

# COMP9417 Project: Forecasting Air Pollution with Machine Learning

November 1, 2025

## Project Description

Air pollution forecasting is a major environmental and public health challenge. In this project, you will design, implement, and evaluate machine learning models that predict air quality indicators using time-series sensor data. The goal is to model how pollutant concentrations evolve and forecast future pollution levels using historical observations and environmental factors.

You will work with the *Air Quality* dataset available from the UCI Machine Learning Repository<sup>1</sup>. This dataset contains **9,358 hourly records** collected by an air quality monitoring device located at road level in a heavily polluted urban area of Italy. The device includes an array of **five metal oxide chemical sensors** measuring gaseous pollutants, along with meteorological variables such as temperature and humidity. The data spans from **March 2004 to February 2005** and represents one of the longest publicly available on-field sensor recordings for air quality monitoring.

Your task is to build machine learning models that learn temporal dependencies and environmental patterns from this dataset. You will explore both regression and classification approaches to forecasting pollutant levels, assess different feature representations, and evaluate model robustness under realistic deployment scenarios.

## Description of the Data

Table 1 summarises the main variables included in the dataset. The features consist of both **sensor responses** (PT08.S1--S5) and **reference pollutant concentrations** for carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene (C<sub>6</sub>H<sub>6</sub>), nitrogen oxides (NO<sub>x</sub>), and nitrogen dioxide (NO<sub>2</sub>). In addition, the dataset includes **meteorological attributes** such as temperature (T), relative humidity (RH), and absolute humidity (AH), which capture environmental factors influencing pollutant behaviour. Several variables contain missing values represented by the sentinel value -200, which should be treated as missing data and handled appropriately during preprocessing. Together, these variables form a multivariate time series suitable for time-based predictive modelling.

Table 1: Variables in the Air Quality dataset.

Variable	Description
CO(GT)	True hourly averaged concentration of carbon monoxide
PT08.S1(CO)	Hourly averaged response of the CO sensor
NMHC(GT)	True hourly averaged concentration of non-methane hydrocarbons
C6H6(GT)	True hourly averaged concentration of benzene
PT08.S2(NMHC)	Hourly averaged response of the NMHC sensor
NOx(GT)	True hourly averaged concentration of nitrogen oxides
PT08.S3(NOx)	Hourly averaged response of the NOx sensor
NO2(GT)	True hourly averaged concentration of nitrogen dioxide
PT08.S4(NO2)	Hourly averaged response of the NO2 sensor
PT08.S5(O3)	Hourly averaged response of the O3 sensor
T	Ambient temperature (°C)
RH	Relative humidity (%)
AH	Absolute humidity

<sup>1</sup><https://archive.ics.uci.edu/dataset/360/air+quality>

# Main Objectives

The project focuses on developing a data-driven framework for forecasting pollutant concentrations and analysing temporal dynamics in environmental data. Your work should address the following components:

- **Exploratory Data Analysis (EDA):** Examine time patterns, seasonal effects, correlations among pollutants, and relationships between meteorological and chemical variables. Identify missing values and data quality issues.
- **Data Preprocessing:** Handle missing values (e.g., -200), merge **Date** and **Time** fields into a unified timestamp, convert decimal separators, and normalise continuous features. Create derived features such as hour, weekday, and month.
- **Anomaly and Event Detection:** Identify and analyse unexpected pollution spikes or sensor faults by applying residual-based anomaly detection and/or unsupervised methods. Compare detected anomalies with meteorological or calendar features (e.g., temperature extremes, weekends) to interpret likely causes. Evaluate precision–recall trade-offs and assess how anomaly detection can enhance model robustness and data reliability.
- **Feature Engineering:** Design temporal features such as lagged variables and moving averages to capture short-term and long-term dependencies. Investigate how temporal granularity (hourly vs. daily averages) affects prediction accuracy.
- **Temporal Data Splitting:** Since the data are time-oriented, evaluate models using temporal splits. Specifically, implement a **chronological split** (train on 2004 data, test on 2005 data) to simulate model deployment and post-deployment evaluation. If a validation set is needed, use the latest portion of the training data.
- **Model Assessment:**
  1. Use RMSE for regression tasks, accuracy for classification. Visualise residuals and the trends between predicted and observed values over time.
  2. Predict pollutant concentrations for the following horizons: 1 hour, 6 hours, 12 hours, and 24 hours ahead.
  3. Compare results against a **naïve baseline** that uses the concentration at time  $t$  as the prediction for  $t + 1$ ,  $t + 6$ ,  $t + 12$ , and  $t + 24$ .
- **Regression Model Development:** Choose two or more regression algorithms (e.g., Linear Regression, Regression Trees, Random Forest, Gradient Boosting, Neural Networks, or Support Vector Regression) to predict concentrations of the five pollutants: CO, NMHC, C6H6, NO<sub>x</sub>, and NO<sub>2</sub>.
- **Classification Model Development:**
  1. For classification tasks, use CO (GT) as the target variable. Discretise it into the following categories: low ( $< 1.5 \text{ mg/m}^3$ ), mid ( $1.5 \leq \text{CO} < 2.5$ ), and high ( $> 2.5$ ).
  2. Choose two or more classification algorithms (e.g., Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Neural Networks, or Support Vector Machines).
  3. Compare results against a **naïve baseline** that uses the discretised concentration at time  $t$  as the prediction for  $t + 1$ ,  $t + 6$ ,  $t + 12$ , and  $t + 24$ .
- **Discussion:** Analyse which features and modelling strategies perform best. Discuss limitations such as sensor drift, data imbalance, and temporal concept drift that may affect real-world deployment.

## Submission Guidelines

- The deadline to submit the report and code is 23 November 2025 (Sunday) at 6 pm.
- Work in groups of 4–5 students and register your group on Moodle.
- Submit your report (.pdf) and code (.zip) via Moodle — one submission per group.
- Late submissions incur a 5% penalty per day. Submissions more than 5 days late will receive a mark of zero.

## Report Structure

Your final report (maximum 6 pages, 12 pt font, 1.5 spacing) should follow this structure:

1. **Introduction:** Problem motivation, dataset description, and project goals.
2. **Data Analysis:** Summary statistics, data cleaning, and exploration of patterns.
3. **Methodology:** Preprocessing, feature engineering, and model selection rationale.
4. **Results:** Quantitative evaluation and visualisation of model performance.
5. **Discussion:** Interpretation of results, limitations, and improvement strategies.
6. **Conclusion:** Summary of findings and implications.
7. **References:** Include all sources used for models, algorithms, or background theory.

## Peer Review

All group members must complete the peer review survey by 24 November 2025 (Monday) by 6 pm. Failure to do so will result in a 10% penalty to that student's mark. Peer evaluations will be used to adjust individual grades based on contribution.

## Project Support

You may use any open-source libraries such as `scikit-learn`, `pandas`, `numpy`, or `xgboost`. Consult official documentation and lecture materials for guidance. General questions can be posted in the COMP9417 project discussion forum on Moodle.