PROGRESS REPORT: AIR QUALITY DATASET ANALYSIS

1. Dataset Description:

The UCI Air Quality dataset presents a full description of air pollution levels in Rome, Italy, with 9358 hourly records collected from March 2004 to February 2005 (missing value denoted -200). Information was gathered using various multisensory air quality devices that have five metal oxide sensors, operating at road level in a highly polluted urban area. The sensors continuously sample the main pollutants, which include Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Total Nitrogen Oxides (NOx), Benzene (C_6H_6), and Non-Methane Hydrocarbons (NMHC).

Furthermore, the dataset also records other aspects about weather which include parameters such as temperature, relative humidity, and absolute humidity. Each record has 15 variables such as timestamps, concentration (reference) of the pollutants (as measured by certified analyzers), and the response of sensors to the pollutants. Accordingly, CO(GT), NMHC(GT), C6H6(GT), NOx(GT), and NO2(GT) are the true hourly concentration of the pollutants while PT08.S1 to PT08.S5 refers to the sensor readings for specific pollutants.

2. Exploratory Data Analysis (EDA)

1. Achieved EDA:

So far, the dataset has been cleaned and explored using standard techniques. Deleted null columns and rows and formatted variables to follow standards. Missing values have been identified by calculating the total number of missing entries per column and summarizing their distribution across rows using frequency table. A correlation matrix and heatmap visualization have been generated to examine relationships between variables. Additionally, a statistical summary of the dataset covering key metrics such as mean, median, standard deviation, and percentiles.

2. Planned EDA & Predictive Modeling:

Pairwise Correlation (Between sensor responses (PT), environmental variables with true concentrations (GT)), Cross-Sensitivity (how sensors respond to non-target gasses), Hypothesis Testing, AQI Calculation.

Regression Modeling for CO Prediction: Predicting CO(GT) using Temperature, Humidity, and Time and comparing Linear Regression and Random Forest to test linear vs. nonlinear relationships and evaluating model performance using R^2, RMSE, and MAE.

Hybrid Modeling: Combined SARIMAX model for time series forecasting by exogenous covariances (CO, NOx, T, Rh) with outputs from SAPRC Biochemistry simulation model to predict NO2. Approximate SAPRC variable using Linear regression, Non-linear regression.

Machine Learning: Compare the performance of different machine learning techniques for predicting pollutant concentrations based on sensor data and environmental conditions.

3. Tentative Analysis Questions

a. Sensor Calibration + Air Quality Analysis

How do sensor responses and environmental variables correlate with true pollutant concentrations, and do sensors exhibit cross-sensitivities? Is there a significant bias between sensor reading and reference measurements, and do sensor residuals depend on environmental factors? How does the AQI vary over time, what are the dominant pollutants contributing to poor air quality, and how does AQI vary by season, with specific times of year when air quality is consistently worse?

b. Regression-Based CO Prediction

How well do Temperature, Humidity, and Time of Day predict CO(GT) levels? Is there a linear or nonlinear relationship between environmental factors and CO(GT)? Does adding sensor data (PT08.S1(CO)) improve prediction accuracy? How does the performance of Linear Regression compare to Random Forest for CO prediction? How does this model compare to time-series forecasting (SARIMAX) and machine learning models (MLP)?

c. Time Series Model + Knowledge Domain Model

Is the NO_2 series stationary? Which predictors (temperature, humidity, NOx, CO, C6H6) and SAPRC-derived chemical reaction outputs correlate strongly with NO_2? How do NO_2 levels change across seasons? Is there a seasonal autocorrelation pattern (e.g., weekly peaks)? How to approximate SAPRC variable using regression? How well does a hybrid model (SARIMAX + SAPRC simulation outputs) predicting NO_2 levels compare to a SARIMAX only approach?

d. Machine Learning with and without time-series:

Is there an ML technique that outperforms other methods on this dataset? Are time-series approaches better than non-time-series approaches?

4. Planned Methods

a. Sensor Calibration + Air Quality Analysis

- Pearson/Spearman correlations and correlation matrix and visualize using a heat map.
- Paired t-tests and residual analysis depend on environmental factors.
- AQI calculation, categorize AQI into health risk level time-series trends and seasonal analysis.

b. Regression-Based CO Prediction

- Convert time into numerical format (Hour of the day) and normalize data (T and RH).
- Train Linear Regression and Random Forest to compare linear vs. nonlinear models.
- Tune Random Forest hyperparameters to optimize performance.
- Evaluate models using R^2, RMSE, and MAE to measure prediction accuracy.
- Perform residual analysis to check for model errors and potential biases.

c. Time Series Model + Knowledge Domain Model

- Data Preprocessing & Stationarity Transformation for ARIMA model.
- Basic SARIMAX Model Development: SARIMA(p, d, q)(P, D, Q, s) with Exogenous Predictors
- Approximate variable of SAPRC Biochemistry by Linear Regression, Polynomial and Interaction terms in Non-linear Regression.
- Use Hybrid SARIMAX model fitting for trend analysis and forecasting future pollution levels.

d. Machine Learning with and without time-series:

- Predict target variables (pollutant concentrations): CO(GT), C6H6(GT), NOx(GT), NO2(GT).
- Involve normalizing data and PCA to reduce dimension (reduce parameters) before training.
- Machine Learning techniques: Ridge regression / Lasso Regression, Support Vector Regression (SVR), Regression Trees, XGBoost, Nearest Neighbour Regression.
- Use lag features for time-series approach.