Project Proposal

Predicting Carbon Monoxide (CO) Levels Using Machine Learning Models for Air Quality Management

Fundamentals of Data Science - COMP 7/8150

Group Members:

Loknadh Venkata Krishna Sai Kona (U00935840)

Yaswanth Kumar Donthuboyina (U00922408)

Submission Date: November 20, 2025

Objectives

The objective of our proposed project is to develop a **predictive machine learning model for estimating carbon monoxide (CO)** concentrations using environmental and pollutant data. Our goal is to build a system that supports real time air quality monitoring, helping authorities take proactive measures to control pollution and protect public health.

Specifically, we plan to:

- Analyze pollutant and meteorological data to identify key factors affecting CO concentration.
- Preprocess, clean, and transform raw data for accurate modeling.
- Train and evaluate multiple regression models, including **Linear Regression**, **Random Forest**, and **Gradient Boosting**, to predict CO levels.
- Assess model performance using **R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE)**.
- Provide actionable, data-driven insights into environmental decision making and air quality management.

Introduction

Air pollution is a growing environmental and health concern, particularly in urban and industrial regions. Carbon monoxide (CO) is a toxic pollutant that can cause serious health problems even at moderate concentrations. Monitoring and predicting CO levels are therefore critical for reducing health risks and improving public safety.

With the rapid advancement of machine learning techniques, researchers and policymakers can model complex relationships between environmental factors and pollutant levels more effectively. Through this project, we intend to leverage machine learning models to predict CO concentrations using features such as NOx, NO_2, temperature, and humidity.

Our objective is to create a reliable, interpretable, and data-driven model capable of providing timely air quality predictions, thus enabling proactive policy interventions and community awareness.

Background

Recent research has shown that pollutants often have nonlinear relationships influenced by temporal and environmental conditions. Studies have demonstrated the effectiveness of ensemble learning methods such as Random Forest and Gradient Boosting in modeling air quality data.

For example, Gupta and Jain (2018) emphasized that Random Forest handles complex pollutant interactions efficiently, while Jiang et al. (2020) found that incorporating meteorological data enhances CO prediction accuracy.

Building on these insights, our proposed work aims to apply these models to the Air Quality UCI dataset, which contains hourly measurements of multiple pollutants and meteorological variables. We will explore temporal variations such as hourly and seasonal changes and pollutant interactions to gain deeper insights into CO behavior across different conditions.

Methodology

Dataset Information

We plan to use the AirQualityUCI dataset from the UCI Machine Learning Repository.
This dataset contains hourly averaged responses from an array of chemical sensors deployed in a polluted area of an Italian city from March 2004 to February 2005.

Dataset Summary:

- Number of Instances (rows): 9,358
- Number of Features (columns): 15
- Time Span: March 2004 – February 2005
- Location: Urban road level area, Italy
- Pollutants Measured: CO, NOx, NO_2, NMHC, Benzene, and O_3
- Environmental Variables: Temperature, Relative Humidity, Absolute Humidity
- Missing Value Indicator: -200

This dataset provides an excellent foundation for exploring the relationship between pollutants and environmental conditions, allowing us to train models that can predict future CO concentrations under varying circumstances.

Data Preprocessing

1. **Handling Missing Values:**
   We plan to replace values labeled as -200 with NaN and impute them using an **interpolation technique** based on temporal trends.

2. **Outlier Detection:**
   Outliers will be detected and handled using **Z-score** and **IQR filtering** to maintain model accuracy.

3. **Feature Scaling:**
    - We will apply **standardization** for algorithms sensitive to data scale (e.g., Linear Regression).
    - **Normalization** will be used for ensemble models (e.g., Random Forest, Gradient Boosting).
4. **Feature Engineering:**
   We plan to create additional variables, such as pollutant interaction features ($NOx \times NO_2$) and temporal features (hour, day, month), to capture seasonal and daily variations in CO concentrations.

Model Development

We will train and compare three regression algorithms:

1. **Linear Regression:** As a baseline model for interpretability.
2. **Random Forest:** To capture nonlinear dependencies and interactions between pollutants.
3. **Gradient Boosting:** To incrementally improve prediction accuracy by reducing residual errors.
4. **Cross-Validation:** To ensure robust model training and prevent overfitting, all models will be evaluated using 3-fold or 5-fold cross-validation. For each model, we will compute the mean and standard deviation of RMSE, MAE, and R^2 across folds. This multi-fold validation ensures statistical reliability and allows us to compare model performance more accurately.

To enhance model generalization and robustness, we plan to implement k-fold cross-validation during model training and explore hyperparameter tuning using Grid Search or Randomized Search methods for Random Forest and Gradient Boosting models. This will help identify the most optimal model configuration and prevent overfitting.

Model Evaluation

We will assess each model's performance using the following evaluation metrics:

- **R^2 (Coefficient of Determination):** Measures how well the independent variables explain the variance in CO levels.
- **MAE (Mean Absolute Error):** Represents the average magnitude of prediction error.
- **RMSE (Root Mean Squared Error):** Provides a sensitive measure of prediction accuracy, penalizing larger errors.

Once all models are trained, we will compare their performance based on these metrics to determine the most accurate and robust model. We expect ensemble models such as Random Forest and Gradient Boosting to outperform the baseline Linear Regression model due to their ability to capture nonlinear and complex relationships.

- **Statistical Validation:** We will apply statistical tests to validate our results. This includes using paired t-tests, Wilcoxon signed-rank tests, and ANOVA to compare model performance across folds. Additionally, 95% confidence intervals will be reported for RMSE, MAE, and R^2 to ensure the performance differences between models are statistically significant and reliable.

In-Distribution and Out-of-Distribution Testing

To rigorously test model generalization, we will evaluate our models using **two separate test sets**:

1. **Test Set 1: In-Distribution (ID) Samples:** This test set consists of the standard hold-out samples drawn directly from the original AirQualityUCI dataset split.
2. **Test Set 2: Out-of-Distribution (OOD) Samples:** An additional test set will be generated by adding controlled Gaussian noise, perturbing pollutants and meteorological variables, and creating synthetic random samples. This allows us to assess how well the models generalize to data that differs from the training distribution.

Out-of-Distribution (OOD) Detection Module

Before applying the CO prediction model, an OOD detection system will be used to automatically identify and filter out anomalous or unseen patterns.
We will experiment with methods such as:

- Z-score thresholding
- Mahalanobis distance-based detection
- Isolation Forest anomaly detection

Any OOD inputs will be removed *before* making predictions. This ensures higher reliability and robustness under unexpected environmental conditions.

The OOD detection module will be evaluated separately using metrics such as true detection rate, false alarm rate, and AUROC to confirm its reliability. For generating OOD samples, Gaussian noise proportional to each feature's standard deviation (about 1.0σ–1.5σ) will be added to pollutant and weather variables, allowing us to create realistic but clearly different samples for testing model robustness.

Deliverables and Expected Outcomes

Our project will deliver the following outcomes:

1. **Complete Written Report** detailing the methodology, data analysis, and model results.
2. **Python Code Implementation** for preprocessing, model training, and evaluation.
3. **Cleaned and Processed Dataset** derived from the AirQualityUCI repository for reproducibility.
4. **Final Presentation** summarizing findings, model performance, and recommendations for air quality management.
5. **Expected Outcomes:**
    - Development of a machine learning model that accurately predicts carbon monoxide (CO) levels.
    - Identification of key environmental and pollutant variables influencing CO concentrations.
    - Visual insights into policy decision making and environmental awareness.
    - Cross-validated model performance (mean ± standard deviation across folds).
    - Evaluation results for both In-Distribution and Out-of-Distribution **test sets**.
    - Fully implemented **OOD detection module** with performance metrics.
    - Statistical validation results include **t-tests, Wilcoxon tests, ANOVA**, and **95% confidence intervals**.

References

- Chen, H., Wu, X., & Li, R. (2021). *Improving CO Prediction Accuracy Using Temporal and Pollutant Interaction Features in Urban Environments.* Journal of Environmental Monitoring, 12(4), 298–312.
- Gupta, P., & Jain, V. (2018). *Machine Learning Techniques for Air Quality Prediction: A Review.* International Journal of Environmental Research, 14(3), 121–130.
- Jiang, S., Wang, Y., & Li, Q. (2020). *Gradient Boosting Models for Enhanced CO Prediction: Integrating Meteorological Data and Pollutant Concentrations.* IEEE Transactions on Environmental Science, 8(2), 215–223.
- Kaur, N., Patel, M., & Zhang, H. (2022). *Explainable AI for Air Quality Prediction Using SHAP Values.* Environmental Data Science Journal, 9(1), 44–53.
- Singh, A., & Kumar, S. (2019). *Challenges in Air Quality Prediction Using Machine Learning: Data Quality and Model Generalizability.* Environmental Science and Policy, 22, 138–145.