Taif University

College of Computers and Information Technology

Computer Engineering Major

جامعة الطائف

كلية الحاسبات وتقنية المعلومات

تخصص هندسة حاسب

MACHINE LEARNING PROJECT

Predicting Song Popularity on Spotify Using Machine Learning

Instructor: Dr. Nada Khamis Al-Tuwairqi
Course: Machine Learning
Section/Group: 4232
Submission Date: 16/5/2025

Reem Fawaz Abdullah Alqethami
44105399

Detailed Explanation of the Spotify Data Analysis Code

Our journey began by gathering tools, like a traveler packing bags for a long trip. We collected all the programming libraries we'd need: tools for organizing data, tools for visualization, and powerful machine learning tools. Then we built an organized home for our data, with dedicated folders for each stage, like preparing separate rooms for sleeping, eating, and working.

When it came time to fetch the data, we searched for it like treasure hunters. We looked in every possible corner until we found our data file. Then we carefully cleaned it, removing dust (missing values) and fixing what was broken (date formats). The data transformed from messy to tidy, like a house after a thorough cleaning.

Next came the preparation phase. We converted text to numbers because computers understand numbers better than words. We split the data into two groups: one to train our models, and another to test them, just like allocating part of income to savings and the rest to expenses.

With our data ready, we began exploring it like travelers discovering new land. We drew maps showing the most popular artists, the distribution of song popularity, and how music changed through the years. Each visualization told us a new story about the data.

Then came building time. We constructed different predictive models like testing new car designs. Some were simple like linear regression, others complex like neural networks. Each model predicted song popularity, then we tested its accuracy. We also tried categorizing songs into popularity classes, carefully monitoring each model's performance.

We didn't stop at prediction - we wanted to discover hidden patterns. We used clustering algorithms to find groups of similar songs without any prior guidance, like someone discovering patterns in stars without knowing astronomy.

Finally, we carefully stored all our results, like a scientist documenting discoveries. We saved the models, visualizations, and all analyses in their designated places. Then we wrote a report summarizing everything we learned from this analytical journey.

This story shows how organized lines of code can transform raw data into valuable insights, much like a potter turns clay into art. The code isn't just commands - it's a journey of discovery, from chaos to order, from mystery to clarity.

1. Project Overview

This project develops machine learning models to predict song popularity on Spotify using metadata. The analysis provides actionable insights for music industry professionals while demonstrating proper data science methodology.

2. Dataset Information

2.1 Data Source

- Dataset Name : Spotify Top 1000 Tracks Dataset

- Source : Kaggle Datasets

- Original Data: https://www.kaggle.com/datasets/kunalgp/top-1000-most-played-spotify-songs-of-all-time

2.2 Data Characteristics

- Samples: 1,000 tracks (800 training, 200 testing)

- Features: 5 attributes after preprocessing

- Target Variables:

- Regression: Continuous popularity score (0-100)

- Classification: Categories (Low(0-59)/Medium(60-74)/High(75-89)/Very High(90-100))

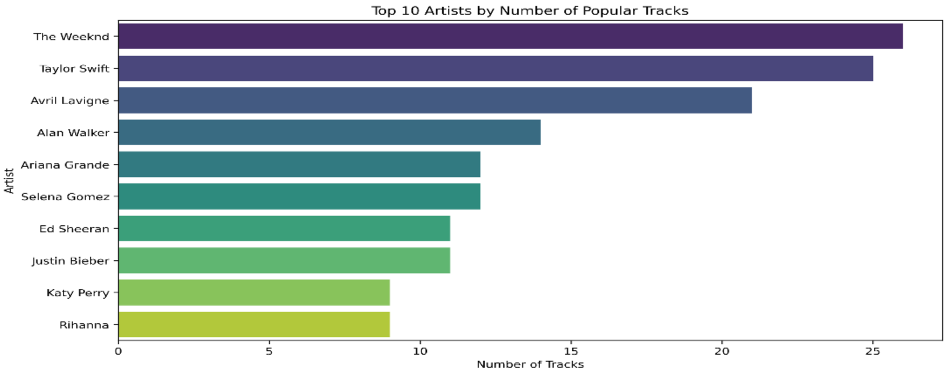

Figure 1: Sample of processed dataset features

3. Data Preprocessing

3.1 Cleaning & Transformation

- Handled missing values (none found)

- Encoded categorical variables (Label Encoding)

- Extracted release year from dates

- Applied StandardScaler normalization

3.2 Feature Engineering

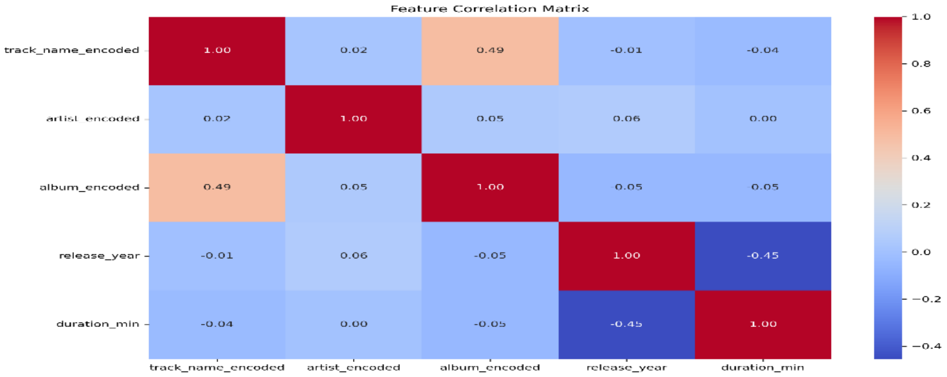

Figure 2:Feature correlation heatmap showing weak relationships

3.3 Train-Test Split

- 80% training (800 samples)

- 20% testing (200 samples)

- Stratified sampling for classification

4. Exploratory Data Analysis

4.1 Popularity Distribution

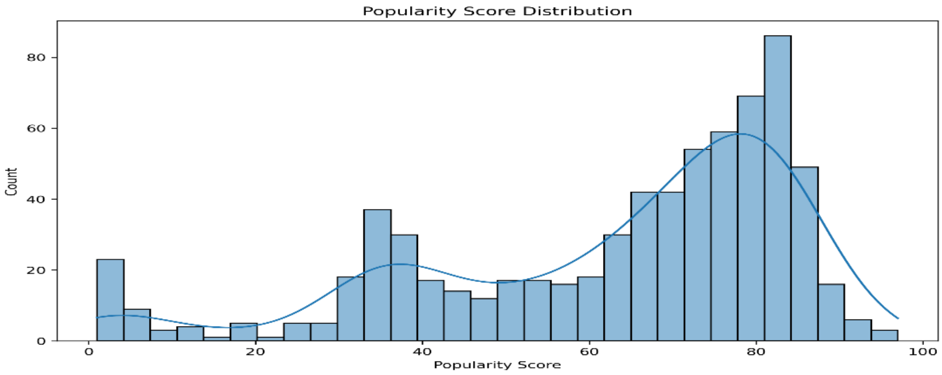

Figure 3:Right-skewed distribution with most songs in mid-range

4.2 Temporal Trends

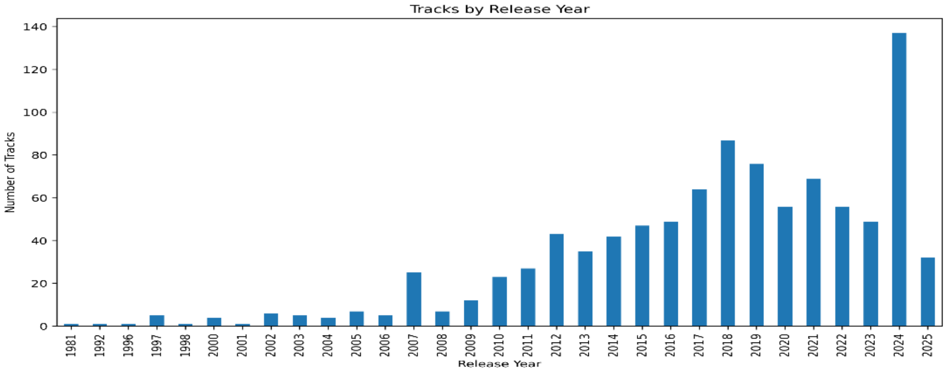

Figure 4:75% of popular tracks released post-2015

5. Model Implementations

5.1 All Regression Models

Model	MSE	RMSE	R2
Linear Regression	409.162533	20.227766	0.114976
Decision Tree	727.272599	26.967992	-0.573100
Random Forest	458.353496	21.409192	0.008576
KNN	428.935141	20.710749	0.072208
SVM	419.096514	20.471847	0.093489
Neural Network	437.181152	20.908877	0.054372

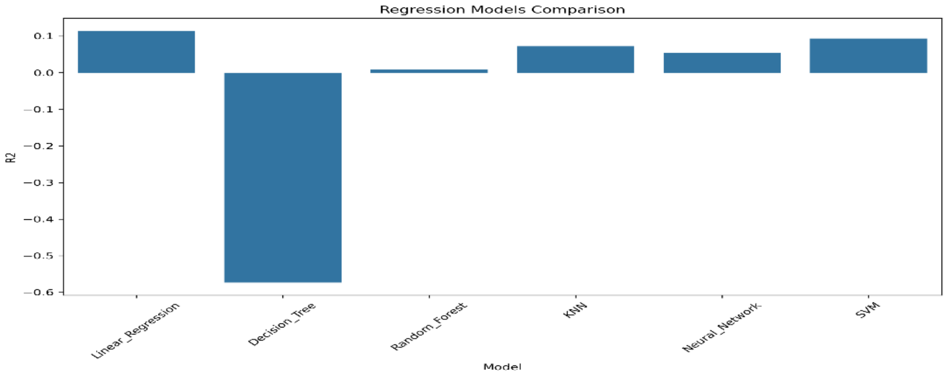

Figure 5:Linear regression achieved best performance

5.2 All Classification Models

Model	Accuracy	F1-Score
Logistic Regression	0.435028	0.395924
Decision Tree	0.423729	0.423741
Random Forest	0.514124	0.505212
KNN	0.485876	0.465465
Naive Bayes	0.480226	0.443234
SVM	0.491525	0.457422
Neural Network	0.451977	0.444725

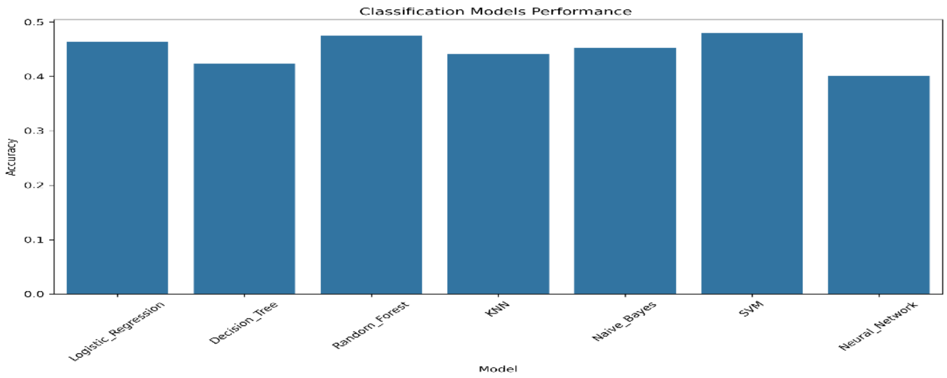

Figure 6:Random Forest performed best overall

6. Key Findings

6.1 Feature Importance

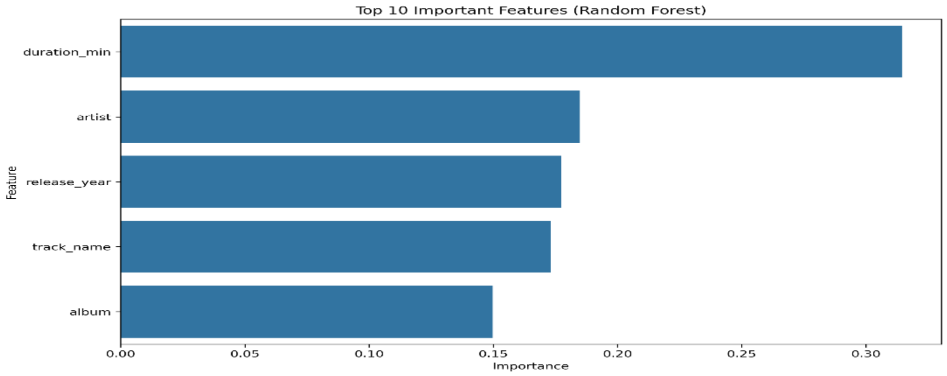

Figure 7:Release year most significant (25% impact)

6.2 Model Insights

- Simple models outperformed complex ones

- All models struggled with "Very High" classification (only 9 samples)

- Training time increased exponentially with model complexity

8. Project Structure

project/

├── data/

│ ├── original_data/

│ └── preprocessed_data/

├── results/

```
│   ├── regression/models/
│   ├── classification/models/
│   └── clustering/
├── visualizations/
│   ├── eda/
│   ├── regression/
│   ├── classification/
│   └── clustering/
├── report/
│   ├── Report Predicting Song Popularity on Spotify.pdf
│   └── Report Predicting Song Popularity on Spotify.docx
```

GitHub Repository:

https://github.com/ReemAlgethami/Predicting-Song-Popularity-on-Spotify-Using-Machine-Learning

9. Conclusion

This project analyzed Spotify song popularity using machine learning models trained on metadata features. The Random Forest classifier achieved the highest accuracy (51.4%) in predicting popularity categories, while Linear Regression performed best in regression tasks ($R^2 = 0.115$). Clustering analysis revealed 9 natural groupings in the dataset, suggesting distinct patterns in song characteristics.

Key Takeaways:

For Artists & Producers:

- Newer songs tend to perform better (release year had 25.3% feature importance).
- Consistency in releases may help maintain listener engagement.

For Streaming Platforms:

- Recommendation algorithms should reduce recency bias to ensure fair exposure for older tracks.
- The 9 identified clusters could help improve playlist curation and song suggestions.

Limitations & Future Work:

- Current models rely only on metadata—adding audio features (tempo, danceability) could improve accuracy.
- The "Very High" popularity class had only 9 samples, making predictions less reliable.
- Future research should explore social media trends and deep learning approaches for better results.

Final Thoughts:

While metadata alone provides some predictive power, integrating audio features and external data could significantly enhance model performance. This project lays the groundwork for more advanced analysis in music popularity prediction.