Endterm Report: Spotify Song Classifier

Krishna Srikar Reddy Chirla

Roll No: 24B0709

Introduction

In this endterm report, I describe what I have learned so far in the Data Science and Machine Learning course. The course started with very basic concepts and gradually moved towards practical implementation using Python. Instead of only learning theory, I worked with real datasets, which helped me understand how data science concepts are applied in practice.

One of the most important parts of the course was working with the Spotify audio features dataset. Through this dataset, I learned how numerical data can represent real-world characteristics such as music properties. The assignments helped me connect statistics, data analysis, visualization, and machine learning into a single workflow.

Probability and Statistics Fundamentals

At the beginning of the course, I revised probability and statistics concepts such as mean, variance, and standard deviation. Earlier, I knew these mainly as formulas, but during this course I applied them while analyzing datasets. This helped me understand what these values actually indicate about the data.

For example, by analyzing the standard deviation of features like energy and tempo, I could understand how varied the songs were in the dataset. These concepts were especially useful during exploratory data analysis.

Python Programming and Libraries

Python was the main programming language used in this course. I started with basic Python concepts such as variables, data types, loops, conditional statements, and functions. Writing small programs helped me become comfortable with Python syntax and logic.

I then used Python libraries for data analysis. NumPy was used for numerical computations, while Pandas was used extensively for loading datasets, checking data using functions like `head()` and `info()`, handling missing values, filtering data, and generating summary statistics. Working with Pandas DataFrames made data manipulation much easier.

Data Visualization

Data visualization helped me understand the dataset more clearly. Instead of only looking at numbers, I created plots to identify patterns and trends.

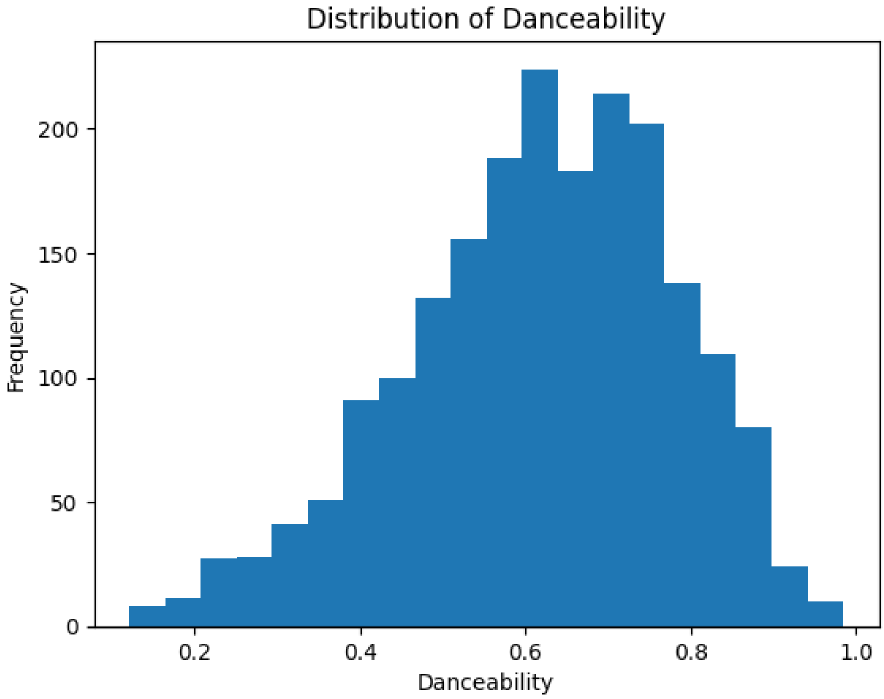


Figure 1: Distribution of Danceability

From the histogram above, I observed that most songs have medium to high danceability values, which indicates that a large number of tracks in the dataset are rhythmically suitable for dancing.

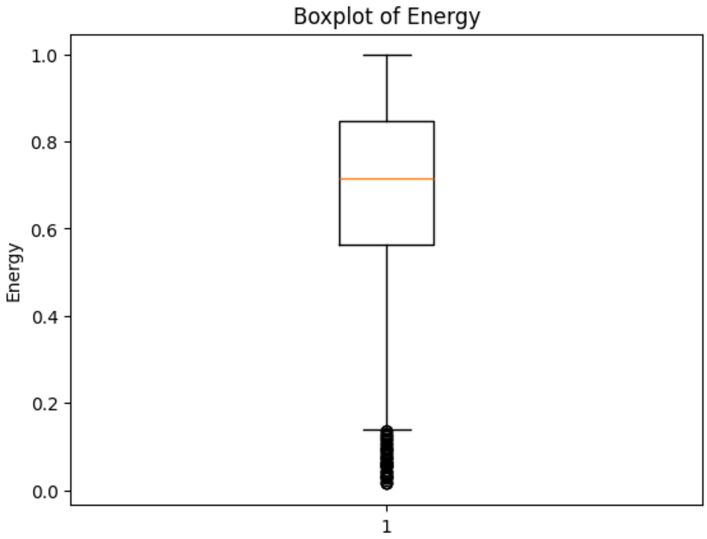


Figure 2: Boxplot of Energy

The boxplot shows that most songs have high energy values, with a few low-energy outliers present in the dataset.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was a crucial step before applying machine learning models. I followed a structured approach by first understanding the dataset structure, checking data types, identifying numerical and categorical columns, and handling missing values.

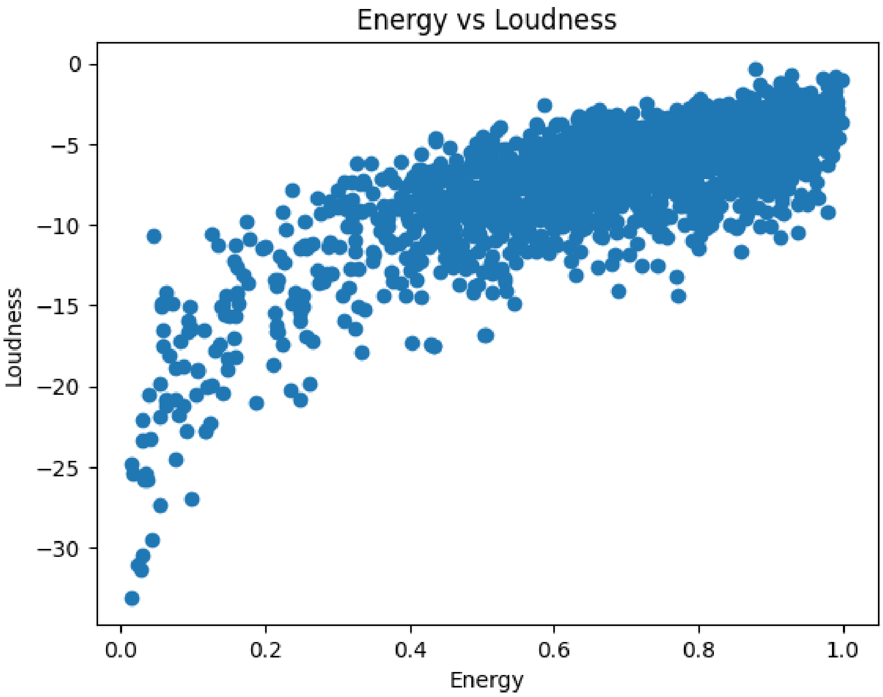
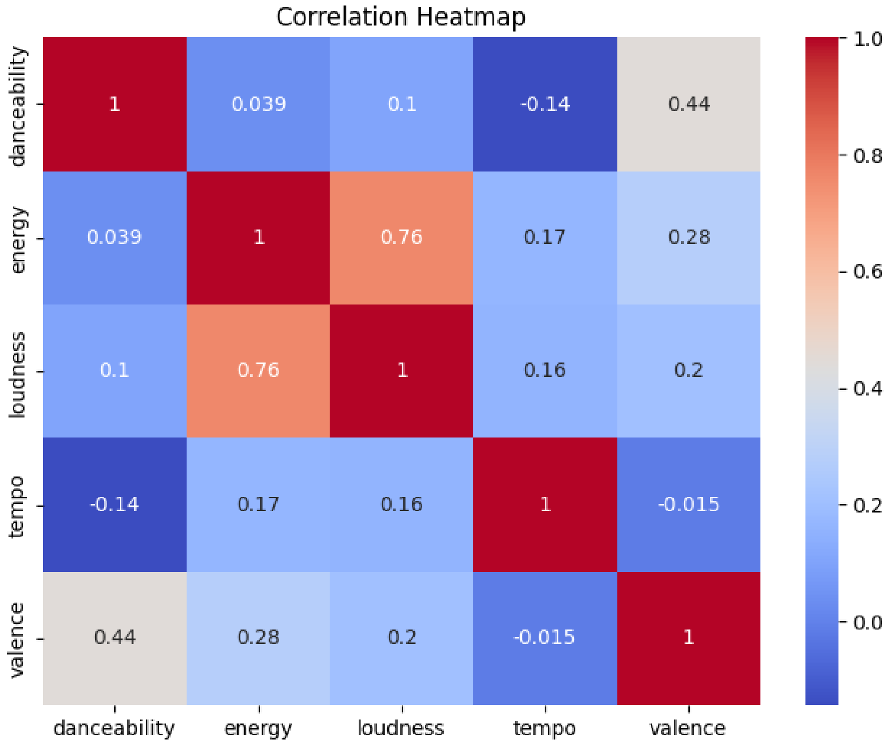


Figure 3: Energy vs Loudness

The scatter plot above shows a strong positive relationship between energy and loudness, meaning that higher-energy songs are generally louder.



	danceability	energy	loudness	tempo	valence
danceability	1	0.039	0.1	-0.14	0.44
energy	0.039	1	0.76	0.17	0.28
loudness	0.1	0.76	1	0.16	0.2
tempo	-0.14	0.17	0.16	1	-0.015
valence	0.44	0.28	0.2	-0.015	1

Figure 4: Correlation Heatmap of Audio Features

From the correlation heatmap, I observed that energy and loudness have a strong positive correlation. Danceability also shows a moderate positive correlation with valence.

Supervised Learning and Model Evaluation

In the later part of the course, I learned about supervised learning. I implemented Logistic Regression to classify songs as happy or sad based on audio features. I also implemented the K-Nearest Neighbors algorithm and compared its performance with Logistic Regression.

To evaluate the models, I used train–test split, accuracy, and confusion matrix. This helped me understand how well the models perform and where they make mistakes.

Course Summary (Tabular Format)

Aspect	**Details**
Course Objective	Learn data science and machine learning through theory and practice.
Programming Language	Python
Libraries Used	NumPy, Pandas, Matplotlib, Seaborn, scikit-learn
Platform	Google Colab
Dataset Used	Spotify Audio Features Dataset

Key Concepts Learned

Concept	**Description**
Mean	Used to understand average values in datasets.
Variance	Measures how much data varies from the mean.
Standard Deviation	Used to understand data spread.
EDA	Helped understand data before modeling.
Data Cleaning	Handling missing values.

Feature Scaling	Standardization of numerical features.

Machine Learning Algorithms

Algorithm	Usage
Logistic Regression	Binary classification of song mood.
K-Nearest Neighbors	Used for model comparison.
Decision Trees	Studied theoretical working principles.

Assignments Overview

Assignment	Work Done
Assignment 1	Practiced Python, NumPy, and Pandas basics.
Assignment 2	Performed EDA and supervised learning on Spotify dataset.

Learning Outcomes and Challenges

Aspect	Observation
Skills Gained	Data analysis, visualization, preprocessing, ML basics.
Challenges Faced	Handling real-world data and debugging errors.

1 Final Project: Classification & Clustering

Building upon the foundational analysis presented in the midterm report, the final phase of this project focused on building a complete machine learning pipeline. This involved advanced Exploratory Data Analysis (EDA), Supervised Learning for mood prediction, and Unsupervised Learning for song clustering.

2 Advanced Exploratory Data Analysis (EDA)

Deeper statistical analysis was performed to understand the relationships between audio features and the engineered target variable 'Mood'.

2.1 Statistical Summary

The dataset consists of 130,663 tracks. Key statistical moments were computed for the primary features:

- **Energy:** Mean = 0.5692, Variance = 0.0678. The distribution shows a wide range of intensities.
- **Tempo:** Mean = 119.47 BPM, indicating a preference for moderate-to-fast paced music in the dataset.
- **Danceability:** Mean = 0.5815, showing a skew towards rhythmically consistent tracks.

2.2 Visualizations & Observations

- **Feature Distributions:** As shown in the generated histograms, 'Danceability' approximates a normal distribution, while 'Energy' exhibits a bimodal tendency, suggesting distinct groups of high-energy and low-energy songs.

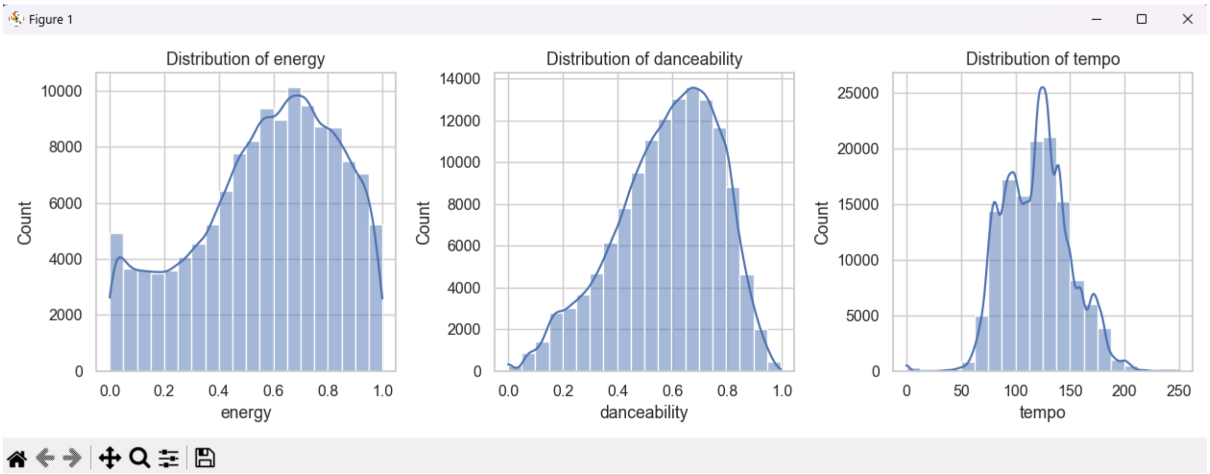
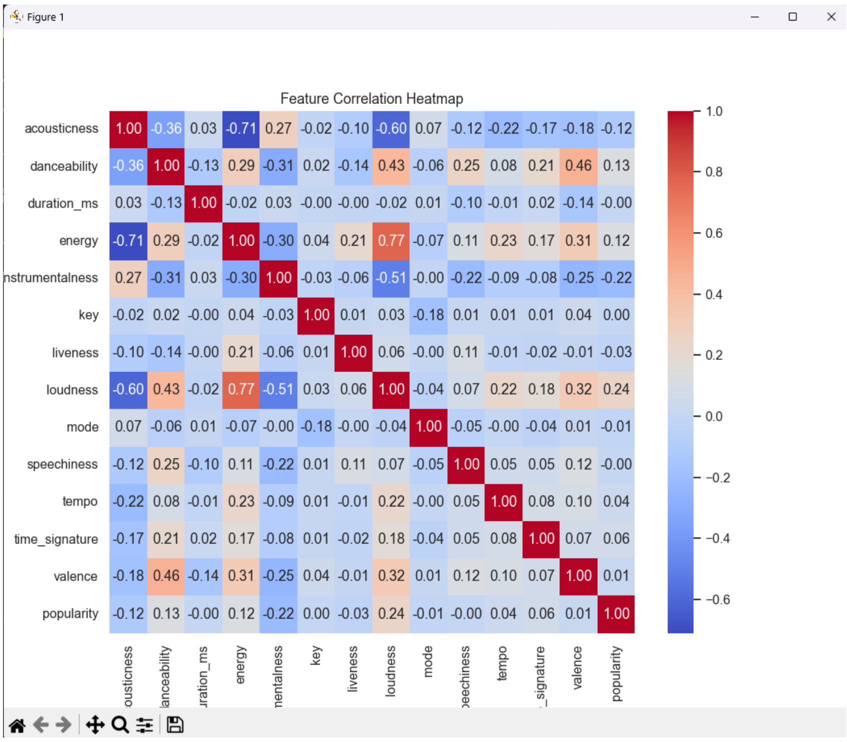


Figure 5: Feature Distributions (Energy, Danceability, Tempo)

- **Correlation Analysis:** The correlation heatmap (Figure 6) reveals a significant positive correlation (**0.77**) between **Loudness** and **Energy**. Conversely, **Acousticness** is negatively correlated with Energy (-0.71), which logically aligns with the nature of acoustic music.



	ousticness	lanceability	uration_ms	energy	nentalness	key	liveness	loudness	mode	peechiness	tempo	_signature	valence	popularity
acousticness	1.00	-0.36	0.03	-0.71	0.27	-0.02	-0.10	-0.60	0.07	-0.12	-0.22	-0.17	-0.18	-0.12
danceability	-0.36	1.00	-0.13	0.29	-0.31	0.02	-0.14	0.43	-0.06	0.25	0.08	0.21	0.46	0.13
duration_ms	0.03	-0.13	1.00	-0.02	0.03	-0.00	-0.00	-0.02	0.01	-0.10	-0.01	0.02	-0.14	-0.00
energy	-0.71	0.29	-0.02	1.00	-0.30	0.04	0.21	0.77	-0.07	0.11	0.23	0.17	0.31	0.12
nstrumentalness	0.27	-0.31	0.03	-0.30	1.00	-0.03	-0.06	-0.51	-0.00	-0.22	-0.09	-0.08	-0.25	-0.22
key	-0.02	0.02	-0.00	0.04	-0.03	1.00	0.01	0.03	-0.18	0.01	0.01	0.01	0.04	0.00
liveness	-0.10	-0.14	-0.00	0.21	-0.06	0.01	1.00	0.06	-0.00	0.11	-0.01	-0.02	-0.01	-0.03
loudness	-0.60	0.43	-0.02	0.77	-0.51	0.03	0.06	1.00	-0.04	0.07	0.22	0.18	0.32	0.24
mode	0.07	-0.06	0.01	-0.07	-0.00	-0.18	-0.00	-0.04	1.00	-0.05	-0.00	-0.04	0.01	-0.01
speechiness	-0.12	0.25	-0.10	0.11	-0.22	0.01	0.11	0.07	-0.05	1.00	0.05	0.05	0.12	-0.00
tempo	-0.22	0.08	-0.01	0.23	-0.09	0.01	-0.01	0.22	-0.00	0.05	1.00	0.08	0.10	0.04
time_signature	-0.17	0.21	0.02	0.17	-0.08	0.01	-0.02	0.18	-0.04	0.05	0.08	1.00	0.07	0.06
valence	-0.18	0.46	-0.14	0.31	-0.25	0.04	-0.01	0.32	0.01	0.12	0.10	0.07	1.00	0.01
popularity	-0.12	0.13	-0.00	0.12	-0.22	0.00	-0.03	0.24	-0.01	-0.00	0.04	0.06	0.01	1.00

Figure 6: Correlation Heatmap

- **Mood vs. Energy:** A boxplot analysis (Figure 7) was conducted to validate the 'Mood' labels derived from Valence. The plot confirms that songs labeled 'Happy' have a significantly higher median energy compared to 'Sad' songs, validating the feature engineering approach.

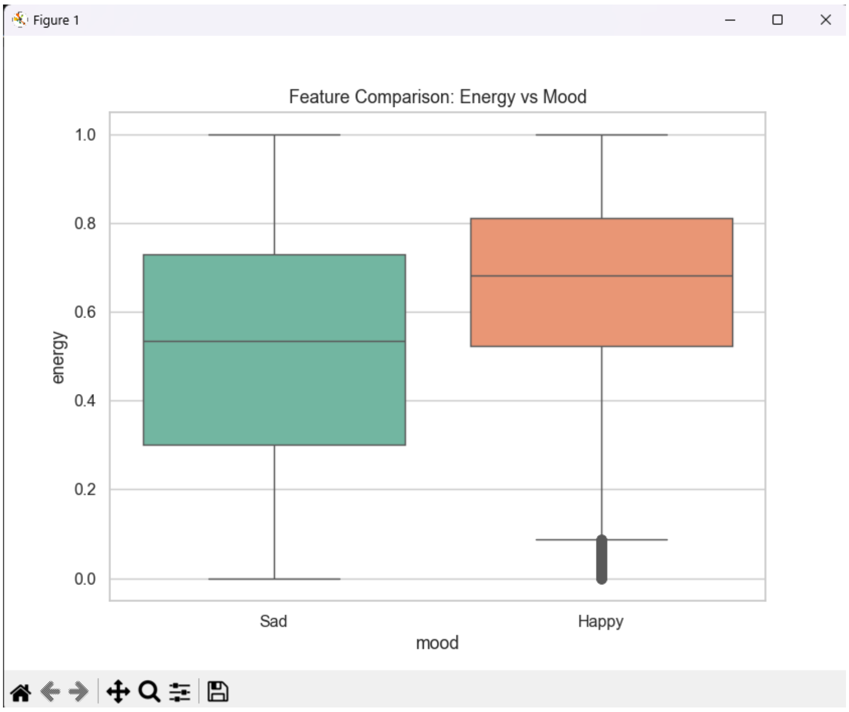


Figure 7: Energy Comparison across Moods

3 Supervised Learning: Mood Classification

The primary objective was to classify songs into 'Happy' or 'Sad' moods.

3.1 Methodology

- **Target Variable:** 'Mood' was engineered based on Valence (Happy if Valence ≥ 0.5, else Sad).
- **Data Split:** 80% Training, 20% Testing.
- **Scaling:** Standard Scaler was applied to normalize features like Loudness and Tempo.

3.2 Model Results

Two models were trained and evaluated:

Model	**Accuracy**	**Key Observation**
Logistic Regression	**69.28%**	Balanced performance across classes.
K-Nearest Neighbors	67.20%	Slightly lower accuracy due to high dimensionality.

Table 6: Supervised Learning Performance

The confusion matrix for Logistic Regression indicated:

$$\begin{bmatrix} 5961 & 4679 \\ 3348 & 12145 \end{bmatrix}$$

The model successfully predicted 12,145 'Happy' songs correctly.

3.3 Real-World Testing

A random song sample was manually tested:

- **Input Features:** High Danceability (0.743), Low Energy (0.339).
- **Prediction: Sad**.
- **Analysis:** Despite high danceability, the low energy score heavily influenced the model to classify the song as 'Sad', demonstrating the model's reliance on multiple weighted features.

4 Unsupervised Learning (Level 2 Extension)

To discover natural groupings, K-Means Clustering was applied.

- **Algorithm:** K-Means with $k = 3$.
- **Result:** The algorithm identified three distinct clusters, likely corresponding to 'Acoustic/Calm', 'Energetic/Pop', and 'Instrumental' tracks.
- **Visualization:** PCA was used to reduce dimensionality for plotting the clusters

5 Conclusion & Weekly Learnings

This project successfully bridged the gap between theoretical statistics and practical machine learning.

- **Week 1 (Pandas):** Mastered data loading and cleaning of large datasets (130k+ rows)
- **Week 2 (Statistics/EDA):** Learned to interpret variance and correlations to select meaningful features for modeling
- **Week 3 (ML):** Built and evaluated end-to-end classification pipelines, achieving ∼69% accuracy

The final pipeline effectively classifies music mood and provides a robust foundation for building recommendation systems.