Machine Learning Final Project Fall 2025

Machine Learning Approach for PM2.5 Prediction and Pattern Discovery in Beijing (2010–2015)

Brian Ma
Student
University of Richmond
Richmond, VA
brian.ma@richmond.edu

I. Project Aim

This project aims to build and evaluate supervised regression models that predict hourly $PM_{2.5}$ levels in Beijing using meteorological and temporal variables. The target series is restricted to **PM_US Post** (the U.S. Embassy monitor), which provides a continuous timeline from 2010 through 2015. This choice preserves informative pre-2013 episodes when official hourly reporting was not yet in place, while remaining compatible with the post-2013 period. The goal is to deliver an interpretable, well-validated predictor that helps anticipate high-PM episodes under similar weather conditions, supporting earlier advisories and practical responses.

II. Overall Plan

A. Introduction

Air pollution remains a major public-health concern in large Chinese cities, with fine particulate matter ($PM_{2.5}$) linked to elevated cardiopulmonary risks. Beijing has experienced recurring haze events driven by seasonal heating, stagnant wind conditions, and regional transport. While China's Ministry of Environmental Protection (MEP) began publishing *hourly* air-quality data for an initial set of 74 cities in early 2013, the **PM_US Post** series from the U.S. Embassy in Beijing offers consistent hourly $PM_{2.5}$ measurements back to 2010. As documented in the dataset description and related AGU/JGR materials, the combined record (embassy PM with airport meteorology such as temperature, pressure, wind, and dew point) enables analyses that span both the pre-2013 and post-2013 regimes.

In this study, only supervised regression is considered. Meteorological and time-of-day/season features are used to estimate $PM_{2.5}$ from the PM_US Post perspective, leveraging the longer time span for model training and validation. Although the data precede the present by about a decade, the atmospheric mechanisms that lead to severe pollution—temperature inversions, weak dispersion, and seasonal heating—remain relevant. Learning from these historical patterns can improve today's ability to anticipate similar conditions and act earlier rather than reactively.

B. Related Work

Prior studies on Beijing and other Chinese regions have applied supervised learning to forecast $PM_{2.5}$ using weather and calendar features, frequently reporting that ensemble tree methods (e.g., Random Forest or Gradient Boosting) and kernel methods (e.g., SVM with RBF kernels) outperform plain linear baselines on non-linear relationships. Dataset notes and atmospheric analyses in the AGU/JGR literature further motivate separating the monitoring provenance (pre-2013 embassy series versus post-2013 official network) when constructing targets and features. Building on these insights, this project benchmarks a focused set of regression models—linear/regularized baselines, SVM, tree-based models, and a shallow neural baseline—under a consistent validation protocol centered on MAE, RMSE, and R^2.

C. Dataset Used

This project uses the *Beijing PM2.5* dataset, restricting the target and primary signal to the **PM_US Post** series (U.S. Embassy monitor in Beijing). The UCI page documents that the dataset contains hourly $PM_{2.5}$ readings from the U.S. Embassy in Beijing along with meteorology from the Capital International Airport, covering 2010–2014 (commonly extended in community mirrors to 2015) [1].

The choice to rely on **PM_US Post** is deliberate: Chinese Ministry of Environmental Protection (MEP) launched nationwide hourly $PM_{2.5}$ reporting only in **January 2013** for an initial set of **74 cities**, later expanding coverage [2], [3]. As a result, Beijing's other local stations in this dataset begin in 2013, whereas the U.S. Embassy monitor provides a continuous time series since 2010. Using PM_US Post preserves informative pre-2013 episodes while remaining compatible with post-2013 official reporting.

This setup leverages a single, long, and internally consistent $PM_{2.5}$ series to learn meteorology–pollution relationships before and after 2013, while avoiding coverage gaps in other local stations.

D. Approach

1) **Data Preparation:** Use the **PM_US Post** series (2010–2015) as the regression target. Align timestamps; handle

TABLE I
KEY VARIABLES (SUBSET USED IN THIS STUDY)

cc

Variable	Description
PM_US Post	Hourly $PM_{2.5}$ from U.S. Embassy monitor (target)
TEMP, DEWP, PRES	Temperature, dew point, pressure (airport)
Iws	Cumulated wind speed (m/s)
cbwd	Combined wind direction (categorical)
year, month, day, hour	Time features for seasonality/diurnal structure

missing values via time-aware interpolation for $PM_{2.5}$ and key meteorology; encode wind direction (cbwd); and standardize numeric features when required by downstream models.

2) **Feature Engineering (with dataset's season):** Build a leakage-safe feature matrix from contemporaneous or lagged information only (no future data). Specifically:
   - *Calendar & categorical:* use the dataset's season column directly (coding: 1=Spring, 2=Summer, 3=Autumn, 4=Winter). Represent it via one-hot encoding (preferred) or as an ordered categorical; include `year`, `month`, `day`, `hour`, and **part-of-day bins** (*morning* 06:00–11:59, *afternoon* 12:00–17:59, *evening/night* 18:00–05:59).
   - *Historical $PM_{2.5}$ aggregates (from PM_US Post only):*
     - **Weekly level:** rolling mean/median/max over the past 7 days; week-of-year indicator.
     - **Monthly level:** rolling mean/max over the past 30 days; month-of-year indicator.
     - **Seasonal level:** current season (1–4) plus rolling summaries over the past 90 days (season-scale window).
     - **Yearly level:** year indicator and year-to-date rolling mean/max up to time t.
     - **Intra-day level:** recent-hour rolling means (e.g., past 3/6/12 hours).
   - *Meteorology:* TEMP, DEWP, PRES, Iws, and encoded cbwd; optionally simple interactions (e.g., Iws×PRES).

   All aggregates are computed with windows ending at the current timestamp to prevent look-ahead leakage.

3) **Supervised Regression & Model Selection:** Train models from the course list: *Linear Regression*, *SGD (gradient descent)*, *Polynomial Regression* (low degree), *SVM* (linear/RBF), *Decision Tree*, *Random Forest* and *Gradient Boosting*, with an optional shallow *Keras MLP* (early stopping). Use k-fold cross-validation; evaluate with MAE, RMSE, and R^2; and select a primary model balancing accuracy and interpretability. Provide concise model cards and an error profile by `season` and part-of-day.

E. Data Analysis Steps

The analysis starts by auditing the PM_US Post time series and the accompanying meteorology: timestamps are aligned; missing values are handled in a time-aware manner; wind direction is encoded; and numeric features are standardized as needed. A compact, leakage-safe feature matrix is then assembled, blending meteorology with seasonality and diurnal structure.

Next, a supervised regression benchmark is established across models from the course topics: Linear/SGD baselines, low-degree Polynomial Regression, SVM (linear/RBF), a Decision Tree, Random Forest and Gradient Boosting, with an optional shallow Keras MLP. Model selection relies on k-fold cross-validation, with MAE, RMSE, and R^2 recorded for comparability. Hyperparameters and random seeds are logged for reproducibility.

Finally, the chosen model is interpreted and stress-tested: coefficient patterns or feature importance are summarized; residuals are profiled by season and wind conditions; and sensitivity checks (e.g., removing a key feature, shifting lags) are used to confirm stability. The deliverables are a concise set of figures (seasonal/diurnal plots, error-by-condition charts) and a brief narrative translating results into practical takeaways for anticipating high-PM episodes when similar conditions recur.

F. Methodology

- **Train/Validation Protocol:** Chronology-aware split with k-fold cross-validation on blocks to reduce temporal leakage; StandardScaler where appropriate.
- **Model Grid (Regression Only):** Linear, SGD (learning rate schedule and early stopping), Polynomial (degree 2–3 with regularization), SVM (linear/RBF with C, γ grid), Decision Tree (depth/min-samples), Random Forest / Gradient Boosting (n-estimators, depth, learning rate), optional shallow Keras MLP (1–2 hidden layers, early stopping).
- **Evaluation & Reporting:** Primary metrics MAE and RMSE; report R^2; error breakdowns by season and wind regime; model card for the selected model.

G. Expected Results

With the enriched temporal feature set (weekly/monthly/seasonal/yearly aggregates from PM_US Post, part-of-day bins, and the dataset's season column), ensemble and kernel regressors are expected to outperform purely linear baselines. Anticipated patterns include:

- **Top predictors:** recent-hour rolling means (e.g., last 3/6/12h), the past-7-day mean/max, and monthly/seasonal summaries should rank among the strongest signals, reflecting short-term accumulation and persistence. Wind speed (`Iws`) is expected to show a negative contribution (dispersion), while higher pressure and lower temperature/dew point often coincide with elevated $PM_{2.5}$.
- **Season and part-of-day effects:** the season indicator (1=Spring, 2=Summer, 3=Autumn, 4=Winter) should capture higher winter levels and relatively cleaner summer periods. Part-of-day bins typically reveal higher

evening/night concentrations versus mid-day, consistent with weaker boundary-layer mixing after sunset.
- **Model ranking:** Random Forest or Gradient Boosting (and SVM with RBF) are expected to achieve the lowest MAE/RMSE, with regularized linear models (Ridge/Lasso) providing interpretable baselines. The addition of temporal aggregates should reduce MAE and RMSE relative to using meteorology + raw calendar features alone.
- **Stability and generalization:** performance should remain stable across the pre-2013 and post-2013 periods; residual analyses by season and part-of-day are expected to show the largest errors during winter evenings/nights under low-wind, high-pressure conditions.

Overall, the final model is expected to deliver reliable short-horizon estimates and clear rules-of-thumb: when recent-hour and weekly aggregates are elevated, winds are weak, and the period is winter evenings/nights, the risk of high $PM_{2.5}$ rises markedly, indicating that earlier advisories are warranted.

TABLE II
PROJECT SCHEDULE

Week	**Task**	**Deadline**
Week 1	Data cleaning and exploratory analysis	Oct 25
Week 2	Feature extraction and visualization	Nov 1
Week 3	Regression modeling and cross-validation	Nov 10
Week 4	Clustering and PCA interpretation	Nov 20
Week 5	Integration, evaluation, and report writing	Dec 5

VII. Optional – Above and Beyond

As an extra, I plan to apply the same analysis on the other four cities' datasets and evaluate a cross-city analysis.

III. References

References

[1] UCI Machine Learning Repository, "Beijing PM2.5," 2017. [Online]. Available: https://archive.ics.uci.edu/dataset/381/beijing%2Bpm2%2B5%2Bdata. [Accessed: Oct. 2025].

[2] Y. Wang, G. Ying, J. Zhang, *et al.*, "Spatial and temporal variations of six criteria air pollutants in China," *Environment International*, vol. 73, pp. 413–422, 2014. (Notes MEP started publishing hourly AQ data for 74 cities in Jan 2013.)

[3] Ministry of Environmental Protection (MEP), "MEP Releases Air Quality of Key Regions and 74 Cities in September and Q3 of 2013," News Release, Oct. 24, 2013. [Online]. Available: https://english.mee.gov.cn/News_service/news_release/201310/t20131024_262189.shtml.

[4] X. Liang, T. Zou, B. Guo, *et al.*, "Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating," *Proc. Royal Society A*, 2015. (Background on $PM_{2.5}$ episodes in Beijing.)