A Machine Learning Approach for Beijing Air Quality Prediction

BANERJEE Rohini (20543577), RAMALINGAM Poojaa (20214528)
The Hong Kong University of Science and Technology (MSc Big Data Technology)

MSBD 5002: Data Mining and Knowledge Discovery (Term Project)

ABSTRACT

The recent rise in air pollution is viewed as an issue of serious concern due to its dreadful impact on the environment and in turn, on human health. It is thus of paramount importance to create, build and implement forecasting techniques to accurately predict the air quality. Machine learning is often seen as an efficient way of handling high-dimensionality data to find underlying patterns in the data or to predict future responses from the data. This project utilizes Random Forest Regressor to accurately predict the hourly concentrations of PM2.5, PM10 and O_3 for a major city in China over a span of two days.

1. INTRODUCTION

Rapid increase in urbanization and industrialization has led to a sharp rise in air pollution levels. Economic development and population upsurge have also contributed to the same. The amalgamation of all the above has led air pollution levels to rise beyond the accepted safety limit. World Health organization (WHO) estimates that 9 out of 10 people in the world live in environments with higher level of air pollutants than the accepted level [1] (figure 1). This mediocre quality of air adversely affects the health of people, making it the leading cause of death among young children [2]. It also accounts for millions of deaths among adults every year [3]. The aforementioned statistics highlight the importance of addressing the issue of air pollution.

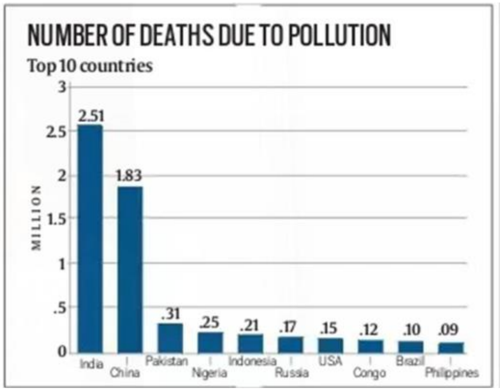


Figure 1: Number of deaths caused due to air pollution

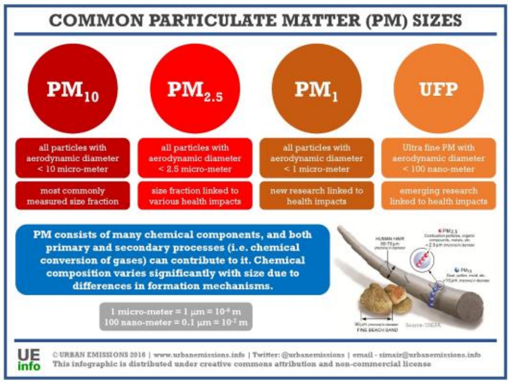

COMMON PARTICULATE MATTER (PM) SIZES
PM10
PM2.5
PM1
UFP
all particles with aerodynamic diameter < 10 micro-meter
all particles with aerodynamic diameter < 2.5 micro-meter
all particles with aerodynamic diameter < 1 micro-meter
Ultra fine PM with aerodynamic diameter < 100 nano-meter
most commonly measured size fraction
size fraction linked to various health impacts
new research linked to health impacts
emerging research linked to health impacts
PM consists of many chemical components, and both primary and secondary processes (i.e. chemical conversion of gases) can contribute to it. Chemical composition varies significantly with size due to differences in formation mechanisms.
1 micro-meter = 1 µm = 10⁻⁶ m
100 nano-meter = 0.1 µm = 10⁻⁷ m
HUMAN HAIR
PM2.5
PM10
FINE BEACH SAND
Source: USEPA
UE info
© URBAN EMISSIONS 2016 | www.urbanemissions.info | Twitter: @urbanemissions | email - simair@urbanemissions.info
This infographic is distributed under creative commons attribution and non-commercial license

Figure 2: Air Particulate Matter Description

The most common pollutants of air are ozone (O_3), carbon monoxide (CO), sulphur dioxide (SO_2), lead, nitrogen oxides (NO_x) and particulate matter (PM) (figure 2). PM is a mixture of both organic and inorganic particles, such as dust, soot, smoke, pollen and liquid droplets. PM2.5 refers to particulate matter that is 2.5 micrometres or less in diameter while PM10 comprises of those which are 10 micrometres or less in diameter. The suspension of these complex amalgamations in the air is immensely detrimental to human health and have been estimated to cause 3 to 7 million deaths every year [4].

For the purpose of predicting the concentration of air pollutants, we have taken past few years' air quality data from Beijing, China. Beijing is one of the most important economic zones in China and has recently seen a significant increase in air pollution levels. Via this project, we predict the hourly concentrations of PM2.5, PM10 and O_3 for 35 air quality stations in Beijing for a period of 2 days in the month of May 2018.

The rest of report is structured as follows. Section 2 presents the related work. Our approach is mentioned in section 3 which includes the exploratory analysis of the data, data cleaning and pre-processing, feature engineering and the details of the machine learning model implemented. Section 4 articulates the model results while section 5 concludes the project report.

2. RELATED WORK

Prediction of air quality through machine learning models has been of profound interest to many researchers. Hence, it is not surprising that ample proposed approaches are present in this domain.

Kalapanidas et al. [5] used case-based reasoning (CBR) system to predict air pollution from meteorological features such as temperature, wind, precipitation, solar radiation, and humidity. Athanasiadis et al. [6] employed the σ-fuzzy lattice neurocomputing classifier to predict and categorize O_3 concentrations into three sub-levels on the basis of meteorological features and other pollutants such as SO_2, NO and NO_2.

Kurt and Oktay [7] modelled the geographic connections into a neural network and then predicted the daily concentration levels of SO_2, CO, and PM10 three days in advance. However, it is to be noted that their process of converting regression tasks to classification tasks is problematic, as it ignores the magnitude of the numeric data and consequently may provide inaccurate results. Corani [8] worked on training neural network models to predict the hourly O_3 and PM10 concentrations on the basis of data from the immediate previous day. Jiang et al. [9] explored multiple models to predict air pollutants and obtained reliable and robust results. Ni, X. Y. et al. [10] compared multiple statistical models for the PM2.5 data surrounding Beijing and suggested linear regression to be a superior model for the studied problem.

This project utilizes both meteorological data and pollutant concentrations to perform predictions of hourly concentrations. We aim to predict the concentrations of PM2.5, PM10 and O_3 in Beijing based on pollutant concentration of neighbourhood areas based on features like temperature, pressure, humidity, wind direction and wind speed.

3. METHODOLOGY

This section focuses on the implementation details of our project. After a brief description of the data, the data processing and feature engineering methods are elaborated. The implemented machine learning model is then discussed.

3.1. Data Description

For this project, we have utilized two types of data – the air quality data and two meteorological (weather) data. All three training datasets are made available for this project. However, only the two weather datasets are available for model testing. Since few parameters in the air quality dataset is also considered as model input features, we have obtained those values from an external data source via the Beijing Meteorological Data [11].

A. Air Quality Data

Air quality data primarily contains hourly concentrations of air pollutants in 35 stations located in Beijing. The air quality data spans from January 1st, 2017 to April 30th, 2018. The attributes of the dataset are the station names and concentrations of PM2.5, PM10, NO_2, CO, O_3 and SO_2 (in $\mu g/m^3$). A separate file containing the latitude and longitude positions for all the air quality stations is also present.

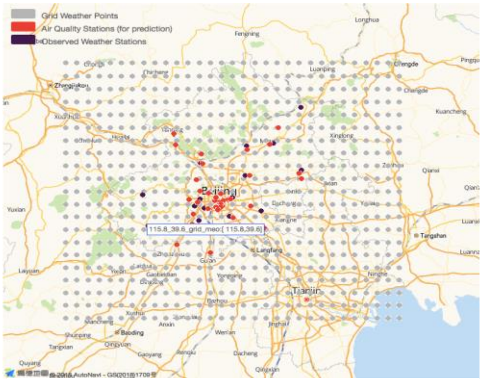

Grid Weather Points
Air Quality Stations (for prediction)
Observed Weather Stations
Beijing
Tianjin
Zhangjiakou
Chengde
Langfang
Baoding
Tangshan
115.8_39.6_grid_meo:[115.8,39.6]
©2018 AutoNavi - GS(2018)1709号

Figure 3: Schematic representation of all air quality and weather stations in Beijing, China

B. Weather Data

Within the meteorological data, two types of data are available – the observed weather data and the grid weather data. Observed weather data is obtained from the measurements of weather stations in Beijing.

The weather in these stations are measured by instruments like thermometer (for temperature) and anemometer (for wind speed). The observed weather data is expected to contain human errors and equipment faults. Hence, there is another data file that gives the grid weather information in Beijing. A quality control process takes the weather station observations, satellite images and other atmosphere data, performs complex calculations and returns a continuous data distribution to produce the aforementioned grid weather data. The data from a total of 18 observed weather stations and 651 grid weather stations are used in this project. Like the air quality data, both the observed weather data and grid weather data also span from January 1st, 2017 to April 30th, 2018.

Table 1: Description of features present in datasets utilized for training

Data File	Attributes
Air Quality Data	Station ID, PM2.5, PM10, NO2, CO, O3, SO2
Air Quality Location	Station ID, latitude, longitude, type
Observed Weather data	Station ID, latitude, longitude, utc-time, temperature, pressure, humidity, wind-direction, wind-speed, weather
Grid Weather Data	Station ID, latitude, longitude, utc-time, temperature, pressure, humidity, wind-direction, wind-speed
Grid Weather Station Location	Station ID, latitude, longitude

3.2. Exploratory Analysis

An initial visualization of the data is performed to capture its underlying characteristics. This is important as it gives an overview of the data distribution and other nuances that further aid in feature engineering and model selection. Figure 4 depicts the correlation between all attributes of the training data. Figure 5 represents the monthly variation of PM2.5, PM10 and O_3. From the visualizations, it is noted that O_3 is highly dependent on temperature and wind speed while PM2.5 and PM10 are dependent on NO_2, CO and SO_2 concentrations. All three air pollutants also vary considerably with the month.

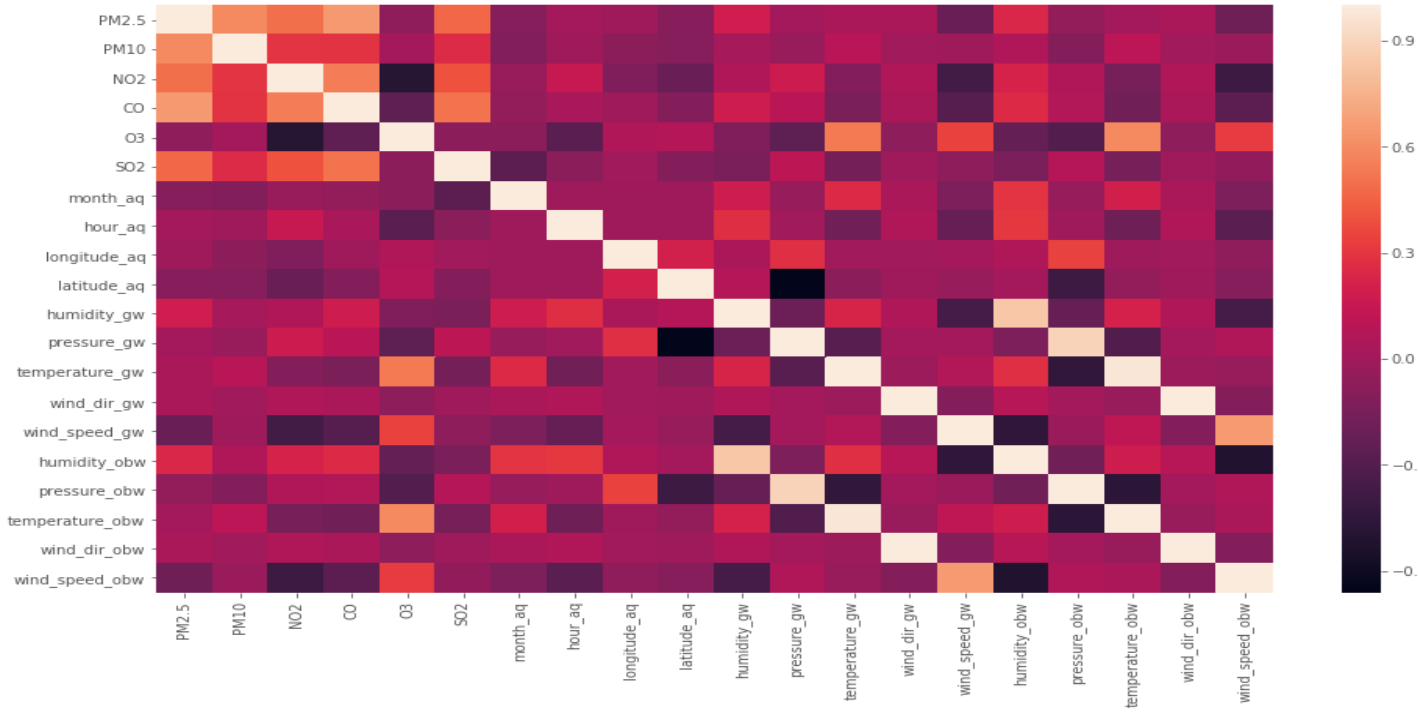


Figure 4: Correlation matrix of all attributes in the training dataset

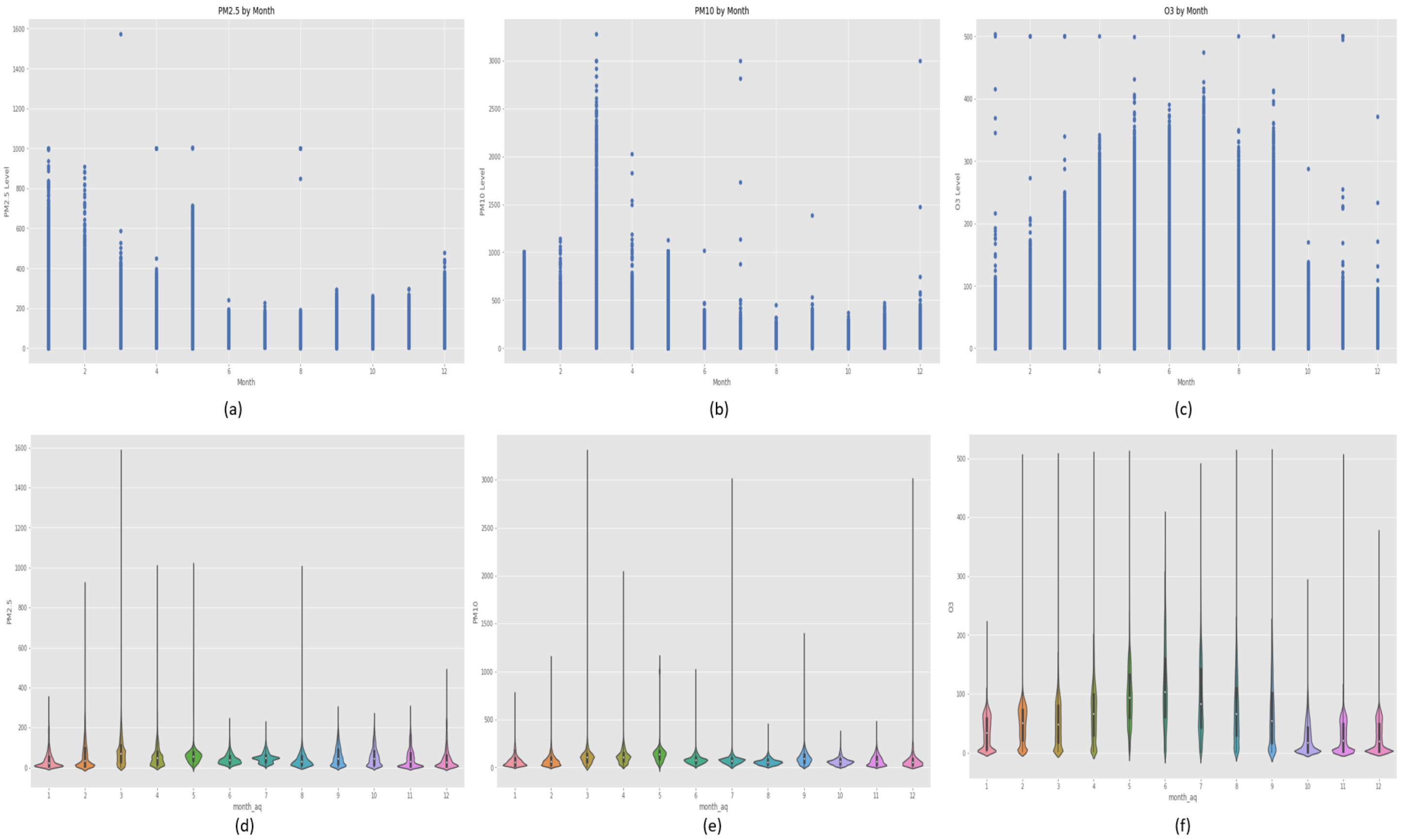


Figure 5: Monthly variation of air pollutants via scatter plot and violin plot for (a)&(d) PM2.5, (b)&(e)PM10 and (c)&(f) O_3

3.3. Individual Data Cleaning

Like all real-world data, the obtained air quality, observed and grid weather data contained noise, mistakes and missing values. Hence, it required multiple stages of data cleaning and pre-processing to make it suitable for feature engineering and model prediction (figure 6). Below is the detailed sequence of steps performed for obtaining individual clean data for air quality, observed weather and grid weather data respectively. Figure 8 describes the steps in a flowchart with reference to the air quality data, but the same steps were performed for the other datasets, including test data, with the necessary changes.

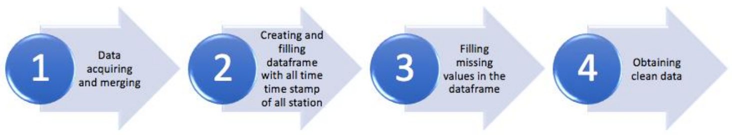

1
Data acquiring and merging
2
Creating and filling dataframe with all time time stamp of all station
3
Filling missing values in the dataframe
4
Obtaining clean data

Figure 6: Schematic overview of the data cleaning process

1. Dataset Merging: For every category (air quality, grid weather and observed weather), the training datasets are spread out over a number of sheets, and hence need to be combined.
2. Create a new data frame containing all the time stamps over the category period range.
3. Find closest neighbouring station ('air quality to air quality' or 'grid to grid' etc.). This is needed to fill the missing values from the dataset with the closest station values from the same time stamp. For this project, all distances are calculated via *Haversine distance* as it provides the accurate distance between two geographic locations based on latitudes and longitudes on a spherical surface (figure 7).

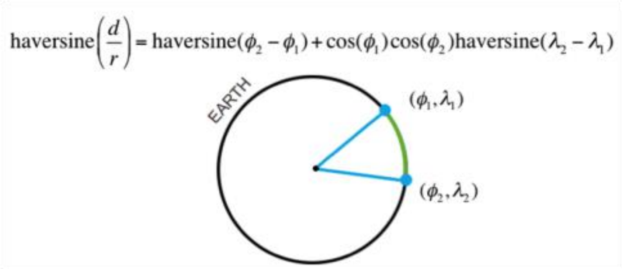

$\text{haversine}\left(\frac{d}{r}\right) = \text{haversine}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\text{haversine}(\lambda_2 - \lambda_1)$
EARTH
(ϕ_1, λ_1)
(ϕ_2, λ_2)

Figure 7: Schematic representation of calculating Haversine distance

4. Any remaining missing value (arising as the nearest station too does not have the value for that particular timestamp) is filled by taking the mean of the value from all the similar timestamps of that particular station within that month.
5. Finally, the cleaned datasets for air quality, grid and weather (train and test) is obtained.

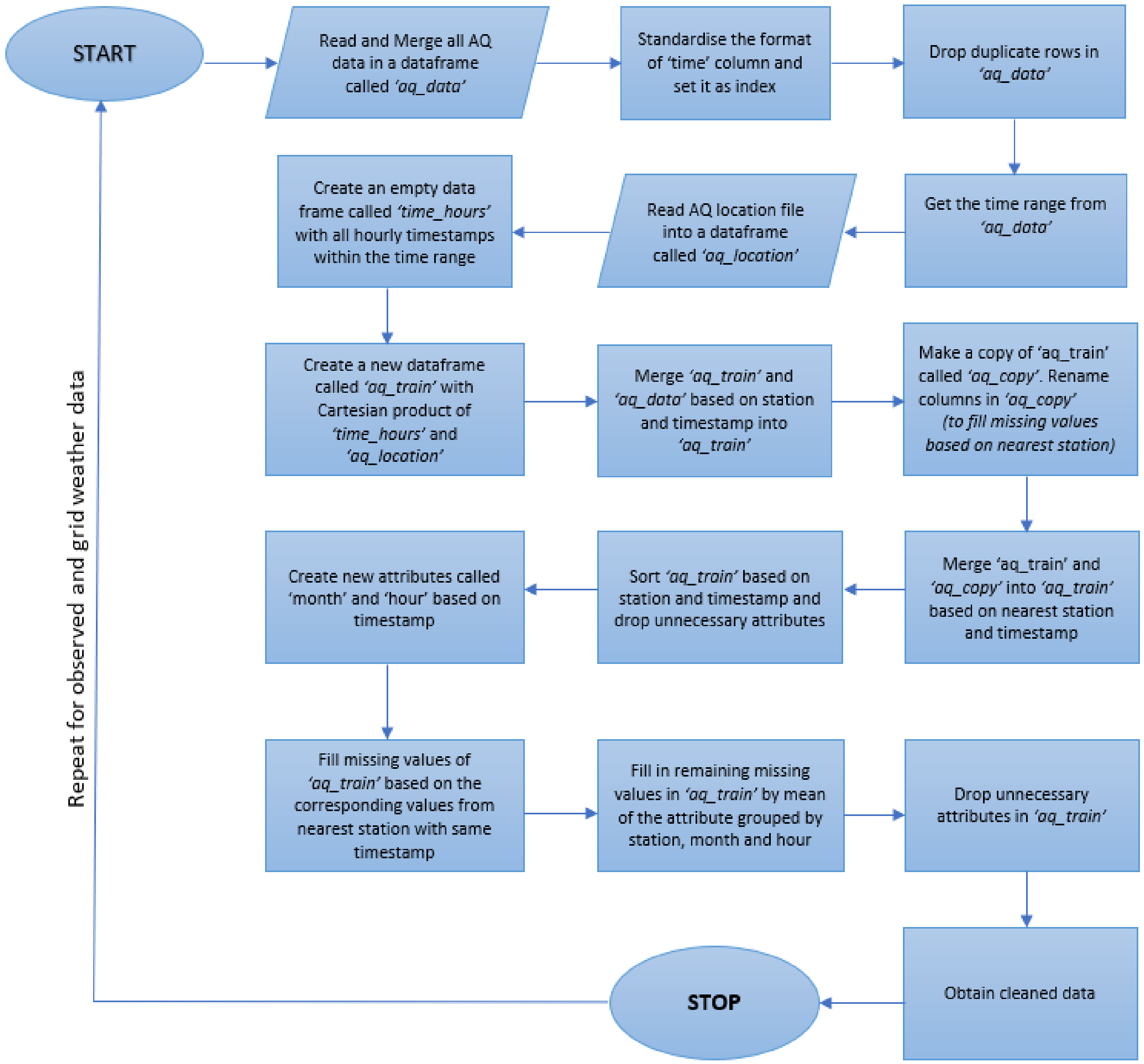

START
Read and Merge all AQ data in a dataframe called *'aq_data'*
Standardise the format of 'time' column and set it as index
Drop duplicate rows in *'aq_data'*
Get the time range from *'aq_data'*
Read AQ location file into a dataframe called *'aq_location'*
Create an empty data frame called *'time_hours'* with all hourly timestamps within the time range
Create a new dataframe called *'aq_train'* with Cartesian product of *'time_hours'* and *'aq_location'*
Merge *'aq_train'* and *'aq_data'* based on station and timestamp into *'aq_train'*
Make a copy of 'aq_train' called *'aq_copy'*. Rename columns in *'aq_copy'* *(to fill missing values based on nearest station)*
Merge 'aq_train' and *'aq_copy'* into *'aq_train'* based on nearest station and timestamp
Sort *'aq_train'* based on station and timestamp and drop unnecessary attributes
Create new attributes called 'month' and 'hour' based on timestamp
Fill missing values of *'aq_train'* based on the corresponding values from nearest station with same timestamp
Fill in remaining missing values in *'aq_train'* by mean of the attribute grouped by station, month and hour
Drop unnecessary attributes in *'aq_train'*
Obtain cleaned data
STOP
Repeat for observed and grid weather data

Figure 8: Detailed flowchart of data cleaning process for air quality training set (to be referenced for all individual cleaning process)

3.4. Group Data Pre-processing

To obtain the final training dataset, all the clean individual datasets need to be combined based on their nearest station ('air quality to grid' and 'air quality to observed') and same timestamp. The following steps are performed to merge the air quality, observed and weather data.

1. Make the starting and ending timestamps the same for all the three datasets.
2. Find the nearest observed and grid weather station for every air quality station using Haversine distance.
3. Merge the air quality data with the nearest observed and grid weather data with the same timestamp.
4. Replace noise (random 999999 values) in the observed weather data with the values from the same timestamp from the nearest grid weather data and vice versa.
5. Convert all wind speed values in the training grid weather dataset from kmph to m/s. All wind speeds in observed weather datasets are in m/s. The wind speed for test grid weather data is already in m/s.
6. Replace noise values (999017) in the wind direction (when wind speed is less than 0.5m/s) attribute of observed weather with the corresponding grid weather wind direction values and vice versa
7. Replace outliers in humidity, pressure, temperature, wind direction and wind speed of observed weather data with the corresponding grid weather data attributes.
8. Sort the rows based on station and timestamp.

3.5. Feature Engineering

Feature engineering is a crucial step for acquiring a comprehensive training data that can further be used for predictive modelling. The data is transformed to add new attributes that increase the prediction accuracy of the machine learning model. In this project, four new features are added to the pre-processed dataset to enhance the training data. Month and hour attributes, derived from the timestamp attribute, are added.

Further, an attribute calculating the ratio of wind direction by wind speed is added along with another attribute containing the overall multiplication of NO_2, SO_2 and CO. The latter attribute is also obtained for the air quality test dataset as the externally used data source contains information about NO_2, SO_2 and CO for all the 35 Beijing air quality stations over the required two day period.

3.6. Feature Selection

Not all features contribute to the accuracy of the model. Hence appropriate feature selection is paramount to achieve optimal model performance. In order to predict the accurate values of PM2.5, PM10 and O_3 we have selected the following essential attributes as part of the final training data. The same attributes are selected in the final test data to get an accurate prediction result.

1. Air quality station
2. Month
3. Hour
4. Temperature of grid weather data
5. AQ (multiplication of NO_2, SO_2 and CO)
6. Humidity of grid weather data
7. Pressure of grid weather data
8. Ratio of wind direction to wind speed (from grid weather values)

Since observed weather data contains high amount of errors (via instruments or humans), it proved to subdue the efficiency of the model's predictions. Thus, we have neglected the observed weather data attributes as part of the final training of the model. However, it is still necessary to utilize the observed data points during data pre-processing as any missing values or noise from the grid weather data is replaced with the nearest observed weather data. In addition, only data samples pertaining to the months of April, May and June are selected for training the model as the monthly distribution of the pollutants are found to be similar over these three months.

3.7. Implemented Model

This project utilizes Random Forest Regressor for the final model. Random Forest is a flexible, easy-to-use machine learning algorithm that produces accurate results without much hyperparameter tuning. Unlike XGBoost and CatBoost regressors, Random Forest Regressor did not overfit the model for this project.

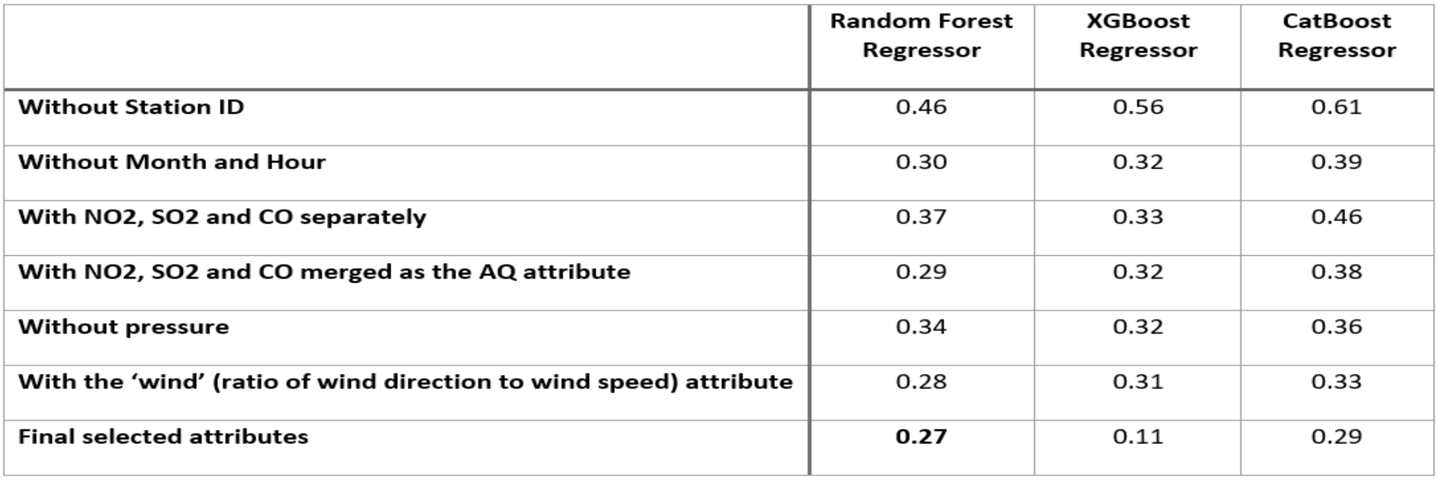
3.8. Performance Evaluation Metric

As an evaluation criterion, we have used Symmetric Mean Absolute Percent Error (SMAPE). An alternative to Mean Absolute Percent Error (MAPE), SMAPE ignores outliers and is invariant to linearly rescaled data. Hence, it is often used as an evaluation metric in time series forecasting.

$$smape = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(A_t + F_t)/2}$$

4. RESULTS AND DISCUSSION

To check the efficiency of the model, we utilized 10 Fold cross validation to find the SMAPE score on the validation set. It was seen that when SMAPE reached 0.11 in the cross validation set, the model was overfitting on the test set (i.e., the output had negative answers). As the main aim of every machine learning model is to reduce test data loss, we chose to opt for the parameters of the Random Forest Regressor that gave an average of 0.27 SMAPE error in the cross validation set. Figure 9 summarizes all SMAPE results obtained via Random Forest, XGBoost and CatBoost models for different combinations of the final attribute list.



	Random Forest Regressor	XGBoost Regressor	CatBoost Regressor
Without Station ID	0.46	0.56	0.61
Without Month and Hour	0.30	0.32	0.39
With NO2, SO2 and CO separately	0.37	0.33	0.46
With NO2, SO2 and CO merged as the AQ attribute	0.29	0.32	0.38
Without pressure	0.34	0.32	0.36
With the 'wind' (ratio of wind direction to wind speed) attribute	0.28	0.31	0.33
Final selected attributes	**0.27**	0.11	0.29

Figure 9: Cross validation SMAPE results for different attributes over three models

5. CONCLUSION

Air pollution poses significant threats to both physical and mental health. Predicting air pollutant levels will help in obtaining suitable measures to curb its effects. Via this project, we have predicted the levels of PM2.5, PM10 and O_3 for all Beijing air quality stations over a span of two days in the month of May 2018. Our model is robust and reliable, with a SMAPE error of 0.27. We have primarily focused on data pre-processing and feature engineering to enhance the training data quality and in turn, the model efficiency. We believe that our proposed approach can be adapted to yield higher forecasting accuracy.

6. INDIVIDUAL CONTRIBUTION

The following table summarizes the individual contributions made towards this project. The details of the work done within each contribution topic can be found under their corresponding headings.

Student Name	Student ID	Contribution
BANERJEE, Rohini	20543577	1. Data Cleaning (Train and Test Data) 2. Data Pre-processing (Train Data) 3. Feature Engineering (Train and Test Data) 4. Feature Selection (Train and Test Data) 5. Model Implementation (Train Data) 6. Report Documentation
RAMALINGAM, Poojaa	20214528	1. Exploratory Analysis 2. Data Pre-processing (Train and Test Data) 3. Model Implementation (Test Data) 4. Report Documentation

7. REFERENCES

[1] World Health Organization News: https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

[2] World Health Organization News: http://www.who.int/en/news-room/detail/06-03-2017-the-cost-of-a-polluted-environment-1-7-million-child-deaths-a-year-says-who

[3] Forbes News: https://www.forbes.com/sites/niallmccarthy/2018/04/18/air-pollution-contributed-to-more-than-6-million-deaths-in-2016-infographic/#7090d11013b4

[4] Rohde RA, Muller RA. Air Pollution in China: Mapping of Concentrations and Sources. PLoS ONE. 2015; 10(8): e0135749. https://doi.org/10.1371/journal.pone.0135749 PMID: 26291610

[5] Kalapanidas, E.; Avouris, N. Short-term air quality prediction using a case-based classifier. *Environ. Model. Softw.* 2001, *16*, 263–272.

[6] Athanasiadis, I.N.; Kaburlasos, V.G.; Mitkas, P.A.; Petridis, V. Applying machine learning techniques on air quality data for real-time decision support. In Proceedings of the First international NAISO Symposium on Information Technologies in Environmental Engineering (ITEE'2003), Gdansk, Poland, 24–27 June 2003.

[7] Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Syst. Appl.*2010, *37*, 7986–7992.

[8] Corani, G. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. *Ecol. Model.* 2005, *185*, 513–529.

[9] Jiang, D.; Zhang, Y.; Hu, X.; Zeng, Y.; Tan, J.; Shao, D. Progress in developing an ANN model for air pollution index forecast. Atmos. Environ. 2004, 38, 7055–7064.

[10]Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. Atmos. Environ. 2017, 150, 146–161.

[11] Beijing Air Quality Historical Data http://beijingair.sinaapp.com/