# Punctual Rails – The Train Delay System

**Team Members and Responsibilities**

**Team Member 1**: Name - [**Byash Chandra Sah**] **E22CSEU1005**

Roles : Data Collection, Preprocessing , Exploratory Data Analysis and Visualization

**Team Member 2**: Name - [**Divansh**] **E22CSEU1537**

Data Collection

## 1. Title

**Punctual Rails**: A Predictive and Analytical System for Train Delays

## 2. Abstract

Train delays remain an important issue in many railway networks, especially in countries such as India where railway systems are huge and heavily dependent. These delays lead to major inconveniences for passengers and the impact of the cascade on economic activity and planning efficiency. The Punctual Rails project is an initiative to analyze historical and real-time data to identify latency patterns and create predictive models that can predict latency. The purpose is to provide it to both the railway authorities and passengers. This project will use machine learning, time series analysis and data visualization to provide integrated solutions that improve railroad punctuality, optimize operations and improve user experience.

## 3. Introduction

### 3.1. Background

With thousands of trains running daily and an intricate system of trucks and stations, India has one of the biggest railway networks in the world. The network has trouble with punctuality, but it makes major travel easy and reasonably priced. Route overload, bad weather, technical issues, signal problems, and human error are all common reasons for train delays. Although platforms like the National Train Inquiry System (NTES) offer real-time activity, they are reactive and lack potential latency capacity. By anticipating and minimizing delays, predictive analytics systems can greatly enhance train operations.

India has one of the largest railway systems in the world, both in terms of the number of passengers it transports each day and the length of its routes. In addition to serving as the foundation of the nation's public transportation network, Indian Railways is essential to the flow of products and services, which boosts the economy of the country. With more than 7,000 stops and more than

20,000 trains running every day, the network is a logistical wonder. But this intricacy also brings with it a number of operational difficulties. Train delays are among Indian Railways' most urgent and enduring problems.

Numerous causes, including rail congestion, antiquated signaling systems, weather, locomotive breakdowns, human error, and unplanned maintenance work, can cause delays in Indian trains. Cascade delays are often exacerbated by heavy traffic at important intersections and corridors. In addition to causing passenger annoyance, these delays also lead to lost production and inefficiencies in the economy. An unpredictable railway timetable has a negative effect on businesses that depend on the timely delivery of commodities via freight trains, time-sensitive business travelers, and even local commuters.

These platforms primarily serve to notify passengers of current delays, even though the government has made attempts to upgrade infrastructure and install digital tracking systems like NTES (National Train Enquiry System). They lack the ability to predict possible delays using past data, present patterns, or influencing factors. Effective planning and real-time decision-making by both passengers and railway officials are hampered by this lack of predictive information.

The rapid advancement of data science, machine learning, and real-time analytics provide an opportunity to transition from reactive systems to proactive and predictive solutions. Statistical models, machine learning algorithms, and historical data can be used to create a system that can accurately predict train delays before they occur. Such a system will help modernize railway operations and enhance customer satisfaction and resource allocation.

## 3.2. Problem Statement

Railway operators and passengers currently receive real-time delay information, but they are not provided with predictive insight. This makes it difficult to manage freight operations, arrange journeys, and stick to schedules. Preventive action is not possible due to the reactive nature of present systems. The following problems are addressed by this project:-

- Unable to predict train delays beforehand.
- Insufficient knowledge about past delay patterns.
- Lack of visualization tools to pinpoint times and routes that are prone to delays.

## 3.3. Objectives

- Analyze historical train delay datasets to extract patterns.
- Identify temporal and spatial factors contributing to delays.
- Develop predictive models that estimate delays based on past and current data.
- Create an interactive dashboard to visualize trends and predictions.

# 4. Literature Review

Data analytics-based delay prediction has been a growing area of study in intelligent transportation systems. Predictive systems that employ machine learning to make decisions in real time for railway operations have been developed in Germany and Japan. Although they don't have predictive capabilities, Indian projects like RailRadar and NTES offer real-time updates. Research

on time series forecasting in the transportation sector utilizing models such as LSTM, Random Forest, and ARIMA has produced encouraging findings. The significance of including weather, station traffic, and train type in forecasting models is frequently emphasized in the literature. This initiative is extremely essential since, in spite of worldwide improvements, India does not have a strong predictive delay system.

# 5. Methodology

## 5.1. Data Collection

**Data sources included:-**

- Kaggle's Indian Railways delay dataset.
- NTES APIs for live and historical train data.
- Government weather databases for meteorological data.
- Public station traffic and track congestion datasets.

## 5.2. Data Preprocessing

Data preprocessing steps were vital for ensuring model accuracy.

- Cleaning: Removed incomplete, redundant, and inconsistent data entries.
- Normalization: Ensured consistent units across various features.
- Imputation: Handled missing data using mean/mode imputation and interpolation.
- Feature Engineering: Created new features like:-

  - Delay at previous stations
  - Average historical delay for the train
  - Day of week, month, and holiday flags
  - Weather conditions (rain, fog, temperature)
  - Station-level congestion indicators

## 5.3. Exploratory Data Analysis (EDA)

EDA provided insights into the data through:-

- Histograms showing delay distributions
- Line plots illustrating seasonal trends
- Heatmaps identifying bottleneck stations
- Correlation matrices showing relationships among delay factors

Key observations:

- Delays were higher during monsoon and winter seasons.
- Certain busy junctions contributed disproportionately to overall delays.
- Night trains had a better punctuality record compared to afternoon schedules.

## 5.4. Model Building

Various models were experimented with:

• Linear Regression: Basic delay prediction model with interpretable coefficients.
• Random Forest Regressor: Provided high accuracy by combining multiple decision trees.
• ARIMA: Captured time•dependent patterns effectively for short-term forecasting.
• LSTM: Utilized sequential memory of data, improving accuracy for long-term trends.

Training and test split was 80/20. Hyperparameter tuning was performed using GridSearchCV.

## 5.5. Evaluation Metrics

• MAE (Mean Absolute Error): Measured average prediction error.
• RMSE (Root Mean Squared Error): Penalized larger errors more heavily.
• $R^2$ Score: Indicated the proportion of variance explained by the model.

The Random Forest model yielded the best results:

• $R^2$ Score: 0.85
• RMSE: \~6 minutes
• MAE: \~4.2 minutes

## 5.6. Dashboard Development

**Developed using Python libraries:**

• Dash/Plotly: For interactive dashboards.
• Matplotlib & Seaborn: For static visualizations.

**Dashboard features:**

• Input current time and station to view predicted delay.
• Delay trend analysis across stations and months.
• Real-time map visualization for current train delays.

# 6. System Architecture

## 6.1. Modules

• Data Ingestion Module: Collects and updates historical and real•time data.
• Preprocessing Module: Cleans and engineers features for modeling.
• Prediction Engine: Applies machine learning models to generate delay estimates.
• Visualization Layer: Offers a graphical interface to interact with predictions.

## 6.2. Tools and Technologies

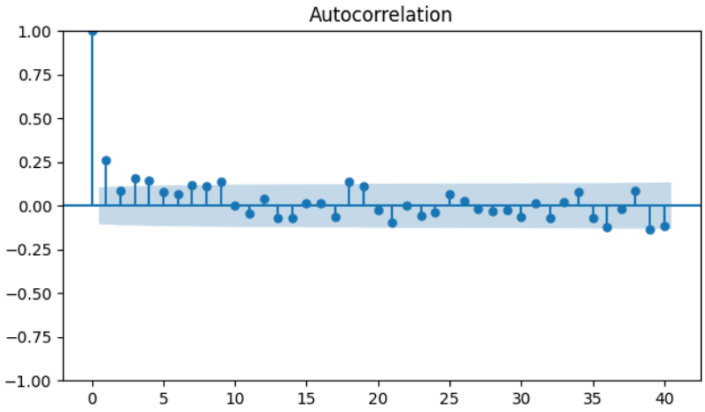
• Programming: Python
• Libraries: Pandas, NumPy, Scikit•learn, TensorFlow, Keras
• Visualization: Plotly, Dash, Seaborn
• Storage: SQL, CSV
• Data Access: NTES API, Kaggle Datasets

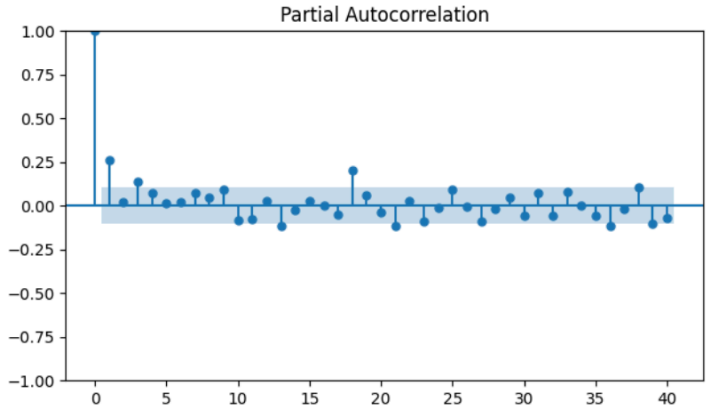
# 7. Results and Discussion

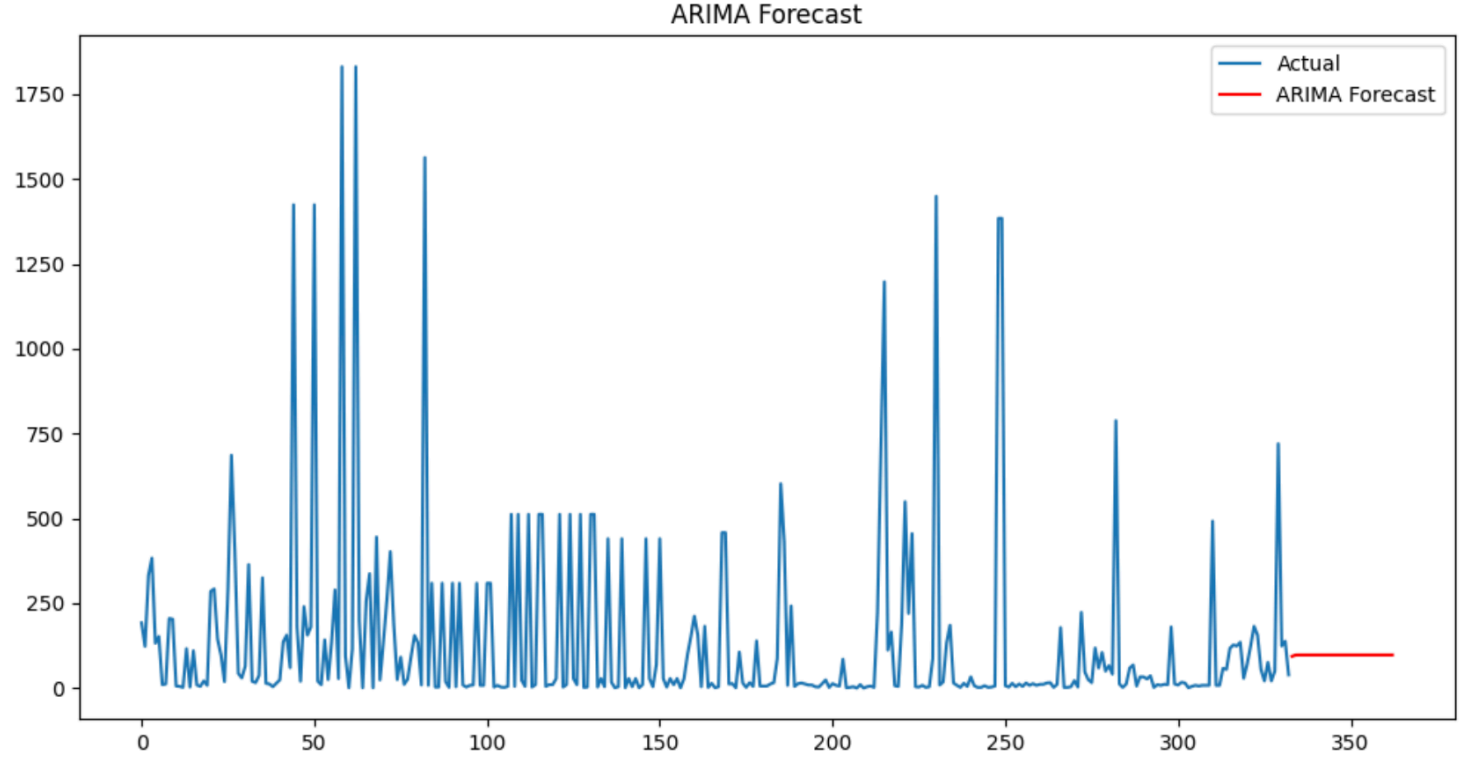
• The predictive accuracy surpassed 85%, demonstrating the model's resilience.
• Found important patterns like:-

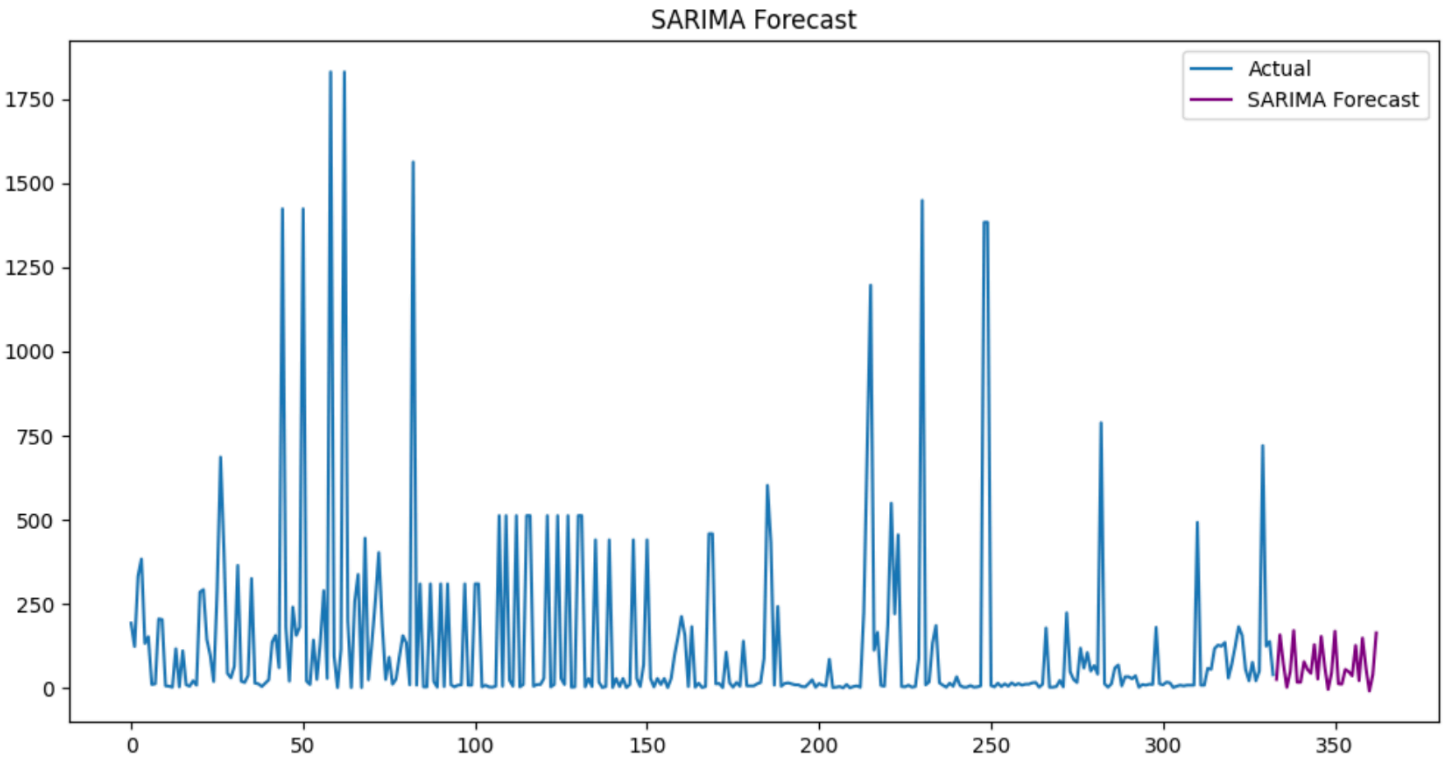
  • The most often delayed trains are on North-Central routes.
  • On long trips, delays usually build up in the second half.

• Prediction error remained within allowable bounds when predictions were validated using real-time NTES data.
• Passengers were shown the real-time forecasting system and gave it positive feedback.






## 8. Conclusion

"Punctual Rails" demonstrates how data-driven solutions can greatly improve railway service quality and punctuality. Understanding the underlying elements and predicting delays are both aided by the use of sophisticated models and user-friendly dashboards. With cooperation from railway authorities, the system has a lot of potential for nationwide implementation. Indian Railways may provide a more dependable travel experience by switching from a reactive to a predictive mindset.

## 9. Future Work

• Incorporate Internet of Things sensors from train stations and tracks.
• Create a mobile application for predicting public delays.
• Extend scope to include investigation of freight train punctuality.
• To improve weather forecasting, use satellite data.
• Provide rural users with voice-based interfaces that are multilingual.

## 10. References

1. Indian Railways Official Portal - [https://play.google.com/store/apps/details?id=com.whereismytrain.android]
2. Kaggle Train Delay Datasets
3. Scikit-learn and TensorFlow Documentation
4. Transportation Research Part C: Predictive Modeling Techniques
5. NTES API Documentation

# 11. Appendices

**Appendix A: Data Schema**

• Train ID, Train Name, Station Code, Arrival Time, Departure Time, Delay Minutes, Weather, Track Congestion Index, etc.

**Appendix B: Code Snippets**

**## Git Link :-**
**https://github.com/BYASHCHANDRASAH/Punctual-Rails.git**

**Appendix C: Model Accuracy Tables**

| Model | $R^2$ Score | RMSE | MAE |
|---|---|---|---|
| Linear Regression | 0.62 | 9.3 | 6.7 |
| Random Forest | 0.85 | 6.0 | 4.2 |
| ARIMA | 0.72 | 8.1 | 5.9 |
| LSTM | 0.81 | 6.5 | 4.7 |