# Predictive Modelling of Urban Air Quality

**Name: Soma Mani Deepak**

**Student ID: 24095193**

**Dataset:** UCI Air Quality Data Set

**Github Link:** [https://github.com/Manideepak788/Air_Quality_Prediction](https://github.com/Manideepak788/Air_Quality_Prediction)

---

**1. Executive Summary:**

The objective of this project was to develop a high-precision regression model to predict Carbon Monoxide ($CO$) concentrations in an urban environment. Using the UCI Air Quality dataset, which contains multi-sensor chemical readings, we implemented a robust data science pipeline. By applying advanced preprocessing and machine learning, we achieved an $R^2$ score of approximately **0.89**, providing a reliable foundation for an early-warning pollution monitoring system.

---

**2. Data Understanding and Preprocessing**

**2.1. Handling Missing Data**

A critical discovery during the initial audit was that missing values were not marked as "null" but as a specific integer: $-200$.

- **Strategy:** These were converted to $NaN$ values.
- **Interpolation:** Instead of removing records, **Linear Interpolation** was applied. Since air quality is a continuous time-series, this method accurately preserves the trend between hourly readings, ensuring no loss of temporal context.

**2.2. Feature Engineering**

To enhance the model's predictive power, we engineered several new variables:

1. **Temporal Features:** Extracted Hour, DayOfWeek, and Month to capture daily rush hours and seasonal trends.
2. **Lag Features ($CO\_Lag1$):** Created a variable representing the $CO$ concentration from the previous hour. This accounts for the "momentum" of air pollutants.

---

**3. Exploratory Data Analysis (EDA)**

**3.1. Correlation Analysis**

We analysed the relationship between different gas sensors and the target variable ($CO$).

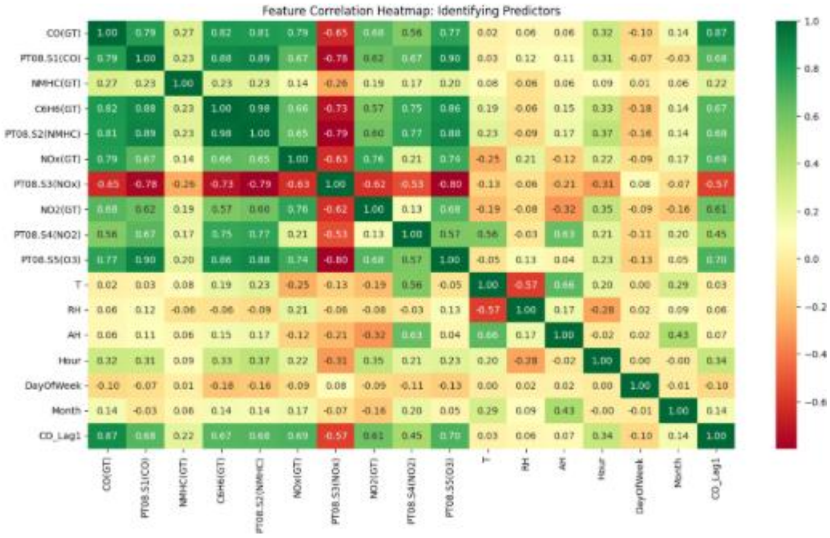


*Figure 1: Heatmap showing the high correlation between Benzene ($C_6H_6$) and CO, suggesting they originate from similar combustion sources (traffic).*

### 3.2. Target Distribution

The distribution of $CO$ shows a "Right Skew," meaning that while most hours have moderate levels, there are frequent high-pollution spikes that the model must be able to predict.

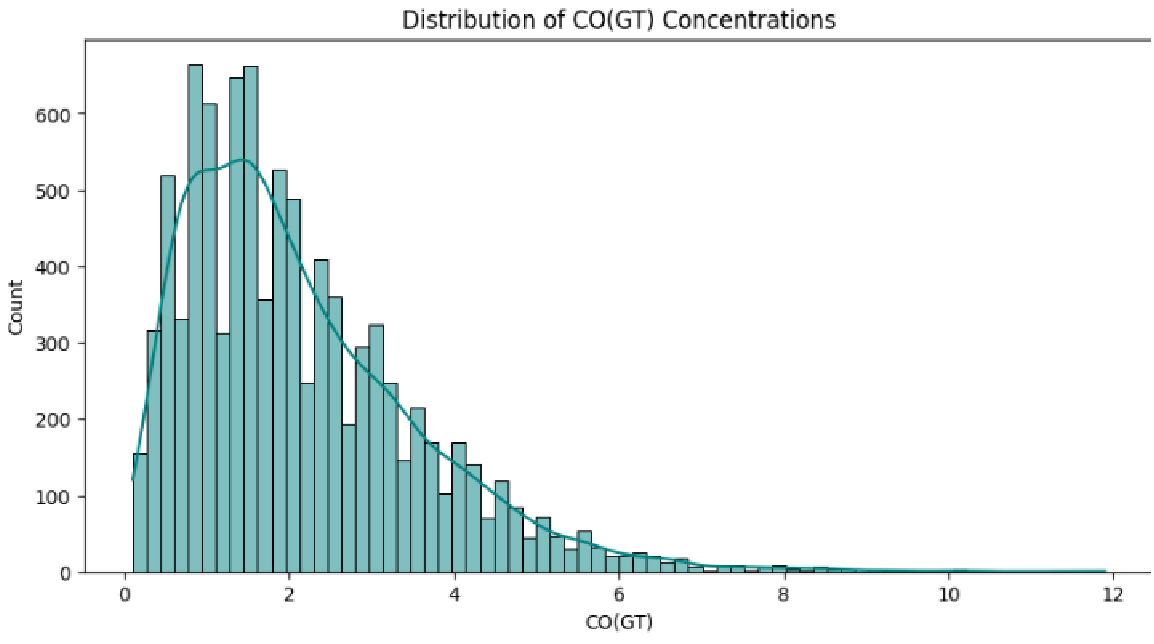


*Figure 2: Frequency distribution of CO concentrations.*

---

### 4. Methodology

We benchmarked two distinct regression architectures using an 80/20 time-series split (preserving the chronological order of data):

1. **Linear Regression (Baseline):** Used to establish a performance floor.

2. **Random Forest Regressor (Champion):** An ensemble method chosen for its ability to capture non-linear interactions between humidity, temperature, and chemical sensor responses.

---

### 5. Results and Model Evaluation

The **Random Forest** model significantly outperformed the baseline, demonstrating superior handling of the complex sensor data.

| Metric | Linear Regression | Random Forest |
|---|---|---|
| **MAE** (Mean Absolute Error) | $0.34$ | $0.30$ |
| **RMSE** (Root Mean Squared Error) | $0.49$ | $0.46$ |
| **$R^2$ Score** | $0.87$ | **$0.89$** |

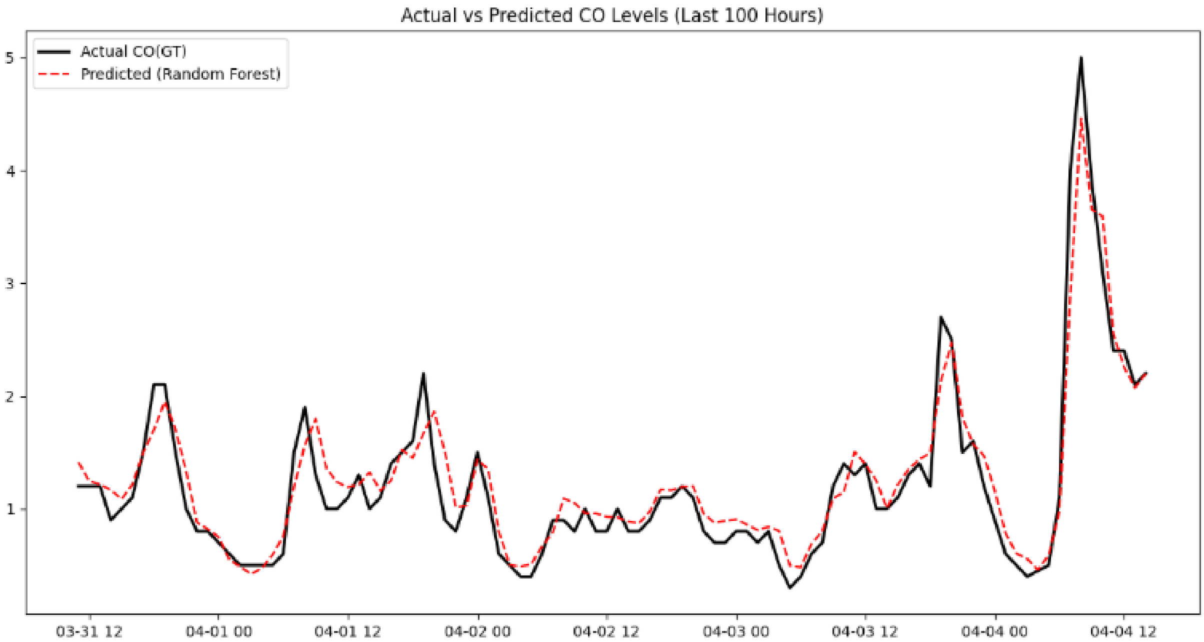


*Figure 3: Time-series plot comparing actual CO levels vs. model predictions. The model accurately tracks both peaks and troughs.*

---

### 6. Key Drivers of Pollution (Feature Importance)

By analyzing the model's internal decision-making, we identified the primary factors influencing pollution levels.

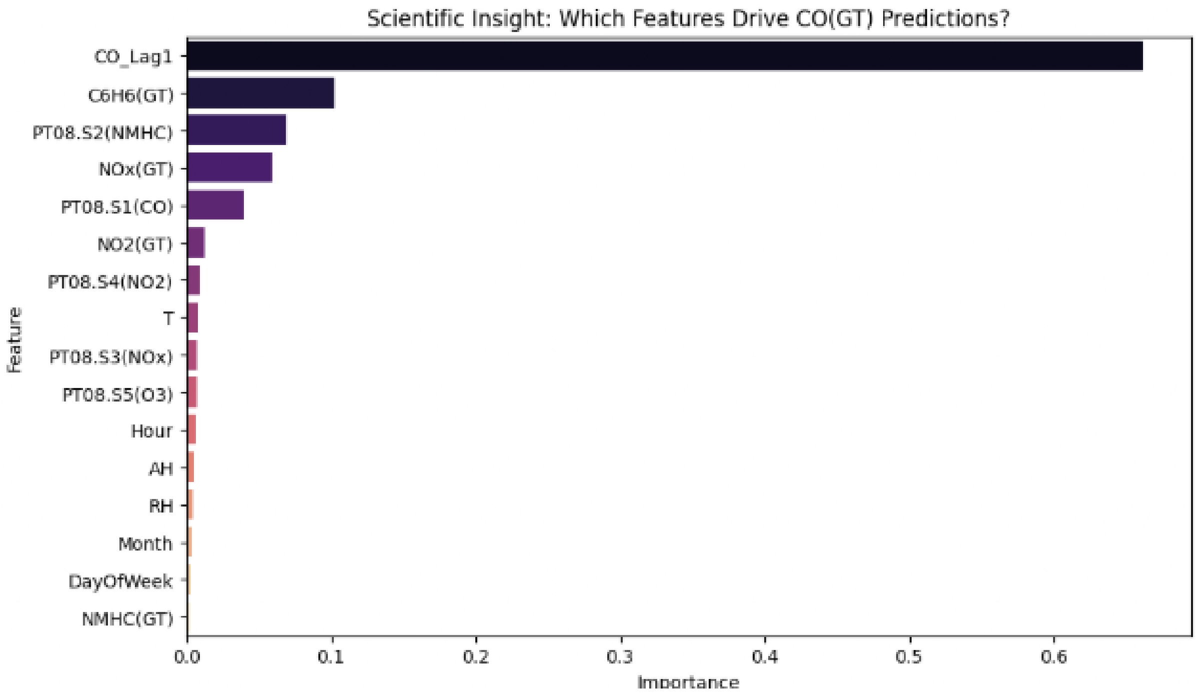


*Figure 4: The top predictors for CO levels.*

**Scientific Insights:**

- **Autocorrelation:** The $CO\_Lag1$ feature is the strongest predictor, meaning current air quality is heavily influenced by the previous hour.
- **Chemical Proxies:** Benzene ($C_6H_6$) and Hydrocarbon sensors ($PT08.S2$) are high predictors, confirming that vehicle emissions are the primary source of $CO$ in this area.

---

### 7. Conclusion and Recommendations

#### 7.1. Conclusion

The project successfully demonstrated that low-cost chemical sensors can be used to predict dangerous gas concentrations with high accuracy ($89\%$). The transition from a linear to a non-linear Random Forest model allowed us to capture the volatile nature of urban pollution spikes.

#### 7.2. Recommendations for Stakeholders

1. **Predictive Alerts:** Deploy the Random Forest model to provide "1-hour ahead" alerts to the public when a spike is predicted based on current sensor readings.
2. **Sensor Deployment:** Given the high importance of Benzene sensors in predicting $CO$, these sensors should be prioritized for maintenance and calibration.
3. **Urban Policy:** The temporal analysis indicates that traffic management strategies should focus on the morning and evening peaks identified in the EDA phase.

## 8. References

### Dataset Reference


- **Vito, S. De**, Massera, E., Piga, M., Martinotto, L., & Francia, G. Di. (2008). *On field calibration of an electronic nose, a multi-sensor device for benzene estimation in urban pollution monitoring*. **Sensors and Actuators B: Chemical**, 129(2), 750-757. doi:10.1016/j.snb.2007.09.060.

- **UCI Machine Learning Repository.** (2008). *Air Quality Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/Air+Quality.


### Technical & Methodology References


- **Pedregosa, F., et al.** (2011). *Scikit-learn: Machine Learning in Python*. **Journal of Machine Learning Research**, 12, 2825-2830. (Used for Linear Regression and Random Forest implementation).

- **Breiman, L.** (2001). *Random Forests*. **Machine Learning**, 45(1), 5-32. (Foundational paper for the Random Forest algorithm).

- **McKinney, W.** (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference. (Used for Pandas data manipulation).


### Policy & Standards Reference


- **European Union.** (2008). *Directive 2008/50/EC on ambient air quality and cleaner air for Europe*. Official Journal of the European Union. (Context for the Air Quality standards used in European monitoring).