Traffic Congestion Prediction

Submitted by:
Purvika Bansal-102317121
Navnidhi-102317298

BE Third Year CSE

Submitted to:

Dr. Anjula Mehto

Assistant Professor

Computer Science and Engineering

Department Thapar Institute of Engineering

and Technology, Patiala

November 2025

TABLE OF CONTENTS

S. No	Topic	Page No.
1	Introduction or Project Overview	3
2	Problem Statement	4
3	Overview of the Dataset used	5
4	Project workflow	6
5	Results	7-9
6	Conclusion	10

Introduction or Project Overview

The dataset used in this project is the Smart Mobility Traffic Dataset**,** sourced from Kaggle. It contains real-world traffic information collected from different urban road segments and includes a wide variety of features representing traffic patterns, environmental conditions, and temporal variations. Because it captures multiple dimensions of traffic behavior, the dataset is highly suitable for developing machine learning models for congestion prediction.

The dataset consists of several traffic-related features**,** such as vehicle speed, road occupancy, traffic flow rate, CO_2 emissions, and lane-specific attributes. These variables describe the operational characteristics of the road network and help identify whether a particular segment is experiencing smooth flow, moderate traffic, or severe congestion. Among these, speed and occupancy act as key indicators for understanding congestion levels.

Apart from traffic attributes, the dataset includes important meteorological and environmental features**,** such as temperature, rainfall, humidity, and overall weather status. Weather conditions have a significant impact on driving patterns, as poor visibility or heavy rainfall typically results in reduced vehicle speed and increased congestion. Incorporating these features allows the model to account for the influence of external environmental factors.

The dataset also contains detailed geospatial information, including latitude, longitude, and location codes. These spatial features make it possible to group similar road segments, identify congestion hotspots, and analyse region-specific traffic patterns. The inclusion of spatial identifiers ensures that the model can generalize across different locations rather than overfitting to a single area.

Another valuable component of the dataset is its time-based features, such as timestamps, hour of the day, day of the week, and weekday/weekend labels. Traffic behaviour shows strong temporal dependencies—for example, morning and evening rush hours display predictable congestion patterns. These temporal attributes allow the extraction of cyclical trends and support the creation of engineered features such as rush-hour flags and time-lag variables.

Overall, the Smart Mobility Traffic Dataset offers a comprehensive multi-dimensional representation of traffic conditions. By combining traffic, weather, spatial, and temporal variables, it provides a strong foundation for developing accurate traffic congestion prediction models. Its richness enables experimentation with both traditional machine learning algorithms and advanced deep learning methods, allowing the project to analyse how different factors collectively influence congestion levels.

Problem Statement

Traffic congestion is a widespread and growing challenge in rapidly urbanizing cities across the world. With increasing population density, rising vehicle ownership, and the expansion of road networks, transportation systems are becoming increasingly stressed beyond their designed capacity. As a result, commuters frequently face long delays, unpredictable travel times, and heightened fuel consumption. These challenges not only inconvenience travelers but also contribute to increased carbon emissions, stress, and significant economic losses due to wasted time and reduced productivity.

Traditional traffic management approaches rely heavily on reactive measures such as manual monitoring, fixed traffic signal timings, and simple rule-based systems. However, these methods are often insufficient because they cannot anticipate traffic conditions in advance. Traffic behavior is dynamic and influenced by multiple non-linear factors including weather variations, accidents, road construction, special events, and time-based patterns. Such complexity makes it difficult for conventional systems to forecast congestion accurately, creating a strong need for data-driven and intelligent prediction techniques.

The central problem addressed in this project is predicting traffic congestion levels using Machine Learning techniques based on historical and real-time features. The dataset incorporates variables such as road occupancy, GPS coordinates, weather conditions, and temporal indicators. These variables undergo preprocessing, encoding, and scaling before being used as inputs to multiple ML models. The goal is to classify congestion into three categories: **Low, Moderate, and High**, with high accuracy and consistency.

This project evaluates a range of Machine Learning and Deep Learning models including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, LSTM, and 1D-CNN—to determine which approach delivers the best predictive performance for time-dependent traffic patterns. Model effectiveness is assessed using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to ensure a comprehensive evaluation. Ultimately, the objective is to develop a robust and efficient congestion prediction system that can support traffic authorities, smart-city platforms, and navigation systems by providing early warnings about potential congestion. A reliable prediction model can enable proactive traffic control, improved route planning, reduced travel delays, and better urban mobility. By addressing this challenge, the project contributes to the advancement of intelligent transportation systems and promotes data-driven decision-making for future smart cities.

Overview of the Dataset used

The dataset used in this project is the Smart Mobility Traffic Dataset, sourced from Kaggle. It contains real-world traffic information collected from different urban road segments and includes **14 features** representing traffic patterns, environmental conditions, and temporal variations. Because it covers multiple aspects of traffic behavior, the dataset is highly suitable for developing machine-learning models for congestion prediction.

The dataset includes several traffic-related features such as road occupancy percentage, traffic flow rate, CO_2 emissions, and lane-specific attributes. These parameters describe the operational state of the road network and help determine whether a particular location is experiencing smooth flow, moderate traffic, or heavy congestion. Among these, occupancy is an especially important indicator of congestion levels.

In addition to traffic metrics, the dataset provides meteorological and environmental information, including temperature, rainfall, humidity, and overall weather status. Weather significantly influences traffic conditions, as rain, fog, or poor visibility generally reduces driving speed and increases the likelihood of congestion. Including these features enables the model to capture the impact of environmental changes on traffic flow.

The dataset also contains detailed geospatial details, such as latitude, longitude, and location-specific codes. These spatial features are useful for grouping similar road segments, identifying congestion hotspots, and analyzing region-based traffic behavior. Incorporating spatial identifiers ensures that the model can generalize to different locations rather than overfitting to a single area.

Another key component of the dataset is its time-based features, including timestamps, hour of the day, day of the week, and weekday/weekend labels. Traffic often follows strong temporal patterns — for example, morning and evening rush hours show predictable congestion peaks. These time-related attributes help extract cyclical trends and enable the creation of engineered features such as rush-hour indicators and time-lag variables.
The dataset contains thousands of records, offering sufficient quantity for splitting into training, validation, and testing sets. Before model development, preprocessing steps are performed, including handling missing values, encoding categorical variables such as weather conditions and traffic status, and scaling numerical features to ensure consistency across all models.

Overall, the Smart Mobility Traffic Dataset provides a comprehensive and multi-dimensional view of real-world traffic conditions. By integrating traffic, weather, spatial, and temporal variables, it forms a strong foundation for accurate congestion prediction.

Project Workflow

The project follows a structured Machine Learning workflow, consisting of data preparation, feature engineering, model development, evaluation, and deployment. Each stage contributes to building a robust and efficient traffic congestion prediction system

1. Data Preprocessing

- **Handling Missing Values:** Filled weather-related gaps and interpolated GPS coordinates wherever necessary.
- **Encoding Categorical Variables:** Converted textual attributes such as weather status, traffic light conditions, and other categorical fields into numerical form.
- **Scaling Numerical Features:** Normalized continuous variables like speed, occupancy, CO_2 emissions, and flow rate to ensure uniform model performance.

2. Feature Engineering

- **Time-Based Features:** Extracted hour of day, day of week, month, weekday/weekend indicator, and rush-hour flags.
- **Lag Features:** Created short-term historical features such as 5–15 minute averages of speed, occupancy, and previous congestion levels.
- **Derived Metrics:** Generated new features like speed variance, accident indicators, and sentiment-based metrics (if available).
- **Spatial Grouping:** Clustered GPS coordinates into meaningful road segments to identify congestion hotspots.

3. Model Building

- **Baseline Models**: Logistic Regression and Decision Tree were used to establish initial performance benchmarks.
- **Advanced ML Models**: Random Forest and Gradient Boosting helped capture complex non-linear patterns for improved accuracy.
- **Deep Learning Models**: LSTM and 1D-CNN were applied to learn temporal and sequential traffic patterns for enhanced prediction capability**.**

4. Training and Validation

- **Time-Based Train–Test Split:** Ensured chronological order is preserved for realistic forecasting.
- **Rolling Window Cross-Validation:** Applied to evaluate model stability over different time intervals.
- **Evaluation Metrics:** Accuracy, F1-score, Precision, Recall, and Confusion Matrix for multi-class congestion prediction

5. Deployment / Extensions

- **Streamlit Dashboard:** Developed to visualize predicted congestion levels across different locations and time periods.
- **Real-Time Integration:** System can connect to APIs providing live GPS and weather data for real-time congestion forecasting.

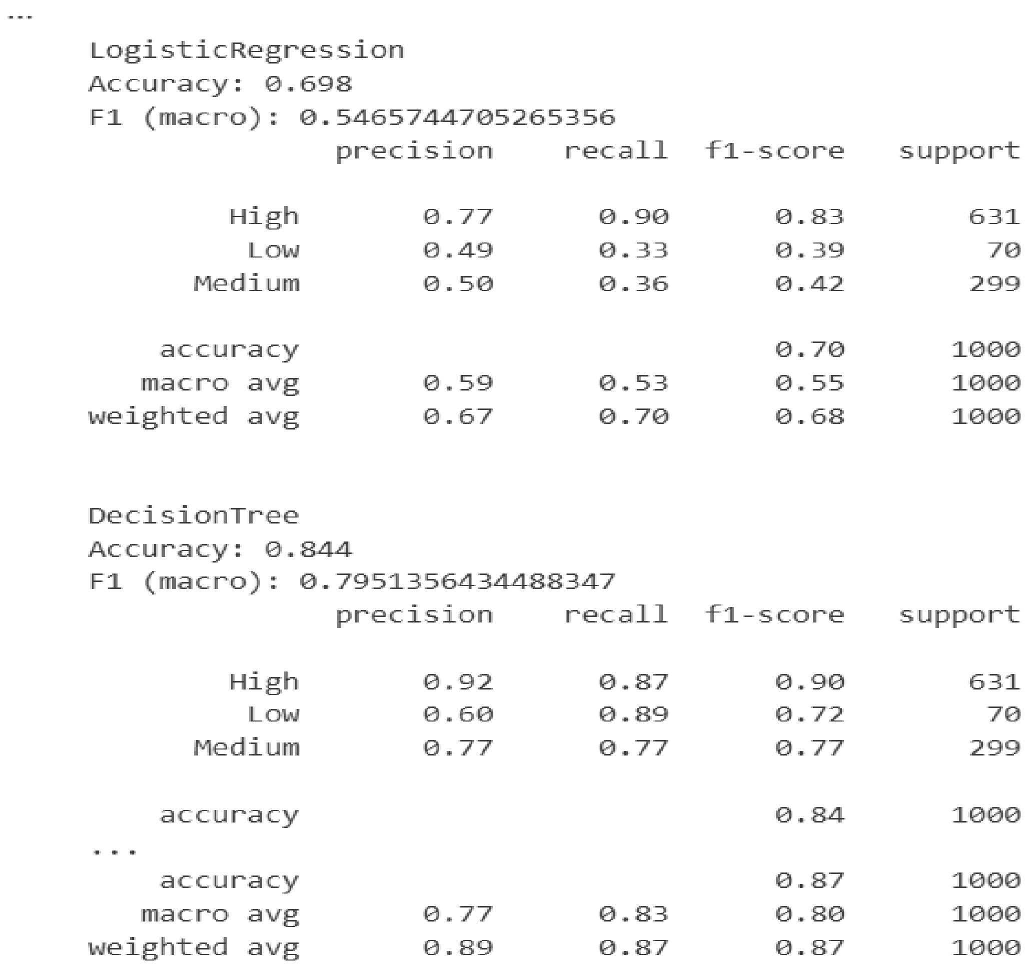
Results

Logistic Regression

The Logistic Regression model achieved an accuracy of 0.698, but struggled to correctly classify Medium and Low congestion levels, resulting in a lower macro F1-score. This indicates that the model performs reasonably well for high congestion cases but lacks balance across all classes.

Decision Tree

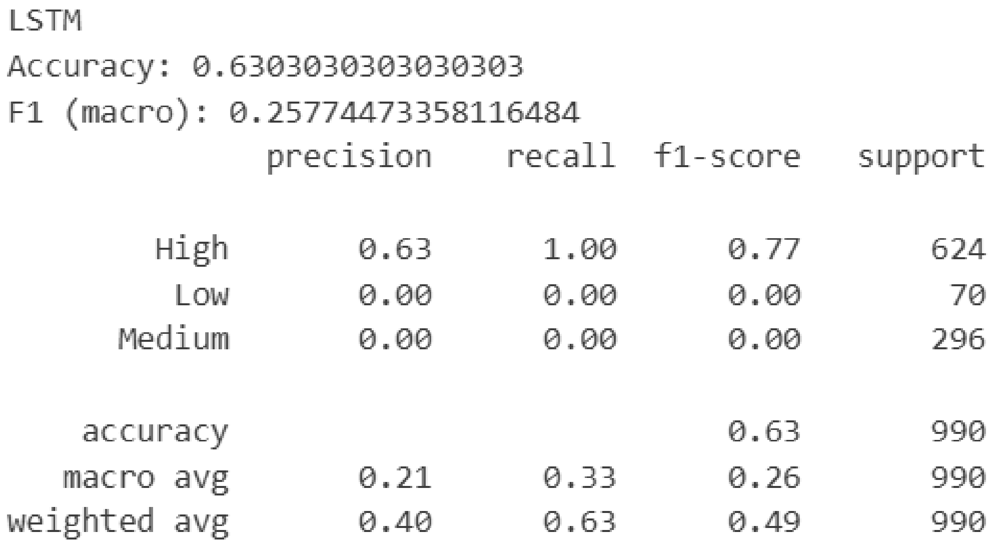
The Decision Tree model performed significantly better with an accuracy of 0.844, showing strong classification performance across all congestion categories. Its higher F1-score reflects improved handling of class imbalance and better overall prediction reliability.



```
...
     LogisticRegression
     Accuracy: 0.698
     F1 (macro): 0.5465744705265356
                   precision    recall  f1-score   support

            High       0.77      0.90      0.83       631
             Low       0.49      0.33      0.39        70
          Medium       0.50      0.36      0.42       299

        accuracy                           0.70      1000
       macro avg       0.59      0.53      0.55      1000
    weighted avg       0.67      0.70      0.68      1000


     DecisionTree
     Accuracy: 0.844
     F1 (macro): 0.7951356434488347
                   precision    recall  f1-score   support

            High       0.92      0.87      0.90       631
             Low       0.60      0.89      0.72        70
          Medium       0.77      0.77      0.77       299

        accuracy                           0.84      1000
     ...
        accuracy                           0.87      1000
       macro avg       0.77      0.83      0.80      1000
    weighted avg       0.89      0.87      0.87      1000
```

LSTM

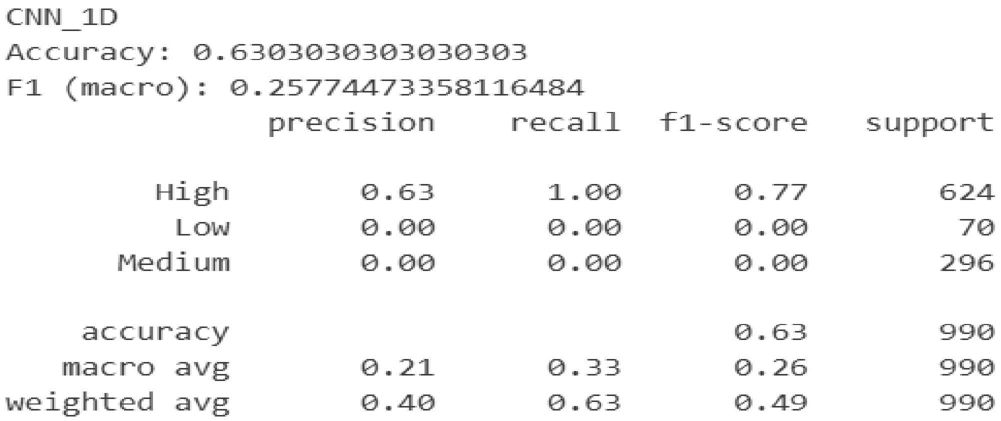
The LSTM model achieved an accuracy of 0.63, but failed to classify Low and Medium congestion levels, predicting mostly the High class. This indicates the model struggled with class imbalance and temporal dependencies in the dataset.



```
LSTM
Accuracy: 0.6303030303030303
F1 (macro): 0.25774473358116484
              precision    recall  f1-score   support

        High       0.63      1.00      0.77       624
         Low       0.00      0.00      0.00        70
      Medium       0.00      0.00      0.00       296

    accuracy                           0.63       990
   macro avg       0.21      0.33      0.26       990
weighted avg       0.40      0.63      0.49       990
```

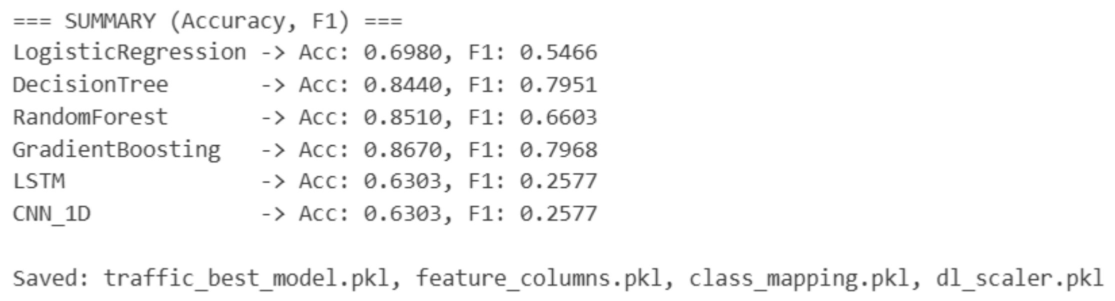
1D-CNN

Similar to LSTM, the CNN model reached an accuracy of 0.63, but showed poor generalization across all congestion categories except High. The model heavily overfitted and was unable to differentiate between Low, Medium, and High congestion levels.



```
CNN_1D
Accuracy: 0.6303030303030303
F1 (macro): 0.25774473358116484
              precision    recall  f1-score   support

        High       0.63      1.00      0.77       624
         Low       0.00      0.00      0.00        70
      Medium       0.00      0.00      0.00       296

    accuracy                           0.63       990
   macro avg       0.21      0.33      0.26       990
weighted avg       0.40      0.63      0.49       990
```

Among all evaluated models, Gradient Boosting performed the best, followed by Random Forest and Decision Tree, while deep learning models like LSTM and CNN underperformed due to class imbalance and limited sequential patterns in the dataset.



```
=== SUMMARY (Accuracy, F1) ===
LogisticRegression -> Acc: 0.6980, F1: 0.5466
DecisionTree       -> Acc: 0.8440, F1: 0.7951
RandomForest       -> Acc: 0.8510, F1: 0.6603
GradientBoosting   -> Acc: 0.8670, F1: 0.7968
LSTM               -> Acc: 0.6303, F1: 0.2577
CNN_1D             -> Acc: 0.6303, F1: 0.2577

Saved: traffic_best_model.pkl, feature_columns.pkl, class_mapping.pkl, dl_scaler.pkl
```

Conclusion

This project successfully explored the use of Machine Learning techniques to predict traffic congestion based on real-world mobility, environmental, and temporal data. Multiple models were trained and evaluated, ranging from traditional algorithms to deep learning architectures. The results demonstrated that classical machine learning models—specifically Gradient Boosting, Random Forest, and Decision Tree—performed significantly better than deep learning approaches like LSTM and CNN for this dataset.

The best-performing model was Gradient Boosting, achieving the highest accuracy and macro F1-score. This shows that tree-based ensemble models are more effective at handling mixed feature types, class imbalance, and non-linear traffic behavior. In contrast, deep learning models struggled due to limited sequential patterns and data imbalance, indicating that they may require larger, more balanced, and time-continuous datasets to perform optimally.

Overall, this project highlights that data-driven predictive modeling can greatly support smarter traffic management systems. Accurate congestion forecasting can help optimize route planning, improve urban mobility, reduce fuel consumption, and enhance commuter experience. With further improvements such as real-time data integration, hyperparameter tuning, feature enrichment, and handling class imbalance, the model can be extended into a fully deployable intelligent traffic monitoring solution for smart cities.