

# The ABCs (and DEFUs) of Going Viral: Predicting Attributes of TikTok Viral Songs

**Author:** Lauren Choi

**Course:** CSCI 1951A, Data Science

**Teammates:** Nadya Tan, Trevor Ing, Teddy Kim

**Faculty Sponsor:** Lorenzo DeStefanis

## Abstract

Since TikTok exploded in popularity in 2019, going viral on the platform has offered unparalleled fame for countless singers, videographers, content creators, and influencers. For many artists, having a song go “TikTok viral” is a guaranteed catapult up the world’s top music charts, from Billboard Top 100 to top Spotify playlists. In this project, we predict attributes of songs that will go viral on TikTok by analyzing existing datasets of trending music, running statistical tests to determine significant qualities, and training machine learning models on the data to predict virality.

## Methods

Our project is broadly split into two parts: statistical analysis and machine learning. The statistical analysis portion cleans several Kaggle datasets of TikTok viral songs from 2020-2022, as well as Spotify Top 200 songs from the same timespan. After combining the datasets and classifying songs as “viral” (in the TikTok dataset) or “nonviral” (in the Spotify dataset but not the TikTok dataset), we performed two-tailed T-tests on the features we thought were most relevant to virality: danceability, energy, and valence (aka lyric mood). We found that high energy was not a statistically significant predictor of virality, but high danceability and positive lyrics were.

The machine learning portion creates models for KNN, Logistic Regression, Decision Tree, and Support Vector Machine. Originally, we used songs from 2020 and 2021 as our training dataset, and we used songs from 2022 as our testing dataset. However, because this yielded statistically different musical attributes (presumably because different years had different trends), we randomly shuffled our training and testing datasets to compare the accuracy values. We conducted k-fold cross-validation with sklearn’s GridSearchCV.

## Capstone

To fulfill the capstone requirement, I used the Flask framework to create a web interface where users can run two-tailed T-tests on various musical attributes (eg. liveness, acousticness, etc.) on our TikTok song datasets. The website allows users to select an attribute and view the resulting significance level, thus showing audiences which musical factors contribute to TikTok virality and which do not.