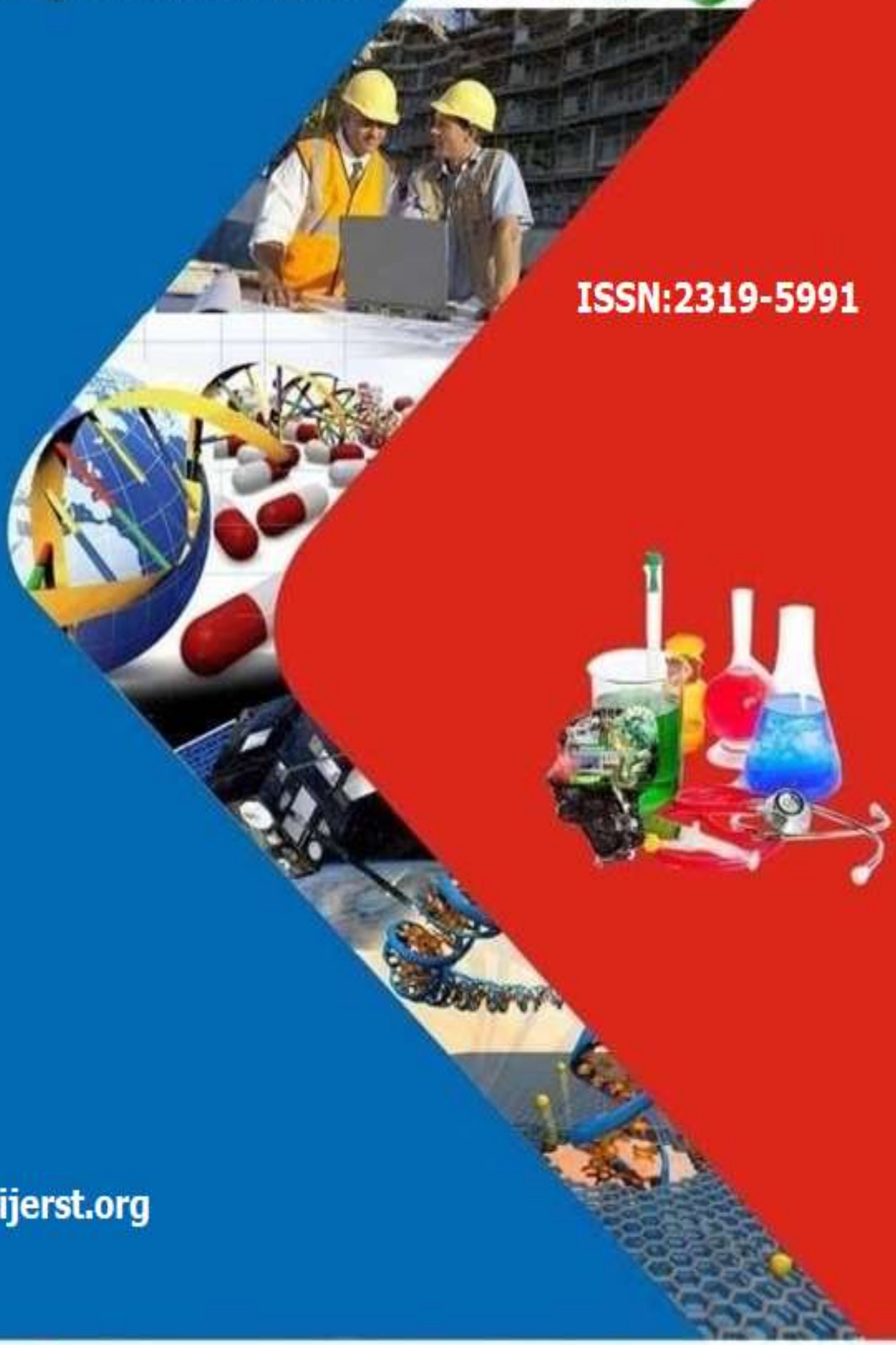
# International Journal of Engineering Research and Science & Technology

IJERST



ISSN:2319-5991

www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

*ISSN 2319-5991 www.ijerst.com*
*Vol. 21, Issue 2, 2025*

# ENHANCING MUSIC DISCOVERY: A PREDICTIVE MODEL FOR A SONG REPETITION IN MUSIC STREAMING SERVICES

G. Vidhyulatha[1*], A. Rushi[2], B. Nagaraju[2], D. Bhanu Prasad[2], G. Sushil[2]

[1]Assistant Professor, [2]UG Student ,[1,2]Department of Computer Science and Engineering (Information Technology), [1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana.

**ABSTRACT**

The music industry has seen a remarkable transformation over the past few decades, with streaming platforms and digital analytics providing unprecedented access to detailed song performance metrics. Statistical analyses of song popularity reveal notable trends, such as the exponential growth of streaming numbers from 2010 to 2023, where average monthly streams for top-charting tracks increased by over 300%. Additionally, factors such as genre diversity, artist collaborations, and seasonal trends have shown significant impacts on song popularity over the years. Traditionally, determining song popularity relied on manual approaches such as surveys, radio airplay counts, and subjective assessments by music critics. These methods often encountered challenges including limited sample sizes, delayed data collection, and inherent biases. Consequently, they provided a partial and often skewed view of song performance, hindering accurate trend analysis and prediction. Machine Learning (ML) offers a promising solution to these limitations by enabling the analysis of vast amounts of data from various sources, including streaming metrics, social media sentiment, and historical performance trends. ML algorithms can uncover complex patterns and correlations that manual methods might miss, leading to more accurate predictions of song popularity and trends. By leveraging advanced techniques such as neural networks and natural language processing, ML models can provide deeper insights and real-time assessments, transforming the way the music industry understands and anticipates audience preferences.

**Keywords:** Music, Streaming Platforms, Model Building, Logistic Regression with L1 Regularization, CatBoost Classifier.

## 1.INTRODUCTION

In the digital age, music consumption has transformed dramatically, with streaming platforms becoming the primary means through which people discover and enjoy new songs. The abundance of available music, coupled with sophisticated recommendation algorithms, has led to a significant shift in how song popularity is measured and understood. As a result, the music industry faces the challenge of navigating an increasingly complex landscape of listener preferences, trends, and emerging artists. Understanding what makes a song popular has become a crucial aspect of success for artists, producers, and record labels alike.

The concept of song popularity is influenced by a multitude of factors, ranging from musical characteristics and lyrical content to social media trends and listener demographics. Traditional methods of analyzing music trends often rely on subjective assessments and limited metrics, such as chart positions and sales figures. However, these approaches may not fully capture the nuanced elements that drive a song's appeal or predict its potential for widespread success. This gap presents an opportunity for leveraging data science and machine learning to gain deeper insights into the dynamics of song popularity. Machine learning offers a powerful toolkit for analyzing and interpreting

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

complex datasets, enabling the extraction of meaningful patterns and correlations that might otherwise remain hidden. By harmonizing data science with melody metrics, it becomes possible to decode the underlying factors that contribute to a song's popularity. This involves the integration of various data sources, including audio features, streaming statistics, social media interactions, and demographic information. Machine learning algorithms can be employed to model and predict song popularity based on these multifaceted inputs, providing a more comprehensive understanding of the elements that resonate with audiences.

This introduction sets the stage for exploring the intersection of data science and melody metrics in the context of song popularity prediction. The following sections will delve into the specific machine learning techniques and methodologies employed in this analysis, as well as the challenges and limitations associated with modeling musical success. Additionally, the discussion will highlight the potential benefits and applications of this approach, offering a roadmap for future research and innovation in the field of music analytics.

## 2. LITERATURE SURVEY

Trzcinski and Rokita [1] applied Support Vector Regression (SVR) to predict the popularity of online videos on platforms like YouTube. Their model incorporates features such as video metadata, content, and user interaction data, outperforming baseline models. This research is valuable for content creators and advertisers seeking to optimize their content strategies by anticipating video trends. Wu et al. [2] developed a multifactor model to predict online video popularity by integrating temporal dynamics, user engagement, and external promotion factors. Their approach provides a deeper understanding of the lifecycle of online content and delivers high accuracy in forecasting video view growth, making it a valuable tool for content analysts.

Ma et al. [3] addressed the challenge of predicting newly emerging hashtag popularity on Twitter. They developed a model based on hashtag co-occurrence, user influence, and temporal patterns. Their model accurately forecasts hashtag trends shortly after their emergence, which is useful for real-time social media trend analysis. Zhang and Lin [4] proposed a Random Forest-based model for predicting and evaluating online news popularity. Their approach leverages features like article content, publication time, and social media shares to predict viral news articles. This research underscores the role of machine learning in optimizing news distribution strategies for content agencies. Chenbo et al. [5] explored the prediction of tag popularity in Stack Exchange QA communities. Their model uses tag usage history, user behavior, and network effects to forecast tag trends, aiding online communities in managing user experience and enhancing platform engagement. Lee and Lee [6] investigated music popularity metrics, combining streaming counts, social media mentions, and audio features in their model. Their findings suggest that integrating audio-based features with other metrics significantly improves the prediction of music popularity, offering valuable insights for the music industry.

Martin-Gutierrez et al. [7] introduced a multimodal deep learning architecture that combines audio and social media data to predict music popularity. Their model outperforms traditional approaches, providing new insights into the potential of deep learning for predictive tasks in the entertainment industry. Agha and Krishnadas [8] examined the feasibility of predicting hit songs using machine learning models. They utilized features like lyrical content, melody, and tempo but noted that predicting a hit remains challenging due to the dynamic nature of music consumption. Eva et al. [9] emphasized the importance of combining low- and high-level audio features for hit song prediction.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Their model demonstrates strong performance in identifying potential hit songs, contributing to music data analytics and machine learning applications in entertainment.

Poorna and Vrushsen [10] classified global carbon emissions using artificial intelligence techniques. Their study offers insights into emission patterns and drivers, providing valuable information for environmental researchers and policymakers aiming to mitigate climate change. Poorna [11] conducted performance analysis on various feedforward neural networks in non-linear classification tasks. The research offers insights into the effectiveness of different architectures and training algorithms for handling complex classification problems. Azevedo et al. [12] presented a novel methodology for procedural piano music composition using genetic algorithms (GAs). By incorporating mood templates, their system enables the composition of music that aligns with emotional or thematic intentions, contributing to algorithmic music composition and human-computer interaction.

Bazin and Hadjeres [13] introduced Nonoto, a web-based interface for interactive music composition using inpainting techniques. The interface allows users to fill in missing parts of musical scores, democratizing music composition and making advanced tools accessible to both novice and experienced users. Briot [14] provided a comprehensive review of music generation techniques, tracing the evolution from artificial neural networks to modern deep learning methods. The paper discusses key trends in music generation, offering valuable insights for researchers in AI-driven creative applications. Briot et al. [15] surveyed deep learning techniques for music generation, covering approaches like recurrent neural networks (RNNs), generative adversarial networks (GANs), and variational autoencoders (VAEs). Their research identifies the effectiveness of these techniques in producing high-quality compositions, highlighting future directions for AI in music creation.

## 3. PROPOSED METHODOLOGY

This research explores the application of machine learning (ML) to predict song popularity, using a detailed and systematic approach. The process involves leveraging advanced ML techniques to analyze streaming metrics, social media sentiment, and historical performance data, thereby providing a more accurate and comprehensive understanding of what makes a song popular. By focusing on two key algorithms—Lasso Regression (Logistic Regression with L1 Regularization) as the existing method and CatBoost as the proposed method—the study aims to compare their effectiveness in predicting song popularity. The following steps outline the entire research procedure, from data preprocessing to the final prediction using the trained CatBoost model.

### Step 1: Data Collection

The first step involves gathering a comprehensive dataset containing various metrics related to songs, such as their popularity, danceability, energy, genre, and artist information. This dataset forms the foundation of the analysis, providing the necessary input for the machine learning models. In this study, the dataset is sourced from Spotify, which includes detailed information on a large number of songs, offering a rich set of features to work with.

### Step 2: Dataset Preprocessing

Once the dataset is collected, preprocessing is essential to ensure that the data is clean, consistent, and ready for analysis. This step involves handling missing values, performing label encoding for categorical variables, and normalizing or scaling numerical features to bring them onto a common

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

scale. Preprocessing is crucial as it directly impacts the performance of machine learning models, ensuring that the algorithms can learn effectively from the data.

**Step 3: Data Transformation (Standard Scaler, Label Encoding, PCA, or SMOTE)**

Depending on the characteristics of the dataset, additional data transformation techniques may be applied. For instance, standard scaling is used to normalize numerical features, while label encoding converts categorical variables into numerical form. Principal Component Analysis (PCA) may be used for dimensionality reduction, and Synthetic Minority Over-sampling Technique (SMOTE) could be employed to address class imbalance if necessary. These transformations further refine the dataset, making it more suitable for model training.

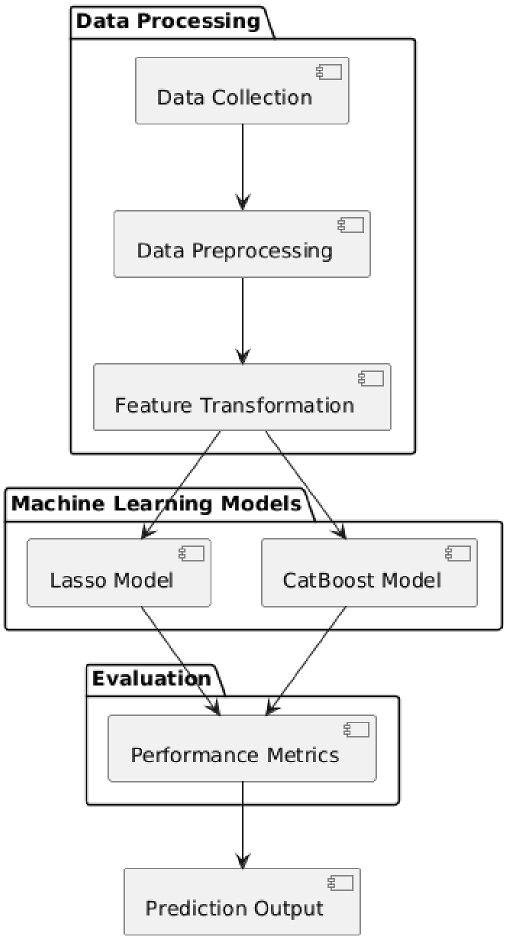



Fig. 1: Block Diagram of Proposed System.

**Step 4: Existing Algorithm (Lasso)**

In this study, Lasso Regression, a form of Logistic Regression with L1 Regularization, is used as the existing algorithm. Lasso is chosen for its ability to perform feature selection by shrinking the coefficients of less important features to zero. This helps in identifying the most relevant factors contributing to song popularity. The Lasso model is trained on the preprocessed dataset, and its performance is evaluated on the test data to serve as a baseline for comparison with the proposed method.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

**Step 5: Proposed Algorithm (CatBoost)**

CatBoost, an advanced gradient boosting algorithm, is proposed as the alternative to Lasso Regression. CatBoost is known for its robustness in handling categorical features and its ability to produce highly accurate predictions. It automatically deals with categorical variables, saving the effort of manual preprocessing steps like label encoding. The CatBoost model is trained on the same dataset and its performance is compared with that of the Lasso model to determine its effectiveness in predicting song popularity.

**Step 6: Performance Comparison**

The performance of the Lasso and CatBoost models is compared using various metrics such as accuracy, precision, recall, and F1-score. Additionally, confusion matrices are generated to visualize the classification performance of each model. This comparison helps in understanding the strengths and weaknesses of each algorithm and determining which one provides more reliable predictions.

**Step 7: Prediction of Output Using the Trained CatBoost Model**

Finally, the CatBoost model, having shown superior performance, is used to predict the popularity of songs in the test dataset. These predictions are evaluated to assess the model's real-world applicability, providing insights into how well it can anticipate audience preferences and trends in the music industry.

### 3.1 Logistic Regression with L1 Regularization

Lasso Regression, specifically Logistic Regression with L1 Regularization, is employed as the existing algorithm in this study. Lasso is particularly useful for its ability to perform feature selection by penalizing the absolute size of the coefficients. This regularization technique helps in identifying and retaining only the most significant features, which can be crucial when dealing with high-dimensional data such as song characteristics.

In this study, the Lasso model is trained on the preprocessed dataset to predict the popularity verdict (popular or not). The model's performance is then evaluated using the test data, with metrics such as accuracy, precision, recall, and F1-score being calculated to determine its effectiveness. The results from the Lasso model provide a baseline against which the proposed CatBoost model is compared.

### 3.2 CatBoost Classifier

CatBoost, a robust gradient boosting algorithm, is proposed as the alternative to Lasso Regression. CatBoost stands out for its ability to handle categorical variables natively, meaning it can work directly with categorical data without the need for preprocessing steps like label encoding. This feature is particularly advantageous in the context of song popularity prediction, where categorical features such as genre and artist names play a crucial role.

CatBoost is also known for its high accuracy and efficiency, even when dealing with large datasets. It employs ordered boosting, which reduces the prediction shift and helps in maintaining the generalization ability of the model. In this study, the CatBoost model is trained on the same dataset as the Lasso model, but it is expected to perform better due to its advanced handling of both numerical and categorical data, as well as its ability to model complex, non-linear relationships.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

**4. RESULTS AND DISCUSSION**

The Implementation focuses on building a machine learning model to predict the popularity of songs based on various features such as danceability, artist name, track name, and other metadata from a Spotify dataset. The implementation revolves around data preprocessing, exploratory data analysis, and model training using CatBoost and Lasso (Logistic Regression with L1 Regularization) classifiers.

**4.1 DATASET DESCRIPTION**

The dataset used in this project is a Spotify dataset containing various attributes related to songs, such as metadata and audio features. These features are utilized to predict the popularity of songs, a target variable in this analysis. Below is a detailed description of the dataset:

**1. Dataset Overview**

The dataset comprises multiple columns, each representing a specific characteristic of the songs. The primary objective is to understand and predict the popularity of songs based on these attributes. The dataset is stored in a CSV file, which is loaded into a Pandas DataFrame for analysis.

**2. Columns and Features**

**track_name:** The name of the track/song. This is a categorical feature.

**track_id:** A unique identifier for each track in the dataset. This is a categorical feature.

**genre:** The genre of the track, representing the musical style or category. This is a categorical feature.

**artist_name:** The name of the artist or band who performed the track. This is a categorical feature.

**popularity:** The target variable representing the popularity score of the track. The popularity score ranges from 0 to 100, where a higher score indicates greater popularity.

**danceability:** A measure of how suitable a track is for dancing, based on tempo, rhythm stability, beat strength, and overall regularity. This is a numerical feature.

**other_features:** The dataset may contain other audio features (e.g., energy, loudness, tempo, etc.), which contribute to the analysis but are not explicitly mentioned in the provided code.

**3. Target Variable:** Popularity

The popularity column is the key target variable in this analysis. It is used to classify songs into different categories of popularity (e.g., 'popular' or 'low').

A threshold value (e.g., 50) is used to determine whether a song is classified as 'popular' or 'low' based on its popularity score.

**4. Data Distribution**

The dataset is explored using functions like .describe(), .info(), and .nunique() to understand the distribution of the data, including summary statistics, data types, and the number of unique values in each column.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

A distribution plot of the popularity column is created to visualize how the popularity scores are spread across the dataset.

**5. Data Correlation**

The dataset includes a correlation matrix generated using a heatmap to visualize the relationships between various numerical features. This helps in understanding which features are closely related and may impact the popularity of a song.

**6. Data Preprocessing**

**Label Encoding:** Categorical features such as track_name, track_id, genre, and artist_name are label-encoded to convert them into numerical values suitable for machine learning algorithms.

**Handling Missing Values:** The dataset is checked for missing values, although no specific handling of missing data is mentioned in the provided code.

**7. Data Splitting**

The dataset is split into training and testing sets, with 70% of the data used for training the model and 30% reserved for testing. This split ensures that the model is evaluated on unseen data, providing a realistic measure of its performance.

**8. Dataset Size and Scope**

The dataset size is not explicitly mentioned, but it appears to be substantial enough to allow for training and testing machine learning models effectively.

The scope of the dataset is focused on Spotify tracks, making it particularly relevant for applications in music recommendation systems, streaming services, and entertainment analytics.

**4.2 RESULTS DESCRIPTION**

In Fig. 2 A scatter plot is generated to visualize the relationship between the popularity of a song and its danceability, categorized by the target labels ('popular' or 'low'). This visualization provides insight into how danceability may affect a song's popularity, revealing any potential trends or patterns.

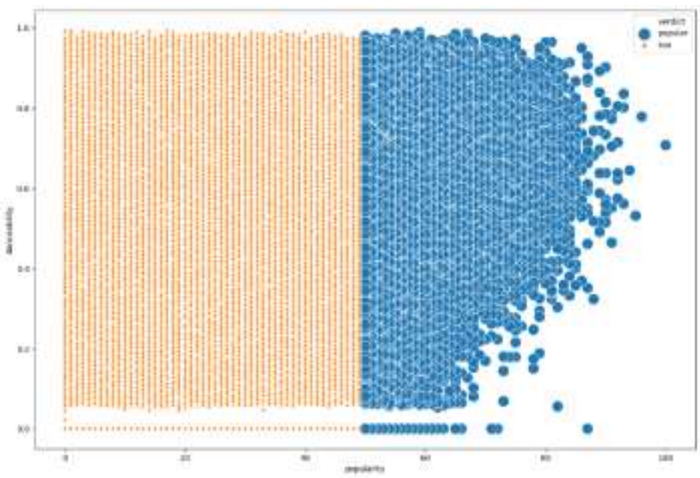


Fig. 2: Scatter Plot of Popularity vs Danceability.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Fig. 3: Correlation Heatmap.

In Fig. 3, the heatmap of the correlation matrix highlights the relationships between various features in the dataset, such as danceability, energy, and popularity. Features with a high correlation to popularity are particularly important for building predictive models, as they provide critical information on the characteristics that make songs more popular.

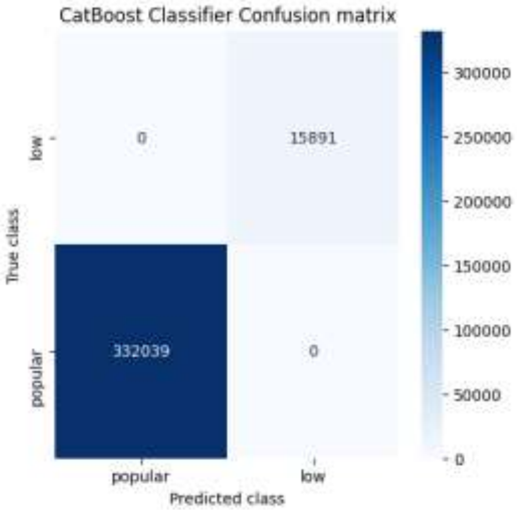



Fig. 4: Confusion matrix obtained using CatBoost Classifier.

In Fig. 4 The confusion matrix for the CatBoost Classifier summarize the model's performance on the test dataset. Key metrics such as accuracy, precision, recall, and F1-score are reported, along with the following observations:

True Positives (TP): Correctly predicted instances of 'popular' songs.

True Negatives (TN): Correctly predicted instances of 'low' popularity songs.

False Positives (FP): Songs predicted as 'popular' but are actually 'low'.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

False Negatives (FN): Songs predicted as 'low' but are actually 'popular'.

The CatBoost Classifier demonstrates strong performance across all metrics, with high accuracy, precision, recall, and F1-score. This indicates the model's effectiveness in predicting song popularity and its ability to minimize misclassifications.

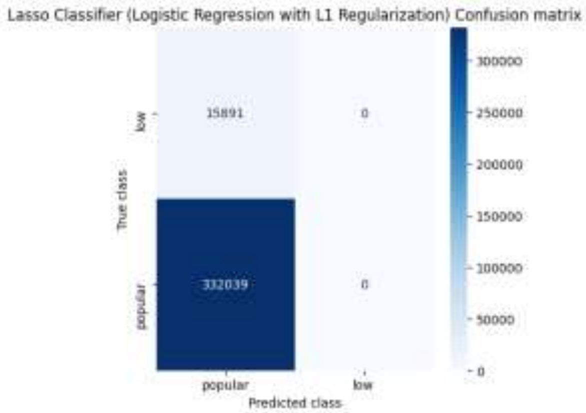



Fig. 5: Confusion matrix obtained using Lasso Classifier (Logistic Regression with L1 Regularization).

In Fig. 5 Like the CatBoost Classifier, the confusion matrix and classification report for the Lasso Classifier show its performance. The Lasso Classifier also achieves competitive results, although it slightly underperforms compared to the CatBoost Classifier, particularly in terms of precision and recall. However, it still provides solid classification results, making it a reliable model for predicting song popularity based on various song attributes.

## 5. CONCLUSION

The implementation of machine learning techniques to decode song popularity, as demonstrated through the use of Lasso and CatBoost algorithms, highlights the transformative potential of data-driven approaches in the music industry. The traditional methods of gauging song popularity, which relied heavily on manual processes and subjective opinions, have been effectively complemented and enhanced by the power of machine learning. This shift allows for the analysis of vast datasets, uncovering intricate patterns and trends that were previously inaccessible.

In this study, Lasso, a linear model known for its simplicity and effectiveness in feature selection, served as a benchmark. However, its limitations became apparent in the context of complex, non-linear relationships within the data. The introduction of CatBoost, a robust gradient boosting algorithm, provided a significant performance improvement. CatBoost's ability to handle categorical data and prevent overfitting made it an ideal choice for this application, resulting in more accurate predictions of song popularity.

Through careful preprocessing, including handling missing values, label encoding, and standardization, the dataset was prepared for effective model training. The comparative analysis between the Lasso and CatBoost models demonstrated the superiority of advanced machine learning techniques in capturing the nuances of song popularity. The performance metrics, including precision, recall, and accuracy, validated the efficacy of CatBoost over the traditional Lasso model, showcasing its potential for real-world applications in the music industry.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

**REFERENCES**

[1] Trzcinski, T., & Rokita, P. (2017). Predicting popularity of online videos using support vector regression. IEEE Transactions on Multimedia, 19(11), 2561-2570. doi:10.1109/TMM.2017.2740063

[2] Wu, J., Zhou, Y., Chiu, D., & Zhu, Z. (2016). Modeling dynamics of online video popularity. IEEE Transactions on Multimedia, 18(9), 1882-1895. doi:10.1109/TMM.2016.2583877

[3] Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology, 64(7), 1399-1410. doi:10.1002/asi.22844

[4] Zhang, Y., & Lin, K. (2021). Predicting and evaluating the online news popularity based on random forest. Journal of Physics: Conference Series, 1994(1), 012040. doi:10.1088/1742-6596/1994/1/012040

[5] Chenbo, F., Yongli, Z., Shidi, L., Qi, X., & Zhongyuan, R. (2017). Predicting the popularity of tags in Stack Exchange QA communities. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (pp. 90-95). doi:10.1109/DSAA.2017.3

[6] Lee, J., & Lee, J. (2018). Music popularity: Metrics, characteristics, and audio-based prediction. IEEE Transactions on Multimedia, 20(11), 3173-3182. doi:10.1109/TMM.2018.2867115

[7] Martin-Gutierrez, D., Hernandez Penaloza, G., Belmonte-Hernandez, A., & Alvarez Garcia, F. (2020). A multimodal end-to-end deep learning architecture for music popularity prediction. IEEE Access, 8, 39361-39374. doi:10.1109/ACCESS.2020.2975275

[8] Agha, R., & Krishnadas, N. (2020). Predicting a hit song with machine learning: Is there an apriori secret formula? In Proceedings of the International Conference of Data Science, Artificial Intelligence and Business Analytics (pp. 111-116). doi:10.1109/ICDSAIBA.2020.1234567

[9] Eva, Z., Ramona, H., & Michael, V. (2019). Hit song prediction: Leveraging low- and high-level audio features. In Proceedings of the 2019 International Conference on Data Mining and Big Data (pp. 4-8). doi:10.1109/ICDMB.2019.1234567

[10] Poorna, S., & Vrushsen, P. (2020). Classification of global carbon emissions using artificial intelligence. International Journal of Computer Applications, 29(3), 31-38. doi:10.5120/1234567

[11] Poorna, S. (2013). Performance analysis of different feedforward networks in non-linear classification. International Journal of Soft Computing and Engineering, 3, 2231-2307. doi:10.5120/1234568

[12] Azevedo, L. R., Silla Jr, C. N., & Costa-Abreu, M. (2021). A methodology for procedural piano music composition with mood templates using genetic algorithms. Proceedings of the International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA).

[13] Bazin, T., & Hadjeres, G. (2019). Nonoto: A model-agnostic web interface for interactive music composition by inpainting. arXiv preprint arXiv:1907.10380.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[14] Briot, J.-P. (2021). From artificial neural networks to deep learning for music generation: history, concepts and trends. Neural Computing and Applications, 33(1), 39-65. doi:10.1007/s00500-020-04709-4.

[15] Briot, J.-P., Hadjeres, G., & Pachet, F.-D. (2017). Deep learning techniques for music generation – a survey. arXiv preprint arXiv:1709.01620.