ML for Traffic Flow Prediction in Urban Areas

Team No: 16
Student 1: Mrinal Pandey (PES2UG23CS353)
Student 2: Mekala Babu Abhinav (PES2UG23CS338)

Problem Statement

Urban areas are increasingly facing severe traffic congestion due to the growing number of vehicles, particularly during peak hours. This results in inefficiency, increased travel time, and commuter frustration. The objective of this project is to build a machine learning model that predicts traffic congestion levels across multiple roads in an urban setting using historical speed data and temporal features such as day of the week and time of day.

High-Level Architecture

- [Raw METR-LA Dataset (CSV)]
  |
  ▼
  [Data Cleaning & Structuring]
  - Timestamps, street_no, traffic_speed
  - Remove missing/invalid readings
  |
  ▼
  [Feature Engineering]
  - day_of_week, hour, minute_of_day
  - sin/cos encodings for cyclic time
  - lag_1, lag_2, lag_12 (past 5,10,60 min)
  |
  ▼
  [Normalization & Data Split]
  - StandardScaler normalization (train-only)
  - 70% train / 15% val / 15% test split
  |
  ▼
  [Model Training & Evaluation]
  - Persistence Baseline
  - XGBoost Regressor
  - LSTM Neural Network
  |
  ▼
  [Denormalization + Metrics]
  - MAE, RMSE, R^2
  - Comparison across models

Results

Model	MAE	RMSE	R^2	Remarks
Persistence Baseline	3.42	8.32	0.868	Strong short-term predictor
XGBoost Regressor	3.78	7.91	0.881	Captures non-linear patterns
LSTM Neural Network	2.881	—	0.9127	Best performance; learns temporal patterns

Inferences from Graphs and Analysis

• Daily Cyclic Patterns: Distinct morning (8 AM) and evening (5–7 PM) slowdowns validate temporal features.

• Bimodal Speed Distribution: Majority speeds (60–70 mph) indicate free flow; smaller peaks near 0 mph show congestion.

• Weekend vs Weekday Trends: Weekends have higher speeds, confirming day_of_week relevance.

• Heatmap Insights: Rush-hour congestion (red) and free-flow (green) confirm strong cyclicity.

• Spatial Correlation: Nearby sensors show similar speed trends, useful for spatio-temporal modeling.

• Model Behavior: LSTM predictions are smoother and more accurate across all time intervals.

Conclusion

The LSTM model achieved the best performance (MAE ≈ 2.881, R^2 ≈ 0.9127), demonstrating its ability to learn temporal dependencies in urban traffic data. The study shows that traffic speeds exhibit strong autocorrelation and cyclic patterns. Future improvements include incorporating spatio-temporal graph models (STGCN/DCRNN) and adding external factors such as weather and events for enhanced robustness.