CONTROL OFFICE APPLICATION
TRAIN DELAY PREDICTION & RAIL INDEX PREDICTION

INTERNSHIP REPORT

Centre for Railway Information Systems (CRIS),

Chennai

Submitted by

Jeeveshraj D, B.E. Final year

Kavin Raam M, B.E. Final year

Electronics and Communication Engineering

SSN College of Engineering, Kalavakkam

Duration:

9th June 2025 to 11th July 2025

ACKNOWLEDGMENT

We take this opportunity to express our sincere and heartfelt gratitude to the **Centre for Railway Information Systems (CRIS)**, an esteemed organization under the Ministry of Railways, for granting us the invaluable opportunity to undertake our summer internship in the **Control Office Application (COA)** department. It has been an enriching experience that allowed us to explore real-world applications of data science, machine learning, and information systems in one of the largest and most complex railway networks in the world.

First and foremost, we would like to express deepest gratitude to our project guide and mentor at CRIS, **Mr. Ramakrishnan KP**, Chief Manager & **Mr. Raghavendra S**, Software Engineer whose technical expertise, thoughtful feedback, and constant encouragement were instrumental in shaping this project. Their clear vision and structured approach helped us align our work with the practical requirements of the COA system, and their willingness to address even the smallest queries fostered a highly conducive learning environment.

This internship provided us with unparalleled exposure to **mission-critical railway information systems**, and we had the privilege of working on real transaction data from the Indian Railways. The project not only enhanced our technical skillset in data wrangling, statistical analysis, and machine learning, but also gave us valuable insights into the logistical, operational, and infrastructural challenges of managing rail transport at a national scale. Working on real train movement data — particularly analyzing delay patterns, predicting delay probabilities using Random Forest models, and building a user-facing dashboard using Streamlit — was a unique blend of research and engineering.

Tabel of Content

ABSTRACT

This report summarizes the work done during the summer internship at the **Centre for Railway Information Systems (CRIS)** under the **Control Office Application (COA)** department. The internship focused on applying data analytics and machine learning to analyse train delays and predict operational performance using real Indian Railways data.

The primary objective was to carry out **Train Delay Pattern Analysis** and build a **Machine Learning-based Delay Prediction System** for **Vaigai Express (Train No. 12635)**, which operates between **Chennai Egmore (MS)** and **Madurai (MDU)**. Ten days of real train movement data were combined with scheduled timetable information to compute station-wise arrival and departure delays. Exploratory Data Analysis (EDA) was performed to identify delay trends, peak hours, and route bottlenecks using statistical summaries and visualizations such as box plots and time-series charts.

A **Random Forest Regressor** was developed to predict arrival delays using features like day of week, hour of travel, halt time, departure delay, and distance. The model achieved an **R^2 score of 0.83**, indicating strong predictive capability. Hyperparameter optimization was done using RandomizedSearchCV.

To make the insights accessible, a **Streamlit-based dashboard** was built with three key sections:

- **Delay Analysis** – to explore average and peak delay patterns across stations and dates.
- **Route Performance** – to compare station-wise metrics and identify consistent delay points.
- **Delay Prediction** – to forecast delay in minutes based on user-specified train conditions.

Additionally, a ML model for **Rail Index Prediction** was developed to assess overall train performance using historical metrics.

This project demonstrated the practical use of data science in the railway domain and provided CRIS with a working prototype for delay intelligence and decision support. The internship offered valuable exposure to real-world transport systems and technical experience in machine learning and dashboard development.

INTRODUCTION

The Indian Railways is one of the largest and most complex railway networks in the world, operating over 13,000 passenger trains daily and transporting more than 20 million people across thousands of stations. Given the vast scale and operational constraints of this system—ranging from infrastructure limitations to weather conditions—train delays are a frequent and critical challenge. Efficient delay management is vital not only for improving passenger satisfaction but also for optimizing resource utilization, reducing cascading operational disruptions, and ensuring the timely movement of goods and services.

The Centre for Railway Information Systems (CRIS) is the organization responsible for designing, developing, and maintaining key information systems used across Indian Railways. One of the most critical systems under its domain is the Control Office Application (COA), which is used by railway control rooms to monitor, plan, and respond to train movement data in real time. The COA plays a pivotal role in assisting dispatchers and railway operators in managing the real-time location and schedule adherence of trains, particularly under disrupted or congested conditions.

As part of our summer internship under the COA division at CRIS, we were assigned a project aimed at building a data-driven system to understand and predict train delays using historical train movement data. The core idea was to analyse delays of the Vaigai Express (Train No. 12635)—which operates between Chennai Egmore (MS) and Madurai Junction (MDU)—over a 10-day period. This project was chosen because the Vaigai Express serves a major route with multiple critical junctions, and analysing its delays could offer useful insights for broader applications.

The project was divided into three major components:

1. Train Delay Pattern Analysis – In this part, actual train transaction data was merged with scheduled referential data to compute arrival and departure delays at each station. Exploratory Data Analysis (EDA) techniques were applied to understand where and when delays were occurring. The objective was to uncover delay trend over time.
2. Machine Learning-Based Delay Prediction – After the delay patterns were explored, a supervised learning approach was used to build a prediction model. A Random Forest Regressor was trained on features such as day of the week, halt time, hour of the day, prior departure delay, and inter-station distance to predict the expected arrival delay at each stop. Hyperparameter tuning was performed using RandomizedSearchCV to improve the model's predictive accuracy.

3. Streamlit Dashboard for Operational Insight – To enable COA staff to use these insights effectively, all data analyses and the predictive model were integrated into an interactive dashboard using Streamlit. This tool allows users to visualize historical delay trends, station-wise performance, and input specific train parameters to generate delay forecasts in real time.

This end-to-end pipeline - from data preprocessing to model training and dashboard deployment - formed the complete **Train Delay Prediction module**. By focusing on the Vaigai Express, we were able to demonstrate how historical movement data can be used to uncover delay trends, anticipate disruptions, and offer actionable insights to COA operators. The model's strong predictive performance and the interactive nature of the dashboard make this solution a powerful decision-support tool for managing real-time operations.

	A	B	C	D	E	F
1	SL NO	TRAIN_NUMB	TRAINDATE	STTN_CODE	ARVL_TIME	DEP_TTIME
2	1	12635	6/15/2025	MS		15/06/2025 13:45:05
3	2	12635	6/15/2025	TBM	15/06/2025 14:13:48	15/06/2025 14:14:41
4	3	12635	6/15/2025	CGL	15/06/2025 14:38:43	15/06/2025 14:41:52
5	4	12635	6/15/2025	VM	15/06/2025 16:03:39	15/06/2025 16:08:49
6	5	12635	6/15/2025	VRI	15/06/2025 16:45:21	15/06/2025 16:47:20
7	6	12635	6/15/2025	ALU	15/06/2025 17:33:01	15/06/2025 17:56:45
8	7	12635	6/15/2025	SRGM	15/06/2025 18:39:59	15/06/2025 18:41:42
9	8	12635	6/15/2025	TPJ	15/06/2025 19:00:26	15/06/2025 19:08:36
10	9	12635	6/15/2025	MPA	15/06/2025 19:35:35	15/06/2025 19:37:05
11	10	12635	6/15/2025	DG	15/06/2025 20:14:17	15/06/2025 20:17:07
12	11	12635	6/15/2025	SDN	15/06/2025 20:59:39	15/06/2025 21:01:13
13	12	12635	6/15/2025	MDU	15/06/2025 21:21:09	

S. No	Station Name & Code	Arrives	Departs	Halt Time	Distance	Day
1	Chennai Egmore - MS	Start	13:45	-	0.0 km	1
2	Tambaram - TBM	14:10	14:12	2m	24.5 km	1
3	Chengalpattu Jn - CGL	14:38	14:40	2m	55.5 km	1
4	Villupuram Jn - VM	15:55	16:00	5m	158.5 km	1
5	Vriddhachalam Jn - VRI	16:40	16:42	2m	213.2 km	1
6	Ariyalur - ALU	17:18	17:19	1m	266.7 km	1
7	Srirangam - SRGM	17:59	18:01	2m	324.8 km	1
8	Tiruchchirappalli Jn - TPJ	18:50	18:55	5m	336.5 km	1
9	Manaparai - MPA	19:24	19:25	1m	373.0 km	1
10	Dindigul Jn - DG	20:03	20:05	2m	430.9 km	1
11	Sholavandan - SDN	20:35	20:36	1m	472.0 km	1
12	Madurai Jn - MDU	21:20			493.2 km	1

Figure 1: Vaigai Timings Dataset Overview

As a secondary task during the internship, we also worked on a **Rail Index Prediction** module aimed at forecasting Lateral Rail Index (LRI) and Vertical Rail Index (VRI) values. This module used a broader dataset of multiple trains to compute rail index that could score performance by region, train type, or time window. This index serves as a high-level summary metric that can support policy decisions and long-term planning.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SECCODE	LINECODE	KMFROM	METFROM	KMTO	METTO	BLOCKNO	PARAM	DATE1	DATE2	RI1	RI2	GMT
2	1071	1	1526	600	1526	800	4	VRI	5/29/2023	6/30/2023	2.41	2.62	3299966
3	1071	1	1526	600	1526	800	4	VRI	6/30/2023	7/12/2023	2.62	2.36	1245875
4	1071	1	1526	600	1526	800	4	VRI	7/12/2023	8/14/2023	2.36	2.33	3530107
5	1071	1	1526	600	1526	800	4	VRI	8/14/2023	8/23/2023	2.33	2.31	1085443

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SECCODE	LINECODE	KMFROM	METFROM	KMTO	METTO	BLOCKNO	PARAM	DATE1	DATE2	RI1	RI2	GMT
2	1071	1	1526	600	1526	800	4	LRI	5/29/2023	6/30/2023	2.32	2.44	3299966
3	1071	1	1526	600	1526	800	4	LRI	6/30/2023	7/12/2023	2.44	2.35	1245875
4	1071	1	1526	600	1526	800	4	LRI	7/12/2023	8/14/2023	2.35	2.34	3530107
5	1071	1	1526	600	1526	800	4	LRI	8/14/2023	8/23/2023	2.34	2.31	1085443

Figure 2: Rail Index Dataset Overview

This report outlines the systematic approach followed during the internship: From data acquisition and preprocessing to visualization, modelling, and deployment. It also highlights key findings, model performance, and future directions for enhancement. The goal of this work was to provide a functional and scalable prototype that could eventually be expanded into CRIS's real-time decision-making tools for railway operations.

PROBLEM STATEMENT

Indian Railways, being one of the largest and most complex railway networks in the world, faces several operational challenges, among which train delays are a significant and recurring issue. These delays lead to schedule disruptions, congestion across control sections, passenger dissatisfaction, and increased operational inefficiencies. Despite the presence of centralized systems such as the Control Office Application (COA), which monitors real-time train movement, the decision-making process largely remains reactive and lacks predictive support. Additionally, the wealth of historical data collected remains underutilized for drawing insights or making data-driven forecasts.

During our internship under the COA division at the Centre for Railway Information Systems (CRIS), we identified the need for an integrated approach to both analyse and predict operational patterns using modern data science techniques. Our work was divided into two key components. The first was a **Train Delay Analysis and Prediction** module based on real-time transaction data for **Vaigai Express (Train No. 12635)**, operating between Chennai Egmore (MS) and Madurai (MDU). We performed delay computations across stations using scheduled vs. actual times, explored station-wise and route-level delay patterns, and developed a Random Forest regression model to predict future delays using features such as departure time, day of the week, halt duration, and distance. These insights were presented through a dynamic Streamlit dashboard that allowed users to visualize delay patterns, examine route performance, and interactively simulate delay predictions.

The second part of our project involved **Rail Index Prediction**. In this task, we worked with a dataset containing block-wise historical index values—specifically the **Lateral Rail Index (LRI)** and **Vertical Rail Index (VRI)**. These indexes are critical in evaluating the performance and risk level of various rail sections over time. Our objective was to build a supervised machine learning model that could predict future LRI and VRI values based on historical patterns, enabling better monitoring and planning for infrastructure and operational reliability.

Together, these two components demonstrate how operational railway data can be transformed into intelligent insights using data analytics and machine learning. Our work not only enhances visibility into current delay patterns but also enables forecasting of critical indicators, contributing to more efficient railway operations.

METHODOLOGY

Our project followed a structured, modular approach to analyse train delays, build a machine learning-based prediction model, and integrate the outcomes into a functional dashboard. In addition, we extended our work to include a separate Rail Index (RI) prediction module using real LRI/VRI data. The overall methodology was organized into the following phases:

1. Data Collection

We utilized two datasets for the Vaigai Express (Train No. 12635):

- **Transaction Data**: Included real-time arrival and departure timestamps at each station over a 10-day period.
- **Referential Schedule Data**: Contained scheduled arrival/departure times, halt durations, and inter-station distances between Chennai Egmore (MS) and Madurai Junction (MDU).

For the Rail Index module, a separate dataset containing **Lateral Rail Index (LRI)** and **Vertical Rail Index (VRI)** values across block sections was used.

2. Data Preprocessing

We standardized column names, removed whitespace, and parsed all timestamp columns into datetime format. Station codes were extracted from text and matched across datasets. Scheduled arrival and departure times were reconstructed as full datetime objects using the train's date of journey.

We also converted halt times into minutes and distances into numeric kilometres for further analysis.

3. Delay Computation

From the merged dataset, we computed the following key metrics:

- **Arrival Delay** = Actual Arrival Time − Scheduled Arrival Time
- **Departure Delay** = Actual Departure Time − Scheduled Departure Time
- **Total Journey Delay** = Delay at the terminal station (MDU)

These delays were filtered to remove missing values and extreme outliers. The cleaned data served as the foundation for both analysis and modelling.

4. Exploratory Data Analysis (EDA)

We used Seaborn and Matplotlib to generate various visual insights:

- **Boxplots** to observe delay variance across stations
- **Bar plots** to show average delays per station
- **Time series plots** to track day-wise delay progression

EDA helped us uncover patterns such as peak-delay stations, worst-performing days, and downstream impact of early delays.

5. Feature Engineering

To train the prediction model, we extracted relevant features:

- Day of the week
- Hour of departure
- Halt duration (in minutes)
- Departure delay at the current station
- Distance from the source station
- Weekend indicator flag

These features were either scaled or encoded where necessary.

6. Delay Prediction using Machine Learning

We used a **Random Forest Regressor** to predict arrival delay (in minutes). After splitting the dataset into training and testing sets, we used **RandomizedSearchCV** to tune hyperparameters for optimal performance.

The final model achieved an **R^2 score of 0.83**, indicating a strong fit between predicted and actual delay values. The model was saved and later integrated into a dashboard.

7. Streamlit Dashboard Deployment

We built a user-friendly dashboard using **Streamlit** that included:

- **Train Delay Analysis**: Visual trend of delay over the period
- **Route Performance**: Per-station delay distributions and delay patterns
- **Delay Prediction**: An interactive module to input journey features and obtain delay predictions

8. Rail Index Prediction (LRI/VRI)

As a parallel task, we developed a machine learning model to predict the **Rail Index**, using **Lateral Rail Index (LRI)** and **Vertical Rail Index (VRI)** data provided by CRIS. These indices reflect track smoothness and ride comfort, typically derived from oscillation sensors in test vehicles.

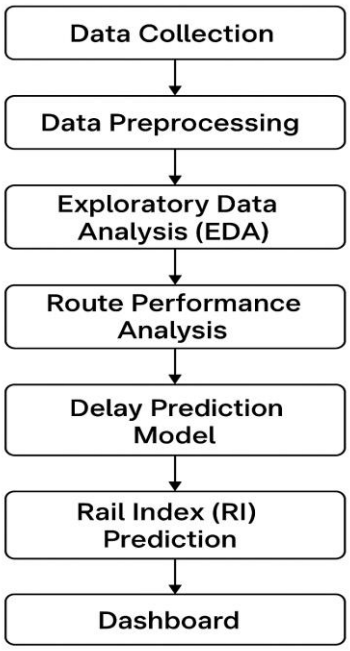
Instead of computing the RI from raw acceleration data, we were given a dataset with LRI/VRI values and operational attributes like section name, line code and zone. We performed the following steps:

- Cleaned the dataset and encoded categorical columns
- Scaled numeric fields such as speed and distance
- Applied **Random Regressor Stacking using Lightgbm and XGBoost** to model LRI and VRI separately

We evaluated the models using:

- **R^2 Score** to assess predictive fit
- **Mean Absolute Error (MAE)** and **Mean Squared Error (MSE)** for accuracy

The models showed strong performance with **R^2 scores around 0.96**, suggesting high reliability in forecasting RI values. This can assist track maintenance teams in identifying poor sections and scheduling targeted inspections.



RESULTS

Our project produced valuable outcomes in both train delay prediction and rail index estimation.

Train Delay Analysis & Prediction:

We analysed 10 days of Vaigai Express data, calculating arrival and departure delays at each station. Exploratory analysis revealed consistent delays at key stations and delay propagation along the route.

Using a Random Forest Regressor with features such as day of week, halt time, distance, and departure delay, we trained a delay prediction model. After tuning with RandomizedSearchCV, the model achieved:

- R^2 Score: 0.83

These results demonstrate strong predictive accuracy. The model was deployed in a Streamlit dashboard for real-time predictions.

Rail Index Prediction (LRI/VRI):

We also developed models to predict Lateral Rail Index (LRI) and Vertical Rail Index (VRI) using operational data. Lightgbm and XGBoost models with Random Regressor achieved:

- R^2 Score: 0.96

These models can assist in identifying rough track sections and aid in proactive maintenance planning.

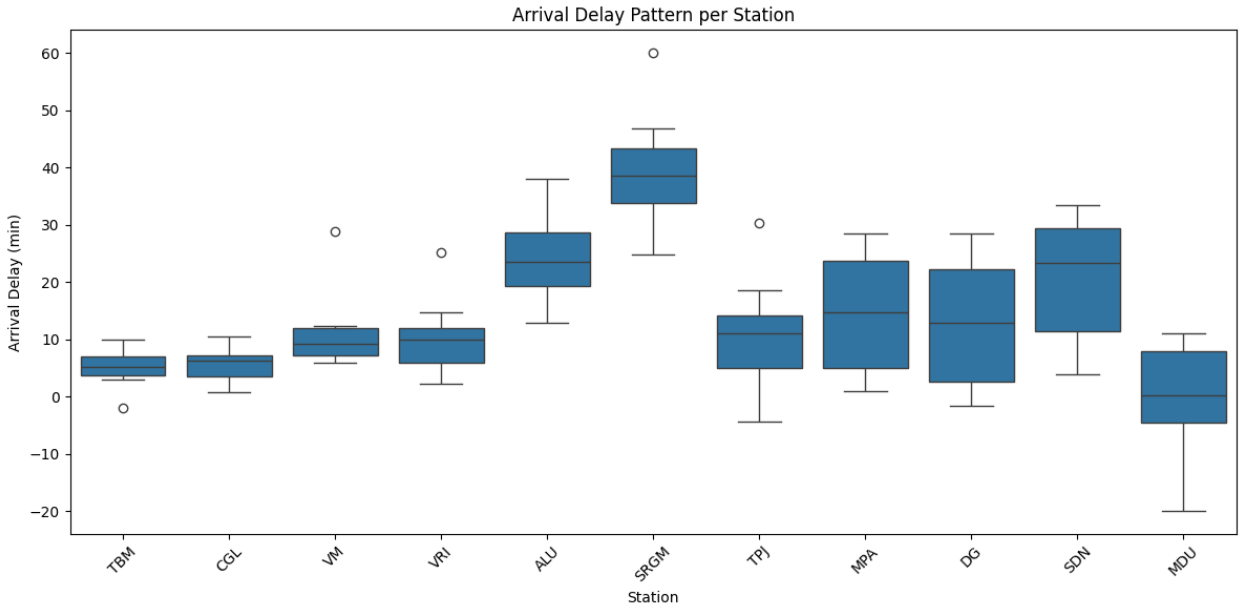


Figure 3: Box Plot of Arrival Delay

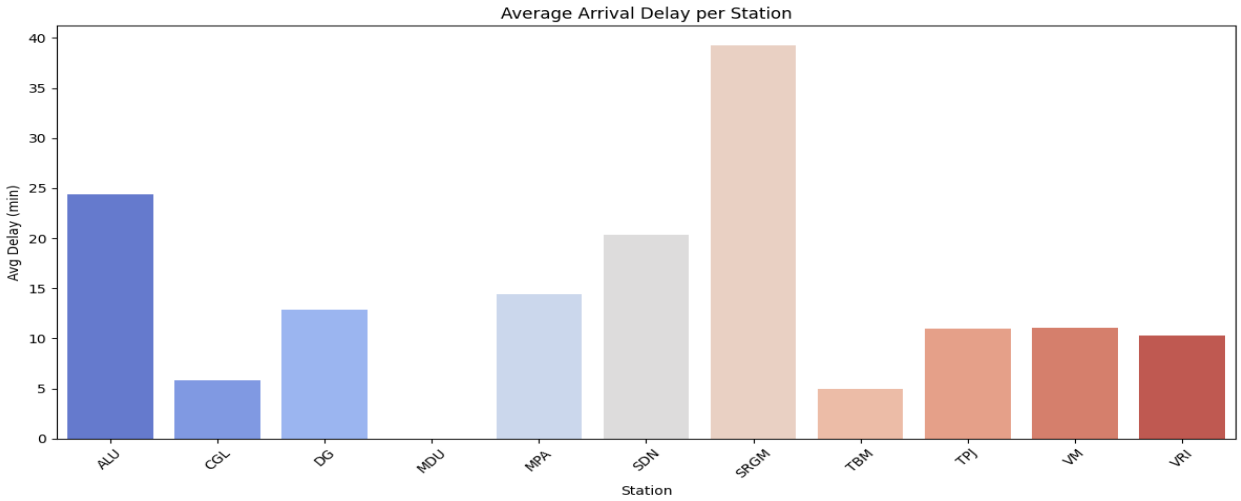


Figure 4: Average Delay per Station

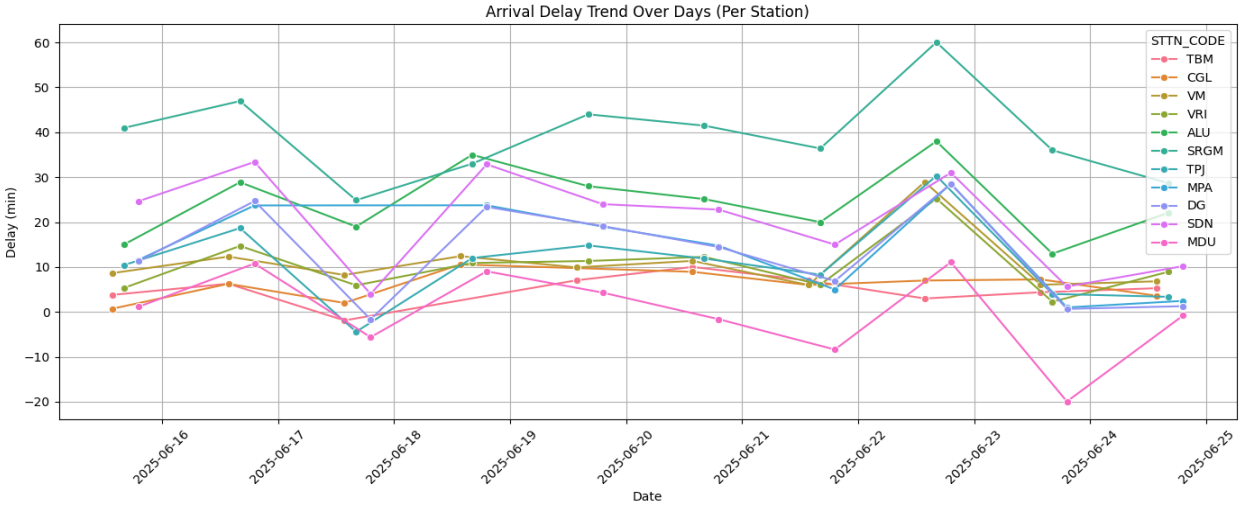


Figure 5: Arrival Delay Trend Over Days

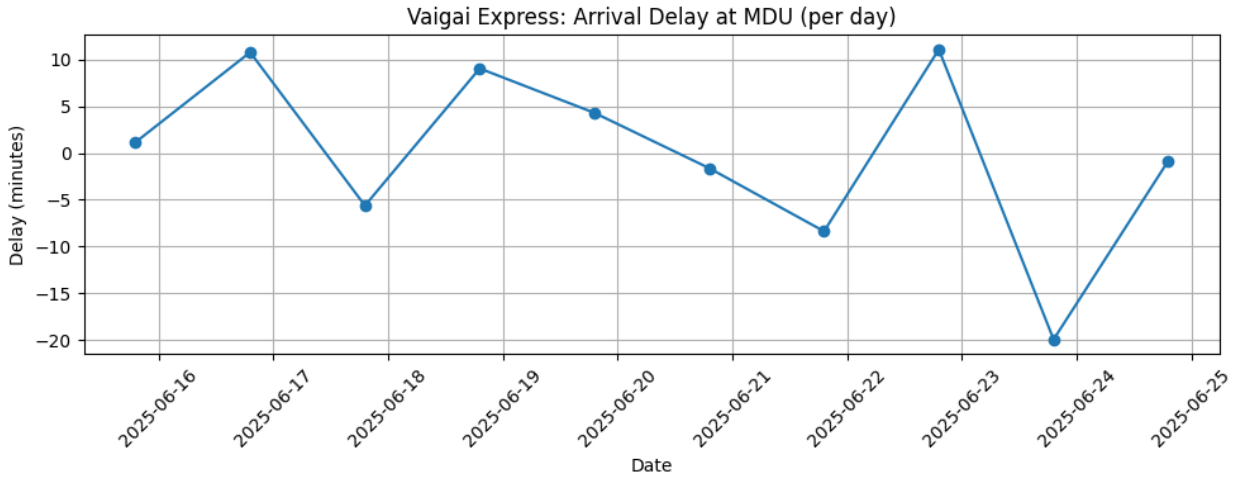


Figure 6: Arrival Delay at MDU Over Days

```
Best Parameters: {'max_depth': 16, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 120}
Optimized R² Score: 0.8305
```

Figure 7: Model Output of Train Delay Prediction

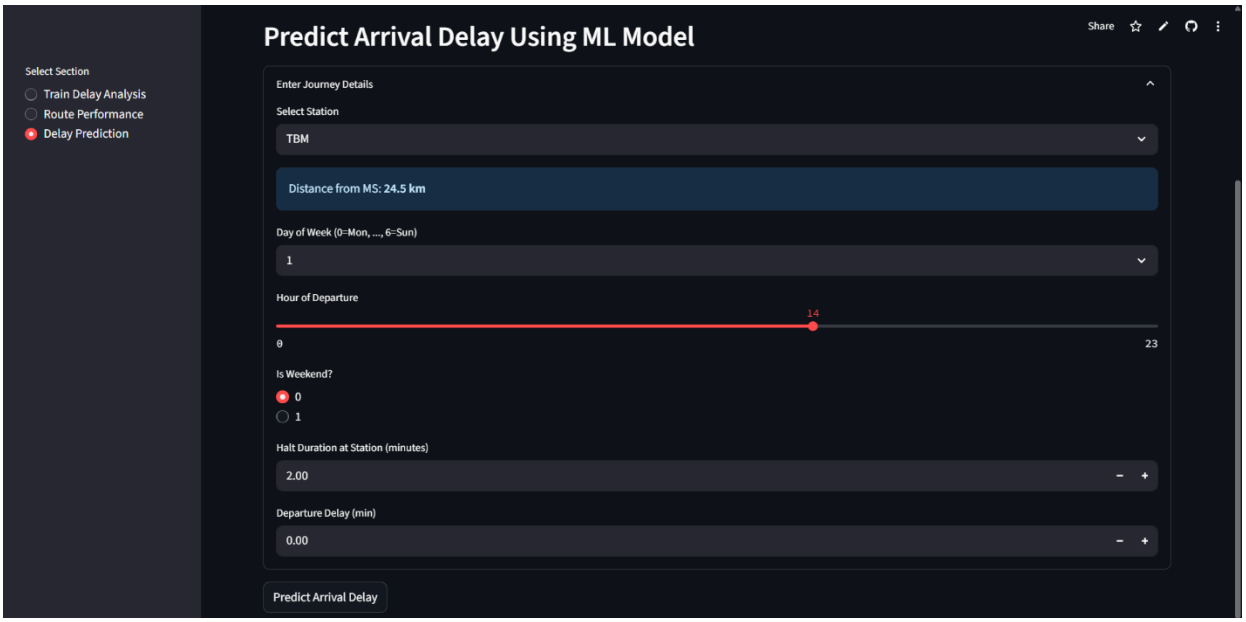


Figure 8: Dashboard Train Delay Prediction

```
print("\nStacking Results (Meta-model: Random Forest Regressor):")
print(f"R² Score : {r2_score(y_test, stacked_preds_rf):.4f}")
print(f"MAE      : {mean_absolute_error(y_test, stacked_preds_rf):.4f}")
print(f"MSE      : {mean_squared_error(y_test, stacked_preds_rf):.4f}")
print(f"RMSE     : {np.sqrt(mean_squared_error(y_test, stacked_preds_rf)):.4f}")
```

```
Stacking Results (Meta-model: Random Forest Regressor):
R² Score : 0.9689
MAE      : 0.0245
MSE      : 0.0013
RMSE     : 0.0355
```

Figure 9: Accuracy of Rail Index Prediction Model

CONCLUSION

Through this internship project at the Centre for Railway Information Systems (CRIS), under the Control Office Application (COA) department, we successfully developed a comprehensive solution for analysing train delay patterns and predicting arrival delays using machine learning. Our work focused on the Vaigai Express (Train No. 12635), covering the route between Chennai Egmore (MS) and Madurai (MDU), and utilized real operational data for a 10-day period.

We began by preprocessing both transaction and referential schedule data, aligning and merging them to compute meaningful metrics such as arrival delay, departure delay, and total journey delay. Exploratory Data Analysis (EDA) helped us identify station-wise delay trends and temporal patterns. This understanding formed the basis for building a feature-rich dataset that was used to train a Random Forest regression model. After hyperparameter tuning, our model achieved a strong predictive performance with an R^2 score of 0.83.

We deployed the insights and model into an interactive Streamlit dashboard, allowing users to explore route performance, visualize delay trends, and receive real-time delay predictions based on selected journey conditions. This dashboard serves as a proof-of-concept for how such tools can assist COA officials in making data-backed operational decisions.

In addition to delay prediction, we worked on a secondary module to predict the Rail Index (LRI/VRI), which reflects track quality and passenger comfort. Using a separate dataset and Lightgbm & XGBoost modelling, we were able to predict LRI and VRI with high accuracy, supporting use cases in maintenance planning and ride monitoring.

Overall, this project strengthened our understanding of real-world railway data systems, and demonstrated how analytics and AI can be practically applied in a critical public infrastructure domain. The experience also gave us valuable exposure to data engineering, model development, and dashboard deployment, and we believe this work can serve as a strong foundation for future innovation in Indian Railways' operational intelligence systems.

FUTURE SCOPE

Our project on train delay analysis, machine learning-based prediction, and rail index forecasting has demonstrated practical potential for data-driven railway operations. However, there are several areas where this work can be expanded to enhance its impact and scalability.

A key opportunity lies in extending the analysis to multiple trains and longer timeframes. Currently, we focused on Vaigai Express, but applying the same methodology to trains across different zones and seasons would help develop a more generalizable and robust prediction system.

Our current model uses temporal and operational features. Incorporating external factors such as weather, track maintenance schedules, and traffic density can improve prediction accuracy and make the system more adaptive to real-world scenarios.

The dashboard developed using Streamlit can also be enhanced into a real-time analytics tool, integrated directly with COA or NTES. This would support control room staff in making proactive decisions based on live delay forecasts and route performance metrics.

For the Rail Index Prediction, future enhancements could involve integrating RI data with GIS-based track visualization, enabling maintenance teams to pinpoint poor-quality sections and plan targeted interventions. This could serve as the foundation for a predictive maintenance system.

We also see value in incorporating model interpretability tools like SHAP to provide transparency on how features influence predictions. This would make the system more explainable and usable by operational staff.

Lastly, the models can be retrained continuously using real-time data, improving their accuracy and ensuring adaptability as conditions change. With further engineering, this project could evolve into a deployable solution contributing to smarter and more efficient railway operations.

REFERENCES

- Indian Railways. *Permanent Way Manual*. Ministry of Railways, Government of India. Retrieved from: https://indianrailways.gov.in
- Centre for Railway Information Systems (CRIS). *Control Office Application (COA) Overview*. CRIS Internal Documentation, 2024.
- Pandas Development Team. (2023). *Pandas: Powerful Python Data Analysis Toolkit*. Version 2.0. Retrieved from: https://pandas.pydata.org
- Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
- Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). *Hyperparameter tuning and nested resampling for random forest construction*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3).
- Streamlit Inc. (2024). *Streamlit Documentation*. Retrieved from: https://docs.streamlit.io
- Seaborn Documentation. *Statistical Data Visualization*. Retrieved from: https://seaborn.pydata.org
- Dataset: Vaigai Express 12635 Transaction Data and Schedule Data (June 2025). *Provided by CRIS COA Department*, Internal Dataset.
- Rail Index Data (LRI/VRI) for Rail Track Evaluation. *Supplied by CRIS*, Engineering Division (Internal Reference Dataset, July 2025).
- National Train Enquiry System (NTES). *Real-time Train Status Platform*. Indian Railways. Retrieved from: https://enquiry.indianrail.gov.in
- Dashboard Link: https://train-delay-prediction-dashboard-hsgxxyfpx2qrrs6ieg5sne.streamlit.app/
- Vaigai Train Delay Prediction Repository: https://github.com/kavinraam/train-delay-prediction-dashboard