International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Analysis of Real-Time Fine Particulate Matter (PM2.5) Concentrations for Air Pollution Using Machine Learning

*Miss. S. Jayasri* [1]*, Mrs. R. Vijayalakshmi*[2]

**[1](M.C.A), Department of MCA, Krishnasamy College of Engineering and Technology.**
**[2]M.C.A, M.Phil., (Ph.D.), Associate Professor, Department of MCA, Krishnasamy College of Engineering and Technology.**

**ABSTRACT**

By utilising machine learning to estimate the air quality index of a certain place, we forecast India's air quality. The air quality index of India is a commonly used indicator of pollution levels (so2, no2, rspm, spm, etc.) through time. We created a model to forecast the air quality index based on historical data from prior years and using regression problem to forecast over a certain forthcoming year. For our predictive problem, we use cost estimation to increase the model's efficacy. When given historical data on pollutant concentration, our model will be able to accurately estimate the air quality index for an entire county, any state, or any contiguous region. We improved performance utilising the conventional regression models in our model by following the suggested parameter-reducing formulations.[1]

**Key Words:** Air Quality Index, Air Pollution

## I. INTRODUCTION

By predicting the air quality index for a certain place, machine learning is used to forecast India's air quality. The air quality index of India is a commonly used indicator of pollution levels (so2, no2, rspm, spm, etc.) through time. Using historical data from previous years and projecting over a certain upcoming year as a Gradient Decent Boosted Multivariable Regression Problem, a model was constructed to predict the air quality index. For our predictive problem, we use cost estimation to increase the model's efficacy. when given historical data on pollutant concentration, is capable of accurately predicting the air quality index for a whole county, any state, or any restricted region.

We improved performance utilising the conventional regression models in our model by following the suggested parameter-reducing formulations. Monitoring of air pollution has been more popular recently because it significantly affects both human health and the ecological balance.[2]

Besides due to the effects of toxic emissions on the environment, health, work productivity and efficiency of energy are also affected by the air pollution. Since air pollution has caused many hazardous effects on humans it should be monitored continuously so that it can be controlled effectively. One of the ways to control air pollution is to know its source, intensity and its origin. Usually, it is monitored by the respective state government’s environment ministry. They keep the cord of the pollutant gases in the respective areas. The data presented by the WHO is warning about the pollutions levels in the country. It tells us it’s high time that we should monitor the air. Air tracking manner to measure ambient ranges of air pollutants inside the air. Monitoring has become a majorjob as air pollution has been increasing day by day. Continuous monitoring of air pollution at a place gives us the levels of pollution in that area[3]. From the information obtained by the device gives us information about the source and intensity of the pollutants in that area. Using that information, we can take measures or make efforts to reduce the pollution level so that we can breathe in a good quality of air. Air pollution affects the ecological balance but also the health of humans. As the levels of gases increases in the air, those gases show a major impact on the human body and lead to hazardous effects. Air pollution also affects the seasonal rainfall too due to an increase of pollutants in the air. The rainfall is also affected.[4]

Hence, continuous monitoring of the air These gases are cannot been seen or noticed which are produced from burning of fossil fuels, wood burnings, industrial boilers and from the explosion of volcano. They may cause the affects in humans and are the main reason for causing cancer, birth defects and breathing-related problems. Air Quality Index- Nowadays pollution levels are increasing due to the PM2.5 gases which affect the heart functionalities, lung cancer and other respiratory and breathing problems. The long-term damage to the liver, kidney, brain, nerve and other organs in the human body system is affected by air pollution. The AQI is a linear feature of the pollutant concentration. The boundaries between AQI there is discontinuous jump between AQI categories unit to other. To calculate the AQI from the concentration the below equation is used.[5]

The Indian Government has declared severe levels of toxic air pollution in Delhi was an "emergency situation." According to global air pollution data and an IndiaSpend analysis of national data, over the Diwali weekend of 2016, India’s air quality was among the world’s worst and 100 % worse in five north Indian cities than at the same time the preceding year. In 2013, a report by Global Burden of Disease (GBD) said that outdoor air pollution was the 5th largest killer in India and nearly one lakh premature deaths happen annually due to airpollution. Recent studies have shown substantial evidence that exposure to atmospheric pollutants has strong links to adverse diseases including asthma and lung inflammation. Approximately 30 million people including children die due to asthma. At the same time particulate materials (PM2.5 or PM10) in the air can cause several kinds of respiratory, cardiovascular diseases and blood diseases.[6]

Medical studies have shown that PM2.5can be easily absorbed by the lung, and high concentrations of PM2.5 can lead to respiratory disease or even blood diseases. Air pollution has both acute and chronic effects on human health, affecting a number of different systems and organs. It ranges from minor upper respiratory irritation to chronic respiratory and heart disease, lung cancer, acute respiratory infections in children and chronic bronchitis in adults, aggravating pre-existing heart and lung disease, or asthmatic attacks. Because of its direct impact on public health, it has gained a lot of attention. In last decade there is a lot of improvement in the techniques of air quality monitoring and forecasting. There are mainly two approaches in which researchers are working. In first approach monitoring of real-time Air Quality Monitoring and another is developing statistical models using historical data.summarizes Air Quality Prediction studies in two categories. The first study is on prediction of PM2.5 / PM10 concentration and on prediction of air pollutants like CO2, O3, NO2 and then inferring Air Quality Index (AQI) using machine learning techniques.[7]

## II. PROBLEM STATEMENT

**Environmental Concerns**: The project may be driven by the growing concern over air pollution and its detrimental effects on public health and the environment. Initiating this project reflects a commitment to addressing and mitigating air pollution issues.[8]

**Public Health Impact:** Air pollution has been linked to various respiratory and cardiovascular diseases, and its impact on public health is a significant concern. Starting this project can contribute to understanding and predicting air quality to better protect the health and well-being of individuals and communities.[9]

**Policy and Decision-making**: Accurate and reliable air quality predictions can provide valuable information for policymakers, urban planners, and other decision-makers. Starting this project can facilitate evidence-based policy formulation and interventions to improve air quality management and regulatory measures.[10]

**Resource Allocation**: Efficient allocation of resources, such as funds, technology, and manpower, requires accurate knowledge of air quality patterns. Starting this project can enable better resource allocation by identifying areas or regions that require more targeted interventions and mitigation strategies.[11]

## III. MODULES

**Data Collection**

- **temp_max:**The maximum temperature for a given day.
- **temp_min:**The minimum temperature for a given day.
- **avg_temp:**The average temperature for a given day, which is typically calculated by taking the mean of the temp_max and temp_min.
- **pressure:**Atmospheric pressure for a given day, typically measured in millibars.
- **humidity:**The amount of water vapor in the air for a given day, typically measured as a percentage of the maximum amount of water vapor that the air can hold at a given temperature and pressure.
- **visibility:**The distance at which objects can be clearly seen in the atmosphere for a given day, typically measured in kilometres or miles.
- **wind:**The speed and direction of the wind for a given day, typically measured in meters per second or miles per hour.[12]

**Data Preprocessing**

**Standard Scaling**

- The standard score of a sample x is calculated as:
- z = (x - u) / s
- the mean of the training samples or zero if with_mean=False, and s is the standard deviation of the training samples or one if with_std=False.
- Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform.[13]

**MODEL TRAINING**

Two different machine learning models:

- Linear Regression
- Random Forest

**Linear Regression**

Linear regression provides interpretable coefficients that can be used to understand the relationship between the predictor variables and the response variable.

In air pollution prediction, these coefficients can be used to identify which pollutants or meteorological factors are most strongly associated with changes in air quality.

Linear regression can be used to model both continuous and categorical predictors, making it a versatile technique for air pollution prediction.

**Rf Regression**

RF regression is robust to outliers in the data, which is important since air pollution levels can be affected by extreme values in the input variables such as sudden weather changes or unexpected emissions from industrial sources.

Handles missing data: RF regression can handle missing data effectively using a technique called imputation, which is useful in air pollution prediction where some sensor data may be missing due to equipment failure or maintenance issues.[14]

**MODEL TESTING**

**Mean Absolute Error**

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where, ŷ—Predicted value of y

ȳ—mean value of y

**Mean Squared Error**

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

**Root Mean Squared Error**

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.[15]

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

**Performance Evaluation**

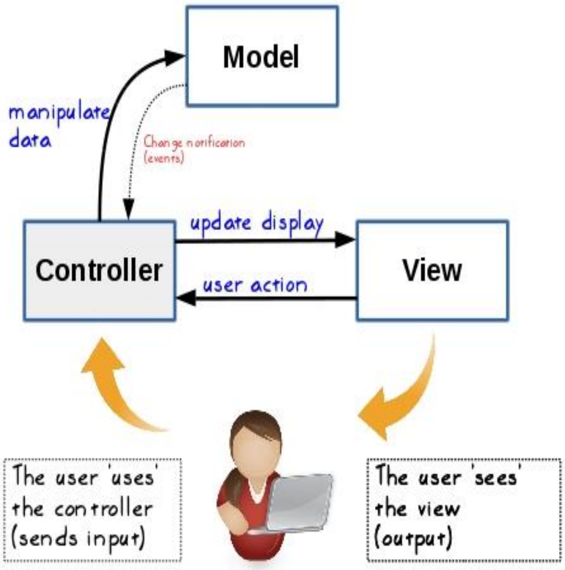
Performance evaluation is a critical step in assessing the accuracy and effectiveness of the air pollution prediction system using the ML model and Flask framework. Here are some common methods for evaluating the performance of the system.

**Accuracy Metrics:** Calculate various accuracy metrics to quantify the performance of the ML model. Common metrics for regression problems include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) value. These metrics measure the deviation between the predicted air pollution levels and the actual values.[16]

**Flask Framework Prediction**

- The front end of a website is the area with which the user immediately interacts. It contains everything that users see and interact with: text colors and styles, images and videos, graphs and tables, the navigation menu, buttons, and colors.

- HTML, CSS, and JavaScript are used in developing the front end. Flask is used for developing web applications using python.
- started by importing the Flask class. We then make an instance of this class.
- The '__name__' argument is passed which is the name of the application's module or package. Flask needs this to know where to look for resources like templates and static files.
- The route () decorator is then used to inform Flask which URL should activate our method. This method returns the message that should be shown in the user's browser.
- Flask is a lightweight web framework for Python that allows you to build web applications quickly and easily. Flask works by Model-View-Controller (MVC) architecture pattern, where the model represents the data, the view represents the user interface, and the controller acts as an intermediary between the model and the view.
- Install Flask
- Define routes: Routes are the URLs that the user can visit in your web application. You can define routes in Flask by using the @app. route () decorator and specifying the URL pattern as a parameter.[17]


Model
manipulate data
Change notification (events)
update display
Controller
user action
View
The user 'uses' the controller (sends input)
The user 'sees' the view (output)

**Output Prediction**

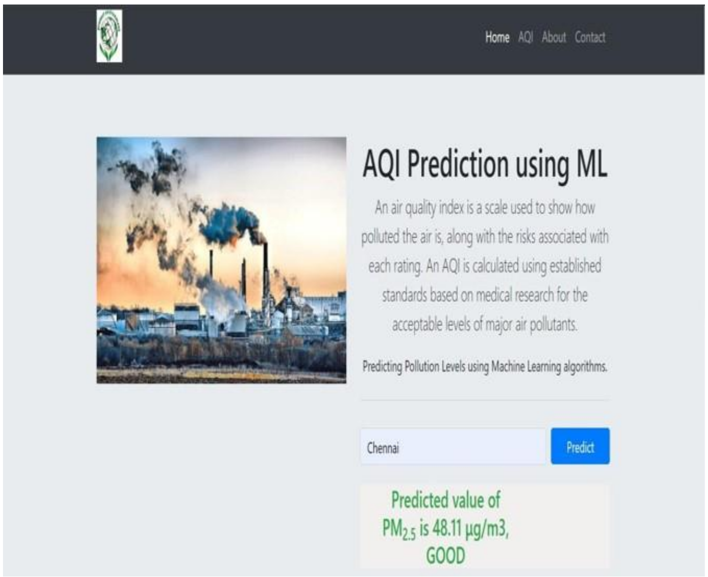
The predicted AQI can be categorized into different air quality levels, such as "good," "moderate," "unhealthy," "very unhealthy," or "hazardous." The specific categorization may vary depending on the air quality index scale being used (e.g., EPA Air Quality Index). The predicted AQI provides an indication of the overall air pollution level and helps users understand the potential health risks associated with the given location.

The output prediction can be presented in various formats within the Flask framework, depending on the user interface design. It could be displayed as a numerical value representing the predicted AQI, along with a corresponding air quality level category. Additionally, the output may be visualized through charts, graphs, or maps to provide a more intuitive representation of the air pollution levels across different locations.

The accuracy and reliability of the output predictions will depend on the quality of the ML model, the data used for training and testing, and the system's overall performance. Continuous monitoring and evaluation of the system's output, as well as incorporating user feedback and updates to the ML model, can help improve the accuracy and ensure the reliability of the predictions over time.[18]

- < 51 Predicted values of $PM_{2.5}$ is GOOD
- < 101Predicted values of $PM_{2.5}$ is MODERATE
- < 151 Predicted values of $PM_{2.5}$ is UNHEALTHY FOR KIDS
- < 201 Predicted values of $PM_{2.5}$ is UNHEALTHY
- < 301 Predicted values of $PM_{2.5}$ is VERY UNHEALTHY
- >301 HAZARDOUS

## IV. RESULT



AQI Prediction using ML

An air quality index is a scale used to show how polluted the air is, along with the risks associated with each rating. An AQI is calculated using established standards based on medical research for the acceptable levels of major air pollutants.

Predicting Pollution Levels using Machine Learning algorithms.

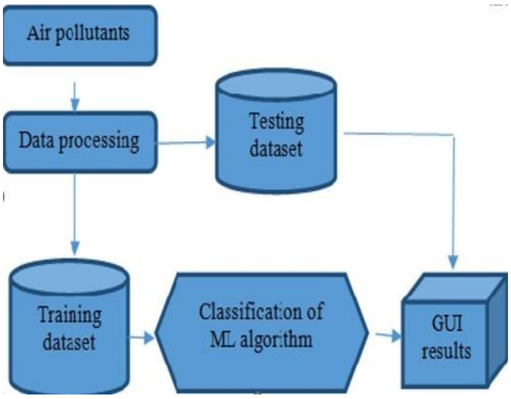
Chennai

Predicted value of $PM_{2.5}$ is 48.11 µg/m3, GOOD

## V. DIAGRAM

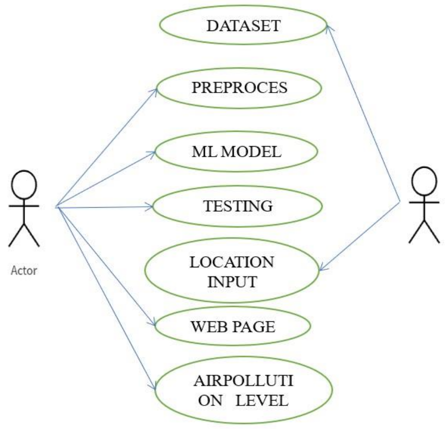
### 1. DATA FLOW DIAGRAM

The air pollutant data preprocessing, classification, and prediction process begins with the collection of historical air pollutant data, including pollutant concentrations and air quality index, from reliable sources. The collected data is then preprocessed by removing incomplete or inaccurate data points and handling missing values. Feature selection and engineering techniques are applied to identify relevant attributes and improve the quality of the data. Next, the preprocessed dataset is divided into a training dataset and a testing dataset. The training dataset is used to train a machine learning algorithm, which learns the patterns and relationships between pollutant concentrations and the corresponding air quality index. Once the model is trained, it can classify or predict the air quality index for unseen data.[19]



### 2. USECASE DIAGRAM

A usecase is a set of scenarios that describing an interaction between a user and a system. A usecase diagram displays the relationship among and usecases. The two main components a user or another system that will interact with the system modelled. A usecase is an external view of the system that represents some action the user might perform in order to complete a task.



## 3. ACTIVITYDIAGRAM

An activity diagram is a visual representation that illustrates the flow of activities or processes within a system. It captures the sequence of actions, decision points, and interactions between different components or actors involved in a specific process. In the context of an air pollution prediction project with a webpage, Flask framework, web scraping, input location, preprocessing, ML algorithm prediction, and air pollution level prediction, the activity diagram would showcase the steps and interactions involved in the overall process.Here's an explanation of the activity diagram for this project:

**User Interaction:**

The activity diagram begins with the user interacting with a webpage or user interface to initiate the air pollution prediction process.

The user provides input, such as the location for which they want to predict air pollution levels.

**Web Scraping:**

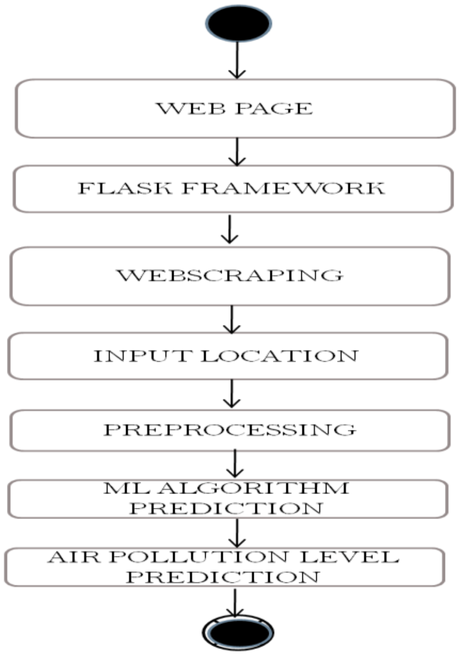
The system utilizes web scraping techniques to gather relevant air pollutant data from reliable sources.

This step involves extracting the necessary data, such as pollutant concentrations, from the web.

**Input Location Processing:**

The system processes the input location provided by the user.

It may involve geocoding techniques to convert the location into geographical coordinates or any other preprocessing required for location-based analysis.[20]



## VI. CONCLUSION

Since our model is capable of predicting the current data with 95% accuracy it will successfully predict the upcoming air quality index of any particular data within a given region. With this model we can forecast the AQI and alert the respected region of the country also it a progressive learning model it is capable of tracing back to the particular location needed attention provided the time series data of every possible region needed attention. The essential perspectives that should be viewed as with regards to gauging of the poison focus are its different sources alongside the components that impact its fixation.The implementation of a machine learning model for air pollution prediction integrated with the Flask framework offers a powerful and practical solution. The ML model, trained on historical pollutant concentration data, can effectively forecast air quality levels. By utilizing Flask, the model can be deployed as a web application or API, allowing users to access air pollution predictions conveniently.

The integration of the ML model with Flask provides several benefits. Firstly, Flask offers a lightweight and flexible framework for building web applications, making it suitable for deploying the air pollution prediction model. Secondly, Flask enables seamless communication between the model and user interfaces, allowing users to input locations or other relevant data for prediction. Thirdly, Flask's scalability and extensibility make it suitable for handling multiple user requests concurrently and accommodating future updates or enhancements to the system.

By combining the ML model with Flask, the air pollution prediction system becomes accessible to a wide range of users, including researchers, policymakers, and the general public. The system can help users make informed decisions regarding their activities, such as planning outdoor events or taking necessary precautions based on predicted air quality levels.

Overall, the integration of a machine learning model with the Flask framework offers a user-friendly and efficient solution for air pollution prediction. It enables users to access accurate and timely information about air quality, contributing to better environmental awareness and potentially improving public health outcomes.

## VII. FUTURE ENHANCEMENT

**Geographical Visualization:** Enhancing the system with geographical visualization capabilities can provide users with a clear understanding of air pollution levels across different regions. Maps, heatmaps, or other visual representations can be incorporated to visualize the predicted air quality index and highlight areas with higher pollution levels.

**Ensemble Models:** Consider implementing ensemble models by combining multiple machine learning algorithms or models. Ensemble methods, such as random forests or gradient boosting, can improve prediction accuracy by leveraging the strengths of different algorithms.

**User Personalization:** Introduce user personalization features by allowing users to create profiles and receive customized air pollution alerts or recommendations based on their preferences and locations. This can enhance user engagement and provide tailored information for individual needs.

**Integration with External Data Sources:** Explore the integration of additional data sources, such as weather data, population density, or traffic patterns, to improve the predictive capabilities of the system. Incorporating these external factors can provide a more comprehensive understanding of the air pollution levels. Mobile Application Development: Extend the system's accessibility by developing a mobile application alongside the web interface. A mobile app would enable users to access air pollution predictions on-the-go, receive notifications, and contribute real-time data through user-generated reports.

**User Feedback and Improvement Loop:** Implement a feedback mechanism to collect user feedback on the accuracy of predictions and incorporate it into model refinement. This iterative feedback loop can help continuously improve the system's performance and address any limitations or inaccuracies.

## VIII. REFERENCES

[1] 'Blacksmith Institute Press Release'. (October 21, 2008). [Online]. Available: http://www.blacksmithinstitute.org/the-2008-top-ten-list-of-world-s-worst-pollution-problems.html

[2] V. M. Niharika and P. S. Rao, “A survey on air quality forecasting techniques,” International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp.103-107, 2014.

[3] NAAQS Table. (2015). [Online]. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table

[4] E. Kalapanidas and N. Avouris, “Applying machine learning techniques in air quality prediction,” in Proc. ACAI, vol. 99, September 1999.

[5] Questioning smart urbanism: Is data-driven governance a panacea? (November 2, 2015). [Online]. Available: http://chicagopolicyreview.org/2015/11/02/questioning-smart-urbanis m-is-data-driven-governance-a-panacea/

[6] D. J. Nowak, D. E. Crane, and J. C. Stevens, “Air pollution removal by urban trees and shrubs in the United States,” Urban Forestry & Urban Greening, vol. 4, no. 3, pp. 115-123, 2006.

[7] T. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations," presented at 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), IEEE, 2016.

[8] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in Proc. the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2267-2276, August 10, 2015.

[9] J. A. Engel-Coxa, C. H. Hollomanb, B. W. Coutantb, and R. M. Hoffc, "Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality," Atmospheric Environment, vol. 38, issue 16, pp. 2495–2509, May 2004.

[10] J. Y. Zhu, C. Sun, and V. Li, "Granger-Causality-based air quality estimation with spatio-temporal (ST) heterogeneous big data," presented at 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2015.

[11] C. J. Wong, M. Z. MatJafri, K. Abdullah, H.S. Lim, and K. L. Low, "Temporal air quality monitoring using surveillance camera," presented at IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2007.

[12] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal, "Classification model for air quality using machine learning techniques," International Journal of Software Engineering and Its Applications, pp. 45-52, 2015.

[13] A. Sarkar and P. Pandey, "River water quality modelling using artificial neural network technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2015.

[14] E. Kalapanidas and N. Avouris, "Applying machine learning techniques in air quality prediction," Sept. 1999.

[15] H. Zhao, J. Zhang, K. Wang, et al., "A GA-ANN model for air quality predicting," IEEE, Taiwan, 10 Jan. 2011.

[16] V. Jagannath. [Online]. Available: https://community.tibco.com/wiki/random-forest-template-tibco-spotfi rer-wiki-page

[17] R. Yu, Y. Yang, L. Yang, G. Han, and, O. A. Move, "RAQ–A random forest approach for predicting air quality in urban sensing systems," Sensors, vol. 16, no. 1, p. 86, 2016.

[18] Machine learning with decision trees. [Online]. Available: https://blog.knoldus.com/2017/08/14/machine-learning-with-decisiontrees/

[19] S. Deleawe, J. Kusznir, B. Lamb, and D. J. Cook, "Predicting air quality in smart environments," J Ambient Intell Smart Environ., pp. 145-152, 2010.

[20] W. F. Ip, C. M. Vong, J. Y. Yang, and P. K. Wong, "Least squares support vector prediction for daily atmospheric pollutant level," in Proc. 2010 IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS), pp. 23-28, IEEE., August 2010.