Work Plan

Project Title: Develop a predictive model to accurately forecast hourly traffic volumes at different road junctions based on historical traffic data
Prepared by: Virendra
Date: 08.12.2025

1. Project Overview

The goal of this Mentorship project is to build a robust, production-ready predictive model to forecast hourly traffic volumes for multiple road junctions using historical traffic data. These forecasts will help in traffic management, planning, and decision-making by providing short-term (next 24–168 hours) predictions for each junction.

2. Objectives

- Assess dataset quality, coverage, missingness, anomalies

- Perform EDA to understand temporal patterns, seasonality, junction behavior

- Engineer time-series features (lags, rolling stats, calendar features)

- Train and compare ML models (LightGBM/XGBoost, plus optional sequence models)

- Evaluate using time-series cross-validation (MAE, RMSE, MAPE)

- Produce final hourly forecasts in required CSV format

- Prepare documentation and final report

3. Scope & Deliverables

Deliverables include:

- **Work_Plan_Virendra.pdf** (this document)

- **eda_notebook.ipynb**

- **feature_engineering.py**

- **train_model.py**

- **model_lgbm.pkl** (best model)

- **forecast_next24.csv**

- **evaluation_report.pdf**

- **README.md**

Upload Requirements:

- Document must be **<10 MB**
- Use filename format: **title_name**
- If PDF is mentioned, upload **only PDF**
- You may upload any additional research material

4. Methodology & Steps

Phase A – Data Understanding & Cleaning

- Inspect columns, parse timestamps
- Verify hourly frequency, resample if needed
- Handle missing values and outliers

Phase B – EDA

- Visualize hourly, daily, weekly patterns
- Create heatmaps, seasonal plots
- Compare junction-level behavior
- (Optional) Analyze external data like weather/holidays

Phase C – Feature Engineering

- Calendar features: hour, day_of_week, month, is_weekend
- Lag features: lag_1, lag_24, lag_168
- Rolling windows (3h, 6h, 24h)
- Junction encoding (label/target)
- Fourier terms for seasonality (optional)

Phase D – Modeling & Validation

- Baselines: naïve, last-week
- Main model: **LightGBM**
- Alternatives: XGBoost, RandomForest, LSTM/Transformer
- Time-series cross-validation
- Evaluate with MAE, RMSE, MAPE

Phase E – Forecasting & Packaging

- Implement multi-step forecasting
- Generate final CSV with timestamp, junction_id, forecast_count
- Create validation graphs

5. Timeline (7-day plan)

Date	Tasks
08.12.2025	Finalize Work Plan; inspect dataset
09.12.2025	Data cleaning + EDA; create eda_notebook. ipynb
10.12.2025	Feature engineering + baseline models
11.12.2025	Train ML models; hyperparameter tuning
12.12.2025	Generate final forecasts; evaluate holdout window
13.12.2025	Prepare evaluation_report.pdf, README, packaging
14.12.2025	Final review and submission

6. Tools & Environment

- Python 3.8+
- Libraries: pandas, numpy, scikit-learn, lightgbm, xgboost, matplotlib, seaborn, joblib
- Jupyter Notebook / Google Colab
- GitHub or Drive for version control and sharing

7. Risks & Mitigation

- **Sparse data:** Use global models + imputation
- **Overfitting:** Track per-junction errors, try alternative models
- **Large file size:** Compress PDFs, include only required outputs

8. Submission Checklist

- File name: **Work_Plan_Virendra.pdf**

- Size <10 MB

- If sharing via Drive → Access: **Anyone with link → Viewer**

9. Next Steps

- Begin EDA and feature engineering

- Set up modeling pipeline

- Prepare notebooks and scripts in a reproducible structure