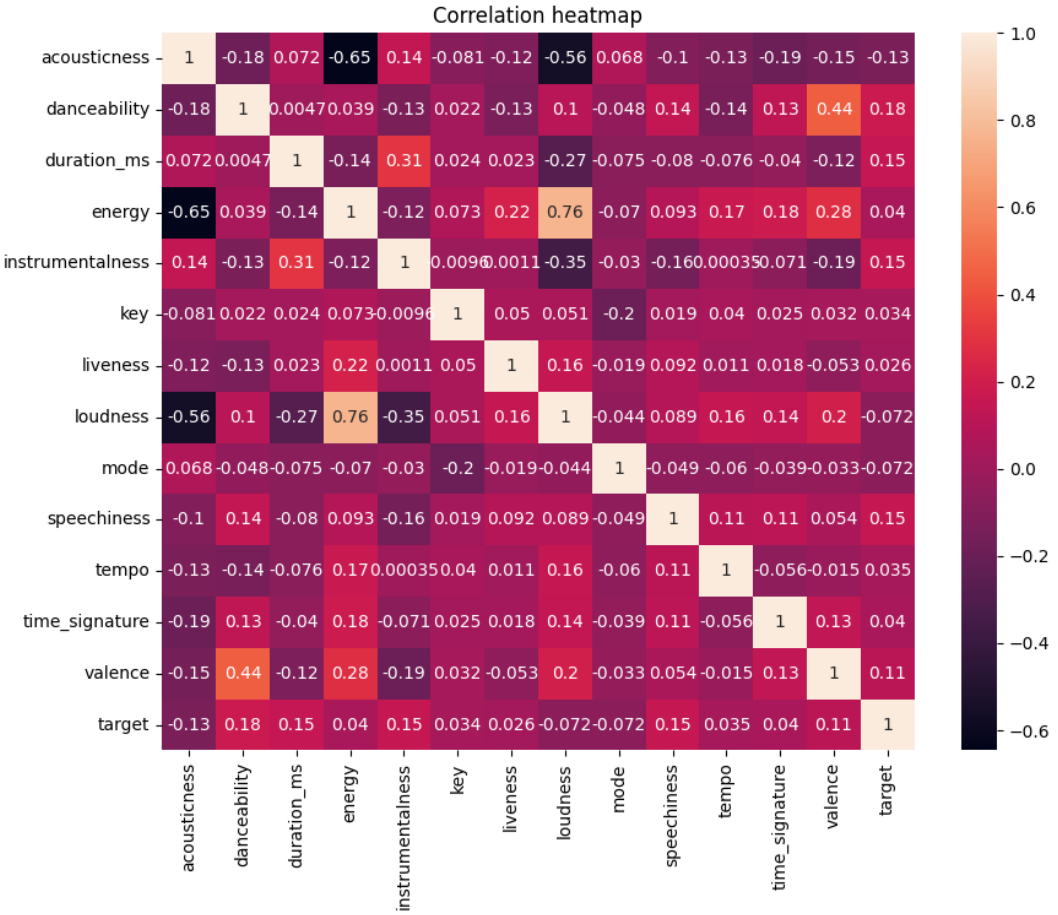






As we can see in the histograms, certain features, such as danceability, energy, and valence, have noticeable differences between the liked and disliked songs. Accousticness and liveness seem to have less strong of a correlation.

I then produced a correlation heatmap to understand which values were numerically the most important.



In the heatmap, we see that acousticness (previously determined to be not so relevant) actually has a score of -0.13, which is high enough to warrant being included in the dataset.

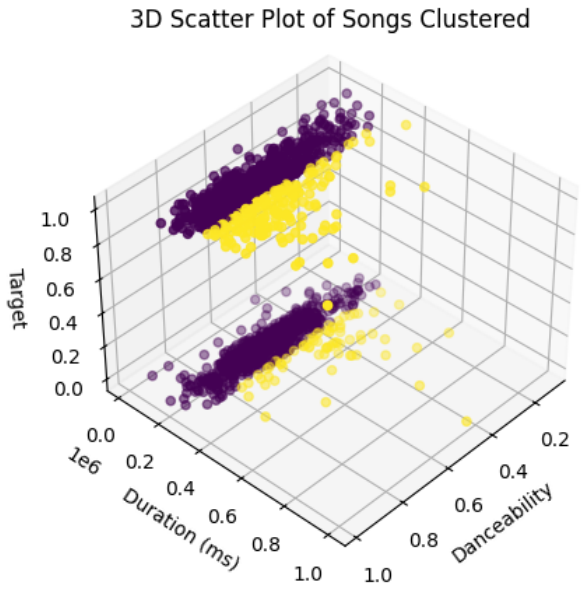
Based on the correlation heatmap and the histograms shown above, I determined that the most relevant attributes with regard to the target variable were acousticness, danceability, energy, instrumentalness, speechiness, and valence.

Yaara Kaplan
CSCI 1420: Capstone

May 15, 2025

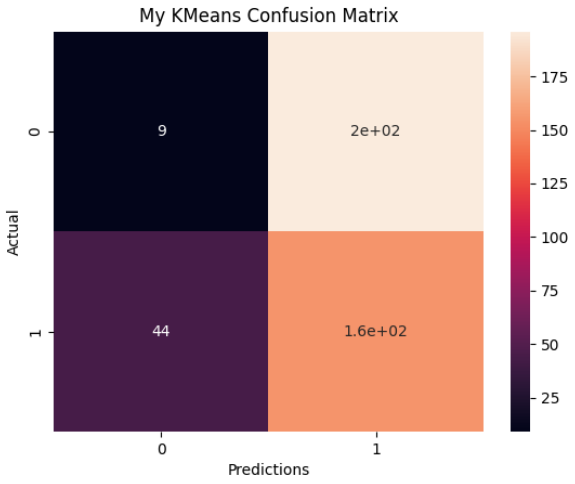
Results

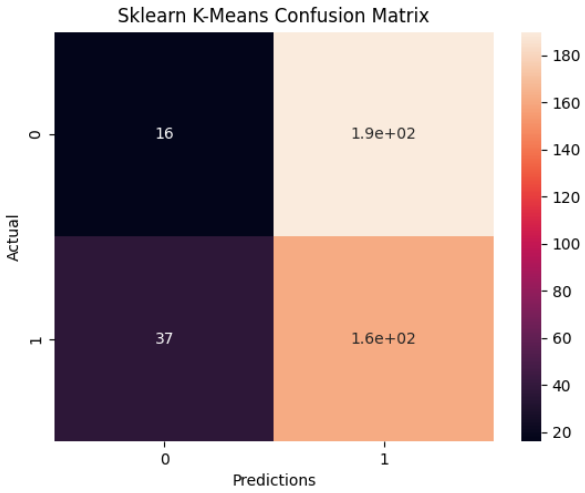
I used a K-means algorithm to cluster the data, choosing attributes that seemed most important given the heatmap produced in the EDA.



This scatterplot shows two distinct clusters that seem to be split based on the duration (normalized) of the song. Notably, this data also shows the distribution of the data between liked and non-liked songs; likely the larger number of liked songs impacted the success of the algorithms.

I then used the results of the K-means algorithm and compared them with those of the K-means algorithm that we wrote in class, and produced the following two confusion matrices:

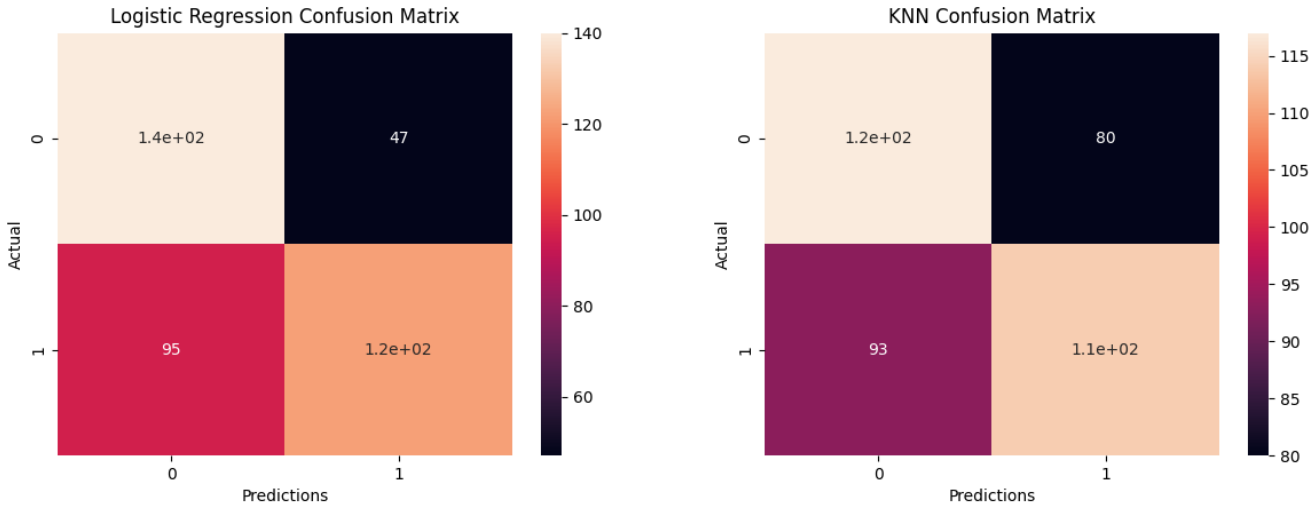




The results of the algorithms were very similar, but what I found most interesting about the results is how low the number of negative predictions there were overall. Because of this result, I understood that the data I was working with was less than ideal for this project, as it is clearly biased towards “liked” songs, which would influence the algorithms’ predictions.

Due to time constraints, however, I continued working with this data, with the knowledge that the results will be somewhat skewed.

Performing logistic regression resulted in a score of 0.65, which is low but not altogether surprising given the size and quality of the data. The K-Nearest Neighbors algorithm output a score of 0.57, with similar results to logistic regression. The confusion matrices for both of these machine learning algorithms can be seen below:



Conclusion and Next Steps

Even with limited data, the results of the project did show a correlation between the song data and the user's preference of the songs. Given this, I believe that with additional time to improve the data, I would be able to find much better results.

If I was able to continue working on this project, I would re-scrape the data, either from Spotify or from another service. This task is doable but would take much more time due to the lack of necessary API endpoints, which is why I was not able to do so during this project.

I also chose here to work with only numerical data, as it is mathematically simpler, however, attributes such as genre and artist can be very informative to predicting if a song is liked by the user – especially because "liking" a song in this context refers to the song being favorited within the Spotify application, rather than enjoying the sound itself. The dataset used in this project did not include genre data, but did include the artist as an attribute, which would be helpful in this context.

Yaara Kaplan
CSCI 1420: Capstone
May 15, 2025

Sources

"Get Track's Audio Features." Spotify for Developers, Spotify, https://developer.spotify.com/documentation/web-api/reference/get-audio-features. Accessed 15 May 2025

Geomack. Spotify Song Attributes. Kaggle, https://www.kaggle.com/datasets/sgeomack/spotifyclassification. Accessed 15 May 2025.