Beijing Air Pollution Study of Machine Learning on Meteorological Factors

Hanyang Shang

Received September 02, 2024
Accepted October 15, 2024
Electronic access October 31, 2024

Air pollution, particularly fine particulate matter (PM2.5), poses significant environmental and public health challenges worldwide. This study focuses on predicting PM2.5 levels in Beijing, China, by analyzing the relationship between various meteorological factors and pollution levels. Using historical data from 2010 to 2014, we apply machine learning models, including linear regression, Multi-Layer Perceptron (MLP), Decision Tree, and Random Forest regressors, to forecast pollution levels. The models are evaluated using Mean Square Error (MSE) to determine the most effective approach. Additionally, we investigate the individual impact of factors like dew point, temperature, and wind speed on pollution levels, compared to their collective influence. The motive in analyzing individual factors is to validate empirical perceptions data on major meteorological contributors to Beijing's air pollution, and possibly make predictions in the future. The findings reveal that while single factors offer some predictive power, humidity emerges as a dominant contributor when all factors are considered together. After assessing the effectiveness of each model, the study proposes a combined model that integrates the strengths of individual models to mitigate erratic performance from any single model called the multiplicative weight update method. This study underscores the importance of using machine learning to analyze complex environmental data, offering valuable insights that could guide future pollution control measures and public health interventions. Although this research is focused on a single dataset from Beijing, it highlights the potential of AI in advancing our understanding of air pollution dynamics and their broader implications for human health and urban planning.

Keywords: Air Pollution Prediction, Machine Learning Models, Meteorological Factors

Introduction

Air pollution is a significant environmental issue that affects the health and well-being of millions of people worldwide. It is primarily caused by the release of harmful substances into the atmosphere, including particulate matter, of which those with diameters less than 2.5 microns, as measured in PM2.5 index, are of particular interest[1]. These pollutants permeate most filtering devices such as masks and cause serious health issues. In certain regions, such as Beijing, China, which has been historically exposed to its northern deserts and windy winters believed to be responsible for some calamitous days when PM2.5 exceeds average by multiple folds[2]. Enormous efforts have been put into forestation and wind control for decades. This study aims to dissect the multiple meteorological factors in their correlation to PM2.5 air pollution. Several studies have been conducted in other regions, such as Vietnam in south-east Asia, which has drastically different climates, and possibly different PM2.5 profiles as related to meteorological factors[3]. A study focused on Beijing aims to corroborate and provide analytical justification of Beijing's air pollution control policy and practices.

The application of artificial intelligence (AI) and machine learning to process large amounts of meteorological data makes it possible to analyze and predict air pollution patterns on fine-granular data, such as the data sampled on hourly basis through an extended period of time[4]. While machine learning provides a host of tools at our disposal, it's important to understand their applicability in processing a particular dataset. Fitting machine learning models is a heuristic process. This study will pay equal attention to direct results from the models as the heuristic insights the process provides, which may be just as important in informing our study.

Using machine learning to analyze and predict air pollution patterns offers several key benefits. First, it provides a more accurate and timely prediction of pollution levels and allows for quick response. This can be particularly beneficial for urban areas, such as Beijing, where pollution levels can fluctuate rapidly[5]. Second, predictive models can help identify pollution hotspots and the key contributing factors to high pollution levels. This information is invaluable for policymakers and public health officials in devising effective strategies to improve air quality. Third, AI-driven insights can empower communities with the knowledge to take preventive measures, such as planning to avoid outdoor activities during high pollution periods or advocating for cleaner technologies and practices[6].

Model Building and Data Processing

Dataset Description

The meteorological and pollution input dataset is obtained from Kaggle[7]. It consists of hourly observations from Jan 1, 2010 - Dec 31, 2014, on several meteorological factors. For machine learning, the training dataset consists of time series between Jan 1, 2010 - Dec 31, 2013. The testing dataset for prediction test consists of time series between Jan 1, 2014 - Dec 31, 2014.

For machine learning model training, we further split the training dataset into two parts: a) 80% of the training dataset is used for model fitting. b) 20% of the training dataset is reserved for evaluating model fitting and adjusting training parameters to minimize the likelihood of overfitting. In order to preserve comparable data distribution within the data's timeframe (i.e. for seasonality consideration), the training dataset and evaluation dataset are split randomly from the time series using the train_test_split function from the Python sklearn package. The dataset partition is illustrated in the Python code below
Load full dataset: df = pd.read_csv(<datafile>)
Training data by date: df_train = df[(df['date'] < '2014-01-01')] Testing data by date: df_test = df[(df['date']>= '2014-01-01')] Randomly split training data into model fitting (80%) and evaluation (20%) parts: df_tr, df_ev = train_test_split(df_train, test_size=0.2, random_state=42)

A sample portion of the input dataset is shown in Table 1.

Our study examines the relationship between wind speed (wnd_spd), dew point (dew), temperature (temp), pressure (press), and PM2.5 pollution concentration (pollution). Other factors in the dataset, such as wind direction (win_dir), snow (snow) and rain (rain) falls, are not considered in this study.

Machine Learning Models

We use several machine learning models to fit them with the training dataset and heuristically find the best model. This process is repeated in studying individual and combined meteorological factors, as each dataset partition may yield different characteristics and fitting for optimal results. For the experiments to be repeatable, we use a control variable (e.g. ramdom_state = 42 for the Python library) to remove the randomness in model fitting when the models are rerun.

Linear Regression Model: The linear regression model aims to establish a linear relationship between input features (such as dew point, temperature, pressure, and wind speed) and air pollution levels. It accomplishes this by fitting a linear equation that predicts pollution levels based on weighted combinations of the input features. This model is interpretable, making it useful for understanding direct correlations between the input factors and pollution. However, its assumption of linearity limits its ability to capture complex, non-linear relationships present in real-world data.

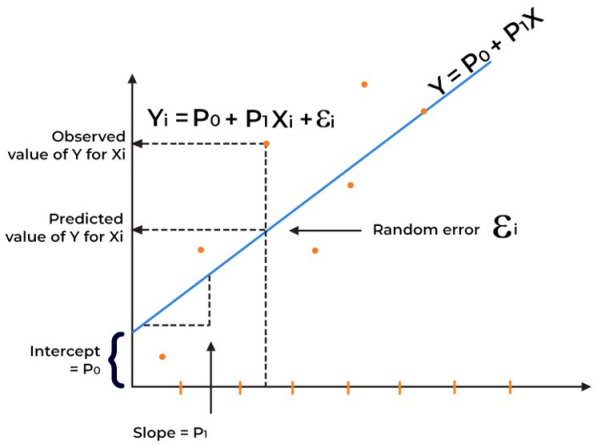

Fig. 1 Demonstration of Linear Regression Model

Multi-Layer Perceptron (MLP) Regressor: The Multi-Layer Perceptron (MLP) regressor is a type of artificial neural network capable of learning complex patterns and non-linear relationships in data. Composed of input, hidden, and output layers of neurons, each connected by weighted edges and activated by non-linear functions, the MLP learns to predict air pollution levels based on input features. Despite its power, MLPs require careful tuning, such as the max_iter parameter to control its iteration levels. We will experiment with a wide range of tentative max_iter values to find where the results converge and more iterations yield no perceivable benefit.

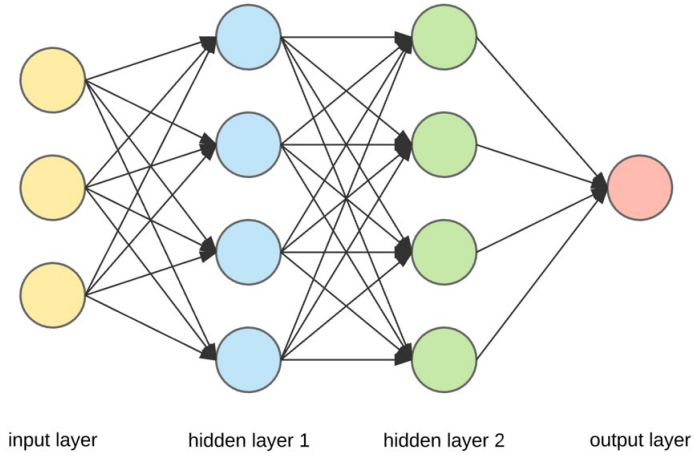

Fig. 2 Demonstration of MLP Regressor

Decision Tree Regressor: The Decision Tree regressor predicts air pollution levels by recursively partitioning the data into subsets based on feature thresholds, optimizing splits to minimize variance or mean squared error. Decision trees offer interpretability, illustrating how features influence predictions,

date	pollution	dew	temp	press	wnd_dir	wnd_spd	snow	rain
2014-01-01 0:00:00	24	-20	7	1014	NW	143.48	0	0
2014-01-01 1:00:00	53	-20	7	1013	NW	147.5	0	0
2014-01-01 2:00:00	65	-20	6	1013	NW	151.52	0	0
2014-01-01 3:00:00	70	-20	6	1013	NW	153.31	0	0

Table 1 Sample Hourly Observation from Input Dataset

but may overfit if the tree becomes too deep or complex. We will experiment with a range of max_depth values to find the optimal tree depth and avoid overfitting. This model excels in capturing interactions between variables and are robust to outliers and irrelevant features, which is useful to examine the cross-influence of some meteorological factors and provides intuitive insights.

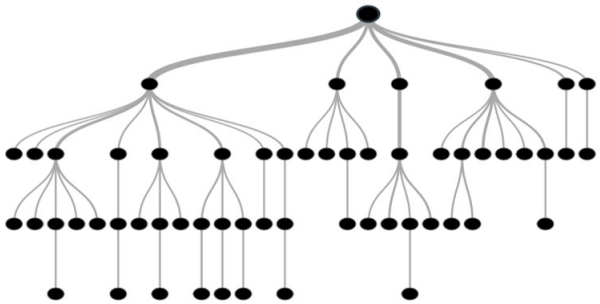

Fig. 3 Demonstration of Decision Tree Regressor

Random Forest Regressor: The Random Forest Regressor is an ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and robustness. In a random forest, each tree is trained on a different random subset of the data, both in terms of samples and features. This randomness helps to reduce the correlation between the individual trees and, consequently, the overall variance of the model. Once all the trees in the forest have made their predictions, the final prediction is typically obtained by averaging the predictions of the individual trees. Similar to Decision Tree Regressor, we will experiment and find optimal depth of the forests to avoid overfitting of the model.

In this study, we define model overfitting as the training MSE being over 10% better than test MSE.

Model Parameters, Training and Prediction

In this study, we perform the data processing on several meteorological factors and their combinations (i.e. input parameter selection) to examine how they individually and collectively correlate to the pollution levels. We use four distinct parameter selections as below, which correspond to four distinct hypothetical correlations between meteorological factors and air pollution levels to examine:

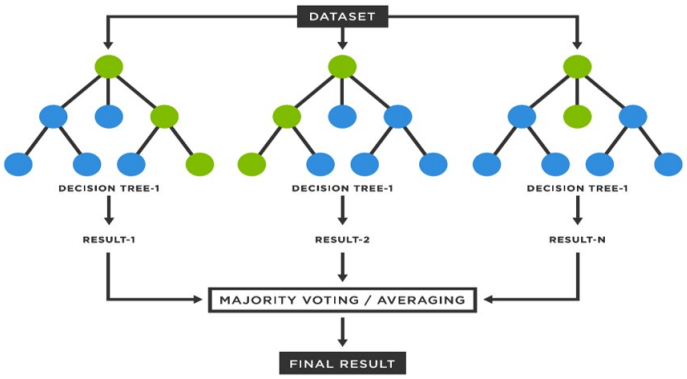

Fig. 4 Demonstration of Random Forest Regressor

1. Wind Speed (Wind) - Wind speed is generally thought to have a positive influence on air pollution, on the hypothesis that wind helps to dissipate pollutants and reduce air pollution level. However, in Beijing, the strong wind in winter also brings in dust from the northern deserts. We want to examine how wind speed influences air pollution in Beijing.

2. Temperature (Temp) - Temperature is thought to play an important role in air turbulence and its effect on air pollutant distribution.

3. Temperature and Dew Point (Humi) - We use this combination as a proxy for humidity. When air temperature meets the dew point, the humidity reaches its maximum level. High humidity has the effect of coalacing smaller pollution into larger particles, which are more subjective to effective filtration and reduction in PM2.5 pollution level.

4. Temperature, Dew Point, Wind Speed and Pressure (All) - This is a combination of all four meteorological factors we examine in this study. It gives a more comprehensive view on the overall correlation between meteorological factors and pollution level. This parameter selection can help mitigate bias of other simpler parameter selections where nonlinear relationships are more dominant.

For each input parameter selection, we apply the four regression models as described in the previous section by using

Python sklearn for model fitting. For each model, we measure MSEs as training (MSEtr), training evaluation (MSEev) and test data subset (MSEts), and check the model for overfitting as the following:

Model is non-overfitting in test *iff*:

$$\text{MSE}_{\text{tr-ts Diff}} = abs(\frac{\text{MSE}_{\text{tr}} - \text{MSE}_{\text{ts}}}{\text{MSE}_{\text{tr}}}) < 0.10$$

Overfitting most likely happens in Decision Tree and Random Forest models, when the max_level model control parameter exceeds certain values. We use experiments to test and find the optimal max_level values for Decision Tree and Random Forest models for each input parameter selection during model training, under the condition that the model is not overfitted by checking the following:

Model is non-overfitting in training *iff*:

$$\text{MSE}_{\text{tr-ev Diff}} = abs(\frac{\text{MSE}_{\text{tr}} - \text{MSE}_{\text{ev}}}{\text{MSE}_{\text{tr}}}) < 0.10$$

From Table 2, we have observed that even Decision Tree and Random Forest models are not overfitted during training, these models demonstrate overfitting once run on three of the four test datasets. For these models, we may need to use different overfitting criteria to accommodate their model characteristics.

For the Multi-Layer Perceptron (ML) model, the key training control parameter is max_iter which denotes the number of iterations the model runs through the training data. It has been observed that after certain iterations, the model will converge to a certain MSE and not improve its accuracy further. We use a heuristic max_iter = 1,000 for all input parameter selections, which also satisfies the non-overfitting check as described above.

Multiplicative Weight Update Method

To create a more accurate and robust predictive model, this study further refine the process of model training and evaluation. Our strategy is to create a combined model prediction results by aggregating prediction results of all four models, while adjusting the model's individual contribution weights dynamically based on their performance after each training iteration. Specifically, after each round of training, we can reduce the weight of each model by a factor of $2^{(error^2)}$, where "error" represents the difference between the predicted and actual values. This approach penalizes models that perform poorly, reducing their influence in subsequent predictions. The formula for a model's contribution weight is:

$$\text{Model Weight}_{\text{adjusted}} = \frac{\text{Model Weight}_{\text{original}}}{2^{\text{error}}}$$

where $\text{error} = abs(\text{Model Prediction}[i] - \text{Actual Pollution Level}[ii])$,

for each data point [i] in the timeseries

For the four models, we normalize so their sum of contribution weights equals to 1. This is done by dividing the weight of each model by the total weight of all models combined, effectively scaling the weights so that their sum equals 1.

For each input parameter selection, the models exhibit different error distributions and thus different contributions to the combined model, as shown in the diagram below. We have observed that the Random Forest model has in general dominated the weight contribution for three of the four input parameter selections, and emerged as the more robust and accurate model among the four. However, for some datasets or segments thereof, other models have shown their strength in accuracy as well. This observation leads us to use combined results for comparison against actual observed data in the next section.

Results

We present our experiment results on the testing data for each model input parameter selection in the form of combined results from weighted contributions of individual models, and compared with the actual observed pollution data for the time period of Jan 01, 2014 to Dec 31, 2014 at hourly intervals and indexed in calendar dates for seasonality analysis, as shown in the diagram below.

From the comparisons, we have noticed that the predictions demonstrate more moderate magnitude changes versus the observed data. This can be explained by the machine learning model's tendency to smooth the impact of outliers in model fitting, and to aim for predicting long-term trends.

For individual model input parameter selection, we have the following observations and analysis:

1. Wind Speed (Wind) - Wind speed shows positive effects on reducing pollution levels during the winter and early spring months, which feature prevailing strong wind in Beijing. The effect of strong winter wind bringing in large amounts of pollution particles from the northern desert into Beijing, as commonly perceived, is not demonstrated in the prediction. This may be explained by the continuous forestation and its effectiveness in air pollution control by the government.

2. Temperature (Temp) - Lower temperature in the winter and spring seasons results in more stagnant air flows, thus higher pollution levels. In contrast, higher temperature in summer and fall seasons facilitate air flows to dissipate pollutants and reduce pollution levels.

3. Temperature and Dew Point (Humi) - As a proxy to measure humidity, the prediction shows higher volatility and magnitude of pollution levels in winter and spring, which

Input Data	Model	Model Control	MSEtr	MSTev	MSEtr-ev Diff	MSEts	MSEtr-ts Diff	Overfitting
Wind	Linear	N/A	7911	8189	3.51%	8350	5.55%	No
Wind	MLP	max_iter = 1000	7584	7943	4.73%	8070	6.41%	No
Wind	DT	max_level = 41	7034	7601	8.06%	7866	11.83%	Yes
Wind	RF	max_level = 46	7055	7759	9.98%	7848	11.24%	Yes
Temp	Linear	N/A	8358	8629	3.24%	8597	2.86%	No
Temp	MLP	max_iter = 1000	8292	8579	3.46%	8529	2.86%	No
Temp	DT	max_level = 10	8226	8563	4.10%	8496	3.28%	No
Temp	RF	max_level = 11	8226	8563	4.10%	8497	3.29%	No
Humi	Linear	N/A	6859	6629	3.35%	7359	7.29%	No
Humi	MLP	max_iter = 1000	6152	6427	4.47%	6640	7.93%	No
Humi	DT	max_level = 8	5918	6327	6.91%	6680	12.88%	Yes
Humi	RF	max_level = 7	5924	6268	5.81%	6601	11.43%	Yes
All	Linear	N/A	6631	6955	4.89%	7187	8.38%	No
All	MLP	max_iter = 1000	5998	6268	4.50%	6581	9.72%	No
All	DT	max_level = 7	5581	5911	5.91%	6544	17.25%	Yes
All	RF	max_level = 7	5355	5810	8.50%	6398	19.48%	Yes

Table 2 Model Training Parameters and Overfitting Test

Fig. 5 Model Contribution to Combined Predictions

are drier seasons in Beijing. In more humid summer and fall, the predicted pollution levels show less changes, which may be explained by higher humidity's effect in coalacing small pollutant particles and reducing PM2.5 levels, as described in our hypothesis.

4. Temperature, Dew Point, Wind Speed and Pressure (All) - It's worth noting the prediction using all meteorological factors as model input tracks the results from humidity extremely close in change patterns, with pollution levels showing slight difference. This may indicate the dominant factor in pollution level change could be humidity in Beijing. The prediction tracks the pattern of observed data more appreciably in different seasons of the year than on individual time series data points. This may reflect the machine learning model's tendency in favoring long-term predictions.

From the analysis above, we can see air pollution is subject to cross-influences of different meteorological factors. Single factors such as wind speed or temperature alone don't show trends as close to the observed data, but correlate well with their empirical effect on air pollution levels. Humidity plays a dominant role when all factors are considered, and the models are more consistent in predicting seasonal or long term patterns than individual days. While this study is limited to one dataset collected

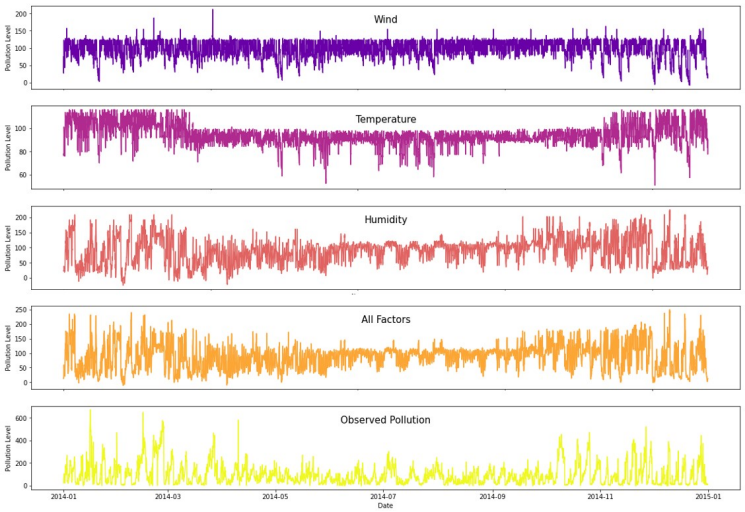

Fig. 6 Comparison of Model Predictions and Observed Pollution

from a single location in Beijing, it demonstrates the great potential of using big data and machine learning in understanding the complex factors in environmental study.

Discussion

Research Question

The primary object of this study is to examine the effectiveness of different machine learning models in predicting air pollution levels and examines the individual impact of specific meteorological factors on pollution. After learning the effectiveness of different models, the study proposes a combined model to mitigate erratic performance from any single model. Additionally, the research investigates how specific meteorological factors like dew point, temperature, and wind speed influence pollution levels, compared to considering all factors collectively. We hypothesized temperature and dew point would affect pollution more than wind would.

Analysis of Results

In this study, we evaluated multiple machine learning models to predict air pollution levels, with data partitioning to select training data and test the models for predictions. We use several techniques to mitigate training biases and improve prediction accuracy, to highlight and respond to the complex relationship between dataset and modeling methodology. We finally present model prediction results in comparison with actual observed data, analyze and explain the results in empirical contexts. This approach shows that the process of machine learning itself is also subject to improvement, and they can give best guidance to the users when combined and correlated with other empirical methods.

In machine learning, overfitting is one of the common issues which favor the training data, but underperforms in prediction with new data. This study uses the MSE threshold of 10% to gauge overfitting, which may also be subject to improvement considering each model's individual characteristics. It's conceivable that different models can require different training parameters for best fitting, including how to measure overfitting. In this study, a fixed MSE threshold of 10% is used to reduce the number of variables. Experimenting with various methods is crucial in machine learning for getting closer to the truth.

Future Direction

Artificial intelligence and machine learning provides new tools to environmental study, particularly in processing large amounts of data. Environmental study generates a tremendous amount of data from a vast array of sources. With the continuous development of machine learning, we look forward to its wide applications to benefit environmental study and public health.

In future studies, we could develop models that predict individual exposure levels based on personal factors (e.g., proximity to pollution sources, time spent outdoors, or pre-existing health conditions) could offer personalized health risk assessments. This approach could guide individuals on when to limit outdoor activities or take protective measures based on predicted pollution levels.

We aim to address several limitations highlighted in our initial research. One critical area for improvement is the choice of machine learning models. While our study utilized linear regression, MLP, Decision Trees, and Random Forest, we recognize the potential benefits of incorporating more advanced techniques such as gradient boosting and LSTM networks. These sophisti-

cated models could offer improved accuracy and robustness in our predictions. Additionally, the complexity and optimization challenges of MLPs were noted, particularly our limited exploration of hyperparameter tuning, stopping primarily at adjusting max_iter. Future research will delve deeper into comprehensive hyperparameter optimization, employing techniques such as grid search or Bayesian optimization to fully harness the potential of MLPs and other advanced models. We also intend to expand our evaluation metrics to provide a more comprehensive assessment of model performance. While our initial research focused on Mean Square Error (MSE), we acknowledge the importance of incorporating additional metrics such as R-squared and Mean Absolute Error (MAE). These metrics will offer a broader perspective on the effectiveness of the models, capturing various aspects of predictive accuracy and model fit. Utilizing multiple evaluation metrics will allow for a more nuanced comparison of models, ensuring that we fully understand their strengths and weaknesses. This holistic approach will enhance the robustness of our conclusions and improve the overall quality of our research.

We plan to strengthen the integration of scientific explanations with our machine learning model predictions by providing more robust data to support our claims. While our initial research discussed how each model relates to scientific phenomena, we recognize the need for empirical evidence to substantiate these connections. We will include detailed visualizations and statistical analyses to explicitly show these correlations and model outputs. By presenting comprehensive data alongside our scientific explanations, we aim to enhance the credibility and depth of our findings, ensuring that our claims are well-supported and transparent.

Relevance

By comparing different machine learning models, the study aims to identify the most accurate methods for predicting air pollution levels. Accurate predictions are crucial for timely interventions, such as issuing health advisories or implementing pollution control measures, which can reduce harmful exposure. The study's focus on individual meteorological factors, like dew point, temperature, and wind speed, also provides a deeper understanding of the specific contributors to pollution levels. This knowledge can guide targeted mitigation strategies, such as adjusting traffic flow during high-pollution periods or optimizing urban planning to reduce pollution hotspots[8].

Methods

This study uses historical meteorological data of one observation spot in downtown Beijing. The dataset consists of meteorological factors believed to be key contributors of PM2.5 pollution, such as dew point, temperature, and wind speed, etc. The time series includes data collected at hourly intervals from Jan 1, 2010 to Dec 31, 2014. The input data we provided were key meteorological factors including dew points, wind speed, and temperature indexes to predict levels of PM2.5 pollution.

In this study, we develop and evaluate several machine learning models to find the most appropriate one for predicting air pollution levels. The data is split into a training and a testing set, where the training set is used to fit the model so it's tuned to correlate meteorological factors with pollution levels . We use 80% of the timeseries (i.e. from Jan 1, 2010 to Dec 31, 2013) for training and 20% (i.e. from Jan 1, 2014 to Dec 31, 2014) for testing. Comparing the predicted output against actual observed data gives us insight into the models' effectiveness.

Considering that the meteorological factors often don't work in isolation, but contribute to PM2.5 pollution with cross-influence, such as when dew points meet the air temperature, they give the highest humidity level which is believed to be one of the most important factors in impacting air pollution levels. To account for such cross-influence, we take two approaches in this study.

Firstly, we will examine several machine learning models which respond to cross-reference with unique characteristics: a) linear regression model, which aims to learn the relationship between the input features and the pollution levels by fitting a linear equation to the training data; b) Multi-Layer Perceptron (MLP) regressor, which is a type of neural network capable of capturing non-linear relationships; c) Decision Tree regressor, which splits the data into subsets based on feature values to make predictions; and d) Random Forest regressor, which further splits the dataset into multiple decision trees to improve accuracy and prevent overfitting. For each model, the MSE (Mean Square Error) is used to measure model accuracy, which is compared among models to determine the most effective one for predicting air pollution levels. A combined model, consisting of individual models weighted on their discrepancy between predicted and observed data, is also used to alleviate erratic performance of a single model.

Secondly, certain meteorological factors, such as dew point, temperature and wind speed, are singled out for studying their effects on pollution levels, in comparison to studying all the factors as a whole. This approach may provide us more insight into each factor's contribution to pollution levels as compared with their collective contributions.

Acknowledgement

I would like to thank my Fellowship program mentor, Odysseas Drosis, for the passion he has inspired in me for data science and scientific research and all his support throughout this project.

References

1 C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu and T. Chi, *A novel spatiotemporal convolutional long short-term neural network for air pollution prediction*, https://doi.org/10.1016/j.scitotenv.2018.11.086.

2 X. An, T. Zhu, Z. Wang, C. Li and Y. Wang, *A modeling analysis of a heavy air pollution episode occurred in Beijing*, https://doi.org/10.5194/acp-7-3103-2007.

3 *Air pollution prediction by using an artificial neural network model*, https://doi.org/10.1007/s10098-019-01709-w.

4 Q. Wu, Z. Wang, A. Gbaguidi, C. Gao, L. Li and W. Wang, *A numerical study of contributions to air pollution in Beijing during CAREBeijing-2006*, https://doi.org/10.5194/acp-11-5997-2011.

5 M. Panagi, Z. Fleming, P. Monks, M. Ashfold, O. Wild, M. Hollaway, Q. Zhang, F. Squires and J. Hey, *Investigating the regional contributions to air pollution in Beijing: a dispersion modeling study using CO as a tracer*, https://doi.org/10.5194/acp-20-2825-2020.

6 A. Patra, S. Gautam and S. Majumdar, *Prediction of particulate matter concentration profile in an opencast copper mine in India using an artificial neural network model*, https://doi.org/10.1007/s11869-015-0369-9.

7 *Kaggle dataset used in this study*, https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate/data.

8 O. Pokrovsky, R. Kwok and C. Ng, *Fuzzy logic approach for description of meteorological impacts on urban air pollution species: a Hong Kong case study*, https://doi.org/10.1016/S0098-3004(01)00020-6.