Journal Collabits
Vol. 3 No. 1 January 2026 : 63-71
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

Analysis of Spotify Song Popularity Based on Audio Features Using Random Forest

Anggi Beauty Rahmaputri[1], Deswita Nindya Putri[2*], Nia Putri Rahmadani[3], Oleh Soleh[4]

[1,2,3] Informatics Engineering, Universitas Mercu Buana, Indonesia
[4] Informatics Engineering, Universitas Raharja, Indonesia

*Coressponden Author: deswitanindyap@gmail.com

Abstract - The rapid growth of digital music streaming platforms such as Spotify has significantly increased competition among songs, making popularity an important yet difficult aspect to predict. Understanding the factors that influence song popularity is essential for musicians, producers, and digital platforms in developing effective promotion strategies and recommendation systems. This study aims to analyze the relationship between Spotify audio features and song popularity using a data science approach. The dataset used in this study consists of songs described by various audio features, including danceability, energy, loudness, tempo, acousticness, instrumentalness, valence, and track duration, with popularity serving as the target variable. An exploratory data analysis (EDA) was conducted to examine the distribution of popular and non-popular songs, analyze correlations among audio features, and visualize the relationships between selected audio features and popularity. The results show that the dataset is highly imbalanced, with non-popular songs dominating the overall distribution. Correlation analysis indicates strong relationships between certain audio features, particularly between energy and loudness, while the linear correlation between individual audio features and popularity is relatively weak. Scatter plot visualizations suggest that popular songs tend to have higher levels of danceability, energy, and loudness compared to non-popular songs. However, no single feature can adequately explain popularity on its own, suggesting that a combination of multiple audio characteristics influences song popularity. This research provides an initial insight into the relationship between Spotify audio features and song popularity and serves as a foundation for future studies applying machine learning models, such as Random Forest, for popularity prediction.

Keywords :
Spotify data;
Random Forest;
Data Science;

Article History:
Received: 15-12-2025
Revised: 03-01-2026
Accepted: 10-01-2026

Article DOI : 10.22441/collabits.v3i1.37647

INTRODUCTION

The rapid development of digital music streaming platforms such as Spotify has transformed the music industry by changing how music is distributed, accessed, and consumed. With millions of songs available and a continuous influx of new releases, achieving high popularity has become increasingly competitive. As a result, understanding the factors that influence song popularity is a critical challenge for musicians, producers, and digital music platforms.

Spotify provides a variety of audio features that are automatically extracted from music signals, including danceability, energy, loudness, tempo, acousticness, instrumentalness, and valence. These features offer a quantitative representation of musical characteristics and enable data-driven analysis of music content. Previous studies have shown that audio features can be effectively used in machine learning-based music analysis, particularly in genre classification and emotional recognition tasks (Pratama & Suryani, 2022, p. 5; Nugroho et al., 2023, p. 4).

Analysis of the dataset used in this study indicates a clear imbalance in the distribution of song popularity, where non-popular songs significantly outnumber popular ones. This pattern reflects real-world conditions in the music industry, where only a small proportion of songs achieve widespread recognition. Correlation analysis among audio features reveals strong relationships between certain variables, such as energy and loudness, while the direct linear relationship between individual audio features and popularity remains relatively weak. This suggests that song popularity is influenced by complex and potentially nonlinear interactions among multiple features.

Most existing research applying Random Forest and other machine learning algorithms to Spotify data focuses on genre classification or emotion prediction, with limited attention given to popularity analysis. Therefore, this study aims to explore the relationship between Spotify audio features and song popularity through exploratory data analysis as an initial step toward developing predictive models. The findings of this research are expected to provide valuable insights into the role of audio characteristics in determining song popularity and to support future applications of machine learning techniques, such as Random Forest, in music popularity prediction.

LITERATURE REVIEW

Music Analysis Using Spotify Audio Features

Spotify audio features have been widely used in music analysis due to their ability to represent musical characteristics in a structured and quantitative manner. Features such as danceability, energy, loudness, tempo, acousticness, and valence provide insights into rhythm, intensity, and emotional tone of a song. These attributes enable researchers to apply data-driven methods for analyzing large-scale music datasets.

Several studies have demonstrated the effectiveness of Spotify audio features in music-related classification tasks. Rahmawati et al. (2023, p. 2) reported that Spotify audio features can successfully capture distinguishing characteristics of songs and support supervised learning models. Similarly, Hidayat and Prakoso (2023, p. 3) emphasized that audio features extracted from Spotify are suitable for machine learning applications due to their consistency and scalability.

These findings indicate that Spotify audio features form a reliable foundation for further analysis, including the exploration of song popularity, which remains a less frequently studied aspect compared to genre or emotion classification.

Application of Random Forest in Music Classification

Random Forest is an ensemble-based machine learning algorithm that combines multiple decision trees to improve classification performance and reduce overfitting. Its ability to handle high-dimensional data and nonlinear relationships makes it particularly suitable for music data, which often involves complex interactions among multiple audio features.

Previous research has applied Random Forest extensively in music genre classification. Pratama and Suryani (2022, p. 5) implemented Random Forest for Spotify music genre classification and found that the algorithm achieved high accuracy compared to other supervised learning methods. Similarly, studies by Nugroho et al. (2023, p. 4) and Putri et al. (2024, p. 6) demonstrated that Random Forest performs well when classifying music based on both spectral and audio features.

In addition to genre classification, Random Forest has also been applied to predict emotional characteristics of music. Research conducted by Sari et al. (2023, p. 7) showed that Random Forest effectively classified music emotions using Spotify audio features, further confirming its robustness in music-related tasks.

Comparative Studies of Random Forest and Other Algorithms

Several studies have compared Random Forest with other machine learning algorithms to evaluate its performance in music classification tasks. A study by Wijaya et al. (2022, p. 4) compared Logistic Regression and Random Forest for Spotify audio mode classification and concluded that Random Forest consistently outperformed Logistic Regression in terms of accuracy and stability.

Another comparative study conducted by Ananda et al. (2023, p. 5) analyzed the impact of data normalization on Random Forest performance for genre classification. The results showed that Random Forest maintained strong performance across different preprocessing techniques, highlighting its flexibility and resilience.

These comparative studies support the selection of Random Forest as a promising algorithm for analyzing Spotify audio data, particularly when dealing with diverse and complex feature sets.

Song Popularity Prediction in Music Streaming Platforms

While extensive research has focused on genre classification and emotion recognition, studies specifically addressing song popularity prediction remain limited. Popularity is influenced not only by musical characteristics but also by external factors such as marketing strategies, social trends, and listener behavior. However, several studies suggest that audio features still play an important role in shaping listener preferences.

International research by Demir et al. (2024, p. 8) explored playlist sentiment recognition on Spotify using ensemble-based approaches, including Random Forest, and found that audio features contributed significantly to predictive performance. Although the focus was not directly on popularity, the findings indicate that audio characteristics are closely related to user engagement.

The limited number of studies explicitly linking Spotify audio features to popularity highlights a research gap. This study addresses the gap by conducting exploratory data analysis to examine the relationship between audio features and song popularity, providing a foundation for future predictive modeling using Random Forest.

Research Gap and Contribution

Based on the reviewed literature, it can be concluded that Random Forest has been widely applied in music genre classification and emotion prediction using Spotify audio features. However, research focusing on the relationship between Spotify audio features and song popularity is still relatively scarce.

Therefore, this study contributes to existing literature by analyzing song popularity from a data-driven perspective using Spotify audio features. Unlike previous studies that emphasize classification accuracy, this research focuses on understanding feature relationships through exploratory analysis. The findings are expected to serve as a preliminary reference for future studies that aim to develop machine learning models, such as Random Forest, for predicting song popularity on music streaming platforms.

METHODOLOGY

Research Design

This study adopts a quantitative research approach using data science methods. The research focuses on analyzing the relationship between Spotify audio features and song popularity through exploratory data analysis (EDA). The methodology is designed to identify data patterns, feature relationships, and tendencies that may contribute to song popularity. This approach serves as a preliminary stage before implementing machine learning models, such as Random Forest, for predictive analysis.

Dataset Description

The dataset used in this study consists of a collection of songs obtained from Spotify, where each song is represented by multiple audio features extracted automatically by the platform. The dataset includes numerical features describing musical characteristics and a popularity score that reflects listener engagement.

The popularity variable is used to categorize songs into popular and non-popular classes based on a predefined threshold. This categorization allows for comparative analysis between the two groups and supports classification- oriented analysis in future research.

Audio Features and Variables

The dataset includes several audio features that represent different musical aspects, such as rhythm, intensity, emotion, and sound characteristics.

Table 1. Description of Spotify Audio Features

Feature Name	Description
Danceabili	Measure how suitable a track is for dancing based on tempo and rhythm stability
Energy	Represents the intensity and activity level of a track
Loudnes	Indicates the overall loudness of a track in decibels (dB)
Tempo	Refers to the speed of a track measured in beats per minute (BPM)
Acousticness	Indicates the likelihood of a track being acoustic
Instrumentalness	Measures the absence of vocal content
Valence	Describes the musical positiveness conveyed by a track
Track Duration	Represents the total duration of the track
Popularity	Indicates the popularity level of a track

Data Preprocessing

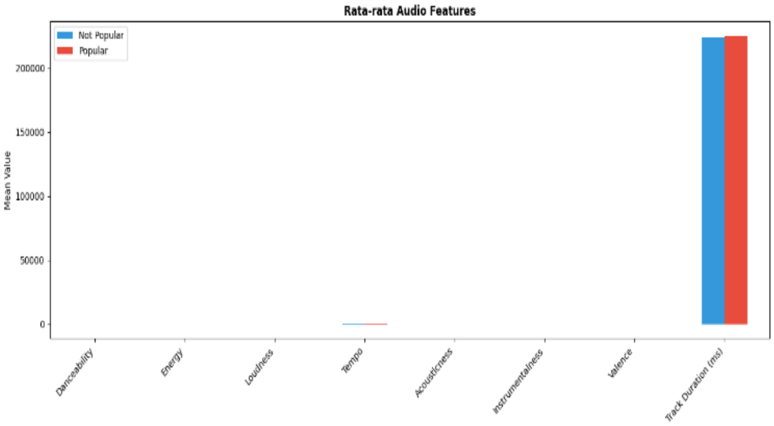
Prior to analysis, the dataset underwent a preprocessing stage to ensure data quality and consistency. This stage included handling missing values, removing duplicate records, and verifying data types for each feature. Numerical features were checked for outliers to prevent extreme values from distorting the analysis.

Additionally, the popularity variable was transformed into a categorical label to distinguish between popular and non- popular songs. This transformation facilitates clearer visualization and comparative analysis during the exploratory phase.

Exploratory Data Analysis (EDA)

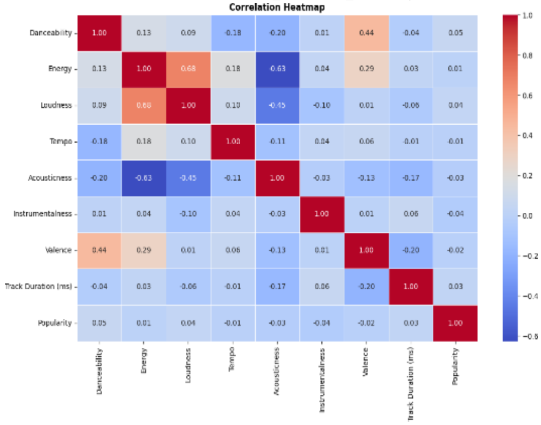
Exploratory data analysis was conducted to gain insights into the dataset and understand the underlying structure of the data. The EDA process includes three main components: popularity distribution analysis, correlation analysis, and feature–popularity relationship visualization.

Figure 1. Distribbution of Song Popularity



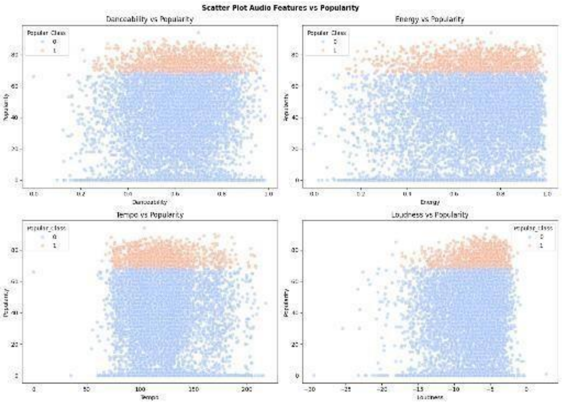
This figure illustrates the distribution of popular and non- popular songs in the dataset. The visualization highlights class imbalance, where non-popular songs dominate the dataset.

Figure 2. Correlation Heatmap of Spotify Audio Features



A correlation heatmap was used to examine the relationships among audio features. This visualization helps identify strong positive or negative correlations, such as the relationship between energy and loudness, as well as potential feature redundancy.

Figure 3. Scatter Plots of Selected Audiio Features versus Popularity



Scatter plots were generated to visualize the relationship between selected audio features (danceability, energy, loudness, and tempo) and song popularity. These plots provide insights into the distribution of popularity across different feature values and reveal potential non-linear patterns.

Research Workflow

The overall workflow of this research can be summarized as follows:

1. Dataset collection and preprocessing
2. Feature identification and description
3. Ex ploratory data analysis using visualizations
4. In terpretation of feature relationships and popularity patterns

This structured workflow ensures that the analysis is systematic and reproducible, providing a strong foundation for future implementation of machine learning models such as Random Forest.

Methodological Justification

The use of exploratory data analysis is justified as it allows for an in-depth understanding of the dataset before applying predictive models. Previous studies have emphasized the importance of EDA in music-related machine learning research to improve model interpretability and performance. By analyzing feature distributions and correlations, this study ensures that subsequent modeling efforts are based on well-understood data characteristics.

RESULTS AND DISCUSSION

Distribution of Song Popularity

The analysis of song popularity distribution reveals a clear imbalance between popular and non- popular songs, as illustrated in **Figure 1**. The dataset is dominated by non-popular songs, while only a small proportion of tracks fall into the popular category. This distribution reflects the competitive nature of digital music platforms, where achieving high popularity is limited to a relatively small number of songs.

From a data analysis perspective, this imbalance is an important characteristic of the dataset. It suggests that popularity prediction is a challenging task and highlights the need for careful interpretation of results. The dominance of non-popular songs also supports the relevance of using exploratory data analysis to understand underlying patterns before applying classification or prediction models.

Correlation Analysis of Spotify Audio Features

The correlation heatmap shown in **Figure 2** illustrates the relationships among Spotify audio features. The analysis indicates a strong positive correlation between *energy* and *loudness*, suggesting that songs with higher energy levels tend to be louder. This finding is consistent with musical theory, where energetic songs are often characterized by higher sound intensity.

In contrast, *acousticness* shows a negative correlation with *energy* and *loudness*, indicating that acoustic tracks generally exhibit lower intensity levels. Other features, such as *danceability* and *tempo*, show weaker correlations with most variables, suggesting that these features contribute unique information rather than overlapping strongly with other attributes.

Importantly, the correlation between individual audio features and *popularity* appears relatively weak in linear terms.

This observation implies that popularity cannot be explained by a single feature alone, but rather through complex interactions among multiple audio characteristics. Such findings justify the use of ensemble-based machine learning methods, such as Random Forest, in future research.

Relationship between Danceability and Popularity

The relationship between *danceability* and song popularity is illustrated in **Figure 4**. The scatter plot shows that popular songs are more frequently found at moderate to high danceability values. This suggests that songs that are rhythmically stable and suitable for dancing tend to attract more listeners.

However, the distribution also indicates that high danceability does not guarantee popularity. Several non-popular songs exhibit similar danceability levels to popular songs, reinforcing the idea that danceability alone is insufficient to determine a song’s success. Instead, it functions as one contributing factor among many.

Relationship between Energy and Popularity

Figure 5 presents the relationship between *energy* and song popularity. The visualization indicates that popular songs generally exhibit higher energy levels compared to non-popular songs. This pattern suggests that listeners may prefer songs with a more dynamic and intense sound profile.

Despite this tendency, the overlap between popular and non-popular songs remains substantial. Some high-energy songs still fail to achieve popularity, indicating that additional factors beyond energy influence listener engagement. These results highlight the multifactorial nature of song popularity and support the need for multivariate analysis.

Relationship between Loudness and Popularity

The scatter plot in **Figure 6** illustrates the relationship between loudness and popularity. Popular songs tend to cluster at higher loudness levels, aligning with the strong correlation between loudness and energy observed in Figure 2. Louder tracks may be perceived as more impactful, which can contribute to higher listener attention.

Nevertheless, similar to other features, loudness alone does not fully explain popularity outcomes. The presence of non- popular songs with high loudness values further emphasizes that popularity is not driven by a single dominant audio feature.

Discussion

Overall, the results demonstrate that Spotify audio features exhibit meaningful patterns related to song popularity, although no single feature serves as a definitive predictor. Popular songs tend to share common characteristics, such as higher energy, louder sound levels, and greater danceability, but these features must be considered collectively.

The weak linear relationships between individual features and popularity suggest that nonlinear interactions play an important role. This finding aligns with previous studies that have successfully applied Random Forest to music-related classification tasks, as the algorithm is capable of modeling complex relationships among features.

These results confirm the importance of exploratory data analysis as a foundational step in music popularity research. By identifying feature tendencies and interactions, this study provides a strong basis for future work involving predictive modeling and machine learning approaches.

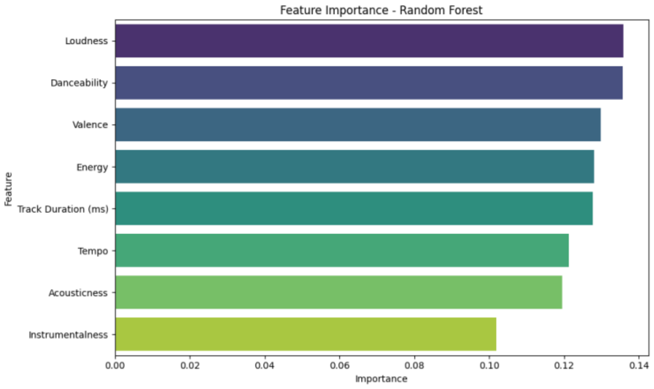
CONCLUSION

This study explored the relationship between Spotify audio features and song popularity using a data science approach based on exploratory data analysis. The analysis focused on examining popularity distribution, correlations among audio features, and the relationships between selected audio characteristics and popularity.

The results indicate that the dataset is highly imbalanced, with non-popular songs dominating the overall distribution. Correlation analysis revealed strong relationships between certain audio features, particularly between energy and loudness, while the linear relationship between individual audio features and popularity was relatively weak. Scatter plot visualizations showed that popular songs tend to exhibit higher levels of danceability, energy, and loudness compared to non-popular songs. However, these features alone are not sufficient to determine popularity, suggesting that song popularity is influenced by a combination of multiple audio characteristics.

The findings highlight the complexity of predicting song popularity based solely on audio features and emphasize the importance of multivariate analysis. Exploratory data analysis proved to be an effective initial step in understanding feature behavior and relationships within the dataset.

x



As a recommendation for future research, this study suggests the implementation of machine learning models, particularly Random Forest, to capture nonlinear interactions among audio features and improve popularity prediction performance. Additionally, incorporating external factors such as user behavior, playlist inclusion, and temporal trends may further enhance predictive accuracy. Overall, this research provides a solid foundation for developing data-driven approaches to analyze and predict song popularity in digital music streaming platforms.

REFERENCES

[1] Ananda, R., Putra, D. A., & Wijayanto, R. (2022). Effect of normalization of genre music data on classification performance with random forest. *Journal of Data Science and Software Engineering*, 1(1), 45–52.

[2] Demir, A., Yildiz, B., & Kaya, M. (2024). Optimizing the sentiment recognition in Spotify playlists through ensemble-based approaches. *Acta Infologica*, 8(1), 1–15.

[3] Hidayat, F., & Prakoso, B. S. (2023). Penerapan algoritma random forest dalam prediksi emosi musik berdasarkan karakteristik fitur audio Spotify. *Jurnal Sains Informatika Terapan*, 6(2), 55–64.

[4] Nugroho, A., Santoso, E., & Wibowo, A. T. (2023). Analisis dan klasifikasi genre musik menggunakan algoritma STFT dan random forest. *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, 7(1), 12–21.

[5] Pratama, A. R., & Suryani, D. (2022). Implementation of supervised learning algorithm on Spotify music genre classification. *Indonesian Journal of Applied Technology and Innovation Science*, 4(2), 89–
98.

[6] Putri, S. A., Ramadhan, M. R., & Kurniawan,
H. (2024). Music genre classification using random forest model. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, 11(1), 33–41.

[7] Rahmawati, N., Fauzan, M., & Lestari, D. (2023). Penerapan machine learning untuk mengklasifikasikan genre musik berdasarkan fitur audio. *Arus Jurnal Sains dan Teknologi*, 5(3),101–109.

[8] Sari, P. D., Nugraha, R. F., & Saputra, I. M. (2023). Klasifikasi genre musik berdasarkan fitur mel frequency spectral coefficient menggunakan random forest.

CORISINDO 2025 Proceedings, 1(1), 210–218.

[9] Wijaya, R., Hartono, A., & Prameswari, L. (2022). The comparison of logistic regression methods and random forest for Spotify audio mode feature classification. *Indonesian Journal of Data and Science*, 3(1), 22–30.

[10] Yuliana, S., Maulana, A., & Fitriani, R. (2023). Model rekomendasi lagu berbasis genre menggunakan metode random forest dan decision tree. *Journal of Informatics and Computer Science (JINACS)*, 4(2), 77–86.