Project 2

COSC 311

Kalyn Howes & Chloe VanCory

12.15.21

**1 Data Summary**

The Spotify Top 200 Charts (2020-2021) dataset from Kaggle, found here, [Spotify Top 200](https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021), includes all of the songs that have been on the Top 200 Weekly (Global) charts of Spotify in 2020 & 2021. Each song has a number of features including: the highest charting position of the song, the number of times that the song has charted, the week when the song had the highest position, the name of the song, the song ID (provided by Spotify), the approximate number of streams the song has, the artist, the number of followers the artist has on Spotify, the genre, the release date, the weeks charted, popularity (between 0-100), danceability (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), acousticness, energy, instrumentalness, liveness (audience during recording), loudness, speechiness (presence of spoken words), tempo, duration, valence (musical positiveness), and finally the main chord of the song. This dataset was mostly gathered from spotifycharts.com, with some features being calculated based on numerous other features.

**2 Classes**

The most obvious “classes'' in this data are the different genres of each song. However, the songs could also be classified by artist, tempo, energy, main chord, etc. Making the genres the classes would allow for easy classification in the learning algorithms. For example, songs with high levels of danceability, energy, tempo, and valence could be classified as the pop genre.

In order to use the genres successfully we needed to simplify them. With some having very specific genres such as ‘Musical Advocacy’ and ‘Dream SMP’ for only one song each, there would be way too many classes. Also, most of the songs were categorized into 2-4 genres which would not work for the machine learning algorithms. Because of these, we developed umbrella genres that are more simple and inclusive.

**3 Machine Learning**

We implemented decision trees and the k nearest neighbors algorithm. The findings of these can be found in the Project2.ipynb notebook in mostly graphical form. These charts are also explained in the presenter notes of the slides.

**5. Feature Visualizations**

The visualizations/plots for the features not used in the machine learning algorithms are included in the Project2Visualizations.ipynb notebook and the best ones are included in the presentation slides.