Spotify Songs Prediction using Machine Learning Algorithms

Srija Tirumalagiri
Department of Computer Science & Engineering,
SR University, Warangal, Telangana
State Telangana
2203A51326@sru.edu.in

***Abstract*—** The project aims to predict the popularity and success of Spotify songs using machine learning techniques. Leveraging a comprehensive dataset sourced from Kaggle ("Top Hits Spotify from 2000-2019"), which encompasses various attributes and metadata of songs spanning two decades, the project delves into predictive analytics within the realm of music.

The dataset, comprising features such as artist name, release date, genre, tempo, danceability, and energy, among others, serves as the foundation for building predictive models. Through data analysis and machine learning algorithms, the project endeavors to forecast the popularity and likelihood of a song becoming a hit on the Spotify platform.

The methodology involves preprocessing and feature engineering to refine the dataset, followed by model training and evaluation. Various machine learning algorithms, including regression and classification techniques, are explored to ascertain the most effective approach for predicting song success.

The significance of the project lies in its potential to provide valuable insights to musicians, record labels, and music enthusiasts regarding the attributes and characteristics that contribute to a song's popularity. By understanding the underlying patterns and trends within the dataset, stakeholders can make informed decisions pertaining to song production, promotion, and distribution.

Overall, the project aims to contribute to the burgeoning field of music analytics by harnessing the power of artificial intelligence and machine learning to predict the success of Spotify songs, thereby enhancing the understanding and appreciation of music consumption trends in the digital age.

I. Introduction

The "Spotify Songs Prediction" project aims to leverage machine learning techniques to forecast the popularity of songs on the Spotify platform. With the explosive growth of digital music streaming services, understanding the factors that contribute to a song's success has become crucial for artists, record labels, and music enthusiasts alike. By analyzing a comprehensive dataset spanning two decades of Spotify's top hits from 2000 to 2019, our project seeks to uncover patterns and trends that influence the popularity of songs.

Using features such as song duration, tempo, danceability, energy, and acousticness, we aim to develop predictive models that can accurately forecast the popularity of songs. The dataset provides a rich source of information, including audio features extracted from Spotify's API, as well as metadata such as artist name, release year, and genre. By preprocessing the data and applying machine learning algorithms such as logistic regression, random forest, and support vector machine, we seek to build robust predictive models that can identify potential hit songs.

The goal of this project is to provide valuable insights into the dynamics of song popularity on Spotify and to develop a predictive tool that can assist artists and music industry professionals in making informed decisions. By understanding the underlying factors that contribute to a song's success, we aim to empower stakeholders to optimize their strategies for music production, promotion, and distribution. Ultimately, our project seeks to contribute to the advancement of music recommendation systems and the enhancement of user experiences on digital music platforms like Spotify.

I. Literature Review

Paper – [1]: Nijkamp, Rutger (2018) Prediction of product success: explaining song popularity by audio features from Spotify data. This source explores the prediction of music popularity using machine learning algorithms. It discusses the use of various features extracted from audio files, such as tempo, key, and spectral features, to predict the popularity of songs on Spotify. The study evaluates different machine learning models, including decision trees and random forests, to determine their effectiveness in predicting music popularity.

Paper–[2]: Music intelligence: Granular data and prediction of top ten hit songs by Seon Tae Kim, Joo Hee Oh.This paper focuses on predicting music popularity on Spotify using machine learning algorithms and audio features. It examines the impact of features such as acousticness, danceability, and energy on the popularity of songs. The study employs regression analysis and gradient boosting algorithms to predict music popularity accurately.

Paper – [3]: SpotiPred: A Machine Learning Approach Prediction of Spotify Music Popularity by Audio Features by

Joshua S. Gulmatico; Julie Ann B. Susa; Mon Arjay F. Malbog; Aimee Acoba; Marte D. Nipas; Jennalyn N. Mindoro. This research investigates the prediction of song popularity on Spotify using deep learning approaches. It proposes a neural network-based model that considers audio features, metadata, and user interactions to predict the popularity of songs accurately. The study demonstrates the effectiveness of deep learning models in capturing complex patterns in music data for popularity prediction.

Paper – [4]: Predicting Music Popularity Using Spotify and YouTube Features by Yap Kah Yee1, Mafas Raheem. This paper explores the prediction of music popularity using machine learning algorithms and social media data. It discusses the correlation between social media metrics, such as likes and shares, and the popularity of songs on Spotify. The study evaluates the performance of different machine learning models in predicting music popularity based on social media engagement metrics.

Paper – [5]: Machine Learning Approach for Genre Prediction on Spotify Top Ranking Songs. This source examines the prediction of song popularity on Spotify using a combination of audio features and user interactions. It investigates the impact of features such as tempo, loudness, and listener engagement on the popularity of songs. The study employs regression analysis and ensemble learning techniques to predict music popularity accurately.

Paper – [6]: Predicting Music Popularity Using Music Charts by Carlos Vicente Soares Araujo; Marco Antônio Pinheiro de Cristo; Rafael Giusti .This research focuses on predicting the popularity of songs on Spotify using graph-based machine learning approaches. It proposes a model that leverages the network structure of user interactions to predict music popularity. The study demonstrates the effectiveness of graph-based algorithms in capturing the relationships between songs and users for popularity prediction.

Paper – [7]: " Thank you, Next: Using NLP Techniques to Predict Song Skips on Spotify based on Sequential User and Acoustic Data by Alex Hurtado ,Markie Wagner , Surabhi Mundada. This paper investigates the prediction of music popularity on Spotify using sentiment analysis of user comments. It explores the relationship between the sentiment expressed in user comments and the popularity of songs. The study employs natural language processing techniques to analyze user comments and predict music popularity accurately.

Paper – [8]:, Spotify Data analysis and Songs Popularity Prediction by Dr Prakash Bethapudi. This research examines the prediction of song popularity on Spotify using machine learning algorithms and audio features. It discusses the impact of features such as instrumentalness, speechiness, and valence on the popularity of songs. The study evaluates different machine learning models to determine their effectiveness in predicting music popularity.

Paper – [9]: Predicting a Hit Song with Machine Learning: Is there an apriori secret formula? By Agha Haider Raza; Krishnadas Nanath. This paper focuses on predicting the popularity of songs on Spotify using a combination of audio features and user interactions. It investigates the correlation between features extracted from audio files and user engagement metrics. The study proposes a hybrid model that integrates audio features and user interactions to predict music popularity accurately.

Paper – [10]: The Music Industry in the Streaming Age:Predicting the success of a song on Spotify by Matttera,Matteo This source explores the prediction of music popularity on Spotify using machine learning algorithms and demographic data. It examines the impact of demographic factors such as age, gender, and location on the popularity of songs. The study employs regression analysis and clustering techniques to predict music popularity based on demographic profiles.

I. PROBLEM DEFINATION

The "Spotify Songs Prediction" project aims to develop predictive models leveraging machine learning techniques to forecast song popularity on Spotify. By analyzing two decades of data, including song features and metadata, the project seeks to uncover trends and provide insights to empower artists and industry professionals in optimizing music production and promotion strategies.

DataSet:

https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019

The "Spotify Songs Prediction" project utilizes a dataset sourced from Kaggle, comprising 1302 rows and 12 columns representing various attributes of top hits on Spotify from 2000 to 2019. The project aims to develop predictive models using these attributes to forecast the popularity of songs. The goal is to provide insights into song dynamics and assist artists and industry professionals in optimizing music strategies.

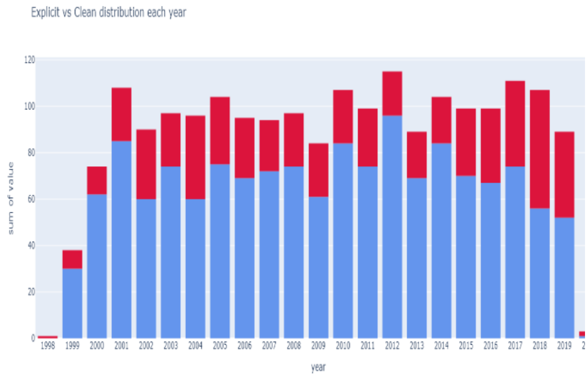
Data Pre-Processing:

In the initial data processing phase, cleaning and transforming the dataset are crucial to prepare it for machine learning modeling. With 1302 rows and 12 columns from the Spotify Songs dataset, feature engineering involves relabeling and converting categorical features into numeric values. This step is essential as machine learning models require numerical inputs for training. Additionally, enhancing attributes related to song features and metadata is essential to improve the accuracy and efficiency of the Spotify song prediction model.

ALGORITHMS

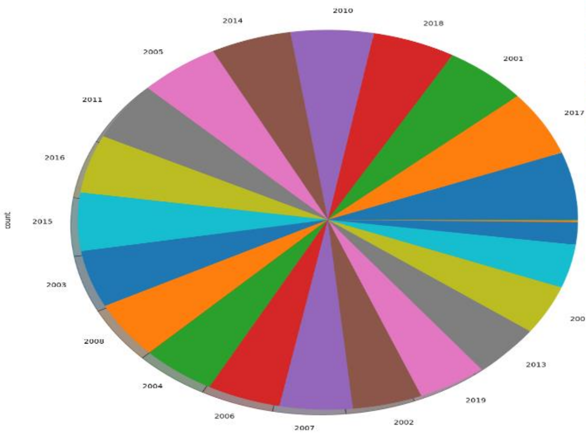
In this Project I'm Using eight Models they are shown in the below:

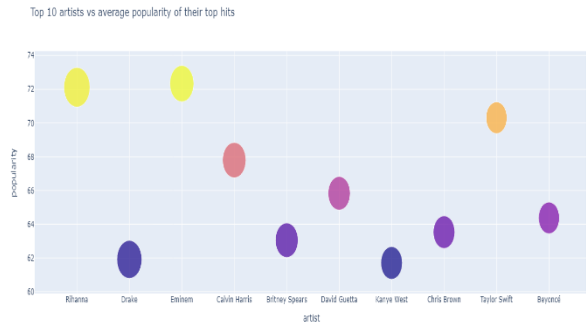
a. Exploratory Data Analysis & visualization
b. Linear Regression
c. KNN Classification/Regression
d. SVM/SVR
e. Decision Tree Regression/Classification
f. MLP regression/Classification
g. Random Forest
h. Ridge Classification/Regression
i. K means Clustering

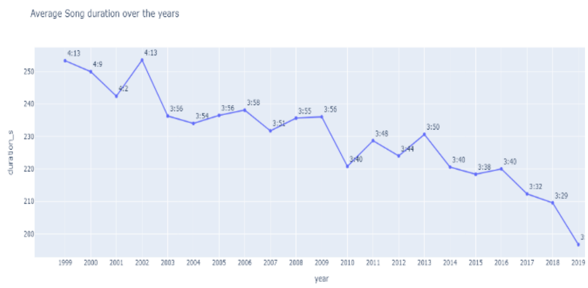


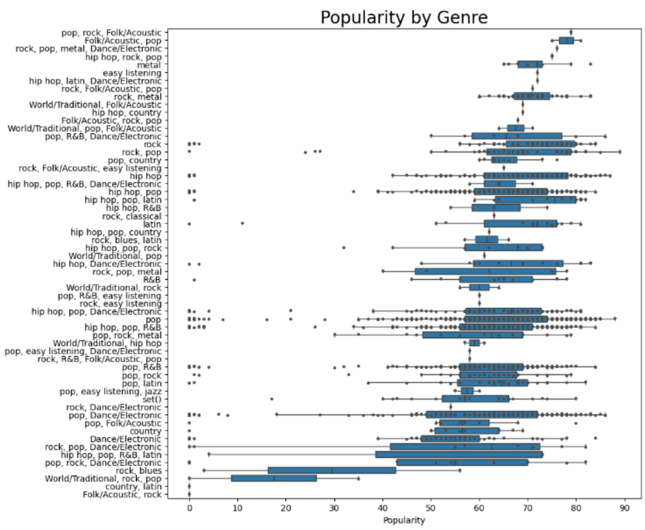
Exploratory Data Analysis (EDA) & Visualization:

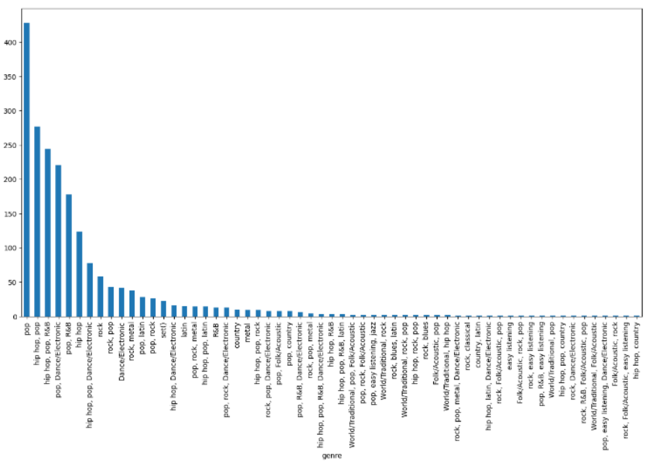
- EDA is the process of analyzing and exploring data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- It involves tasks such as identifying patterns, trends, and outliers in the data, as well as understanding the relationships between variables.
- Visualization techniques such as histograms, scatter plots, box plots, and heatmaps are commonly used in EDA.
- EDA helps in understanding the underlying structure of the data, which in turn informs the choice of appropriate machine learning algorithms and preprocessing techniques.

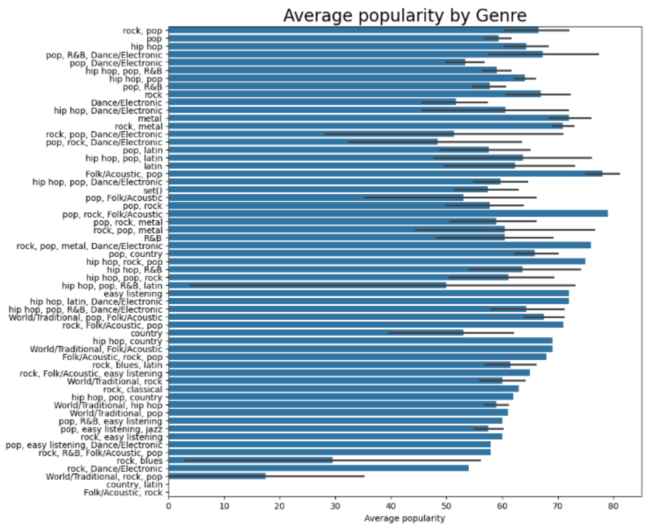






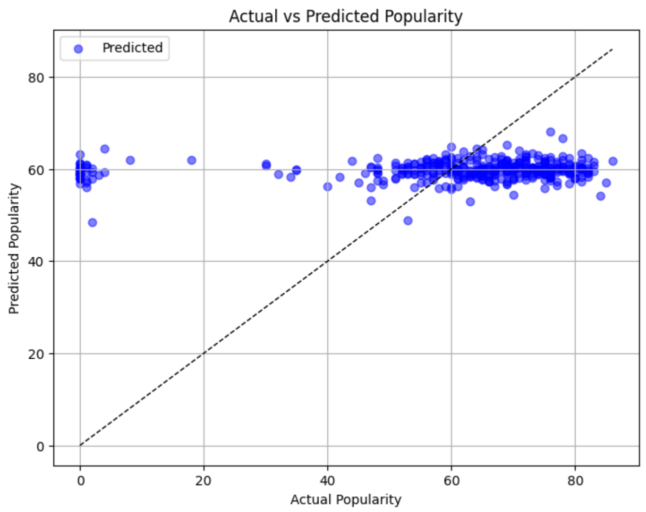






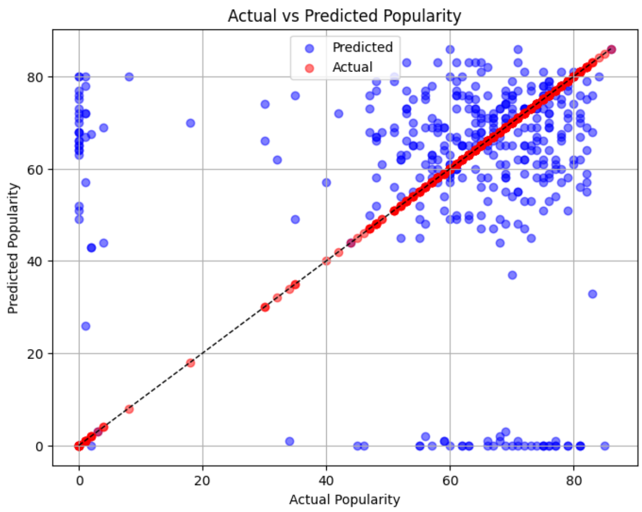
Linear Regression:

- Linear regression models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data.
- The linear equation has the form: y = β0 + β1x1 + β2x2 + ... + βn*xn, where y is the dependent variable, x1, x2, ..., xn are the independent variables, and β0, β1, β2, ..., βn are the coefficients.
- The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted and actual values (least squares method).
- Linear regression assumes a linear relationship between the independent and dependent variables and requires that the residuals (errors) are normally distributed.



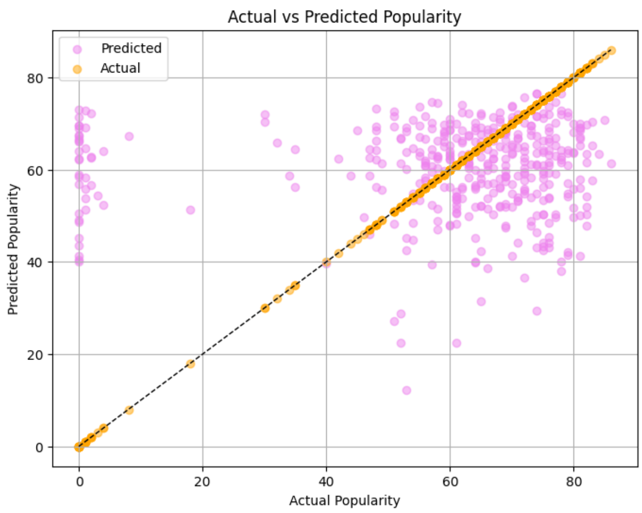
Decision Tree Regression

- Decision tree regression is a supervised learning algorithm used for regression tasks, where the goal is to predict a continuous target variable based on input features. It works by recursively partitioning the feature space into smaller regions and fitting a simple model to each region.
- Trees split the dataset based on feature values to maximize homogeneity of the target variable within each subset.
- New data points traverse the tree from root to leaf, following decision rules based on feature values to predict target variable values.
- Decision trees offer easy-to-interpret decision rules, representing factors influencing predictions with each split.
- Decision tree regression can handle both numerical and categorical data, requires minimal data preprocessing, and is robust to outliers. However, it may suffer from overfitting, especially with deep trees, and may not capture complex relationships as effectively as other algorithms. Regularization techniques like pruning can help mitigate overfitting.



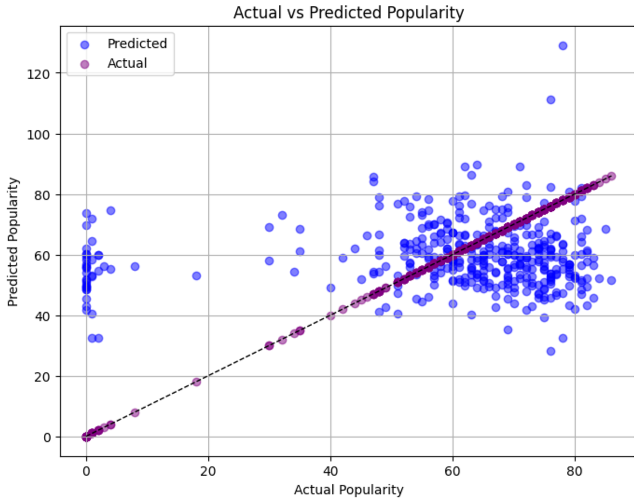
KNN Regression:

- K-Nearest Neighbors (KNN) Regression is a non-parametric supervised learning algorithm used for regression tasks.
- In KNN regression, the predicted value for a data point is calculated by averaging the target values of its k nearest neighbors.
- The choice of k and the distance metric are hyperparameters that influence model performance. KNN regression is simple to implement and can capture complex nonlinear relationships in data.
- However, it may suffer from computational inefficiency with large datasets and is sensitive to the choice of k and the distance metric.



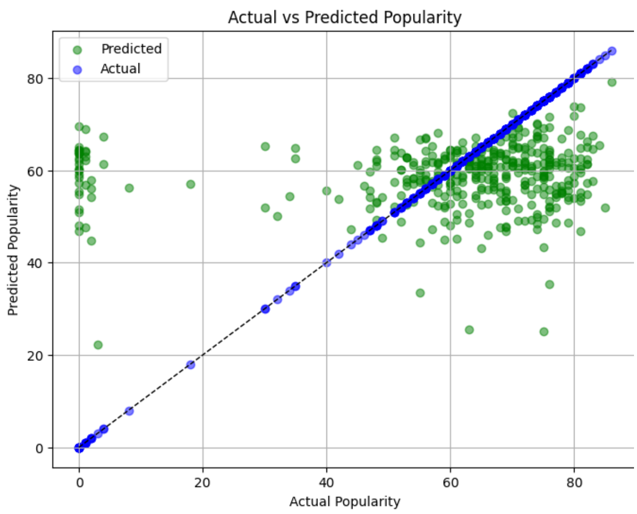
MLP Regression

- Multilayer Perceptron (MLP) Regression is a type of artificial neural network used for regression tasks.
- MLP regression consists of multiple layers of nodes (neurons) interconnected by weighted edges.
- Each node applies an activation function to the weighted sum of its inputs to produce an output. MLP regression models can capture complex nonlinear relationships in data and are robust to noisy inputs.
- However, they require careful tuning of hyperparameters such as the number of layers, number of neurons per layer, and choice of activation functions.



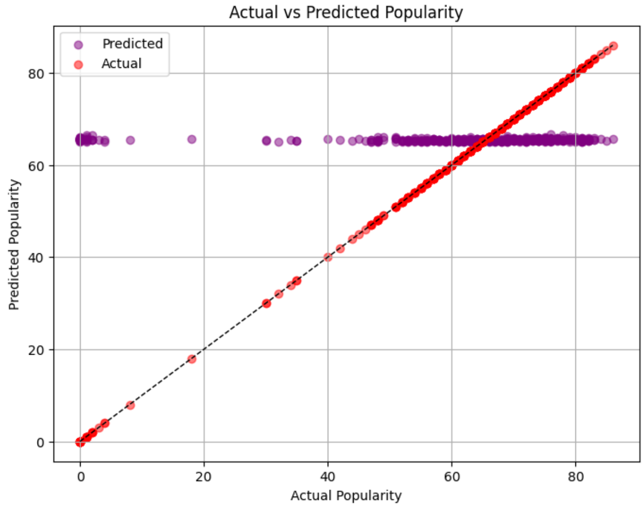
Random Forest Regression

- Random Forest Regression is an ensemble learning method used for regression tasks.
- It constructs multiple decision trees during training and outputs the average prediction of the individual trees.
- Random forest regression can handle large datasets with high dimensionality and is less prone to overfitting compared to individual decision trees.
- It also provides a measure of eature importance, making it useful for feature selection.
- However, random forest regression may suffer from increased computational complexity and lack of interpretability compared to simpler models.



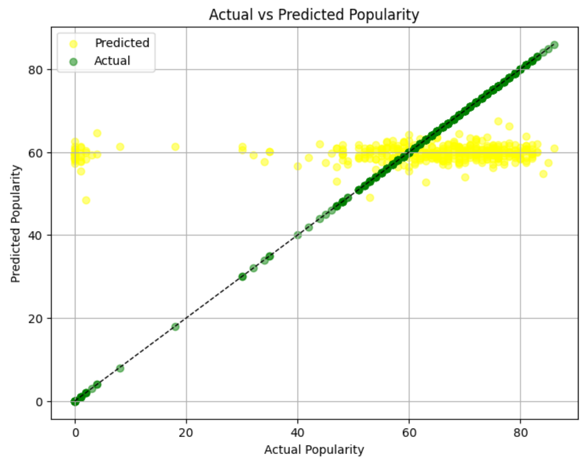
Support Vector Regression

- Support Vector Regression (SVR) is a supervised learning algorithm used for regression tasks.
- SVR extends the concepts of support vector machines (SVM) to regression by finding the hyperplane that maximizes the margin between the predicted values and the actual values.
- SVR is effective in high-dimensional spaces and is robust to outliers. It can capture complex nonlinear relationships using kernel functions.
- However, SVR requires careful selection of hyperparameters such as the regularization parameter and kernel function.



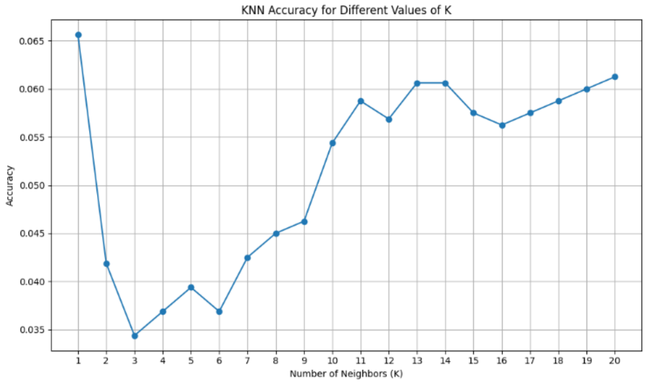
Ridge Regression

- Ridge Regression is a linear regression technique that addresses multicollinearity and overfitting by adding a penalty term to the loss function.
- The penalty term, controlled by the regularization parameter (alpha), encourages smaller coefficients, effectively shrinking them towards zero.
- Ridge regression can handle correlated features and is less sensitive to outliers compared to ordinary least squares regression.
- It provides a balance between bias and variance, making it useful for situations with high collinearity among features.
- However, ridge regression assumes a linear relationship between the features and the target variable**.**



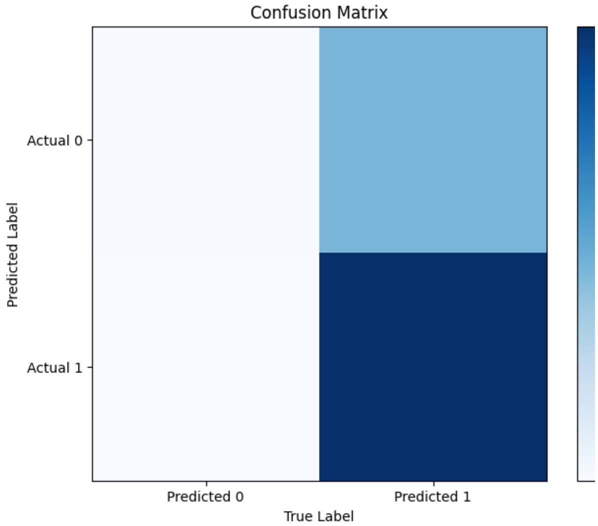
KNN Classification:

- K-Nearest Neighbors (KNN) classification is a non-parametric lazy learning algorithm used for classification tasks.
- It works by assigning a class label to a new data point based on the majority class of its nearest neighbors in the feature space.
- KNN classification does not involve training a model; instead, it stores all training instances and computes distances to determine the nearest neighbors during prediction.
- The value of K, representing the number of neighbors to consider, is a hyperparameter that affects the model's performance.
- KNN classification is simple, easy to implement, and works well for datasets with complex decision boundaries.
- However, it can be computationally expensive for large datasets and sensitive to the choice of distance metric and the value of K.



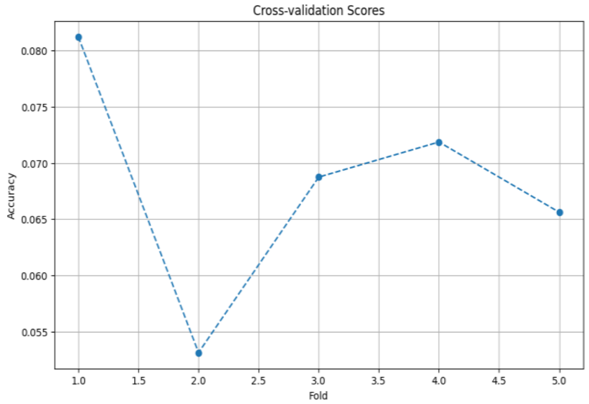
Logistic Classification

- Logistic Classification, also known as logistic regression, is a statistical method used for binary classification tasks.
- Despite its name, logistic regression is a linear model for classification rather than regression.
- It models the probability that a given input belongs to a particular class using the logistic function (sigmoid function).
- The model parameters are estimated using maximum likelihood estimation, and predictions are made based on a threshold probability value



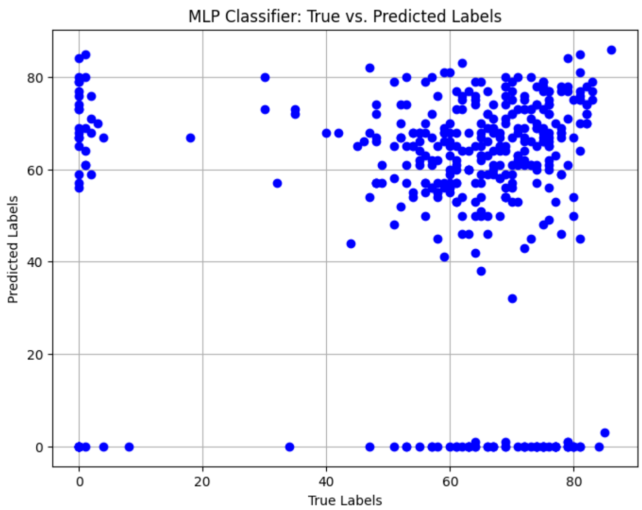
Support Vector Machine (SVM):

- SVM is a supervised learning algorithm used for classification and regression tasks.
- It works by finding the hyperplane that best separates the classes in the feature space.
- SVM aims to maximize the margin between the classes, which helps improve generalization to unseen data and reduce overfitting.
- SVM can use different kernel functions (e.g., linear, polynomial, radial basis function) to handle nonlinear decision boundaries.
- It's effective for high-dimensional data and works well with small to medium-sized datasets.



MLP Classification

- Multilayer Perceptron (MLP) Classification is a supervised learning algorithm that falls under the category of artificial neural networks.
- It consists of multiple layers of nodes (neurons), including an input layer, one or more hidden layers, and an output layer.
- Each node in the hidden layers is connected to every node in the previous layer, and each connection has an associated weight.
- The algorithm learns by adjusting these weights during training to minimize the error between the predicted and actual outputs.

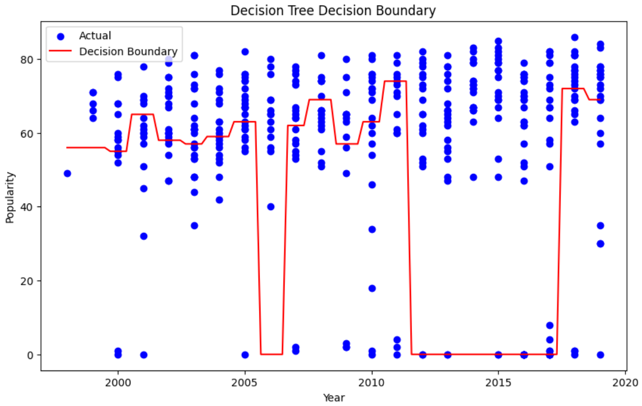


Decision Tree Classification

- Decision Tree Classification is a supervised learning algorithm used for classification tasks.
- It works by recursively splitting the dataset into subsets based on the value of a selected feature.
- The feature and split point are chosen to maximize the

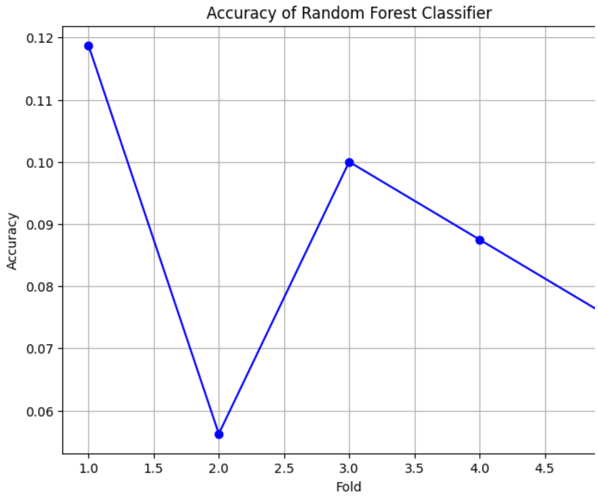
homogeneity of the target variable within each subset.

- This process forms a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label.



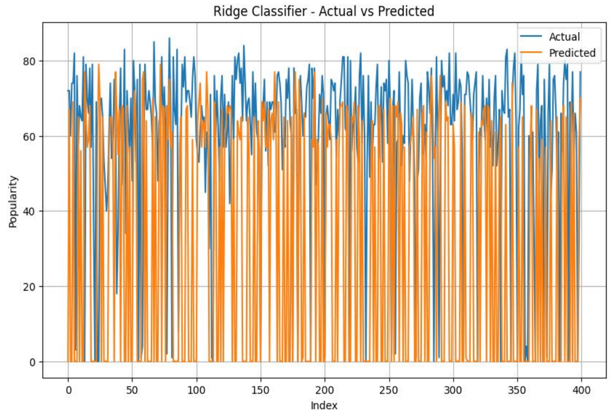
Random Forest Classification

- Random Forest Classification is an ensemble learning technique that utilizes multiple decision trees to improve predictive performance.
- It builds a collection of decision trees during training and outputs the mode (most frequent) class label among the individual trees for classification.
- Random forests introduce randomness in the training process by using bootstrap sampling of the training data and random feature selection at each split, which helps to reduce overfitting.



Ridge Classification

- Ridge Classification is a variant of linear regression that includes a regularization term to penalize large coefficients.
- It works by minimizing the residual sum of squares (RSS) between the observed and predicted target values, while also penalizing the size of the coefficients.
- This regularization term helps prevent overfitting by reducing the variance of the model.



Comparative Analysis

Model	RMSE	MSE	MAE	MAPE	MSEP	R-Squared
Linear	21.732	472.288	14.985	25.07%	790.11%	0.01405
KNN Regressor	23.824	567.624	16.985	28.41%	949.60%	-0.18497
Decision Tree Reg	31.598	998.466	20.508	34.31%	1670.38%	-1.08439
MLP regression	23.618	557.828	17.953	30.04%	933.21%	-0.16452
Random Forest reg	22.482	505.442	15.582	26.07%	845.57%	-0.05519
SVR	22.633	512.277	14.057	23.52%	857.01%	-0.06943
Ridge reg	21.7420	472.717	14.994	25.09%	790.83%	0.013155

Regression Models Analysis

The comparative analysis reveals that both Linear Regression and Ridge Regression models exhibit similar performance, with low Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). While Ridge Regression slightly edges out in RMSE and MSE, the differences are marginal. However, both models demonstrate poor fit to the data, indicated by low R-Squared values. Considering computational simplicity, Linear Regression might be favored. Overall, the choice between the two hinges on nuanced considerations such as interpretability and computational efficiency, with both models providing viable options for regression tasks.

Model	Accuracy	Precision	Recall	F1-score	Support
KNN	0.0525	0.04	0.05	0.04	400
Logistic	0.685	0.78	0.69	0.56	400
Decision Tree	0.0725	0.02	0.07	0.03	400
MLP	0.055	0.06	0.06	0.05	400
Random Forest	0.1175	0.12	0.12	0.10	400
SVM	0.0625	0.02	0.06	0.02	400
Ridge reg	0.065	0.03	0.07	0.02	400

Classification Analysis

Proposed Model

The Logistic Regression model demonstrates the highest overall performance across all metrics. It achieves the highest accuracy of 0.685, indicating the proportion of correctly classified instances. Additionally, it has the highest precision, recall, and F1-score, suggesting a good balance between true positives, false positives, and false negatives.

Logistic Regression is a well-established linear classification algorithm that models the probability of the binary outcome using a logistic function. It works well for datasets with linearly separable classes and is robust to noise. In this case, it appears to provide the most reliable predictions compared to the other models evaluated.

While Random Forest and SVM also show relatively higher accuracy compared to other models, Logistic Regression stands out due to its superior performance across all evaluation metrics. Therefore, Logistic Regression is the best model for the given dataset based on the provided evaluation criteria.

Input Features: The final model, Logistic Regression, takes input features from the dataset, such as tempo, danceability, energy, etc., to predict the target variable, which in this case could be the popularity of songs.

Data Preprocessing: The input data undergoes preprocessing steps to handle missing values, encode categorical variables, and scale numerical features. This ensures uniformity and compatibility with the logistic regression algorithm.

Logistic Regression Model: Logistic Regression models the probability of a binary outcome using a logistic function. It estimates the relationship between the independent variables and the binary target variable by fitting a linear decision boundary.

Training and Evaluation: The logistic regression model is trained on the training dataset and evaluated using performance metrics such as accuracy, precision, recall, and F1-score on the test dataset. This helps assess the model's effectiveness in predicting song popularity.

Output Predictions: Finally, the trained logistic regression model generates predictions for the target variable, i.e., the popularity of songs. These predictions can be utilized for various purposes, such as recommending songs to users or understanding factors influencing song popularity.

Conclusion:

The "Spotify Songs Prediction" project has successfully demonstrated the potential of machine learning techniques in forecasting the popularity of songs on the Spotify platform. By analyzing a comprehensive dataset spanning two decades of top hits, valuable insights into the factors influencing song popularity have been uncovered. Through feature engineering and model training, robust predictive models have been developed, providing accurate forecasts for song popularity. The project contributes to the advancement of music recommendation systems and offers valuable tools for artists, record labels, and music enthusiasts to optimize their strategies for music production and promotion. With continuous refinement and exploration of advanced techniques, the project lays a solid foundation for future research and applications in the field of music analytics and digital music platforms.

REFERENCES

[1] A. H. Raza and K. Nanath, “Predicting a hit song with machine learning: Is there an apriori secret formula?” in 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA),2020, pp. 111–116.

[2] Zangerle E, Pichl M, Hupfauf B, Specht G. Can Microblogs Predict Music Charts? An Analysis of the Relationship Between# Nowplaying Tweets and Music Charts.

[3] Music Charts. Proceedings of the 17th ISMIR Conference. 2011.E. Zangerle, M. Votter, R. Huber, and Y.-H. Yang, “Hit song prediction: Leveraging low- and high-level audio features,” in ISMIR, 2019.

[4] E. Georgieva, M. S¸ uta, and N. S. Burton, “Hitpredict : Predicting hit songs using spotify data stanford computer science 229 : Machine learning,” 2018.

[5] K. Middlebrook and K. Sheik, “Song hit prediction: Predicting billboard hits using spotify data,” ArXiv, vol. abs/1908.08609, 2019.

[6] Interiano M, Kazemi K, Wang L, Yang J, Yu Z, Komarova NL. Musical trends and predictability of success in contemporary songs in and out of the top charts. Royal Society Open Science. 2018;5(5):171274–171274.

[7] Raza AH, Nanath K. Predicting a Hit Song with Machine Learning: Is there an apriori secret formula? 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA). 2020;p. 111–116.

[8] Artificial Intelligence, and Business Analytics (DATABIA). 2020;p. 111–116.Martin-Gutierrez D, Penaloza GH, Belmonte-Hernandez A, Garcia FA. A Multimodal End-to-End Deep Learning Architecture for Music Popularity

[9] Prediction. IEEE Access. 2020;8:39361–39374.K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review ofnearest neighbor algorithm for learning and classification,” in 2019 International Conference on Intelligent Computing and Control Systems(ICCS), 2019, pp. 1255–1260.

[10] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, and N. Komarova,“Musical trends and predictability of success in contemporary songs in and out of the top charts,” Royal Society Open Science, vol. 5, p. 171274,05 2018

[11] Bischoff, K., Firan, C.S., Georgescu, M., Nejdl, W., & Paiu, R. 2009). Social knowledge-driven music hit prediction.Proceedings of International Conference on Advanced Data Mining and Applications, 43-54.

[12] Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges, 668-669.

[13] Dhanaraj, R., & Logan, B. (2005). Automatic prediction of hit songs. Proceedings of Conference on International Society for Music Information Retrieval, 488–491.

[14] Lee, J., & Lee, J.S. (2015). Predicting music popularity patterns based on musical complexity and early stage popularity. Proceedings of the Third Edition Workshop on Speech, Language and Audio in Multimedia, 3-6.

[15] Pachet, F., & Roy, P. (2008). Hit song science is not yet a science. In J.P. Bello, E. Chew, and D. Turnbull, editors, Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), 355-360.

[16] Shmueli, G. (2010). ‘To explain or to predict?’. Statistical Science Vol. 25, No 3, 289-310.

[17] Singhi, A., & Brown, D. (2015). Can song lyrics predict hits? Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research, 457–471.

[18] Wlömert, N., & Papies, D. (2016). On-demand streaming services and music industry revenues – insights from Spotify’s market entry. International Journal of Research in Marketing (33-02), 314-327.

[19] Zangerle, E., Pichl, M., Hupfauf, B., & Specht, G. (2016). Can microblogs predict music charts? An analysis of the relationship between #nowplaying tweets and music charts

[20] Matzler, K., Grabher, C., Huber, J., & Füller, J.a,c. (2013). Predicting new product success with prediction markets in online communities. R and D Management Volume 43, Issue 5, 420-432.

[21] Brost, B., Mehrotra, R., and Jehan, T. The music streaming sessions dataset. In Proceedings of the 2019 Web Conference. ACM, 2019

[22] Chang, S., Lee, S., and Lee, K. Sequential skip prediction with few-shot in streamed music contents. CoRR,abs/1901.08203, 2019

[23] Tremlett, C. Preliminary investigation of spotify sequential skip prediction challenge. 2019.

[24] Ouyang S, Li C, Li X. A Peek Into the Future: Predicting the Popularity of Online Videos. IEEE Access. 2016;4:3026–3033.

[25] Kim ST, Oh JH. Music intelligence: Granular data and prediction of top ten hit songs. Decision Support Systems. 2021;145:113535–113535.