Midterm Report: Data Science & Machine Learning Foundations

Krishna Srikar Reddy Chirla

Roll No: 24B0709

Introduction

In this midterm 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 func-

tions. 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.

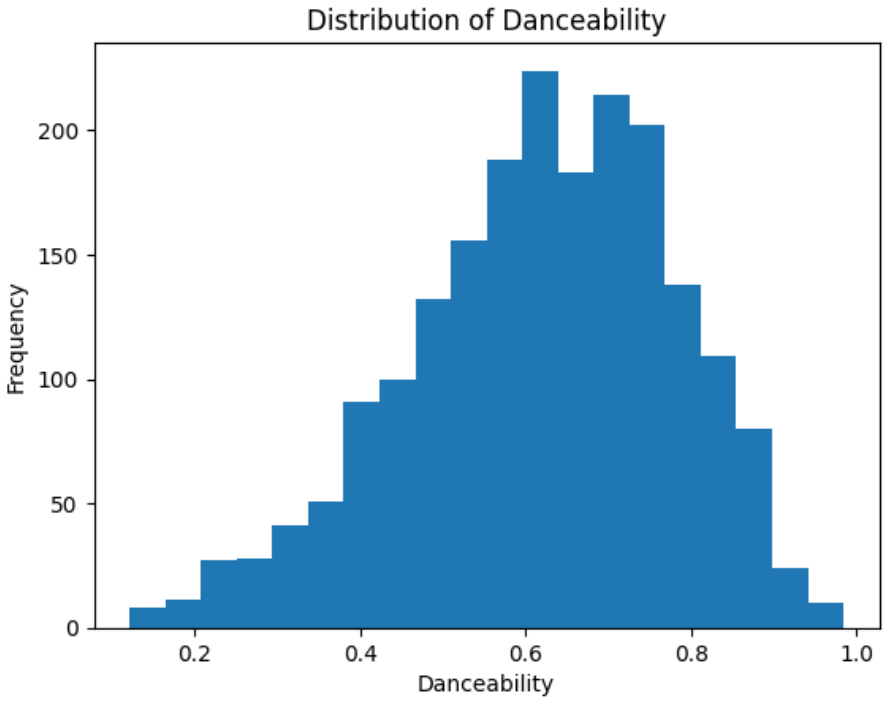

Figure 1: Distribution of Danceability

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.

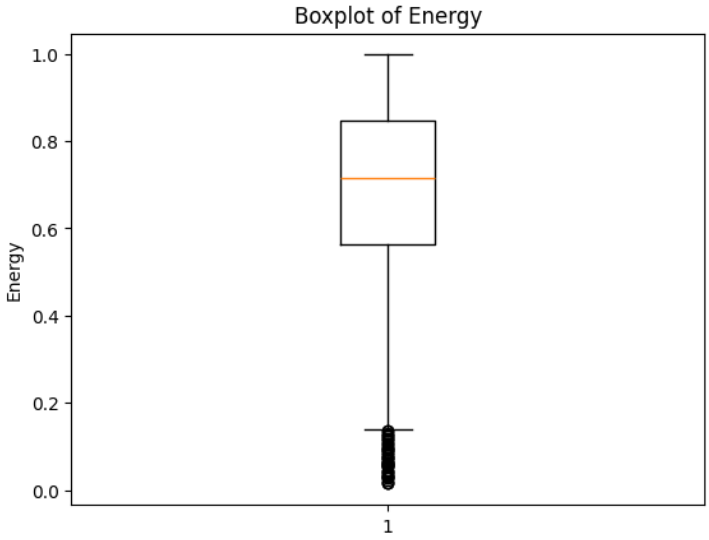

Figure 2: Boxplot of Energy

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.

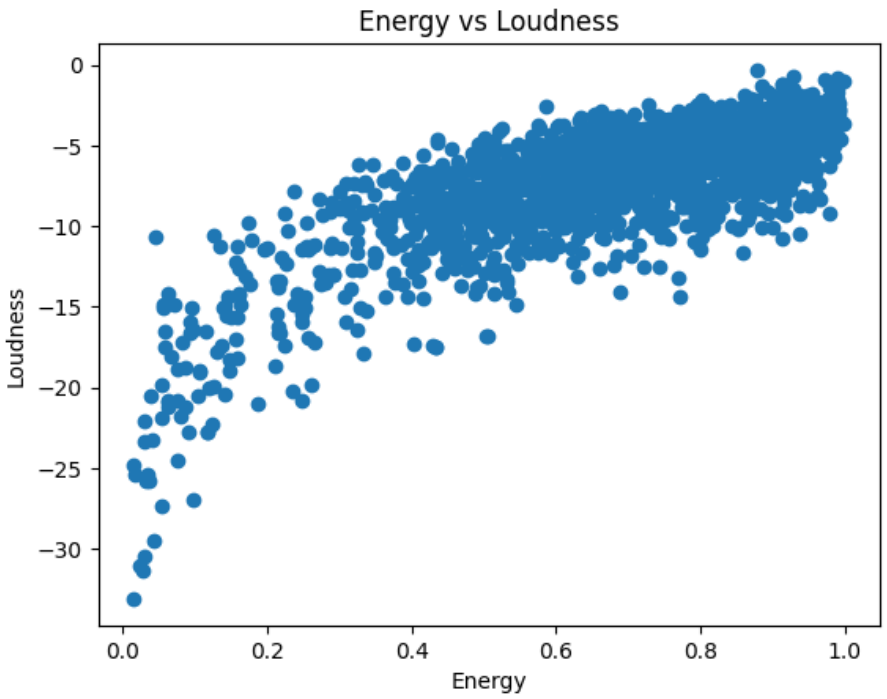

Figure 3: Energy vs Loudness

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

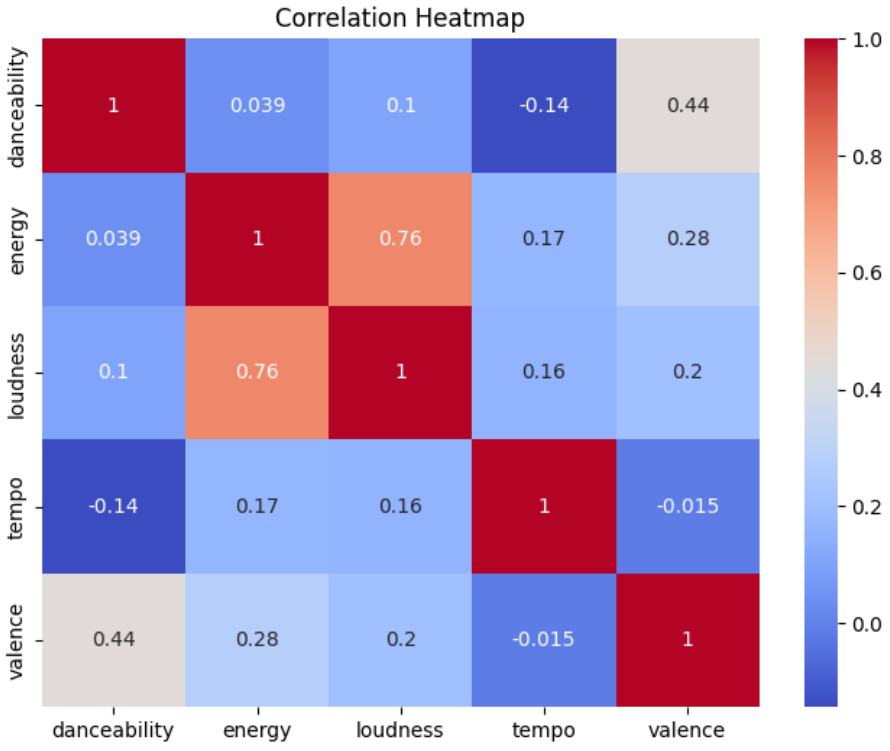

Figure 4: Correlation Heatmap of Audio Features

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.

Future Learning Direction

Area	**Planned Focus**
Unsupervised Learning	Learn clustering techniques in detail.
Advanced ML	Explore more complex models.
NLP	Learn text-based data analysis techniques.

Conclusion

Through this course, I learned how to analyze data step by step, visualize patterns, preprocess datasets, and apply basic machine learning models. The assignments helped me understand how theory is used in practice. Overall, this course provided me with a strong foundation in data science and machine learning.