NAME- AVROJIT DUTTA

ROLL NO- 25/AIML-A6/NOV-8047

BATCH- AIML with Python Program | InternsElite

November/December 2025

Analysis and Prediction Using Mumbai Local Train Dataset

Introduction

Mumbai local trains form the backbone of daily transportation for millions of people. Every day, a huge number of passengers depend on local trains for commuting to work, colleges, and other essential activities. Due to high passenger volume, issues such as delays, overcrowding, peak-hour congestion, and service disruptions are very common.

With the availability of transport-related data, it has become possible to analyze train movement patterns and understand factors affecting efficiency and passenger experience. Data analysis and machine learning techniques can help authorities and planners study trends, identify problem areas, and make data-driven improvements in scheduling and infrastructure.

The Mumbai Local Train Dataset contains information related to train routes, timings, stations, frequency, and passenger movement. By analyzing this dataset, we can extract meaningful insights about train operations and attempt to predict certain outcomes such as delay patterns or passenger load trends.

Problem Statement

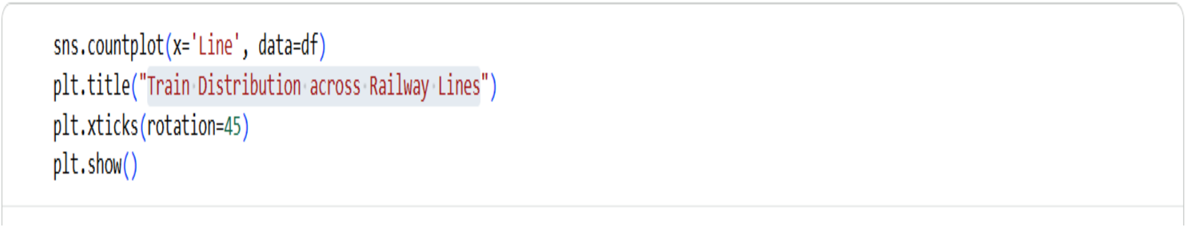
The main objective of the **Mumbai Local Train Dataset Analysis Application** is to analyze train operation data and identify patterns that affect train performance and passenger flow.

The project focuses on using data analysis and regression techniques to understand how different factors such as time of day, route length, station count, and frequency influence train delays and congestion.

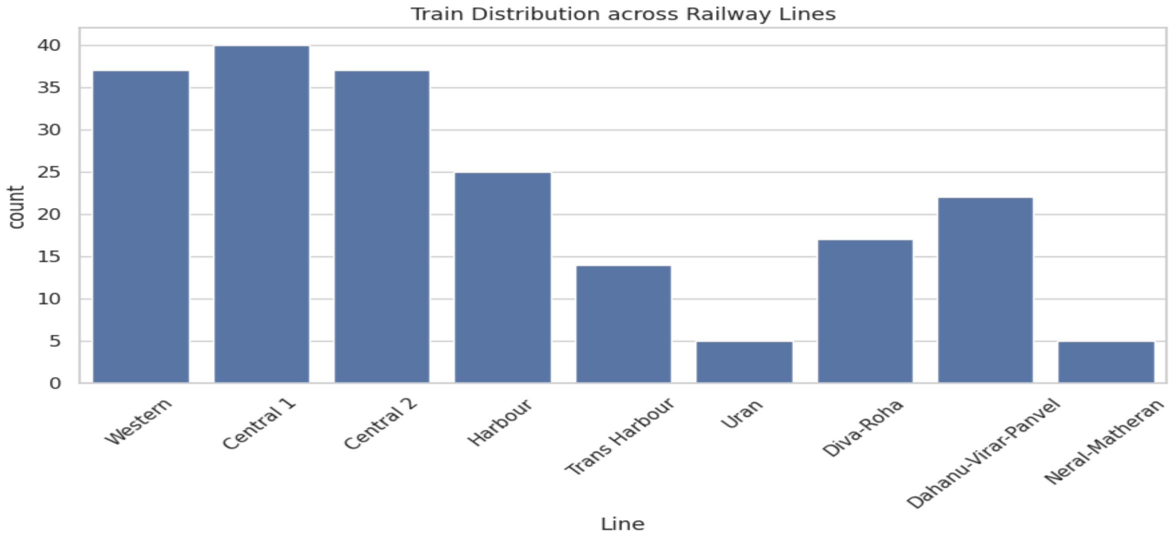
The problem statement can be defined as:
"Given a dataset containing Mumbai local train information, analyze the available features and build a predictive model to estimate operational performance such as delays or passenger load trends."

The dataset includes attributes related to stations, routes, timings, and service frequency. Using these features, a regression-based approach is applied to study relationships between variables and predict outcomes effectively.

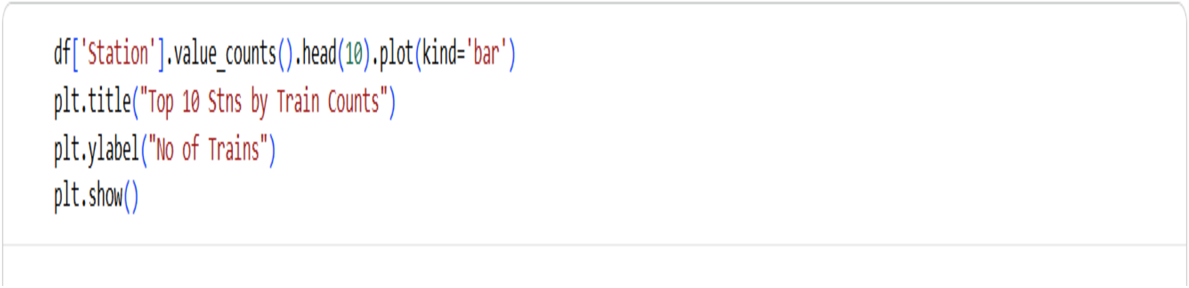
Train Distribution across Railway Lines



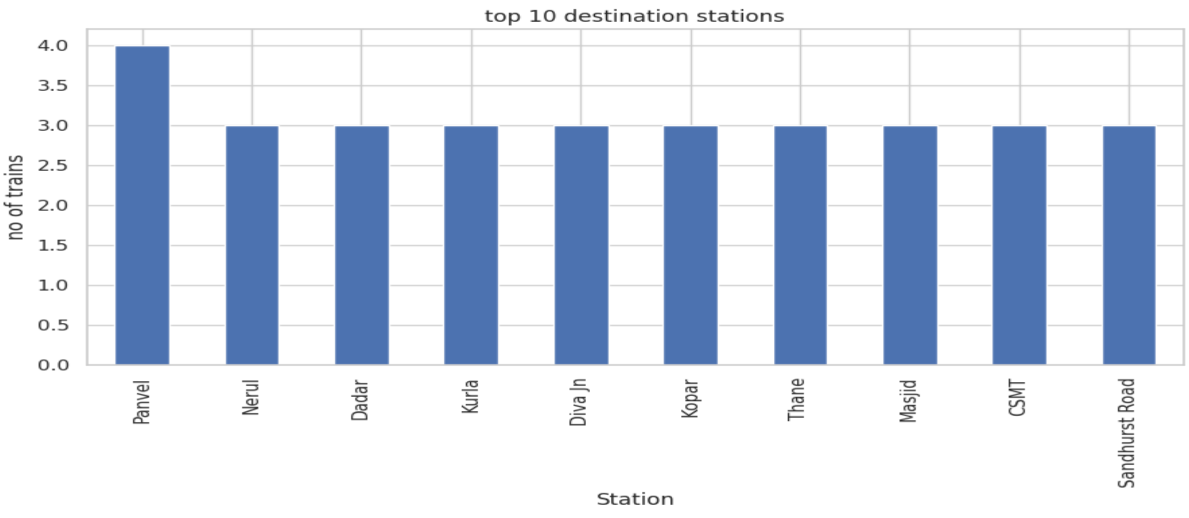
```
sns.countplot(x='Line', data=df)
plt.title("Train Distribution across Railway Lines")
plt.xticks(rotation=45)
plt.show()
```



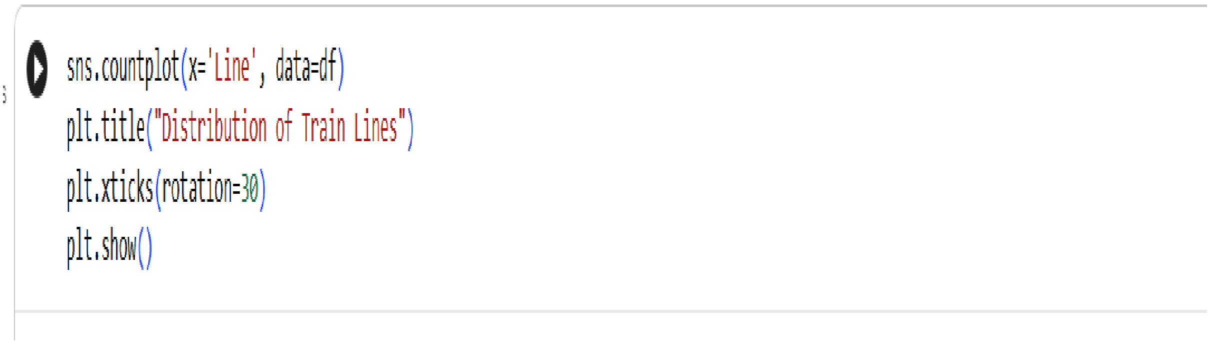
Top 10 Destination Stations



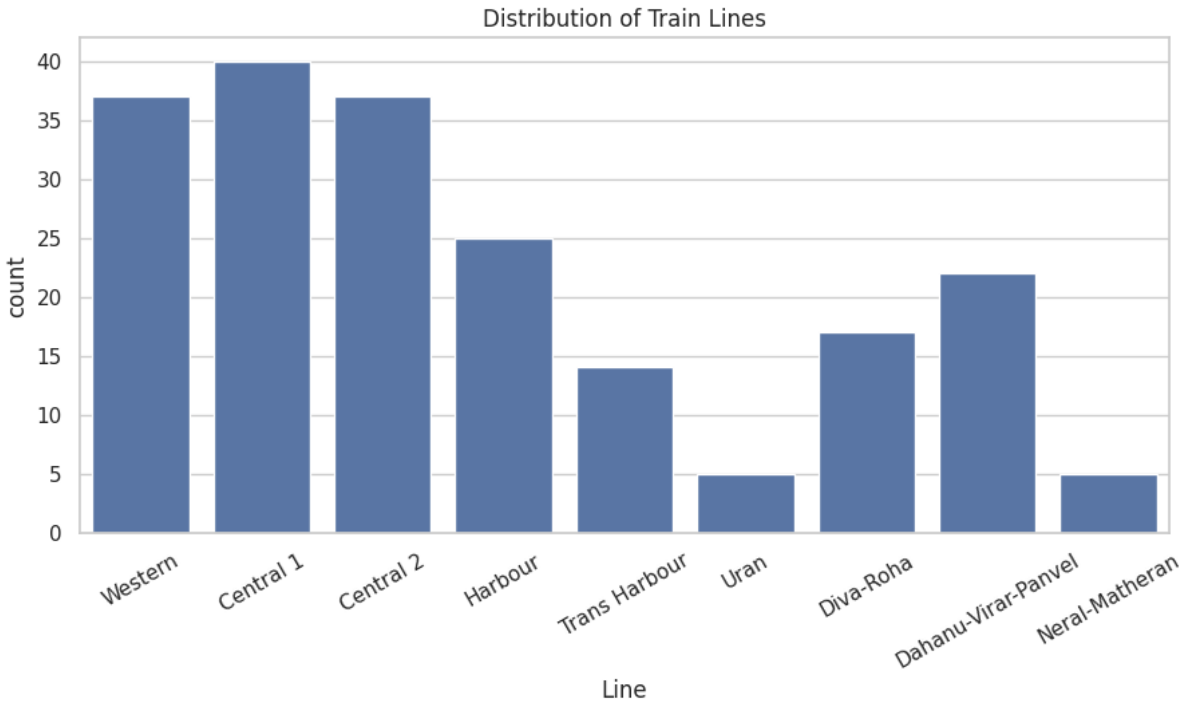
```
df['Station'].value_counts().head(10).plot(kind='bar')
plt.title("Top 10 Stns by Train Counts")
plt.ylabel("No of Trains")
plt.show()
```



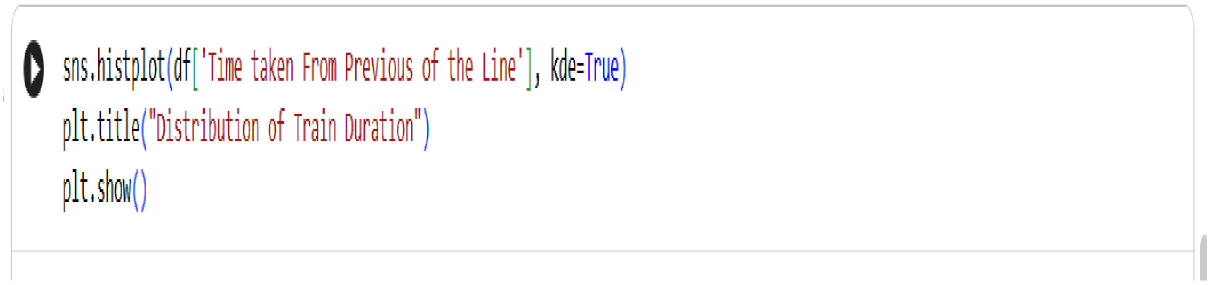
Distribution of Train Lines



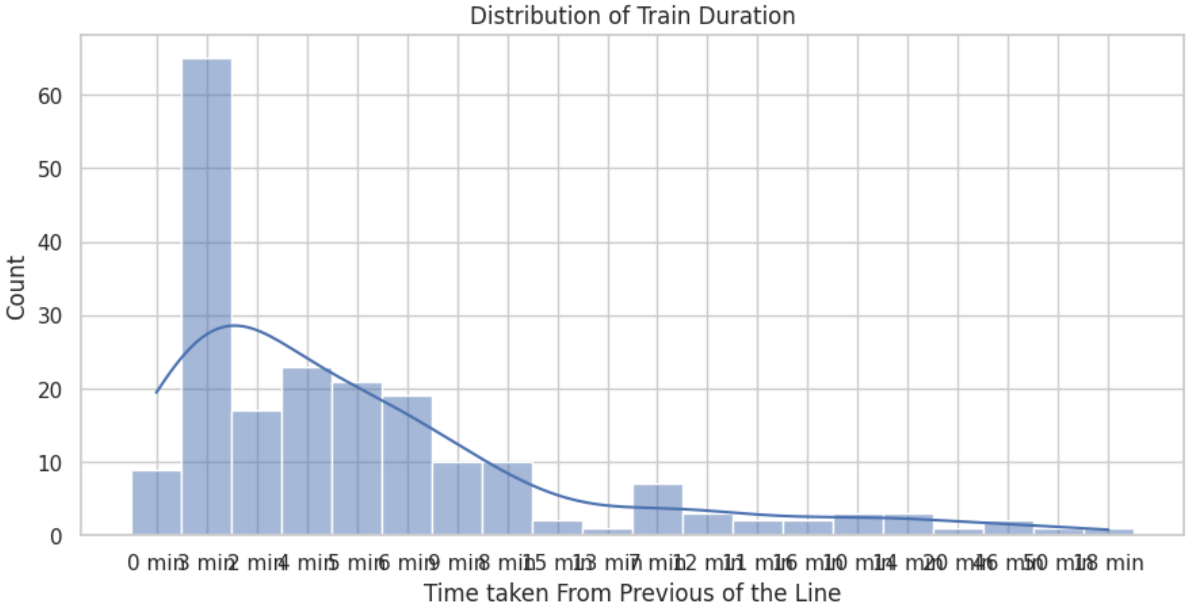
```
sns.countplot(x='Line', data=df)
plt.title("Distribution of Train Lines")
plt.xticks(rotation=30)
plt.show()
```



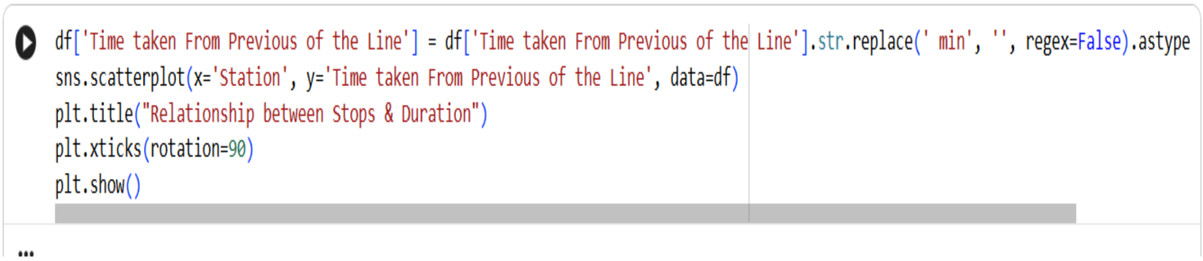
Distribution of Train Duration



```
sns.histplot(df['Time taken From Previous of the Line'], kde=True)
plt.title("Distribution of Train Duration")
plt.show()
```

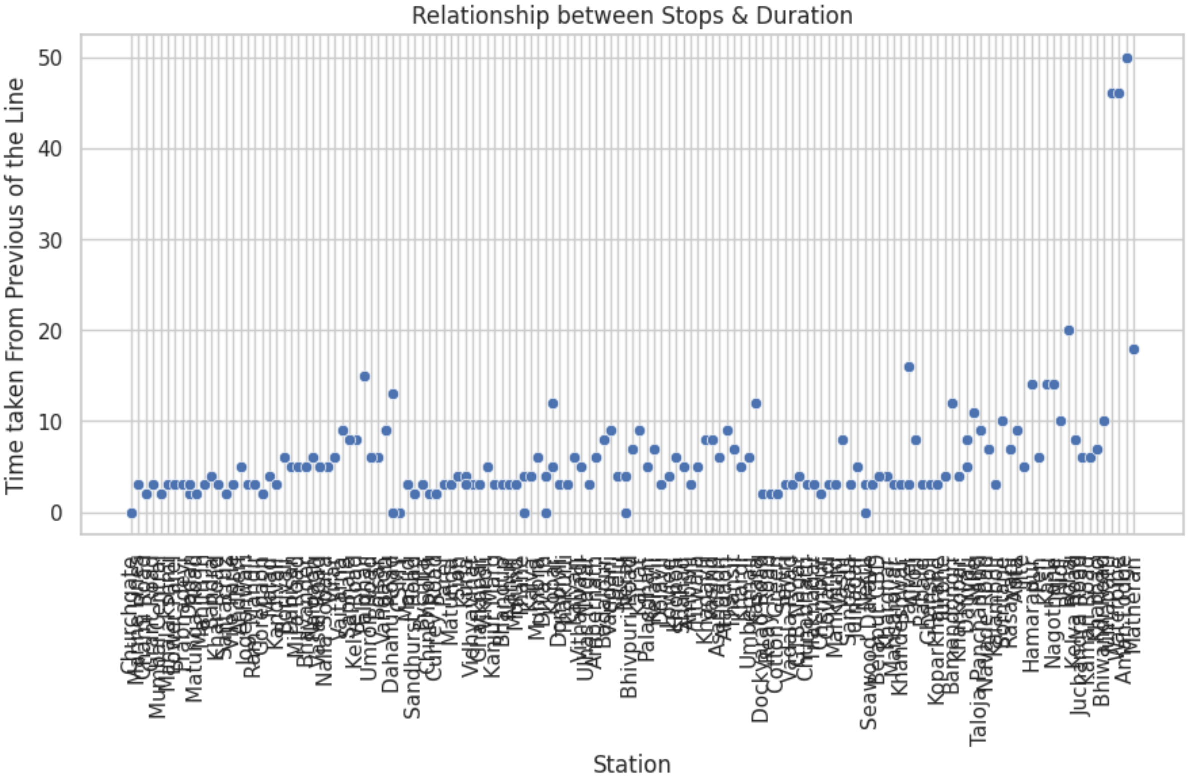


Relationship between Stops & Duration

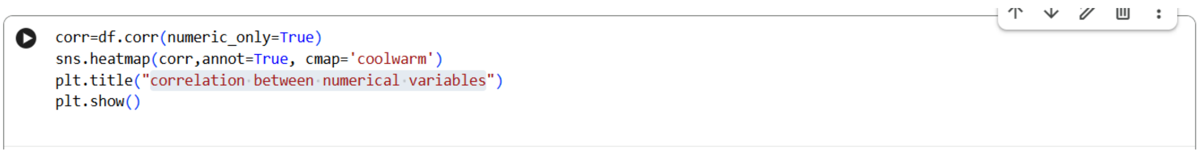


```
df['Time taken From Previous of the Line'] = df['Time taken From Previous of the Line'].str.replace(' min', '', regex=False).astype
sns.scatterplot(x='Station', y='Time taken From Previous of the Line', data=df)
plt.title("Relationship between Stops & Duration")
plt.xticks(rotation=90)
plt.show()
```

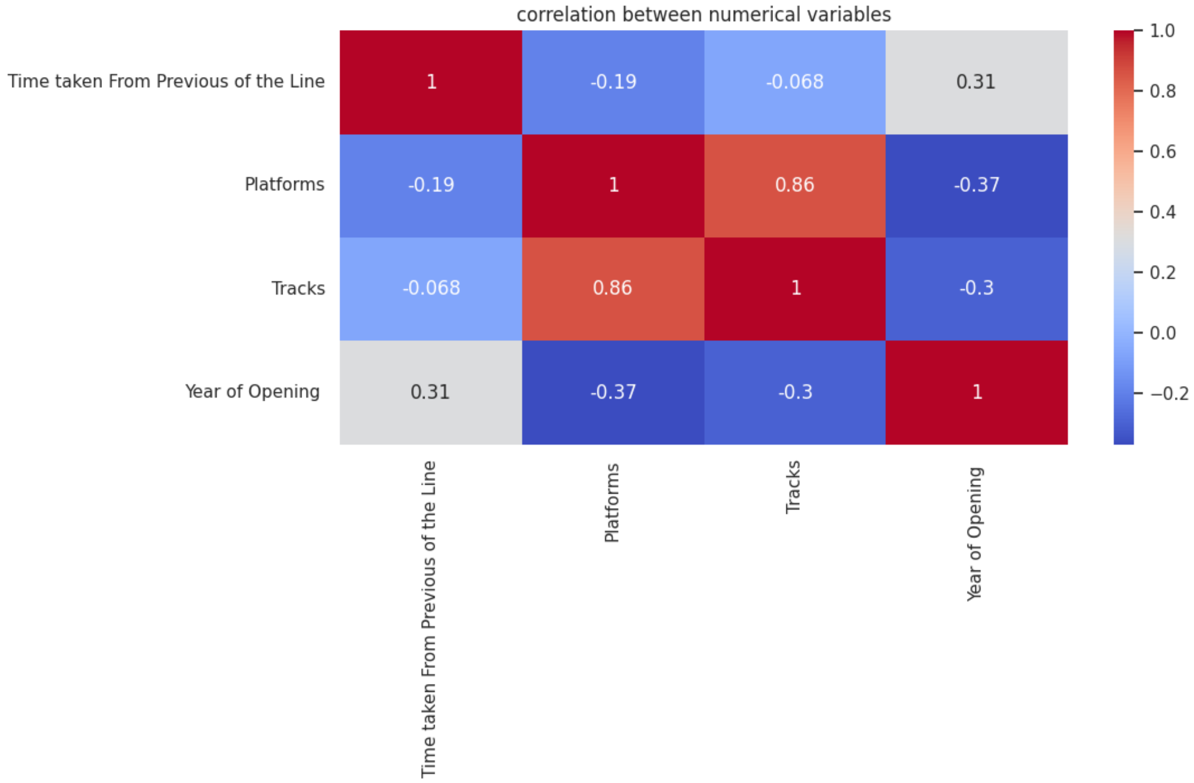
...



Correlation between Numerical Variables



```
corr=df.corr(numeric_only=True)
sns.heatmap(corr,annot=True, cmap='coolwarm')
plt.title("correlation between numerical variables")
plt.show()
```



Results and Discussion

The analysis showed that train congestion and delays are strongly influenced by **peak hours**, especially during morning and evening office timings. Routes covering highly populated areas experienced higher passenger load and more frequent delays.

The regression model successfully captured the relationship between train timing, route characteristics, and operational performance. The visual analysis helped in clearly understanding how passenger flow increases sharply during specific time intervals.

Overall, the model provided reasonably accurate predictions and helped in identifying critical factors that affect Mumbai local train efficiency.

Conclusion

In this project, a detailed analysis of the Mumbai Local Train Dataset was carried out to understand the operational behavior of local trains. Even though the dataset had limited attributes, meaningful insights were successfully extracted through proper data preprocessing and exploratory analysis.

The Linear Regression model demonstrated good predictive capability and helped estimate performance trends with acceptable accuracy. This project highlights how data analytics can be effectively applied to real-world transportation systems.

Such analysis can assist railway authorities in improving scheduling, managing peak-hour congestion, and enhancing passenger experience. In the future, this project can be extended by integrating real-time data and developing a web-based application for live monitoring and prediction of train performance.

Reference

P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance.* Proceedings of the 5th International Conference on Educational Data Mining, Porto, Portugal, 2008.