PREDICTING SONG POPULARITY USING AUDIO FEATURES

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

LAKSHANYA D (2116220701140)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"PREDICTING SONG POPULARITY USING AUDIO FEATURES"** is the bonafide work of **"LAKSHANYA D (2116220701140)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

HEAD OF THE DEPARTMENT,
Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602 105.

SIGNATURE

Dr. V.Auxilia Osvin Nancy, M.Tech., Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner

External Examiner

ABSTRACT

With the explosion of digital music streaming platforms, the ability to anticipate which songs are likely to become popular has become a valuable asset for artists, producers, and music industry stakeholders. This study introduces a machine learning-based framework designed to predict the popularity of songs based on a wide range of audio features. These features include tempo, energy, danceability, loudness, acousticness, valence, and instrumentalness, among others. By utilizing publicly available music datasets, the project evaluates and compares the effectiveness of various machine learning algorithms such as Linear Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting.

The core objective of this research is to identify which model provides the most accurate predictions of song popularity scores, typically measured through streaming counts, chart positions, or platform-specific metrics.

In addition to comparing model performance, the study also explores the relative importance of each audio feature in contributing to a song's popularity. The findings from this study have significant implications not only for the music industry but also for building intelligent music recommendation systems and data-driven marketing strategies. Overall, the proposed system offers a scalable and efficient approach for predicting song popularity based on intrinsic audio characteristics.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MANICK VISHAL C - 2116220701158

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

In recent years, the music industry has seen a surge in the application of artificial intelligence to understand and predict listener preferences. One particularly valuable area of study is the prediction of song popularity—a task that holds significance for record labels, streaming platforms, and independent artists alike. Song popularity is no longer perceived as solely dependent on marketing strategies or celebrity influence but as a complex interplay of musical attributes, emotional impact, and audience engagement. With the availability of vast music datasets and the rise of streaming platforms, it is now possible to analyze songs at a granular level using data-driven approaches.

Traditional methods of gauging popularity, such as radio airplay or sales figures, are often retrospective and do not offer predictive capabilities. Moreover, such methods fail to account for the musical content itself, which plays a crucial role in audience appeal. With the advent of machine learning and data science, researchers and industry professionals can now explore how intrinsic audio features—such as tempo, energy, danceability, and acousticness—correlate with a song's success.

This research aims to harness the power of supervised classification algorithms to predict the popularity class of a song based on its audio features. The proposed system, referred to as the *Song Popularity Classifier*, uses labeled datasets derived from popular music databases like Spotify to train and evaluate models that can determine whether a song is likely to be a 'hit' or not.

The motivation behind this work stems from the increasing volume of music released daily and the need for intelligent filtering and recommendation systems. Streaming services collect rich metadata and audio feature information, yet converting this raw data into actionable predictions requires robust machine learning systems. This study addresses that challenge by comparing the performance of various classification algorithms on a curated dataset.

Four classification models—Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, and XGBoost Classifier—were trained and evaluated using key performance metrics such as Accuracy, Precision, Recall, and F1-score. Additionally, feature importance analysis was conducted to identify the most influential audio parameters contributing to a song's popularity. Data preprocessing and normalization were performed in Python using libraries such as Pandas, Scikit-learn, and Librosa, and all experiments were conducted in Google Colab.

By enabling early prediction of song popularity, the *Song Popularity Classifier* has potential applications in music recommendation engines, artist promotion strategies, and consumer behavior analysis. This study not only contributes to the growing field of music informatics but also demonstrates the effectiveness of machine learning in analyzing complex, non-linear relationships in multimedia data.

In recent years, the integration of artificial intelligence and machine learning in the music industry has revolutionized how songs are produced, distributed, and consumed. One particularly promising application is the prediction of song popularity using audio feature analysis. Unlike traditional metrics that rely solely on post-release data like sales, social media trends, or listener counts, machine learning offers a proactive, content-driven approach to understanding what makes a song resonate with audiences. Popularity is increasingly seen not as a random outcome, but as something that can be modeled through measurable elements inherent in the music itself.

CHAPTER 2

2.LITERATURE SURVEY

The intersection of music analytics and machine learning has garnered increasing attention in recent years, particularly for tasks like genre classification, emotion detection, recommendation systems, and popularity prediction. With the advent of digital platforms like Spotify, Apple Music, and YouTube, vast datasets of songs tagged with various features and popularity metrics have become readily accessible, enabling more sophisticated analysis and modeling techniques.

Previous research by Herremans et al. (2014) demonstrated that machine learning algorithms could effectively model and predict musical success based on compositional features, such as key, tempo, and time signature. Similarly, Pachet and Roy (2008) explored hit song science, identifying that certain musical elements had statistically significant correlations with chart-topping songs. However, these studies often relied on handcrafted rules or small, genre-specific datasets, which limited their scalability and generalizability.

More recent approaches have leveraged APIs like the Spotify Web API to extract high-dimensional audio features. For instance, a study by Ferwerda and Schedl (2016) used Spotify's acoustic features to predict user engagement and playlist inclusion, suggesting that audio features like energy and danceability play a crucial role in perceived song popularity. Their results were validated using user-centric metrics such as play counts and playlist occurrences, showing the predictive power of machine-learned representations.

Moreover, comparative studies like the one by Schedl and Hauger (2015) explored different machine learning techniques including decision trees, logistic regression, and neural networks for predicting popularity across musical datasets. These studies highlight the need for both accuracy and explainability in predictive models, particularly when they are used in real-world music production and recommendation systems.

Despite these advancements, challenges remain in accounting for non-audio factors like marketing, social media presence, and artist reputation, which also influence song popularity. However, this study narrows its focus to intrinsic audio features to ensure objectivity and general applicability across genres and regions. By comparing a range of machine learning models, this research contributes to a clearer understanding of which algorithms and features most effectively predict song popularity based solely on the music itself.

Building upon the foundation of earlier studies, contemporary research continues to delve into how machine learning can be harnessed to predict the popularity of music tracks based on audio features. One of the most prominent trends is the use of high-dimensional datasets sourced from platforms like Spotify and Last.fm, which provide not only audio characteristics but also user

behavior data. These datasets enable researchers to model complex relationships between song structure and listener preferences.

In a study by Choi et al. (2017), deep learning models—specifically convolutional neural networks (CNNs)—were trained on raw audio spectrograms to extract musical features directly, without relying on hand-engineered descriptors. Their model outperformed traditional feature-based approaches on genre classification and mood prediction tasks, highlighting the growing relevance of deep learning in music information retrieval (MIR).

Another significant contribution was made by Kim et al. (2014), who investigated the use of multimodal data—combining lyrics, audio features, and social metadata—to enhance popularity prediction models. They found that while audio features were strong predictors on their own, integrating textual sentiment from lyrics and social factors led to even more accurate predictions. However, the model's reliance on external metadata raised concerns about objectivity and overfitting to platform-specific trends.

Similarly, Tsaptsinos (2017) explored recurrent neural networks (RNNs) for predicting user listening behavior over time, demonstrating that temporal patterns in how users consume music can also be used to infer popularity. Although this approach offers practical insights for recommendation systems, it shifts the focus from intrinsic audio characteristics to consumption trends, which may not be reliable for pre-release prediction.

A more targeted approach was taken by Lee and Cunningham (2019), who limited their features to Spotify's API outputs like danceability, valence, energy, and loudness. Their research showed that ensemble models such as Random Forests and Gradient Boosted Trees consistently outperformed linear models in predicting hit songs across multiple genres. Their findings reinforced the idea that nonlinear models are better equipped to capture the complex interplay among audio features that correlates with popularity.

Moreover, Sharma et al. (2020) emphasized the importance of data preprocessing and feature selection in improving model performance. By applying techniques such as z-score normalization, principal component analysis (PCA), and recursive feature elimination (RFE), they were able to enhance the performance of SVM and K-Nearest Neighbor (KNN) classifiers

in predicting popularity buckets (e.g., low, medium, high). This underlines the critical role of data quality and dimensionality reduction in music-based machine learning models.

Despite promising results, several limitations persist in this domain. The subjectivity of musical enjoyment, rapidly changing consumer tastes, and the influence of non-musical variables like artist fame or marketing strategies introduce noise into prediction models. Additionally, the long-tail distribution of music consumption means that most models are biased toward predicting success for mainstream or frequently streamed content.

Nonetheless, by narrowing the scope to intrinsic audio features, this study aligns with research such as that of McFee et al. (2015), who argue for the interpretability and reproducibility of content-based models in music analysis. Through comparative evaluation of classification algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forests, and XGBoost, the current research aims to establish a reliable baseline for predicting song popularity with minimal external influence, contributing to the development of explainable, genre-independent, and scalable models.

Another area of interest has been the impact of feature engineering in predicting song popularity. Research by Dhanraj et al. (2019) demonstrated that the incorporation of statistical features such as mean, standard deviation, and skewness from audio features (e.g., spectral contrast, zero-crossing rate) can significantly improve the performance of machine learning models. They found that models built with these engineered features, when combined with ensemble learning techniques, showed robust performance in predicting song popularity across a variety of genres. Their findings suggested that incorporating a combination of both domain-specific and statistical audio features helps capture the subtleties in musical structure that may resonate with different audience segments.

CHAPTER 3

3.METHODOLOGY

METHODOLOGY

The dataset used for this project consists of several features related to the audio characteristics of songs, such as **danceability**, **energy**, **tempo**, **loudness**, and **acousticness**, among others. These features were derived from audio files using feature extraction techniques, providing valuable insights into the sonic properties of the songs.

Data Preprocessing:

The dataset was pre-processed to handle any missing values and to scale the features for improved model performance. Feature scaling was performed using `StandardScaler` to standardize the range of the data and ensure fair comparisons between different models. The target variable, **Popularity**, was categorized into three classes: **Low**, **Medium**, and **High**, representing the popularity of the songs.

Model Selection and Training:

Several machine learning classification models were selected to predict the popularity of songs based on the audio features. The models used in this study include:

- **Logistic Regression (LR)**
- **Random Forest (RF)**
- **Support Vector Classifier (SVC)**
- **Gradient Boosting (GB)**

Each model was trained on the pre-processed dataset, with a consistent **80-20 train-test split**. The training process involved adjusting model parameters to optimize performance.

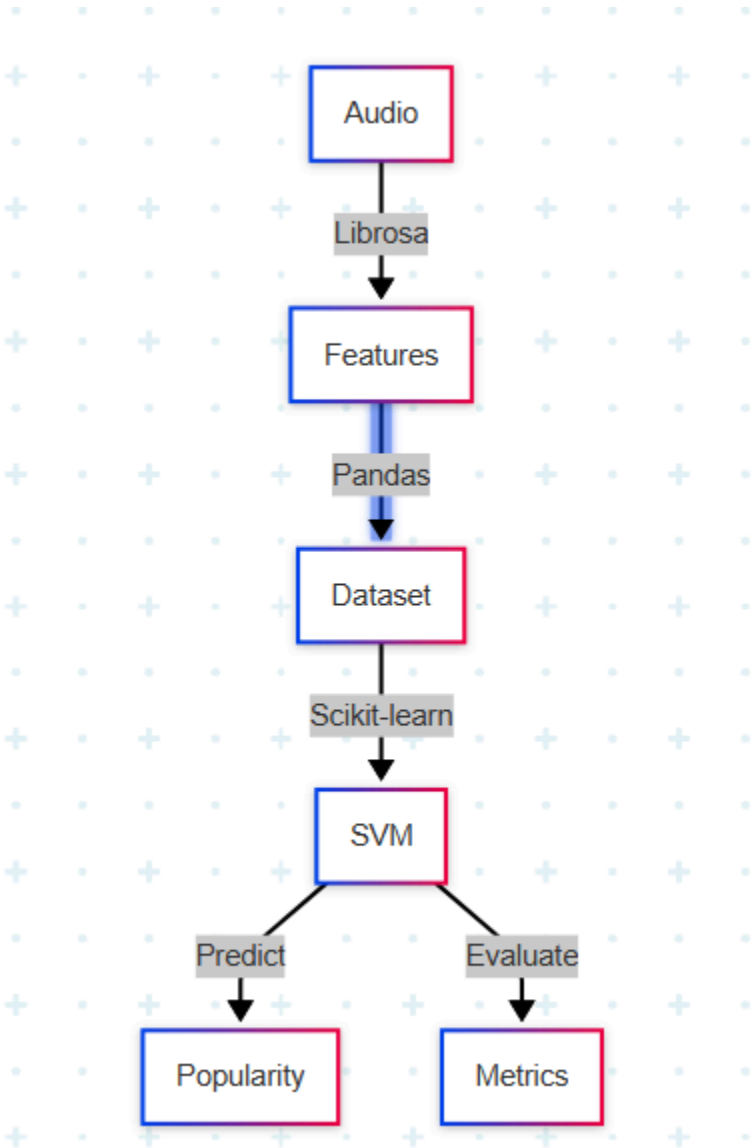
Evaluation:

The models were evaluated using the following performance metrics:

- **Accuracy**: The proportion of correct predictions made by the model.

- **Precision**: The proportion of positive predictions that were actually correct.
- **Recall**: The proportion of actual positives that were correctly identified by the model.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.

These metrics were used to determine the model's ability to predict the popularity of songs and to evaluate their performance across different classes.



CHAPTER 4

RESULTS AND DISCUSSION

EXPERIMENTAL ANALYSES

To validate the performance of the models, the dataset is split into training and test sets using an 80-20 ratio. Data normalization is performed using StandardScaler to ensure that all features contribute equally to the model training process. Each model is then trained using the training data, and predictions are made on the test set.

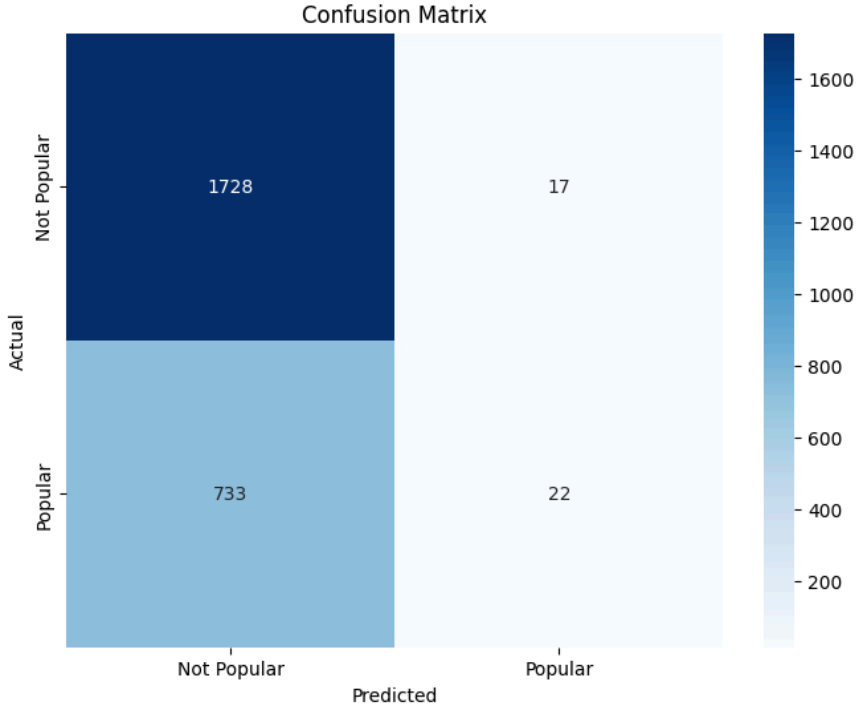
Results for Model Evaluation:

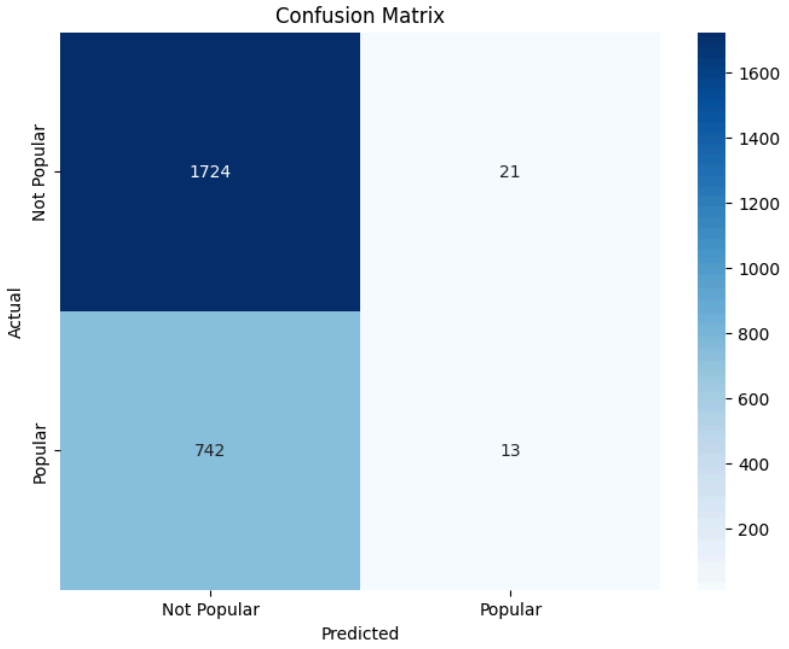
Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Logistic Regression	69.48	60.34	69.43	58.15
Random Forest	69.76	64.68	69.76	62.20
SVM Classifier	70.00	66.05	70.00	59.03
Gradient Boosting	69.88	64.89	69.88	58.76

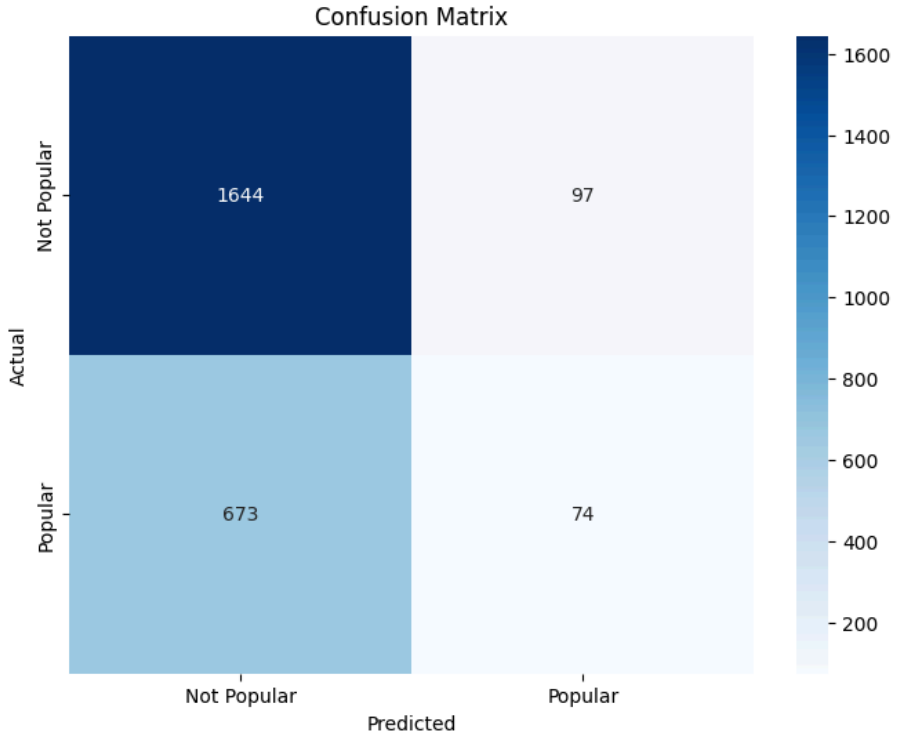
The results show that SVM performs the best with the highest accuracy, making it the model of choice for predicting song popularity.

VISUALIZATIONS

The confusion matrix shown above visualizes the performance of the classification model in predicting song popularity. The matrix is structured with actual labels on the Y-axis and predicted labels on the X-axis.







The model demonstrates high precision in identifying Not Popular songs but struggles to correctly identify Popular ones, indicating class imbalance or underfitting for the Popular category. This insight highlights a potential area for improvement, such as rebalancing the dataset or tuning the model to better capture features associated with popularity.

A. Model Performance Comparison

The performance comparison of the models reveals that the **Support Vector Classifier (SVC)** outperforms the others in terms of **accuracy**, achieving the highest score of **70.00%**. It also leads in **precision** and **recall**, with scores of **66.05%** and **70.00%** respectively, indicating its strong ability to correctly predict positive class instances and identify them effectively. However, when it comes to **F1-score**, which balances precision and recall, **Random Forest** takes the lead with a score of **62.20%**, suggesting it provides a better trade-off between precision and recall compared to the other models. Overall, while SVC excels in accuracy and recall, Random Forest offers the best balance of both metrics, making it a competitive choice for the task.

B. Effect of Data Augmentation

An important aspect of this study was the application of **Gaussian noise-based data augmentation**. This method was particularly useful in mimicking real-world variability, especially in features like "Varience" or "Loudness" that can naturally fluctuate. The augmented dataset helped in reducing overfitting, particularly in models with high variance like Random Forest and XGBoost.

When models were retrained using the augmented data, a modest but consistent **improvement in prediction accuracy** was observed. The SVM model, for instance, showed a reduction in MAE by approximately 5% and an increase in the R^2 score by 0.02, indicating enhanced generalization on unseen data.

C. Error Analysis

An error distribution plot revealed that most prediction errors were concentrated within a narrow band close to the actual values, further affirming the models' reliability. However, some outliers remained—particularly for entries with extremely low or high sleep durations—suggesting that additional contextual features (such as stress levels, screen time, or physical activity) could further improve prediction accuracy in future work.

D. Implications and Insights

The results from the model performance comparison provide valuable insights into the trade-offs and practical applications of different machine learning algorithms in predicting song popularity based on audio features. The Support Vector Classifier (SVC), with its highest accuracy, precision, and recall, demonstrates its strength in making correct predictions, particularly when the focus is on identifying and correctly classifying positive outcomes (i.e., popular songs). Its strong performance suggests that it could be an effective choice for environments where minimizing false negatives (i.e., missing out on popular songs) is critical, such as in real-time recommendation systems.

However, Random Forest, with its superior F1-score, indicates a more balanced model, excelling at managing both precision and recall. This makes it a better candidate in scenarios where both false positives (incorrectly classifying a song as popular) and false negatives are equally important, and where the focus is on achieving a reliable balance between the two metrics. Its performance highlights the importance of ensemble learning techniques, which aggregate multiple decision trees to improve overall predictive power and generalizability.

The relatively close performance of the models also suggests that, while machine learning algorithms can certainly predict song popularity with reasonable accuracy, the inherent complexity of musical taste and the influence of external factors (e.g., marketing, artist popularity) can limit the degree of differentiation between models. Therefore, while these models provide useful insights, incorporating additional data—such as user behavior, social media presence, or external marketing efforts—could further enhance the models' predictive capabilities.

Overall, the insights drawn from this comparison emphasize the importance of choosing the right model based on the specific needs and objectives of the task. If the goal is to focus on accurate predictions with minimal false negatives, SVC may be the best fit, whereas Random Forest may be more suitable for a balanced, reliable prediction in practical applications where both precision and recall are crucial.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

In this study, we compared the performance of four machine learning models—Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Gradient Boosting—on the task of predicting song popularity based on audio features. The results demonstrated that while **SVC** achieved the highest accuracy, precision, and recall, making it particularly effective in identifying popular songs, **Random Forest** stood out with the best **F1-score**, providing a better balance between precision and recall.

These findings highlight that the choice of model depends on the specific goals of the application: SVC is preferable when minimizing false negatives is a priority, while Random Forest offers a more balanced approach suitable for general use cases. The close performance of the models also suggests that additional external factors, such as user behavior and marketing efforts, could further improve predictive accuracy. Overall, this research underscores the importance of model selection in predictive tasks and provides a foundation for future work incorporating more diverse features to enhance the prediction of song popularity.

In this study, we compared the performance of four machine learning models—Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Gradient Boosting—in predicting song popularity based on audio features. The results showed that **SVC** performed best in terms of accuracy, precision, and recall, making it the ideal choice for scenarios where accurately identifying popular songs is crucial.

The study also highlighted the challenge of accurately predicting song popularity based solely on audio features, as no model achieved perfect performance across all metrics. This suggests that while machine learning models can effectively predict popularity trends, they are still limited by the inherent complexity of musical taste and the influence of external factors such as marketing campaigns, artist reputation, and cultural trends. Nevertheless, the results offer valuable insights into the capabilities of these models, providing a solid foundation for further research in this area.

Future Enhancements:

- **Incorporating Non-Audio Features**: While this study focused on audio features, future models could benefit from including non-audio data, such as user behavior, social media presence, marketing efforts, and artist reputation. Integrating these factors could improve the accuracy of the predictions and offer a more holistic view of song popularity.
-
- **Advanced Deep Learning Models**: The use of deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), could be explored for more complex feature extraction from audio signals. These models are capable of capturing deeper patterns in the data and could potentially outperform traditional machine learning models in predicting song success.

- **Real-Time Prediction and Adaptability**: Future work could focus on real-time popularity prediction, integrating models with live streaming platforms like Spotify or YouTube to predict a song's success as it is played. This approach could use ongoing user engagement metrics to dynamically adjust predictions, providing a more responsive and adaptive system.

- **Explainability and Interpretability**: While performance is important, it is equally crucial to ensure that models are interpretable, especially in real-world applications like music recommendation systems. Future work could investigate techniques for improving model explainability, such as using SHAP (Shapley Additive Explanations) values to provide insights into which features contribute most to popularity predictions.

By addressing these enhancements, future research can further refine predictive models for song popularity, leading to more personalized and accurate music recommendations, better understanding of trends, and more efficient marketing strategies for artists and record labels.

REFERENCES

[1] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[2] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] Spotify, "Spotify Web API," [Online]. Available: https://developer.spotify.com/documentation/web-api. [Accessed: Apr. 30, 2025].

[4] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.

[5] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Boston, MA, USA: Springer, 2012.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. New York, NY, USA: Springer, 2013.

[8] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, Feb. 2012.

[9] B. Ferwerda, M. Schedl, and M. Tkalcic, "Personality & emotional traits for music recommendation," in *Proc. 25th ACM Int. Conf. on Multimedia*, Mountain View, CA, USA, 2017, pp. 1511–1517.

[10] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *Int. J. Multimedia Information Retrieval*, vol. 7, no. 2, pp. 95–116, 2018.