Analyzing the relationship between musical features and the popularity of a song using Machine Learning models

Reeyad Ahmed Ornate, Farah Binta Haque, Ehsanur Rahman Rhythm, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
kha-208, 1 Bir Uttam Rafiqul Islam Ave, Dhaka 1212, Bangladesh
reeyad.ahmed.ornate@g.bracu.ac.bd, farah.binta.haque@g.bracu.ac.bd, ehsanur.rahman.rhythm@g.bracu.ac.bd, and 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. 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 prompt meticulous pre-processing steps. Feature normalization and the creation of a binary output column, "Popularity," based on mean average values, further refine the dataset. 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. Results reveal varying accuracy, 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. 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, preprocessing, feature scaling, model training, and a comparative analysis of machine learning models, providing a comprehensive understanding of this intersection of music and data science.

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].

One of the papers addresses the growing need for efficient and diversified music retrieval systems in the context of evolving music styles and changes in public aesthetics. It critiques traditional music classification systems for their reliance on perfect music samples at the training stage, which limits their adaptability to new music samples. The paper introduces the two-stage process of audio music classification, involving feature extraction and classification. Various musical features, such as short-time energy, zero-crossing rate, bandwidth, spectral centroid, and MFCC coefficients, are discussed. The task of music style classification is presented as a means to efficiently manage music databases and aid users in finding relevant music styles. The paper then focuses on the application of the Support Vector Machine (SVM) algorithm for music style classification, presenting experimental results on UCI standard datasets. It highlights the impact of dataset size on SVM's generalization ability. It discusses the limitations of traditional text-based multimedia retrieval methods and emphasizes the need for content-based music information retrieval technologies [9].

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. Subsequently, four robust machine learning algorithms—KNN, Decision Tree, SVM, and Logistic Regression—were adeptly trained on an 80% split of the dataset. 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*

The dataset employed in this study was meticulously selected from Kaggle, a renowned platform for datasets. This particular dataset, rich in diverse musical features and popularity labels, aligns with our research objective of understanding the relationship between these features and a song's popularity. The inclusion of ten numeric musical attributes provides a nuanced exploration of various facets of music composition.

B. *Data Preprocessing*

Robust data preprocessing is imperative to ensure the integrity and efficacy of subsequent analyses. The dataset underwent careful scrutiny to address challenges such as null values within musical features, views, and streams. Rows containing null values were judiciously removed, while unnecessary categorical variables, including 'Url_youtube,' 'Description,' 'Uri,' 'Track,' 'Channel,' 'Title,' 'Album_type,' 'Url_spotify,'

'Artist,' and 'Album,' were excluded to streamline the dataset for machine learning model training.

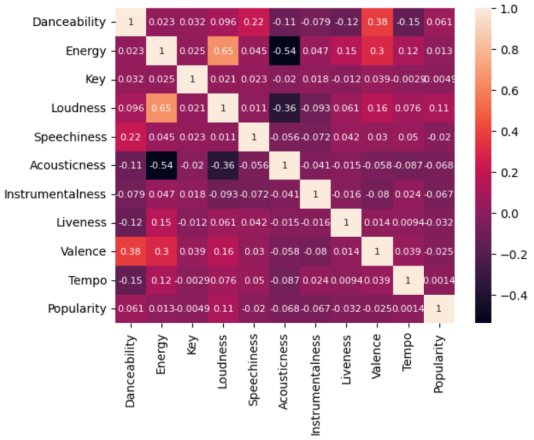

Fig. 1. Feature Relation

In Fig-1, we can notice the relation between all the feature through a heatmap. Feature normalization using the interquartile range (IQR) was applied to eliminate outliers, and a binary output column, "Popularity," was created based on mean average values.

C. *Model Training*

To unravel the complex relationship between musical features and song popularity, four machine learning algorithms were employed: K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. These models were chosen for their suitability in binary classification tasks and their potential to capture intricate patterns within the dataset. The training process involved utilizing 80% of the dataset, with careful consideration given to feature scaling to mitigate the impact of varying feature scales on model performance.

D. *Evaluation and Prediction*

The trained models were subjected to rigorous evaluation using a dedicated test set comprising 20% of the dataset. Model accuracy, error rates, and comparative analyses were conducted to assess the efficacy of each algorithm in predicting song popularity. Predictions were made on this independent test set, allowing for a comprehensive understanding of each model's performance in real-world scenarios.

E. *Ensemble Model*

Recognizing the potential for enhanced predictive performance through model ensemble techniques, future work may explore the integration of multiple models into an ensemble framework. Ensemble models, combining the strengths of individual algorithms, have demonstrated efficacy in achieving higher predictive accuracy. This avenue could serve as a promising extension, further refining the predictive capabilities of our analysis and potentially overcoming limitations observed in individual model performances.

V. Result Analysis

The culmination of our methodological approach involved training four distinct machine learning models—K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression—to unravel the intricate relationship between musical features and song popularity. Each model was rigorously evaluated, and the results provide valuable insights into their individual performances.

A. *K-Nearest Neighbor (KNN)*

The KNN algorithm, which identifies the K-nearest neighbors in the training set to classify new data points, exhibited an accuracy of 73.12%. The model's strength lies in its ability to capture localized patterns within the dataset, leading to accurate predictions. However, it is essential to note that KNN is sensitive to outliers and the choice of the number of neighbors (K), warranting careful consideration in model parameterization.

B. *Decision Tree*

The Decision Tree model, which recursively splits the dataset based on the most informative features, achieved an accuracy of 67.12%. While Decision Trees are adept at capturing complex relationships within data, the observed accuracy suggests potential limitations in their ability to generalize patterns. Further tuning of hyperparameters may be explored to enhance predictive performance.

C. *Support Vector Machine (SVM)*

Employing a hyperplane to maximize the margin between classes in high-dimensional space, SVM achieved an accuracy of 73.12%. SVM is particularly effective in capturing non-linear relationships, showcasing its robust performance in our analysis. The observed accuracy aligns with KNN, indicating a comparable ability to discern patterns in the dataset.

D. *Logistic Regression*

As a binary classification algorithm measuring the relationship between independent variables and the dependent variable, Logistic Regression exhibited an accuracy of 73.09%. Logistic Regression is known for its simplicity and interpretability, making it a valuable tool in binary classification tasks. The observed accuracy places it in line with the performance of KNN and SVM.

E. *Comparative Analysis*

A comparative analysis of the models reveals that KNN and SVM achieved the highest accuracy, both recording 73.12%. Logistic Regression closely follows with an accuracy of 73.09%, while Decision Tree lags slightly behind at 67.12%. The error rates for each model—KNN (26.88%), Decision Tree (32.88%), SVM (26.88%), and Logistic Regression (26.91%)—reflect the models' ability to make accurate predictions, but also highlight potential areas for improvement.

In Figure-2, all the models are displayed based on the accuracy comparison of their performance.

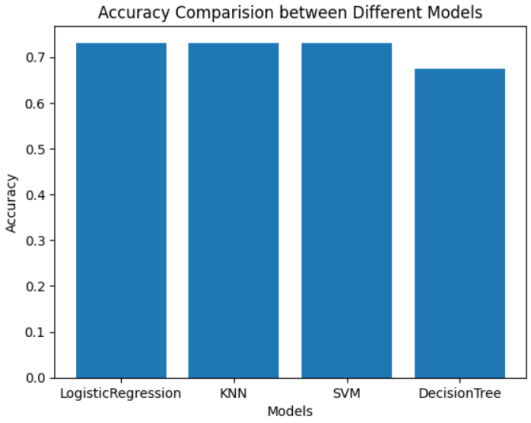

Fig. 2. Model accuracy Comparison Table

To visually represent the comparative performance of the models, a bar chart was constructed. The chart clearly illustrates the accuracy scores of each model, with KNN and SVM emerging as the top performers, followed by Logistic Regression and Decision Tree. This visual aid enhances the accessibility of the results, facilitating a quick understanding of the relative strengths of each model. In conclusion, the results of our analysis provide valuable insights into the predictive capabilities of machine learning models in discerning the relationship between musical features and song popularity. While all models demonstrated commendable accuracy, the observed variations warrant further investigation into hyperparameter tuning and potential ensemble modeling for improved predictive performance.

VI. Future Work

The current analysis, while providing valuable insights into the interplay between musical features and song popularity, opens avenues for future research and enhancements [3]. Several aspects warrant exploration to further refine our understanding and improve predictive accuracy.

A. *Hyperparameter Tuning*

Future work could delve into a more granular exploration of hyperparameters for each machine learning algorithm [3]. Fine-tuning parameters such as the number of neighbors in KNN, the depth of Decision Trees, and the regularization parameter in SVM and Logistic Regression might lead to improved model performance. This meticulous parameter optimization may uncover latent patterns within the data that were not fully captured in the initial analysis.

B. *Ensemble Modeling*

The exploration of ensemble modeling techniques presents a promising direction for future work [2]. Combining the strengths of multiple models through methods like bagging or boosting could potentially enhance overall predictive accuracy. Ensemble models are known for mitigating overfitting and improving generalization, providing a robust framework to harness the strengths of diverse algorithms.

C. *Feature Engineering*

Further refinement of feature engineering techniques may contribute to enhanced model performance [6]. Experimentation with different combinations of musical features, creating new composite features, or exploring non-linear transformations could reveal additional dimensions influencing song popularity. Feature selection methods can be employed to identify the most impactful variables for predictive modeling [1].

D. *Deep Learning Approaches*

The application of deep learning techniques, such as Convolutional Neural Networks (CNNs) or recurrent neural networks (RNNs), stands as an intriguing avenue for future research [3]. These architectures have demonstrated success in capturing intricate patterns in various domains, and their application to music analysis could uncover more complex relationships within the dataset.

E. *Larger and Diverse Datasets*

Expanding the dataset size and diversity could provide a more comprehensive understanding of the factors influencing song popularity [6]. Incorporating data from various genres, cultures, or time periods may unveil patterns that are not discernible in the current dataset. A larger and more diverse dataset would contribute to a more robust and generalizable model.

F. *Cross-Validation Techniques*

Implementing advanced cross-validation techniques, such as k-fold cross-validation, would provide a more rigorous assessment of model performance [6]. This approach involves partitioning the dataset into multiple folds for training and testing, ensuring that each data point is part of the test set at least once. This can offer a more reliable estimate of the model's performance and generalization capabilities.

The proposed enhancements aim to refine model accuracy, uncover latent patterns, and contribute to a more nuanced understanding of the intricate relationship between musical features and song popularity. As the field of music analytic evolves, these avenues for future research hold the potential to further advance our insights and applications in this dynamic intersection of music and data science.

VII. Conclusion

In the pursuit of unraveling the complex relationship between musical features and song popularity, this study has presented a comprehensive analysis employing machine learning models. The exploration of a diverse dataset, encompassing ten numeric musical attributes, offered valuable insights into the predictive capabilities of K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression algorithms [1].

The results showcased commendable accuracy with KNN and SVM leading at 73.12%, closely followed by Logistic Regression at 73.09%. Decision Tree, while slightly behind

at 67.12%, contributed valuable perspectives on the dataset's complexity. The comparative analysis, supported by a visually informative bar chart, highlighted the strengths and areas for improvement in each model.

Future work has been outlined, emphasizing the importance of hyperparameter tuning, ensemble modeling, deeper feature engineering, exploration of deep learning approaches, larger dataset sizes, and advanced cross-validation techniques [3] [6]. These avenues aim to refine predictive accuracy, uncover latent patterns, and contribute to a nuanced understanding of the intricate dynamics between musical composition and audience preferences.

While the current analysis has shed light on the interplay of musical features and song popularity, the exploration remains dynamic and open to continuous refinement. The intersection of music and data science presents an evolving landscape, and further research endeavors promise to deepen our insights and applications within the realm of music analytic.

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

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, 2001, doi: https://doi.org/10.1109/cvpr.2001.990517.

[6] C. V. Soares Araujo, M. A. Pinheiro de Cristo, and R. Giusti, "A Model for Predicting Music Popularity on Streaming Platforms," Revista de Informática Teórica e Aplicada, vol. 27, no. 4, pp. 108–117, Dec. 2020, doi: https://doi.org/10.22456/2175-2745.107021.

[7] Y. K. Yee and M. Raheem, "Predicting Music Popularity Using Spotify and YouTube Features," Indian Journal Of Science And Technology, vol. 15, no. 36, pp. 1786–1799, Sep. 2022, doi: https://doi.org/10.17485/ijst/v15i36.2332.

[8] J. Lee and J.-S. Lee, "Music Popularity: Metrics, Characteristics, and Audio-Based Prediction," IEEE Transactions on Multimedia, vol. 20, no. 11, pp. 3173–3182, Nov. 2018, doi: https://doi.org/10.1109/tmm.2018.2820903.

[9] C. Wang and X. Geng, "Simulation of music style classification model based on support vector machine algorithm," in Proceedings of the Sixth International Conference on Intelligent Computing, Communication, and Devices (ICCD 2023), June 16, 2023, article number 127030Z, doi: 10.1117/12.2682982.