Traffic Congestion Prediction and Flow Optimization

Iteration 04

Team Members: Jainil Desai, Adit Vakil, and Kalpan Shah

Course: DS5110 – Data Science Project

1. Dataset Description

Our team is using the **Traffic Inventory 2024 Dataset** from the **U.S. Department of Transportation (US DOT)**. The dataset can be accessed at: `https://data.transportation.gov/`

The dataset includes traffic volume counts, roadway classifications, average daily traffic, congestion patterns, and vehicle composition across major U.S. highways. Each record includes information such as the route number, station ID, geographic coordinates, vehicle counts, and temporal data (day, month, and year).

We chose this dataset because it provides high-quality, publicly available data collected by a reliable government source. It aligns directly with our project goal of predicting traffic congestion levels and analyzing the causes behind congestion in urban areas. The structure of the dataset, which combines both spatial and temporal dimensions, makes it ideal for data-driven modeling and trend analysis.

2. Tools and Methodologies

Our project will use a combination of **Python**, **Pandas**, and **NumPy** for data processing, along with **scikit-learn** for predictive modeling. Visualization and exploratory data analysis will be conducted using **Matplotlib**, **Seaborn**, and **Plotly**.

For model development, we plan to start with regression-based approaches (Linear and Random Forest Regression) to predict congestion levels based on traffic volume and temporal features. Later, we will experiment with deep learning models using **TensorFlow** for higher accuracy.

This toolset offers strong community support, scalability, and compatibility with our workflow. Compared to alternatives like R or MATLAB, Python provides greater flexibility and integration with machine learning frameworks and visualization libraries.

3. Preliminary Timeline

- **Week 1 (Nov 10–Nov 12):** Clean and preprocess dataset – handle missing values, normalize traffic counts, and prepare time-based features.
- **Week 2 (Nov 13–Nov 15):** Perform exploratory data analysis (EDA); visualize congestion trends across regions and time periods.
- **Week 3 (Nov 16 – Nov 20):** Train baseline models (Linear Regression, Decision Tree) and evaluate using RMSE and MAE.
- **Week 4 (Nov 21 – Nov 25):** Tune hyperparameters; integrate spatial features and traffic patterns into advanced models (Random Forest, XGBoost).
- **Week 5 (Nov 26 – Dec 1):** Build dashboards for visualization; summarize congestion insights and causes.
- **Week 6 (Dec 1 – Dec 5):** Prepare final report and presentation; push all updates to GitHub and finalize submission.

4. Team Member Contributions

Each team member has contributed to distinct but interconnected aspects of the project:

- **Jainil Desai** – Leading data preprocessing, cleaning scripts, and early-stage modeling. Jainil is also coordinating repository organization and documentation.
- **Adit Vakil** – Focusing on exploratory data analysis, visualizations, and identifying key congestion factors using statistical techniques.
- **Kalpan Shah** – Developing predictive models, optimizing performance, and handling report writing and presentation design.

We collaborate via GitHub for version control and Overleaf for documentation. Roles may evolve as we move toward deployment and model evaluation; Jainil will focus more on optimization, while Priya will handle final reporting and visualization integration.

5. Progress and Next Steps

So far, we have completed dataset exploration, schema understanding, and initial cleaning. We have identified missing values and begun early feature engineering (e.g., extracting peak-hour indicators and weekday/weekend flags).

Our next steps include completing preprocessing, running baseline models, and analyzing feature importance to identify key predictors of congestion. Depending on model

results, we plan to experiment with ensemble methods and visualize congestion heatmaps by region.

We will update our GitHub repository with recent scripts, cleaned data files, and our evolving project plan before the next submission.

GitHub Repository: `https://github.com/Adit-Vakil/data_science_project/tree/main`
Overleaf Project Link: `https://www.overleaf.com/project/691284a1b67f8660d04c0ff5`