Predicting Song Popularity Using Machine Learning

Group - 2

NAING NAING (AIU 23102101)
Khalid Ghaleb (AIU 22102089)
MYAT MIN KHANT (AIU 23102103)
MOHAMED NAWRAN (AIU 23102141)

1 Introduction

The music streaming industry generates vast data on songs and user interactions, yet predicting which tracks will achieve popularity remains challenging. This project leverages machine learning to analyze audio features and uncover patterns that drive song success, empowering artists, record labels, and streaming platforms to optimize recommendations, marketing, and curation.

2 Problem Statement

Predicting song popularity is complex due to:

- Unclear relationships between audio features (tempo, energy, danceability) and chart performance
- Limited actionable insights for data-driven decisions in playlist curation and artist promotion
- Inefficient resource allocation in the music industry due to unpredictable hit potential

3 Problem Objectives

1. Identify Domain Behavior
   - Analyze audio features (danceability, energy, valence, etc.) influencing popularity
   - Investigate correlations between features and Billboard chart success
   - Characterize differences between popular vs. non-popular songs
2. Predict Domain Behavior
   - Develop ML models to classify songs as "popular" or "not popular"
   - Forecast potential Billboard rankings using audio features
   - Identify audio feature trends linked to emerging hits

4 Data Preprocessing

- Missing Values
- Feature Selection
  - Supervised: Used Random Forest

Kept top audio features, removed weak ones

  - Unsupervised: Selected 9 numeric audio features

K-Means relies on distance (Euclidean) categorical features like genre can distort results
Focused on sound-based features for meaningful, interpretable clusters

- Feature Scaling

Used StandardScaler (Z-score)
Ensures fair distance calculation

- Encoding

Not needed — no categorical features used

- Data Splitting

Hot100: 80% training / 20% testing
Spotify: No split (unsupervised)

Summary:

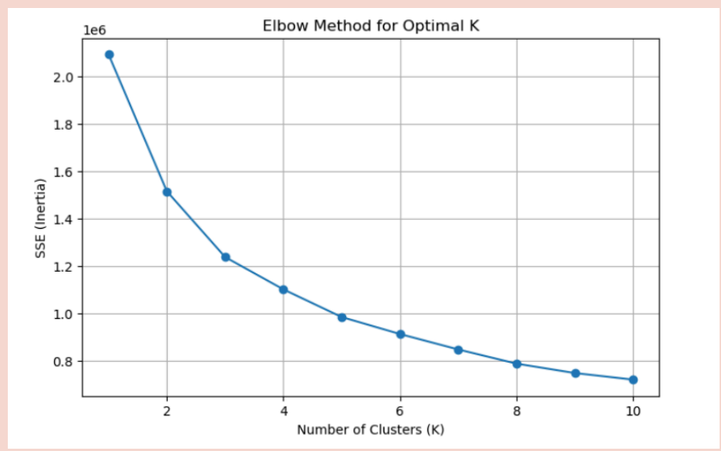
- Supervised learning (Hot100) required feature selection and splitting for prediction.
- Unsupervised learning (Spotify) focused on clean numeric features for clustering.

Features were selected based on domain knowledge, not encoding, to maintain clustering accuracy.

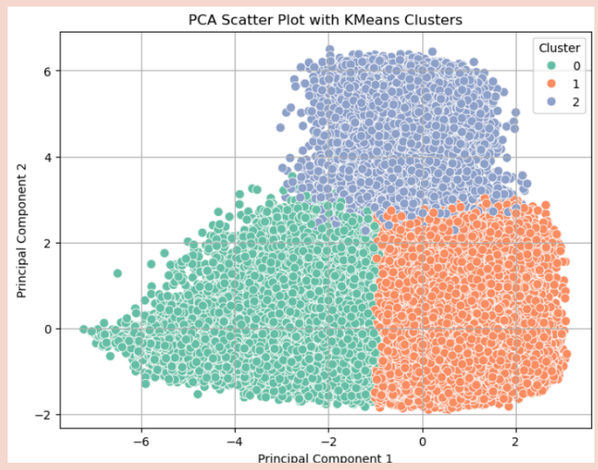
5 Unsupervised Learning

Model Used: KMeans Clustering (k = 3),
Silhouette Score : 0.3689

- Applied on 9 audio features like energy, danceability, loudness
- Used to group similar songs based on audio features.
- PCA used for 2D cluster visualization



Clusters :
Cluster 0 – (Low energy, acoustic) Passive Listeners
Cluster 1 – (Balanced energy and features) Regular Listeners
Cluster 2 – (High energy, loud, danceable) Active Listeners



6 Supervised Learning (Popularity Prediction)

Data set : hot100
Rows: 620 , Columns : 18
Target column : popularity

Models Used:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

All models were trained on 80% of the data and tested on the remaining 20%. The table below summarizes the evaluation results:

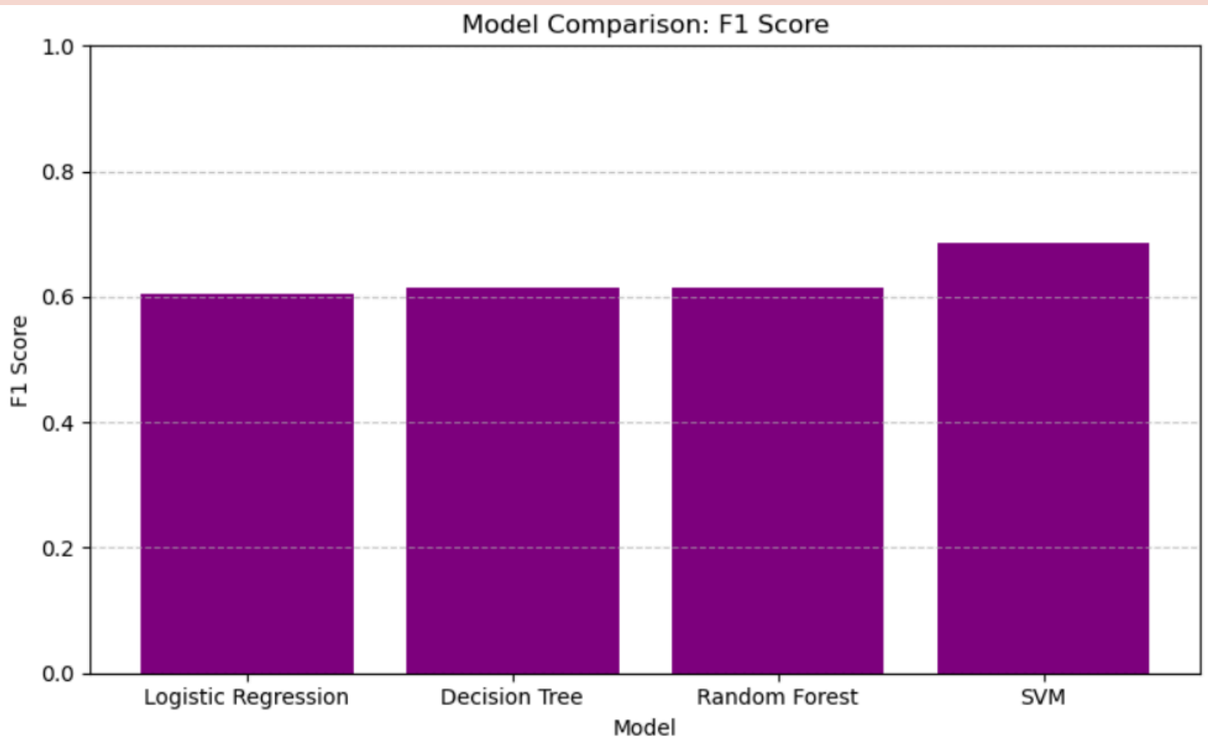
Model Performance Comparison Table:

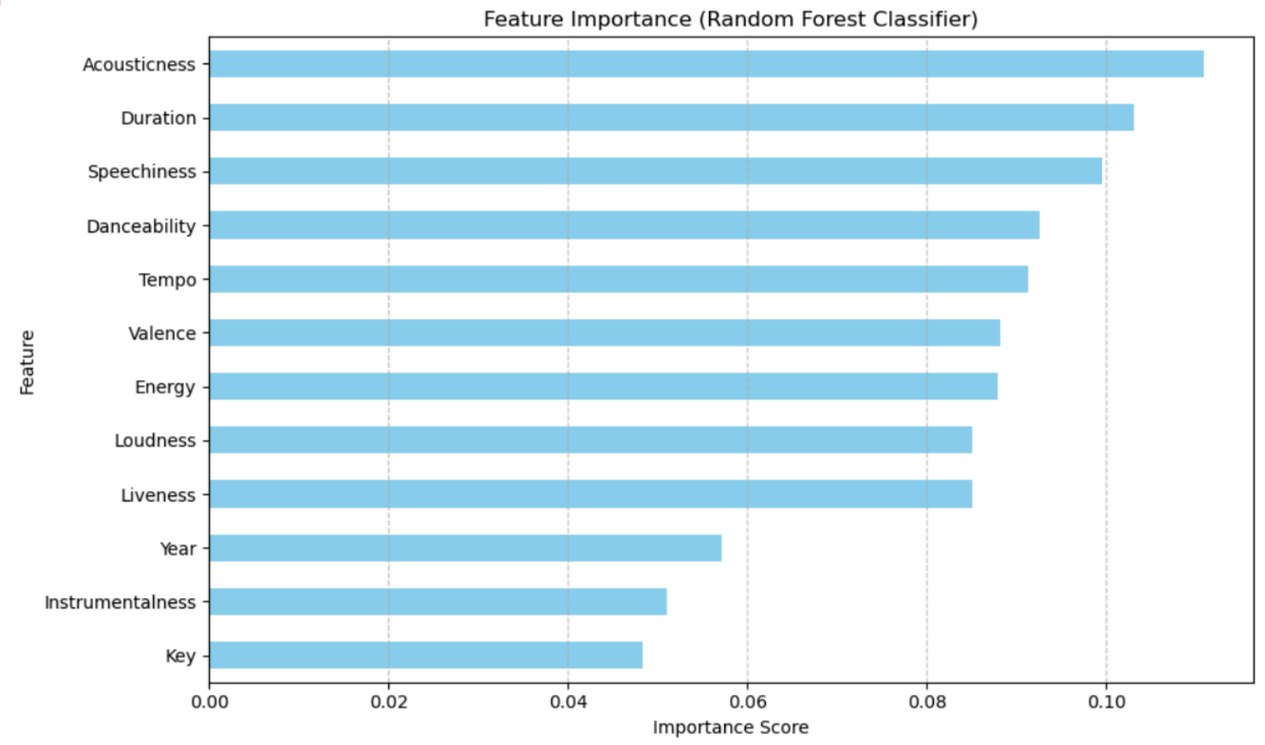
	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	Logistic Regression	0.548387	0.597222	0.614286	0.605634	0.561376
1	Decision Tree	0.564516	0.614286	0.614286	0.614286	0.557143
2	Random Forest	0.564516	0.614286	0.614286	0.614286	0.601323
3	SVM	0.637097	0.671233	0.700000	0.685315	0.656878

Best Model:
Support Vector Machine (SVM)

- Highest performance across all evaluation metrics
- Strong balance of precision and recall
- Most reliable for predicting song popularity

We trained a Random Forest Classifier using 12 audio features to identify which ones were most important in defining listener clusters. The top-ranked features included acousticness, duration, and speechiness.





7 Conclusion

- **Machine learning successfully predicted song popularity using audio features like energy, danceability, and loudness.**
- **Unsupervised learning (KMeans) revealed 3 distinct user/listener types based on song characteristics.**
- **Supervised learning models were tested to classify songs as popular or not popular.**