Analyzing the relationship between musical features and the popularity of a song, as measured by the number of Spotify streams or YouTube views using Machine Learning models

Reeyad Ahmed Ornate, Ehsanur Rahman Rhythm and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
reeyad.ahmed.ornate@g.bracu.ac.bd, ehsanur.rahman.rhythm@g.bracu.ac.bd,
annajiat@gmail.com

***Abstract*—In this research, we dive into the detailed interplay between musical features and the popularity of songs, gauged by Spotify streams or YouTube views. By utilizing a comprehensive dataset with diverse musical attributes, we employ statistical analysis and machine learning methodologies to unveil the impact of these features on a song's popularity which offers crucial insights for the music industry [6]. Motivated by the ubiquity of music in daily life and the subjective nature of personal taste, we aim to determine whether specific musical elements directly contribute to a song's popularity or if intangible factors elude current machine learning capabilities. The dataset presents challenges such as null values and unnecessary categorical columns that prompts meticulous pre-processing steps. Feature normalization and the creation of a binary output column, "Popularity," based on mean average values, further refine the dataset [2]. Feature scaling addresses varying scales, mitigating the risk of dominant features during model training. The dataset is split into training and testing sets, with K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression algorithms employed for model training [1]. Results reveal varying accuracies, with KNN and SVM reaching the highest at 73.12%. Despite falling short of anticipated levels, these findings offer valuable insights into the potential influence of musical features on song popularity, prompting future explorations with alternative machine learning algorithms and Convolutional Neural Networks (CNNs) to enhance predictive performance [3]. This research contributes to music analytics, providing a nuanced understanding of the dynamic relationship between musical composition and audience preferences.**

***Index Terms*—Music Analytics, Popularity Prediction, Machine Learning, Feature Analysis, KNN, Decision Tree, SVM, Logistic Regression, Dataset Preprocessing, Binary Classification.**

I. Introduction

The contemporary music landscape is marked by a dynamic interplay between artists, audiences, and evolving preferences. This research delves into the intricate relationship between specific musical features and song popularity, measured by Spotify streams or YouTube views. Utilizing a robust dataset with diverse musical attributes, advanced statistical analyses, and machine learning models, our aim is to unravel the nuanced dynamics that underlie a song's resonance with listeners [6]. Motivated by the ubiquity of music in daily life and the subjective nature of preferences, we seek to determine whether quantifiable musical elements directly correlate with a song's popularity or if intangible aspects elude current machine learning frameworks [1]. The dataset, a rich reservoir of musical features, presents a unique opportunity to decipher patterns that transcend genres and resonate across diverse listener demographics. Despite challenges like null values and superfluous categorical columns, meticulous pre-processing techniques are employed to handle missing data, remove unnecessary columns, and normalize features for robust model training [2]. A key aspect lies in creating a binary "Popularity" column, enabling a dichotomous classification of songs into popular and less popular categories [1]. Beyond academic curiosity, the implications of our findings extend to strategic decision-making in the music industry. Musicians can gain nuanced insights into audience preferences, record labels can optimize promotional strategies, and stakeholders can leverage data-driven approaches for music release success [2]. This research unravels the intricate relationship between musical features and song popularity, merging artistic expression with quantitative analysis. Subsequent sections will delve into the dataset, pre-processing, feature scaling, model training, and a comparative analysis of machine learning models, providing a comprehensive understanding of this intersection of music and data science.

Identify applicable funding agency here. If none, delete this.

II. Literature Review

In line with our exploration into how different musical features affect a song's popularity, the insights from the feature selection study seamlessly connect with our main analysis. While we're mainly looking at how various musical traits impact popularity, this particular research focuses on a specific

aspect: how we choose which features to consider. It suggests that the number of features not only affects accuracy but also how quickly our models can process information. This understanding adds a layer of insight into finding the right balance for our models to perform optimally in predicting music popularity[1].

Complementing our investigation into what makes songs popular, the findings from the study on music track popularity on streaming platforms fit well into our analysis of Spotify data. While we're broadly exploring different musical aspects, this study narrows down to factors influencing a song's popularity and how long it stays popular on streaming services. Integrating insights from this research refines our understanding of non-musical factors, like an artist's reputation and social information, giving us a more complete picture of what influences music popularity. The study's method, combining computer-based and statistical approaches, provides a practical guide for improving our methods in assessing music popularity[2].

In harmony with our focus on how musical features affect a song's popularity, this research introduces a new way of looking at popular music based on emotional content. While our primary interest is in features related to popularity, adding emotional recognition, especially within choruses, brings a more profound understanding of what makes songs click with people. The study suggests that considering emotions in our analysis could uncover deeper connections between how people feel about a song and its popularity. This approach might enrich our insights into what truly resonates with audiences[3].

As we dive into the world of music analytics, the thorough review on music emotion recognition integrates seamlessly with our main exploration of song popularity. While we're primarily focused on what makes songs popular, this review provides a broader context by delving into how music can evoke emotions. By understanding the challenges in associating emotions with music, the review enriches our grasp of the complex nature of people's responses to music. Integrating insights from this review helps us see our findings within the larger context of how people perceive and connect with music emotionally, addressing the challenges posed by the ever-changing and subjective nature of musical moods[4].

III. Dataset Analysis

The dataset constitutes the cornerstone of our investigation into the intricate relationship between musical features and song popularity [2]. Comprising 20,718 data points, each representing a distinct song, the dataset encapsulates a diverse spectrum of ten numeric musical attributes, including danceability, energy, key, loudness, acousticness, instrumentalness, liveness, valence and tempo. This rich compilation of musical features facilitates a comprehensive exploration across genres and styles. Integral to our analysis is the binary popularity label assigned to each song, serving as a pivotal variable for dichotomous classification into popular and less popular categories [1]. Despite its richness, the dataset presents challenges, notably in the form of null values within musical features, views, and streams. Additionally, categorical variables, although providing context, are omitted during machine learning model training for streamlined analysis [2]. The subsequent sections will delve into the nuanced pre-processing steps undertaken to address these challenges and refine the dataset for robust model training and analysis.

IV. Methodology

Initiating our approach, we collected a diverse dataset from Kaggle, focusing on essential musical features to explore the dynamics of song popularity. Through rigorous data preprocessing, we addressed null values, removed extraneous categorical variables, and applied feature normalization [2]. Subsequently, four robust machine learning algorithms—KNN, Decision Tree, SVM, and Logistic Regression—were adeptly trained on an 80% split of the dataset [1]. Evaluation metrics, including accuracy and error rates, meticulously gauged model performance on a dedicated test set. Looking ahead, our methodology envisions the exploration of ensemble models, aiming to elevate predictive accuracy and underscore the adaptability inherent in our analytical framework.

A. Dataset Selection

B. Data Preprocessing

C. Model Training

D. Evaluation and Prediction

E. Ensemble Model

References

[1] F. Khan et al., "Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms," Electronics, vol. 11, no. 21, p. 3518, Oct. 2022, doi: https://doi.org/10.3390/electronics11213518.

[2] T. Mulla, "Assessing the factors influencing the adoption of over-the-top streaming platforms: A literature review from 2007 to 2021," Telematics and Informatics, vol. 69, no. 101797, p. 101797, Apr. 2022, doi: https://doi.org/10.1016/j.tele.2022.101797.

[3] C.-H. Yeh et al., "Popular music representation: chorus detection & emotion recognition," Multimedia Tools and Applications, vol. 73, no. 3, pp. 2103–2128, Sep. 2013, doi: https://doi.org/10.1007/s11042-013-1687-2.

[4] Y. Kim et al., "MUSIC EMOTION RECOGNITION: A STATE OF THE ART REVIEW," 2010. Available: https://archives.ismir.net/ismir2010/paper/000045.pdf