AI Project Assignment Proposal

Title: Predictive Analysis of Public Transportation Delays Using a Dirty Real-World Dataset

1. Project Overview

Public transportation systems generate large amounts of operational data, often recorded under real-world constraints that lead to errors, missing values, noise, and inconsistencies. In this project, students will work with a **deliberately "dirty" dataset** representing bus schedule information. The dataset contains:

- Missing timestamps
- Inconsistent time formats
- Incorrect GPS points
- Categorical noise (e.g., "sunny," "SUN," "clody")
- Outliers in passenger counts
- Mixed route ID formats

Students are expected to clean, preprocess, analyze, and model this dataset using AI and machine learning techniques. This assignment reinforces key concepts from the Introduction to AI course, such as data processing, feature engineering, predictive modeling, and model evaluation.

2. Objective

The primary objective of this project is to **develop an AI-based predictive system for estimating bus delay behavior** using imperfect real-world data.

Students will:

1. Perform comprehensive data cleaning and preprocessing
2. Conduct exploratory data analysis (EDA)
3. Engineer meaningful features from the cleaned data
4. Build and evaluate predictive ML models
5. Interpret results using explainability tools
6. Present findings in an academic-style report

3. Dataset Description

Students will use the provided dataset:
dirty_transport_dataset.csv
(Contains ~300 bus trip records)

Dataset Attributes

Column Name	Description
route_id	Noisy categorical bus route identifiers
scheduled_time	Scheduled arrival time (clean format)
actual_time	Actual arrival time (multiple inconsistent formats; missing values)
weather	Noisy weather condition (misspellings, case differences)
passenger_count	Passenger count with outliers and missing values
latitude	GPS latitude, includes impossible coordinates (e.g., 999)
longitude	GPS longitude, includes missing values

“Dirty” Characteristics Students Must Fix

- Missing actual arrival times
- Time format inconsistencies (e.g., "12:45", "12.45PM", "1245")
- Noisy weather labels (“clody”, “CLOUDY”, “sunny”)
- Incorrect passenger values (0, >200, negative numbers)
- GPS errors (impossible coordinates, null values)
- Mixed route identifiers such as “R1”, “3”, “Route-4”

This dataset is intentionally unclean to ensure students practice real data preparation skills, not just model training.

4. Methodology Requirements

Students must follow a structured and fully documented pipeline:

4.1 Data Cleaning & Preprocessing

Required tasks:

1. **Missing Data Treatment**

- Choose appropriate imputation strategies for time, weather, and passenger count
- Justify your choices

2. **Standardization & Formatting**
   - Convert all timestamps into **consistent ISO format (YYYY-MM-DD HH:MM:SS)**
   - Normalize weather labels (lowercase, corrected spelling)
   - Create a unified route ID format
3. **Outlier Detection**
   - Use IQR or Z-score to detect extreme passenger counts and delay values
   - Decide which outliers to keep, fix, or remove **with justification**
4. **GPS Validation**
   - Identify impossible lat/long values
   - Decide whether to remove or replace those records
5. **Feature Engineering**
   Students should at minimum create:
   - Delay duration in minutes (scheduled vs. actual)
   - Time-of-day category (morning/afternoon/evening)
   - Day type (weekday vs weekend)
   - Weather severity index (light/moderate/heavy)
   - Route frequency features

4.2 AI/ML Modeling

Students must test **at least two ML models**, such as:

- Linear Regression
- Random Forest Regression
- XGBoost / Gradient Boosting
- kNN Regression

4.3 Explainability Techniques

Students must include:

- SHAP values *or* feature importance
- Discussion of key features impacting prediction

5. Expected Challenges & Limitations

Students must reflect on:

- Bias introduced during imputation
- Correlations between time-related features
- Noise affecting model stability
- Weather inconsistencies
- Possible overfitting in high-complexity models
- GPS errors limiting spatial analysis

This reflection should be included in the report.

6. Timeline & Deliverables

3 weeks

Required Deliverables

Students must submit:

1. **Cleaned dataset**
2. **Python Jupyter Notebook** (well-structured and commented)
3. **Final written report (8–12 pages)**
4. **Slide presentation**
5. **Model evaluation summary**
6. **Interpretability results**

7. Evaluation Criteria

Model Performance Metrics

Students must evaluate models using:

- **MAE** (Mean Absolute Error)
- **RMSE** (Root Mean Squared Error)

- **MSE** (Mean Squared Error)
- **R^2 Score**
- Cross-validation scores

Grading Breakdown

Component	Weight
Data Cleaning & Preprocessing	25%
EDA & Feature Engineering	20%
Model Development	20%
Model Evaluation	15%
Explainability & Interpretation	10%
Final Report & Presentation	10%

8. Final Notes for Students

This project emphasizes **hands-on real-world AI work** rather than theoretical exercises.
Students are expected to justify every step, make data-driven decisions, and demonstrate critical thinking.

Creativity in modeling and analysis is encouraged.