Predicting Popularity of Songs on Spotify

Jacob Knox and Noah Gabryluk

Our project will be utilizing the "Top Hits Spotify from 2000-2019" dataset from Kaggle (https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019) to create models to predict the popularity rating of songs on Spotify based on roughly 14 parameters to include the song's duration, release year, key, tempo, loudness, mode, danceability, energy, speechiness, acousticness, instrumentalness, liveness, and valence and whether or not the song is explicit. There is debate on whether or not to include the artist given that some artists only have one or two songs represented in the data set. There is also debate on whether or not to include the genre since a song can belong to multiple genres, which could possibly complice the models unnecessarily. Since the output is a discrete value representing the predicted popularity rating, performance cannot simply be measured with "right or wrong." As such, we will likely measure performance using the mean absolute error (MEA):

$$\frac{\sum_{i=0}^{k} \left| v_{Ai} - v_{Ei} \right|}{k}$$

where k is the number of samples, v_{Ai} is the ith actual value, and v_{Ei} is the ith estimated value.

Our project will likely involve training and testing, at the least, a decision tree model, neural network, and support vector machine with the data to determine which model best predicts song popularity.

This problem interested us, because we both enjoy music and it seems like a very relevant, modern application of machine learning. It may not be revolutionary or life saving work, but it's something we're interested in nonetheless. Additionally, this problem could potentially be useful for a musical engineer or an artist trying to learn about what kind of music turns out to be popular and tailor their music accordingly.