Modeling Song Popularity: The Impact of Audio Features and Streaming Metadata in the Digital Music Landscape

By Petros Petrou

Abstract— **Various factors contribute to a song's popularity, including musical characteristics and listener engagement metrics. This study explores their impact using a dataset of over 100,000 songs, applying data preprocessing, feature selection, and normalization before training Random Forest Classification and Regression models. Results indicate that both audio attributes and streaming metrics play a role in popularity prediction, with their combined use improving model performance. The evaluation shows that classification performs better than regression, suggesting that song popularity is better represented as a categorical variable rather than a continuous one.**

I. INTRODUCTION

In the constantly evolving digital era, music consumption has shifted dramatically from physical formats and downloads to streaming platforms, where millions of track compete for listener attention and overall appreciation. This study uses data analytics methods to examine how musical attributes influence song popularity, with a structured methodology to learn more about the key features and develop appropriate predictive models.

Aims:

- Analyze the relationship between musical attributes (e.g., danceability, energy, loudness, key, genre) and song popularity.
- Preprocess and clean the dataset, addressing missing values, standardizing numerical features, and encoding categorical data.
- Perform a feature selection process to determine the most important features
- Develop two machine learning models
- Evaluate model performance using key metrics

The insights gained from this study will contribute to a deeper understanding of musical trends and how the constitution of a track affects them along with additional parameters.

II. LITERATURE REVIEW

Understanding the factors that influence song popularity is an essential area of research within music analytics. A reasonable amount of studies have shown that there is a noticeable relationship between audio features and streaming metrics to determine what makes certain tracks more popular and successful than others. With access to large-scale streaming data, researchers gained the ability to apply machine learning models and examine the impact of musical characteristics on listener engagement and overall streaming counts.

Recent studies have focused on song's fundamental audio features - such as danceability, energy, loudness, tempo, valence etc - to assess their role in determining a song's popularity, Zhao et al. (2023) [1] found that songs with higher danceability and positive valence tend to perform better in terms of engagement, reinforcing the idea that rhythm and mood play a crucial role in shaping listener preferences. Similarly, Saragih (2023) [2] analyzed streaming data from the Indonesian market and observed that energy and liveness were strong predictors of a song's success. However, these studies also indicated that while audio features contribute significantly to song popularity, they do not fully explain the variations in streaming performance.

Additional research has revealed that also historical streaming data also provides more accurate indicator of future popularity. In Nijkamp's paper (2018) [3], the main purpose was to investigate whether Spotify audio features alone could predict success and concluded that their predictive power was moderate at best. Instead, the study emphasized that a combination of past streaming trends, playlist placements, and user listening behaviors yields better predictions. This is also supported by Yee & Raheem (2022) [4], who demonstrated that models incorporating historical streaming activity outperformed those relying solely on audio attributes. These findings suggest that while musical properties shape a song's appeal, its placement within algorithmic recommendations plays an equally significant role in determining long-term success.

Various machine learning techniques have been applied to enhance popularity prediction models. Studies have found that Random Forest and Logistic Regression are particularly effective when combining audio features and metadata-based variables such as past chart performance and artist popularity [5]. Saragih (2023) [2] identified that Extra Trees Regressor and Random Forest Classifier were among the most reliable models for predicting listener engagement based on Spotify datasets. Across multiple studies, danceability, valence, and energy consistently affect streaming activity [1][2][5].

Despite these advancements, predicting song success remains a complex challenge due to several external factors that are difficult to quantify. Nijkamp (2018) [3] highlighted that no single model can fully capture the unpredictability of listener behavior, as marketing efforts, artist reputation, and viral trends often play an outsized role in determining a song's performance. Furthermore, cultural differences across global markets mean that feature importance varies between regions, making it difficult to develop a universal prediction model [2][3].

In summary, existing research demonstrates that both audio features and streaming patterns are crucial to understanding song popularity. While audio features consistently correlate with success, the integration of historical streaming metrics and playlist placement significantly enhances predictive accuracy [1][4][5]. Future work in this area should focus on hybrid models that incorporate real-time streaming trends with musical characteristics[5].

III. Data Management

A. Datasource

The dataset used in this study was downloaded from Kaggle[6]. It is a highly large dataset that contains over 100000 songs and their corresponding features, taken from Spotify. In addition, it is worth mentioning that the tracks range between 125 different genres.

B. Feature Description

The features that included in the dataset are mainly the audio features of each song (danceability, energy, valence etc) with additional columns that contain information about the name and the information of the song. A more comprehensive description of the features is explained below:

Type of the Features

- Categorical: Explicit, Mode, Time Signature, Key, Track Genre
- The rest of the features are numerical.

TABLE I: Feature Names and Descriptions from the Spotify Dataset.

Feature Name	**Description**
song_popularity	Popularity score of the track (0-100), assigned by Spotify.
duration_ms	Length of the track in milliseconds.
explicit	Boolean value indicating whether the track contains explicit content.
danceability	Measure of how suitable a track is for dancing (0-1).
energy	Measure of intensity and activity (0-1).
loudness	Overall loudness of the track in decibels (dB).
mode	Indicates whether the track is in major (1) or minor (0) key.
speechiness	Measure of spoken words in the track (0-1).
acousticness	Probability of the track being acoustic (0-1).
instrumentalness	Predicts the likelihood of the track being instrumental (0-1).
liveness	Measure of whether the track was recorded live (0-1).
valence	Musical positiveness of the track (0-1).
tempo	Estimated tempo of the track in beats per minute (BPM).
time_signature	Estimated time signature of the track.
track_genre	Genre classification of the track.

C. Additional Data Collection

Alongside the audio features, two additional numerical features where fetched from the Last.fm API to improve the model training and also to justify even further the whole topic of the study which is the the impact of the audio features alongside the metadata from streaming platforms to the popularity of a song. The features are shown in Table II.

After the successful fetching of those features, the records that have not gotten any data were being removed.

D. Data Preparation/Pre-processing

Before training the machine learning models, preprocessing the dataset is crucial to ensure the data is clean, structured, and free from inconsistencies. The

TABLE II: Names and Descriptions for the Additional Features

Feature Name	Description
playcount	Total Number of Plays per Song (Simulate stream approximation).
listeners	Total Unique Listeners per Song.

following preprocessing steps were performed on this dataset:

1) Data Cleaning:

- **Removing Unnecessary Columns:** 'Unnamed: 0', 'track_id', 'artists', 'album_name', and 'track_name' were removed as they do not contribute to model learning.
- **Checking for Missing Values:** The function df.isnull().sum() was used to identify missing values in each feature.
- **Handling Duplicates:** The dataset was checked for duplicates and any duplicate records were dropped to maintain data integrity.

2) Defining the target variable: The target variable 'song_popularity' was defined as the dependent variable(y), while all other features in the dataset were selected as independent variables for model training and analysis.

E. Exploratory Data Analysis (EDA)

This part covers the data exploration that was conducted to understand the nature of the dataset and determine the most suitable ways to get the maximum possible modeling performance. The analysis helps identify key features that influence song popularity, providing insights into their relationships and significance. Additionally, it aids in feature selection and dimensionality reduction by revealing correlations and patterns within the data.

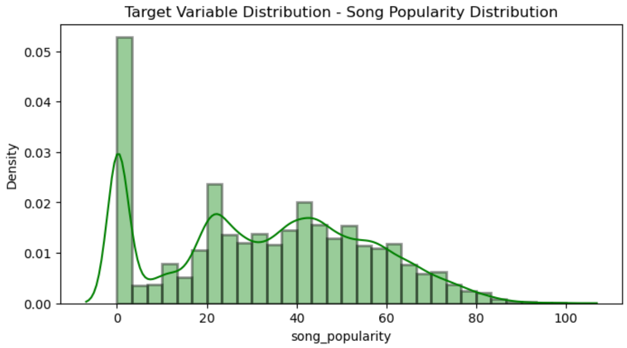

Fig. 1: Target Variable Distribution

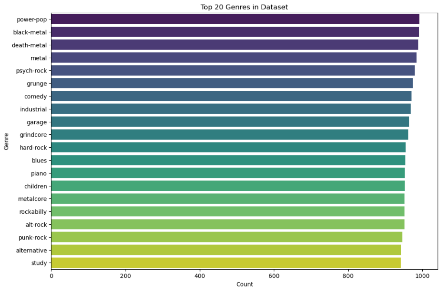

Fig. 2: 'track_genre Feature Plotting'

1) Analysing the distribution of the target variable: The histogram in Figure 1 indicates that 'song_popularity' is right-skewed, with a high concentration of songs having low popularity scores (close to 0). This result is highly reasonable, knowing that a song popularity calculated by Spotify does not have a fixed value and it constantly changes as time passes. Having that in mind one possible explanation is that when the dataset was being created, most of the old songs had zero popularity despite the fact that once were popular. Additionally, the plot suggests that data manipulation and model adjusting is highly needed.

F. Data Visualisation

1) Plotting 'track_genre' Feature: The bar chart in Figure 2 displays the top 20 most common genres, revealing a dominance of "power-pop," "black-metal," "death-metal," and "metal." This genre imbalance could bias model training, favoring well-represented categories while underperforming on others. Addressing this imbalance through data preprocessing can enhance model performance and ensure fair predictions.

2) Plotting the Numerical Features: The distribution plots shown in Figure 3 provide useful insights into the characteristics of the dataset. The numerical features show significant variability, with some, like listeners and playcount, highly skewed, while others, like danceability and tempo, are more balanced. Multimodal patterns in mode and key suggest categorical-like behavior. The varying scales highlight the need for standardization to ensure optimal model performance and interpretability.

3) Plotting the Corelation Heatmap of the Numerical Features: The correlation heatmap is used to analyze relationships between numerical features in the dataset. It helps identify strong positive or negative correlations that may influence model performance and guide feature selection. Highly correlated features may introduce redundancy, while weak correlations to the target variable suggest limited predictive power.

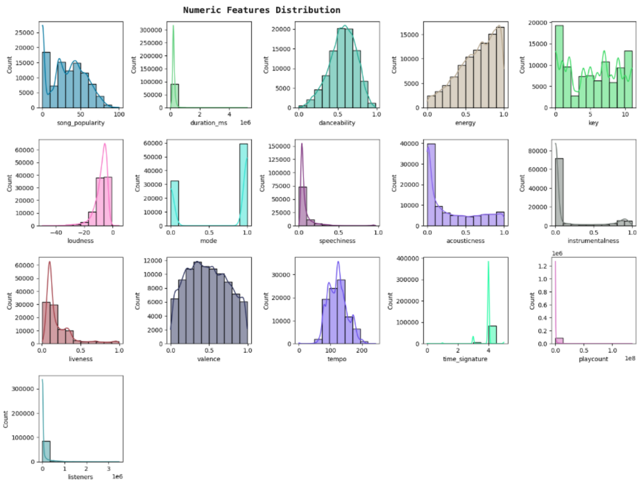

Fig. 3: Numerical Features Plotting

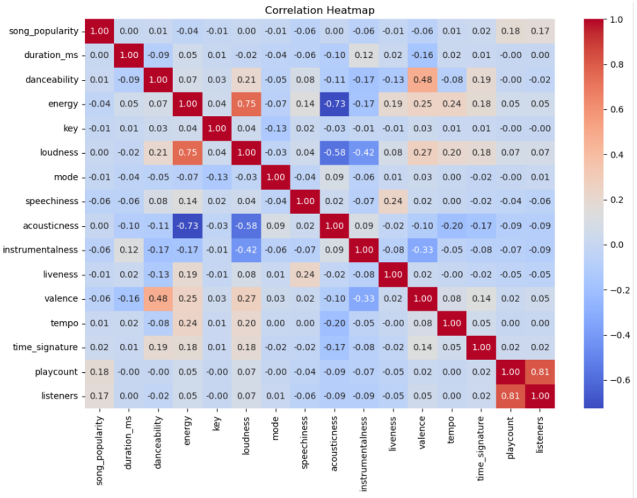

Fig. 4: Corelation Heatmap Plot of the Numerical Features

Interpretation: From the corelation heatmap in Figure 4, playcount and listeners exhibit a strong positive correlation (0.81), indicating that more listeners generally lead to higher play counts. However, their correlation with song popularity is moderate (0.18 - 0.17), suggesting other factors contribute significantly to popularity. Features like loudness and energy show a noticeable correlation (0.75), which aligns with expectations that high-energy songs tend to be louder. On the other hand, features such as acousticness and danceability exhibit weak correlations with popularity, indicating that these attributes alone are not strong determinants of a song's success.

G. Data Manipulation

After completing the Data Exploration Analysis, two main points were detected that with a targeted data manipulation, they can help improve the performance of the following modeling.

1) Encoding the Categorical Feature 'track_genre' using One Hot Encoding: One-Hot Encoding is a method used to convert categorical variables into numerical form by creating binary (0 or 1) columns for each category. Each row has a 1 in the column representing its category and 0 in all others. It was chosen because it avoids assigning arbitrary numerical rankings to categories. It ensures that categorical relationships do not introduce unintended ordinal meaning in the model.

After applying One-Hot Encoding, the track_genre column is transformed into multiple binary columns (e.g., track_genre_techno, track_genre_pop, etc.).

2) Normalisation of the Numerical Features: From the previous visualisations of the numerical features, it was observed that the features had different scales and some of them were skewed while others were balanced. This issue could potentially affect the training process. To address this, standardization was applied to transform the numerical features to have a mean of zero and a standard deviation of one, ensuring that all features contributed equally during model training. The *StandardScaler* method from the *sklearn.preprocessing* library was used for this process. This method standardizes each numerical feature by subtracting the mean and dividing by the standard deviation.

H. Feature Selection Process

After data manipulation, the next step is feature selection, whose main purpose is to identify the most relevant features that can possibly help the model achieve the best possible performance and accuracy. The following feature selection methods were implemented:

1) Random Forest Feature Importance(RFI): This method uses the Random Forest algorithm to rank features based on their contribution to the model's predictions. Features that have a higher impact on reducing impurity in decision trees are assigned greater importance.

Chosen features: ['playcount', 'listeners', 'duration_ms', 'acousticness', 'loudness', 'danceability', 'valence', 'tempo', 'speechiness', 'energy', 'liveness', 'instrumentalness', 'key', 'mode', 'time_signature']

2) Variance Inflation Factor (VIF): VIF measures how much a feature is correlated with other features in the dataset. A high VIF indicates multicollinearity, meaning the feature is redundant and can be removed to improve model stability.

Chosen features: ['duration_ms', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'playcount', 'listeners']

3) Lasso Regression (L1 Regularization): Lasso (Least Absolute Shrinkage and Selection Operator) is a regression technique that applies L1 regularization, which forces some feature coefficients to shrink to zero. This helps in selecting only the most important features while reducing overfitting.

Chosen features: ['duration_ms', 'danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'time_signature', 'playcount', 'listeners']

Based on the feature selection results, RFI and Lasso identified nearly identical sets of influential features, highlighting their strong predictive power, while VIF focused on removing multicollinear variables to improve model stability. This suggests that RFI and Lasso should be prioritized for selecting impactful features, while VIF can serve as a secondary check.

IV. Methodology

This section covers the definition of the two machine learning models that used to conduct the study. It highlights the reasons why those models were chosen and also describes the theoretical background and how those are consisted.

A. Random Forest Classifier

The Random Forest Classifier (RFC) is a supervised learning model used for classification tasks. It is chosen for its ability to handle high-dimensional datasets with both numerical and categorical features while reducing overfitting through ensemble learning. RFC constructs multiple decision trees during training and determines the final class prediction based on majority voting. Mathematically, RFC predicts the class *k* that receives the most votes across *T* decision trees:

$$\hat{y} = \text{mode}\left(h_1(X), h_2(X), ..., h_n(X)\right)$$

where: $h_i(X)$ is the prediction of the i-th decision tree, n is the total number of trees, mode refers to the most frequently predicted class among the trees.

B. Random Forest Regressor

The Random Forest Regressor (RFR) is a supervised learning model used for regression tasks, predicting continuous values instead of discrete classes. It is chosen for its ability to capture complex relationships between variables and reduce variance through ensemble learning. Unlike RFC, which uses majority voting, RFR averages predictions from multiple decision trees to estimate a numerical output. The mathematical formula for RFR is:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

where: $\hat{y}$: Predicted numerical value, T : Total number of decision trees in the forest, $h_t(x)$: Output prediction from an individual

Machine learning models can approach predictive problems in different ways, and the study utilizes both classification and regression to analyze song popularity. The RFC categorizes songs as popular or not popular, while the RFR predicts a continuous popularity score. By employing both models, the study directly compares how classification and regression handle the same dataset, evaluating which approach is more effective for predicting song success based on audio features and streaming metadata.

V. Evaluation and Results

A. Evaluation Metrics for Classification Performance:

1) Accuracy: This is the proportion of correctly classified instances out of the total number of predictions. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- **TP (True Positives)**: Correctly classified popular songs.
- **TN (True Negatives)**: Correctly classified non-popular songs.
- **FP (False Positives)**: Non-popular songs misclassified as popular.
- **FN (False Negatives)**: Popular songs misclassified as non-popular.

2) Precision: It measures how many of the songs predicted as popular were actually popular. It is defined as:

$$Precision = \frac{TP}{TP + FP}$$

3) Recall: It quantifies the model's ability to correctly identify all popular songs:

$$Recall = \frac{TP}{TP + FN}$$

4) F1-Score: It is the harmonic mean of precision and recall, balancing both measures:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5) Receiver Operating Characteristic (ROC) Curve: It visualizes the trade-off between the true positive rate (recall) and the false positive rate across different thresholds. The AUC score measures overall model performance, where 1.0 is perfect classification and 0.5 indicates random guessing. A higher AUC suggests better discrimination ability.

TABLE III: Classification Performance Comparison

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
RFI	0.8989	0.8802	0.8504	0.8636
VIF	0.8924	0.8716	0.8414	0.8547
Lasso	0.8969	0.8774	0.8478	0.8610
Combined	0.8991	0.8805	0.8506	0.8639

6) ***Interpretation:**:* From the results shown in Table III, the best approach to get the maximum possible performance was to combine the features from all the three feature selection methods. This suggests that leveraging multiple feature selection techniques together helps in capturing the most relevant and non-redundant features. The Combined approach has the highest accuracy (0.8991) and F1-score (0.8639).

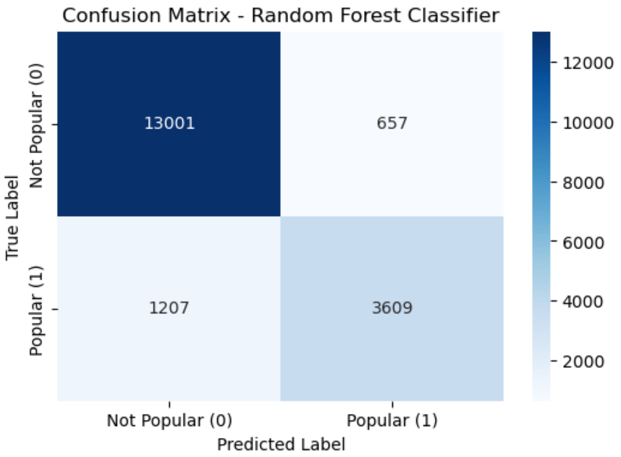

Fig. 5: Confusion Matrix - Random Forest Classifier - Combined Approach

The confusion matrix demonstrates that the Random Forest Classifier effectively distinguishes between popular and non-popular songs, with a high number of correctly classified instances.

ROC Curve Analysis

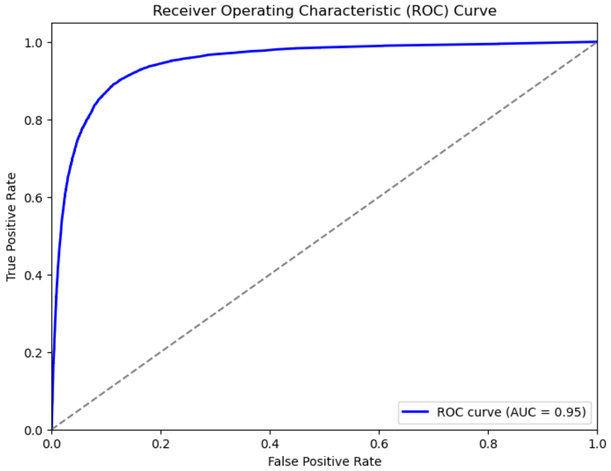

Fig. 6: ROC Curve Plot

The ROC curve shows a high AUC score of 0.95, indicating that the classifier effectively distinguishes between popular and non-popular songs with strong predictive performance.

B. Evaluation Metrics for Regression Performance

1) Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values. Lower values indicate better model accuracy.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2) Mean Absolute Error (MAE): Represents the average absolute difference between actual and predicted values, providing a more interpretable error measurement.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

3) R-Squared Score (R²): Indicates how well the model explains the variance in the target variable. A score closer to 1 suggests a strong predictive capability.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where: y_i - Actual observed value, $\hat{y}_i$ - Predicted value, $\bar{y}$ - Mean of actual values, n - Total number of samples

4) ***Interpretation:**:* From the results shown in Table IV, the best regression performance was achieved the Random Forest Importance (RFI) feature selection method, which resulted in the lowest Mean Squared Error (MSE) of 178.45, the lowest Mean Absolute

TABLE IV: Regression Performance Comparison

Method	MSE	MAE	R^2
RFI	178.45	8.70	0.6523
VIF	182.39	8.85	0.6446
Lasso	178.76	8.70	0.6517
Combined	179.30	8.72	0.6506

Error (MAE) of 8.70, and the highest R² score of 0.6523. In contrast with the Classifier, in this case the combined approach did not outperform RFI alone.

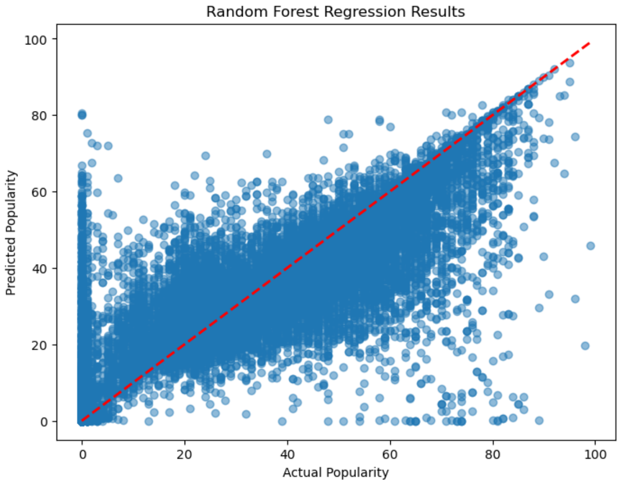

Fig. 7: Scatter Plot - Random Forest Regression

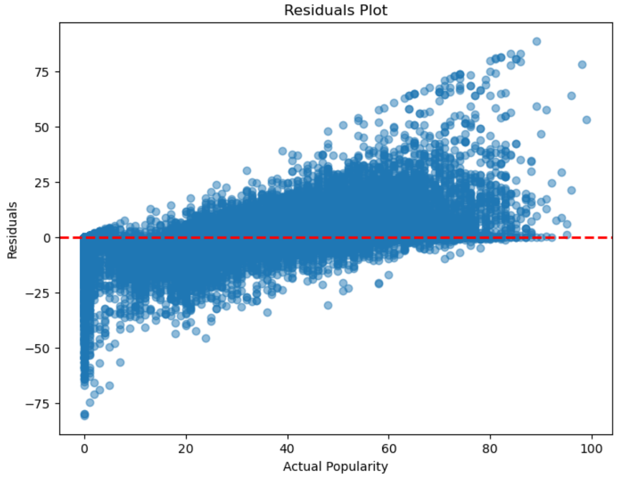

Fig. 8: Residual Plot - Random Forest Regressor

The scatter plot in Figure 7 shows that the Random Forest Regressor captures general trends in song popularity, but variance is evident, particularly at lower popularity levels. The residuals plot in Figure 8 indicates that while errors are centered around zero, they tend to increase for higher popularity values. This suggests that while the model provides reasonable predictions, there may be room for improvement, particularly in handling extreme values.

C. *Additional Comment for Both Models*

Standardization was applied to both classification and regression models. However, after evaluation, it was found that standardization did not improve neither of the models. In contrast, using the original numerical features yielded better results. Consequently, the final models were trained without standardization.

VI. Conclusion

A. *Summary of Findings*

After the completion of the analysis, all initial objectives were successfully achieved. The findings confirmed that song popularity cannot be fully predicted using audio features alone, as streaming metadata—particularly play count and listeners—played a crucial role in improving model performance. The findings also align with the literature, which emphasize that external factors like historical trends and audience engagement significantly impact a song's popularity and success. Furthermore, classification (RFC) outperformed regression (RFR), demonstrating that popularity is best modeled as a categorical outcome rather than a continuous variable.

B. *Future Work and Recommendations*

Future studies should integrate real-time streaming trends, social media sentiment for improved accuracy. Additionally, incorporating playlist placements and user engagement metrics could refine prediction models.

References

[1] M. Zhao, M. Harvey, D. Cameron, F. Hopfgartner, and V. J. Gillet, "An analysis of classification approaches for hit song prediction using engineered metadata features with lyrics and audio features," in *International Conference on Information*, pp. 303–311, Springer, 2023.

[2] H. S. Saragih, "Predicting song popularity based on spotify's audio features: insights from the indonesian streaming users," *Journal of Management Analytics*, vol. 10, no. 4, pp. 693–709, 2023.

[3] R. Nijkamp, "Prediction of product success: explaining song popularity by audio features from spotify data," B.S. thesis, University of Twente, 2018.

[4] 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, 2022.

[5] J. Pham, E. Kyauk, and E. Park, "Predicting song popularity," *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*, vol. 26, 2016.

[6] P. Choksi, "Spotify dataset 114k songs," 2023. Accessed: March 2025.