Mathematical Sciences and Informatics Journal
Vol. 4, No. 1, May 2023, pp. 1-10
http://www.mijuitmjournal.com
DOI : 10.24191/mij.v4i1.21889

Evaluation of Machine Learning in Predicting Air Quality Index

Abdullah Sani Abdul Rahman
Faculty of Sciences and Information Technology, Univesiti Teknologi Petronas, Perak, Malaysia
sani.arahman@utp.edu.my

Aizal Yusrina Idris
Yanbu Industrial College, Kingdom of Saudi Arabia
idrisa@rcyci.edu.sa, aizalyusrina@gmail.com

Suhaimi Abdul Rahman
Yanbu Technical Institute, Kingdom of Saudi Arabia
abulrahmans@rcyci.edu.sa, suhaimi.ar@gmail.com

Article Info

Article history:

Received Feb 15, 2023
Revised Mac 25, 2023
Accepted Apr 12, 2023

Keywords (minimum 5):

Air Quality Index (AQI)
Generalized Linear Model
Decision Tree
Support Vector Machine
Prediction

ABSTRACT

Environmental pollution poses significant health risks, and Malaysia is facing a critical air pollution issue due to the rapid growth of urbanization and industrialization. The Air Quality Index (AQI) is a standard measure of air pollution, and machine learning methods have shown promise in accurately predicting AQI levels. However, there is limited research on the application of intelligent approaches to predict AQI in Malaysia. This research investigates the impact of various AQI components, including Particulate Matter 2.5 (PM2.5), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), and Ozone (O3), using 125 random locations across Malaysia, ranging from the north to the southern regions. Three machine learning algorithms, namely Generalized Linear Model, Decision Tree and Support Vector Machine are used in this research. The results show that PM2.5 has the most significant impact on AQI levels among all components analyzed, and all selected machine learning algorithms exhibit high prediction accuracy, with R^ above 90% and low prediction errors (less than 2 MAE and RMSE). This research provides essential insights into predicting AQI levels using machine learning approaches and highlights the critical role of PM2.5 in determining AQI levels in Malaysia. The findings can aid authorities in obtaining rapid and accurate information to effectively manage air pollution in the country.

Corresponding Author:
Aizal Yusrina Idris
Yanbu Industrial College, Kingdom of Saudi Arabia
email: idrisa@rcyci.edu.sa, aizalyusrina@gmail.com

1. Introduction

Environmental pollution is a critical global issue that has adverse impacts on human health, the natural environment, and the economy[1],[2]. The World Health Organization (WHO) estimates that air pollution causes seven million premature deaths each year worldwide[3]. Malaysia is not exempt from this issue and faces severe air pollution problems due to rapid urbanization and industrialization. According to the IQAir AirVisual 2019 World Air Quality Report[4], several Malaysian cities ranked among the world's most polluted cities in terms of air quality. This highlights the urgent need to develop effective measures to manage and control air pollution in Malaysia.

Air Quality Index (AQI) is a standard measure of air quality that provides information about the air quality in a particular region. The AQI is calculated based on the concentrations of several air pollutants, including Particulate Matter 2.5 (PM2.5), Nitrogen Dioxide (NO2), Carbon monoxide (CO), and Ozone (O3)[5],[6]. PM2.5 is a type of air pollutant with a diameter of 2.5 micrometers or less. PM2.5 can penetrate deeply into the respiratory system, causing adverse health effects, such as heart disease, stroke, lung cancer, and asthma. NO2 is a toxic gas that can irritate the lungs and

PENERBIT PRESS UNIVERSITI TEKNOLOGI MARA

lower resistance to respiratory infections, while CO, a greenhouse gas, can cause global warming and climate change. O3, also known as ground-level ozone, can cause respiratory problems and harm crops and other vegetation. The AQI considers these components and provides an overall rating of air quality.

It is essential to understand the impact of each AQI component on the air quality to address the air pollution issue effectively. Identifying the primary factors that contribute to AQI levels can help authorities take appropriate actions to mitigate air pollution. Moreover, AQI prediction is vital to provide timely information to the public and stakeholders to make informed decisions, such as avoiding outdoor activities during high AQI days. Predicting AQI is a challenging task due to the complex relationship between AQI components and other environmental factors[7]. Machine learning approaches can help address this challenge by analyzing large datasets to identify patterns and relationships between AQI components and other environmental factors. This research aims to evaluate the effectiveness of various machine learning algorithms to predict AQI levels in Malaysia.

The implications of this research are significant, as it provides insights into the use of machine learning approaches to predict AQI levels. Accurate AQI prediction can help authorities take timely and appropriate measures to mitigate air pollution, such as reducing emissions from industries and enforcing vehicle emission standards. This research can also inform policymakers on the importance of taking proactive measures to manage air pollution in the country.

2. Literature Review

Recent advancements in machine learning have shown promising progress in the prediction and control of air quality issues. However, despite the progress made in air quality prediction using machine learning, limited research has been conducted on predicting the Air Quality Index (AQI) in Malaysia using these methods. This is evidenced by a search of the Web of Science (WOS) with 23 evidences and Scopus databases with 14 results with “Air pollution” AND “machine learning” AND “Malaysia” as the search keywords. Several recent studies conducted in different countries have explored the topic of air pollution and AQI. These studies can be summarized as follows.

The study in [8] examines the impact of COVID-19 lockdown on air pollution by comparing pollutants' concentrations before and during lockdown. The study uses hierarchical cluster analysis to highlight the differences in pollutant clusters and finds that particulate matter significantly decreased during lockdown. Additionally, the study predicts the air quality index using various machine learning algorithms based on pollutant concentration forecasts. In [9], The research findings suggest that there are various factors that affect the air quality index (AQI), including meteorological factors and socioeconomic indicators. These factors can be used to predict the monthly AQI of Taiwan and the daily and hourly AQI of specific cities. Deep learning was found to be more effective than machine learning in predicting the hourly AQI. The proposed work in [10] aims to improve air pollution estimation using a deep learning model to classify a region as urban or rural, which is critical in estimating the AQI. Based on Random Forest, the study highlights the importance of considering regional characteristics in air pollution estimation and suggests the use of satellite images for analysis in regions with limited installation of air quality machines. Researchers in [11] examines the use of two machine learning algorithms, neural networks and support vector machines, for predicting air quality index (AQI) based on datasets downloaded from the Central Pollution Control Board. The proposed model is applied to predict AQI in Delhi. A more interesting work is reported in [12] that proposes enhanced machine learning models for predicting air quality to reduce health problems related to air pollution. The models use datasets from different domains and apply k-Nearest Neighbors, XGBoost, Support Vector Machine, and Decision Tree models. The authors optimize hyperparameters for each model and find that XGBoost produces the best results with an error rate of 1.6, outperforming other models.

To highlight some important studies for Malaysia context including[13] that proposes an enhanced long short-term memory (ELSTM) model to investigate the association between cardiorespiratory hospitalization and air pollution and predict hospitalization based on air pollution. The study was conducted in seven study locations in Klang Valley, Malaysia. Researcher in [14] evaluated the performance of Bayesian Model Averaging (BMA) in predicting next-day PM10 concentration in Peninsular Malaysia. The study utilized 17 years' worth of air quality monitoring data from nine monitoring stations in Peninsular Malaysia, and eight air quality parameters. In [15], the researchers aimed to undertake a spatial hazard assessment of the air quality index using particulate matter with a diameter of 10μm or lesser (PM10) in Selangor, Malaysia, by developing four machine

learning models: eXtreme Gradient Boosting (XGBoost), random forest (RF), K-nearest neighbour (KNN), and Naive Bayes (NB). The method in [15] involved applying tree different machine learning algorithms (decision tree, boosted regression tree, and random forest) to predict PM10 concentrations in Kota Bharu, Kelantan. The data used for the study was the maximum daily data from January 2002 to December 2017. Different in [16] that implemented machine learning algorithms to predict particulate matter (PM2.5) air pollution in smart cities of Malaysia. The study used the Malaysia Air Pollution dataset and tested two machine learning algorithms; Multi-Layer Perceptron (MLP) and Random Forest.

Given the increasing global concern surrounding air pollution and the effectiveness of machine learning techniques in predicting and mitigating its effects, it is crucial to explore this issue further using a variety of machine learning methods. However, despite the growing body of research on this topic, there remains a significant lack of exploration of machine learning approaches in the context of air pollution in Malaysia.

3. Methodology

3.1 The dataset

The dataset was a collection of 125 Air Quality Map provided by eLichens technology from https://map.elichens.com ranging from the Malaysia north to the southern regions consists of AQI, Particulate Matter 2.5 (PM2.5), Nitrogen Dioxide (NO2), Carbon Dioxide (CO), and Ozone (O3). Figure 1 shows the main interface of the eLichens webpage and some of the map with different color that representing the level or the air quality. Based on the standard AQI level, the category of each air quality component can be set to each of the data as seen in Figure 2, based on Python codes.

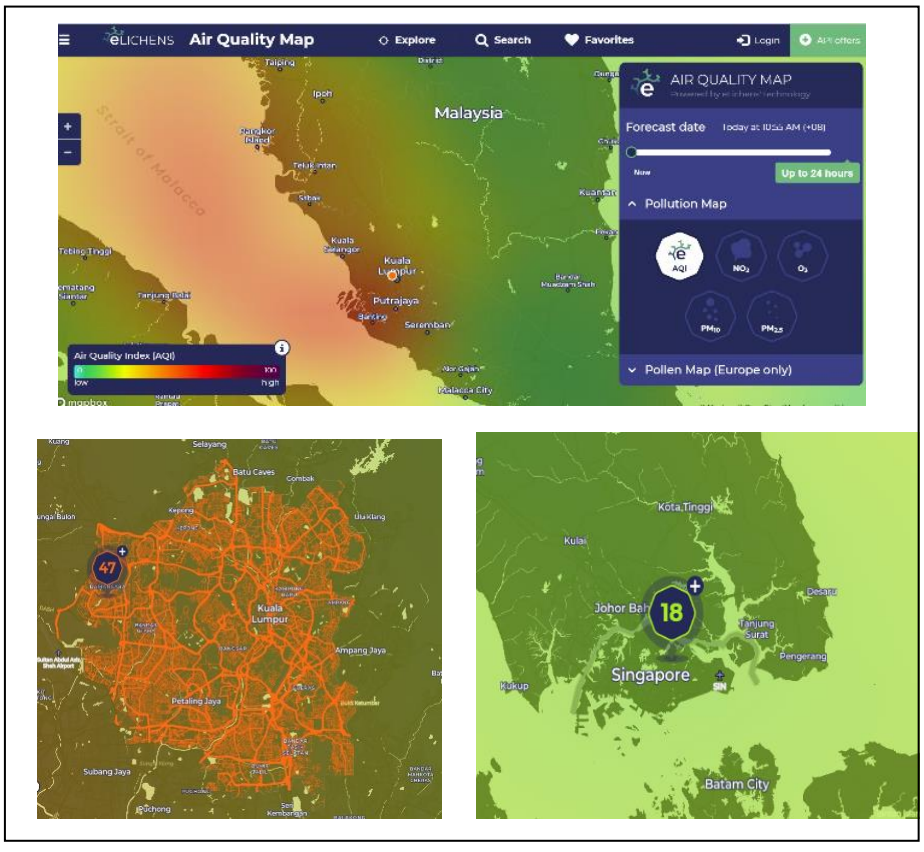

Figure 1. Air Quality Map by eLichens

```
df[df['Country']=='Malaysia']
```

[16]:

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
271	Malaysia	Ayer Keroh	59	Moderate	1	Good	34	Good	0	Good	59	Moderate
280	Malaysia	Bandar Penggaram	76	Moderate	1	Good	39	Good	0	Good	76	Moderate
287	Malaysia	Batu Gajah	92	Moderate	2	Good	57	Moderate	1	Good	92	Moderate
576	Malaysia	Limbang	95	Moderate	9	Good	95	Moderate	0	Good	46	Good
741	Malaysia	Masjid Tanah	79	Moderate	2	Good	36	Good	0	Good	79	Moderate
...	...	...	...	...	...	...	...	...	...	...	...	...
23104	Malaysia	Kudat	40	Good	1	Good	23	Good	0	Good	40	Good
23185	Malaysia	Taiping	131	Unhealthy for Sensitive Groups	3	Good	24	Good	10	Good	131	Unhealthy for Sensitive Groups
23301	Malaysia	Teluk Intan	90	Moderate	2	Good	57	Moderate	1	Good	90	Moderate
23322	Malaysia	Rawang	163	Unhealthy	6	Good	8	Good	25	Good	163	Unhealthy
23462	Malaysia	Marang	70	Moderate	1	Good	38	Good	0	Good	70	Moderate

125 rows × 12 columns

Figure 2. Some samples data

The formula to categorize the general AQI and it sub-AQIs can be referred to the US EPA's Air Quality Index (AQI) at https://www.airnow.gov/aqi/aqi-basics/. The general AQI is calculated based on the highest concentration among the AQI components. For example, if the PM2.5 level is 50 and the O3 level is 70, the AQI would be reported as 70 because it is the higher value of the two pollutants. The AQI is reported on a scale of 0 to 500, where higher values indicate greater levels of air pollution and increased health risks associated with breathing the air.

3.2 Machine learning algorithms

Based on AutoModel RapidMiner, three best machine learning algorithms out of the six suggested were selected to be compared namely Generalized Linear Model, Decision Tree and Support Vector Machine. The remaining three algorithms, namely Deep Learning, Random Forest and Gradient Boosted Trees exhibited less precise with higher prediction error as depicted in Figure 3. The Maximal Depth for Decision Tree is 10 to achieve the lowest relative error rate (5.6%) as seen in Figure 4. The optimal hyper-parameters for Support Vector Machine to achieve the lowest error rate (1.7%) can be seen in Figure 5, where Kernel Gamma is 0.005 and C is Y is 1000 (plotted as X and Y respectively).

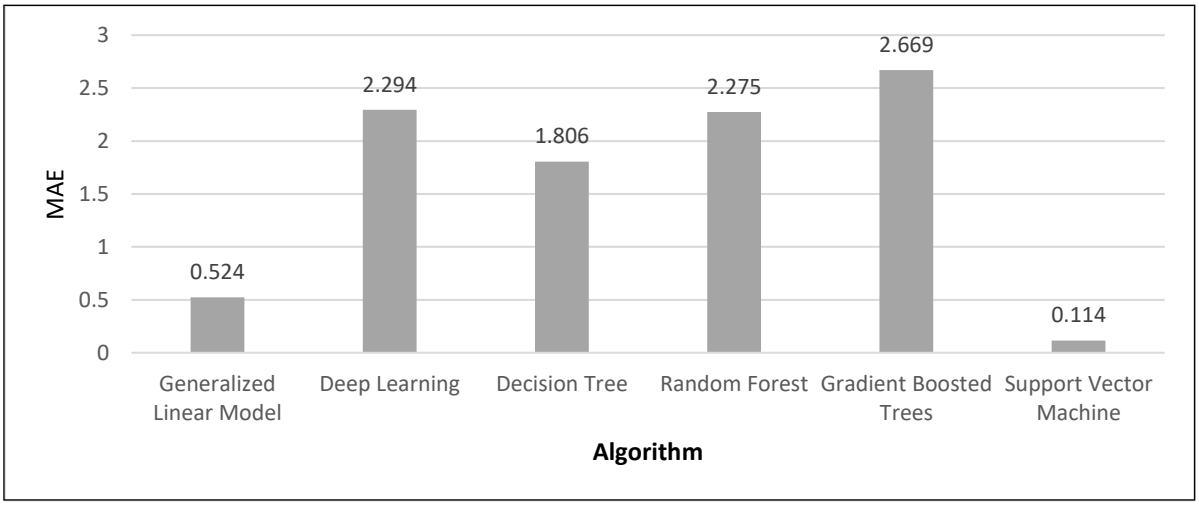

Figure 3. Preliminary Accuracy Results

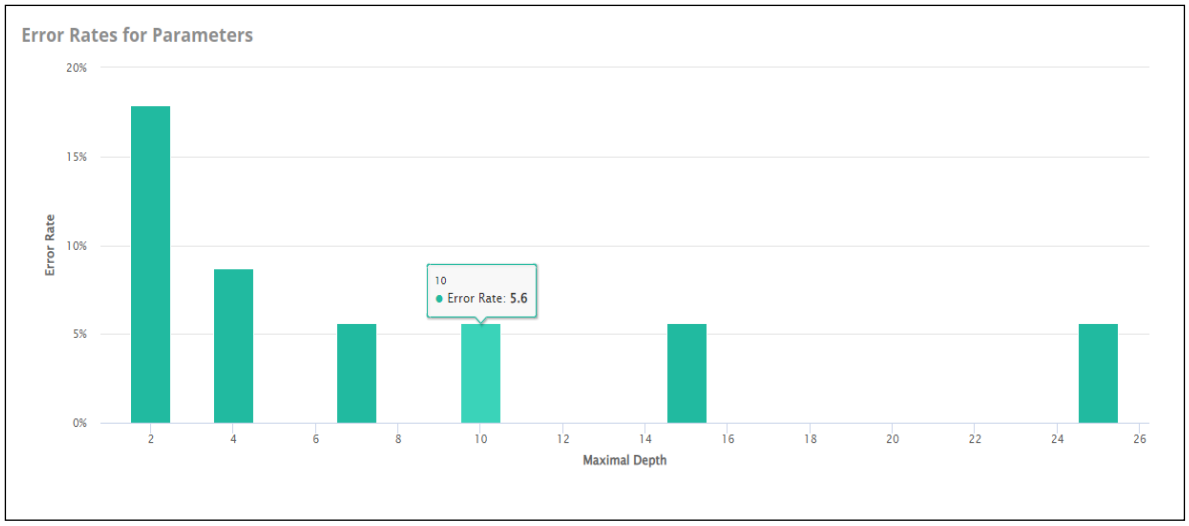

Figure 4. Different error rates of Decision Tree by different maximal depth

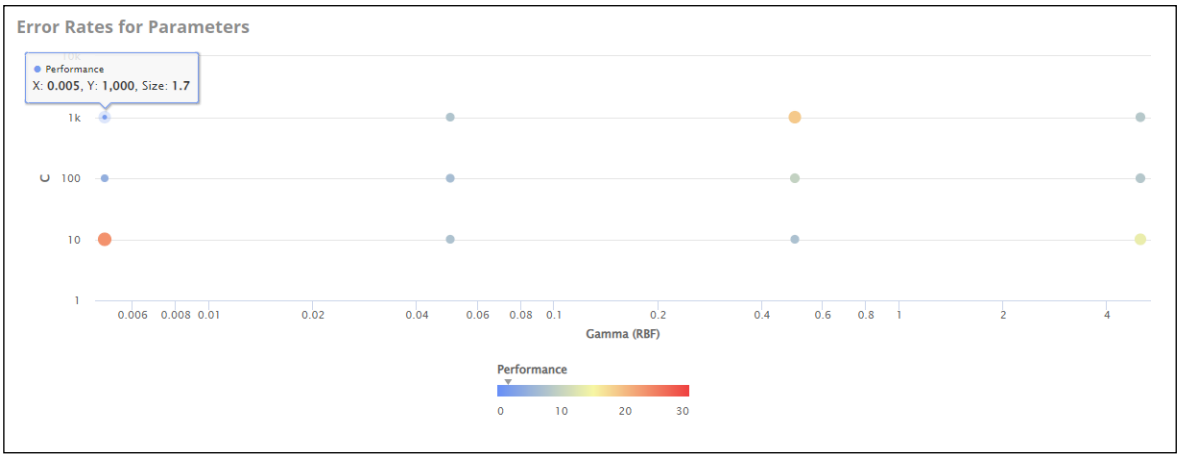

Figure 5. Different error rates of Support Vector Machine by different Kernal Gamma(X) and C(Y)

3.3 Evaluation

The dataset was split into two sets, training and testing, using a 60:40 ratio. The training set contained 75 records, which were used to train the machine learning algorithms. The remaining 50 records were allocated to the testing set, which was used to evaluate the performance of the trained models. By splitting the dataset in this way, the model can be trained on a subset of the data, and the performance can be tested on the remaining data.

To measure the accuracy of the model, Root Mean Square Error (RMSE) and Absolute Error and Relative Error were used. Additionally, R-squared (R^2) was used as an additional metric to indicate the total importance of the features in the AQI prediction models. By using these metrics, a comprehensive assessment of the model's accuracy and feature importance can be obtained.

4. Results and Discussion

The first part of the analysis presents the performance results of the machine learning models in predicting AQI, which are summarized in Table 1. In the second part of the analysis, the importance of AQI components in each machine learning algorithm will be described after exploring the dataset distributions and correlations.

Table 1. The performance results

Algorithm	R^	RMSE	Absolute Error	Relative Error (%)	TTC (s)
Generalized Linear Model	0.999	0.79	0.52	0.7	0.068
Decision Tree	0.996	2.44	0.81	2.3	0.039
Support Vector Machine	1	0.19	0.14	0.1	2

As listed in Table 1, Decision Tee has the lowest performance among the three algorithms with the highest accuracy errors (RMSE 2.44, Absolute error 0.81, Relative error 0.81%), but it has the shortest time to complete, which is 39ms. On the other hand, the Support Vector Machine has the highest accuracy among the three algorithms, but it took 2000ms to complete. The Generalized Linear Model falls in between for both accuracy and time to complete performances.

All the algorithms have achieved very good R-squared results at above 90%. Thus, it is important to understand the features importance of the AQI prediction model. Figure 6 displays the parametric distribution of the dataset, highlighting the PM2.5 concentration with the highest likelihood of being used to determine the overall AQI values. Figure 7 illustrates that Malaysia's air pollution status is generally categorized as "Unhealthy for Sensitive Groups" based on the AQI, with particular emphasis on PM2.5. However, the level of NO2 pushes the category further into the "Unhealthy" range. Despite NO2 having the worst AQI, the data distribution is not as extensive as PM2.5. Carbon Dioxide (CO) has the smallest amount of data and is distributed sparsely across all AQI categories. The amount of data for Ozone AQI is less than PM2.5, and it is mainly categorized as Unhealthy for Sensitive Groups. Figure 8 illustrates that the only component that has a linear relationship with the general AQI is PM2.5, and therefore, it is considered as the most important feature for all the machine learning algorithms in the AQI prediction models. Figure 9 compares the weights of correlations each component AQI value to the general AQI in the different machine learning algorithms. As expected, PM2.5 is the weight of correlation from Pearson Correlation test in AutoModel RapidMiner. Despite having a larger amount of data than CO and NO2 beyond machine learning (refer Figure 6), Ozone exhibits the lowest correlation coefficient among all the tested machine learning algorithms (refer Figure 9).

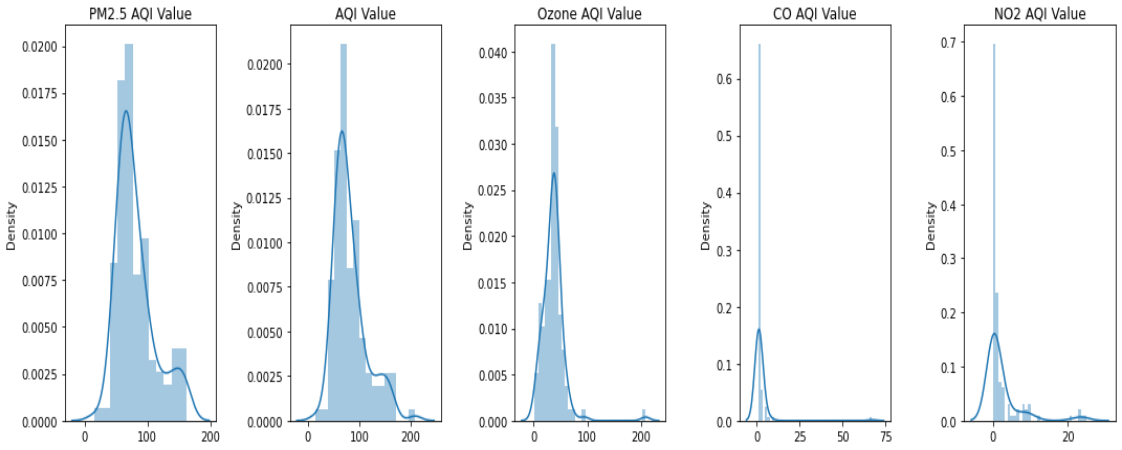

Figure 6. Parametric distribution of AQIs

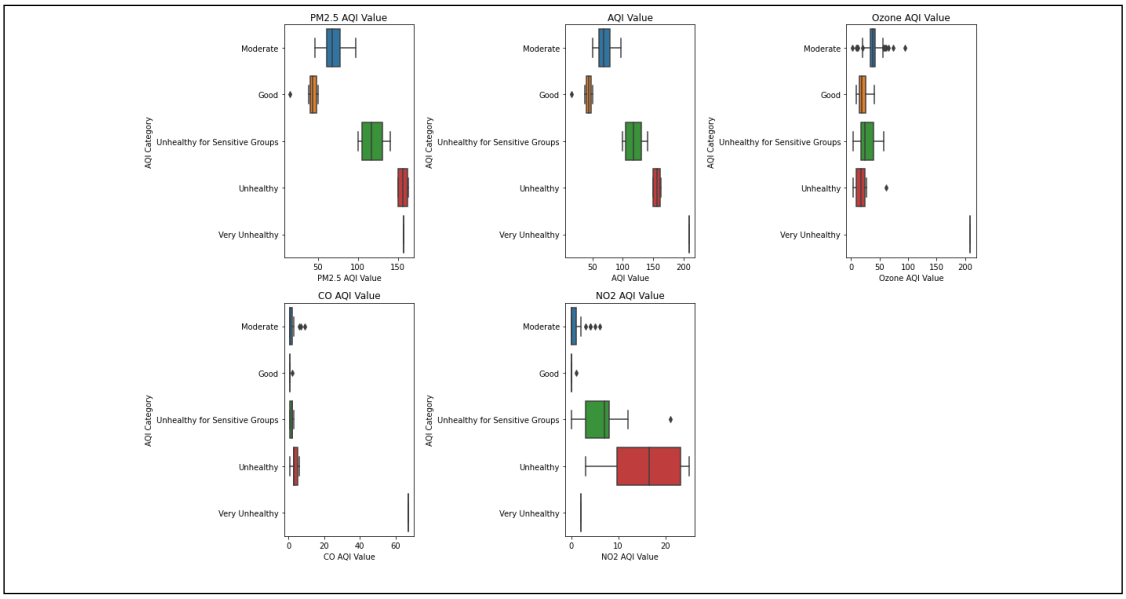

Figure 7. AQI category for each air pollution component

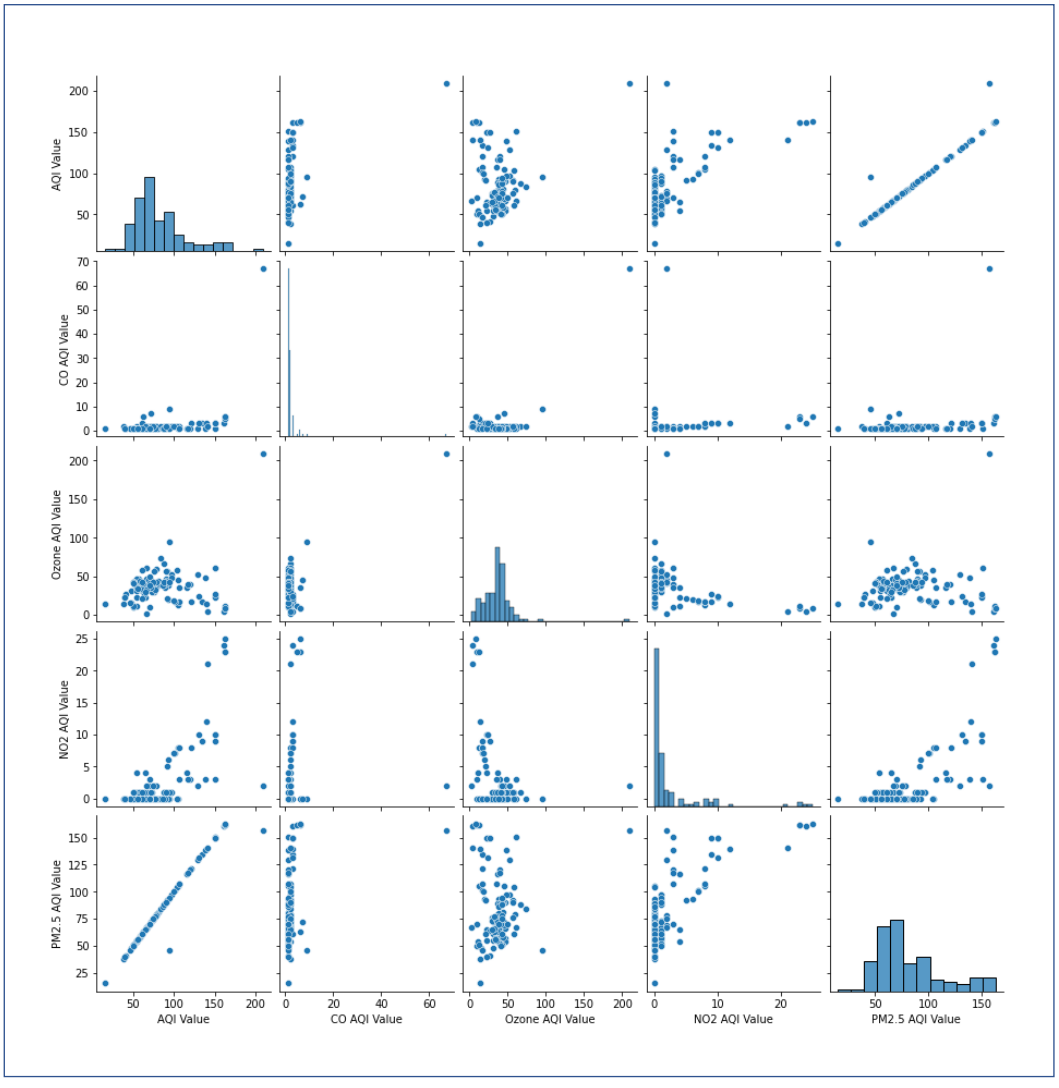

Figure 8. Relationship among the AQI components

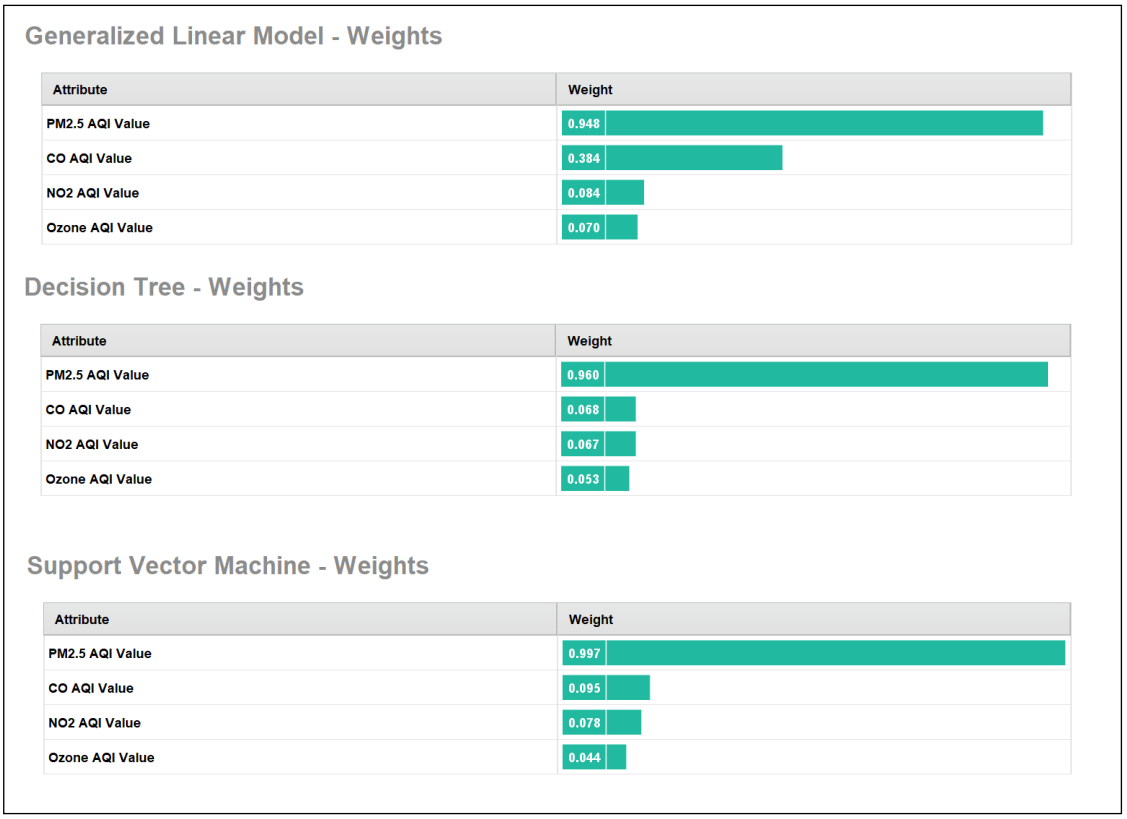

Figure 9. Comparison of weights between correlation coefficient

5. Conclusion

In conclusion, this research has explored the Air Quality Index components and evaluated the performance of machine learning algorithms for predicting the AQI. The benefits of this research are significant, as accurate prediction of the AQI can help to inform decision-making related to air quality and public health. Machine learning algorithms can provide a more efficient and effective way to analyze and predict air quality data, which can lead to better decision-making by policymakers and stakeholders. Additionally, understanding the relationship between AQI components can help to identify the sources of pollution and develop targeted strategies to improve air quality. In addition to evaluating machine learning algorithms for predicting AQI in Malaysia, this research has highlighted the importance of PM2.5 as the major contributor to AQI. The findings suggest that PM2.5 should receive more attention and efforts should be made to control its emissions in order to improve air quality in the country. By identifying the most important AQI components and evaluating machine learning algorithms for prediction, this research can provide valuable insights for policy makers, researchers, and other stakeholders in improving air quality management strategies.

Acknowledgements

The authors gratefully acknowledge the Universiti Teknologi Petronas for the support of this research under the International Collaborative Research Fund, application Number: RG2022-1393(cost center: 015ME0-321).

Conflict of Interest

The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] R. Fuller *et al.*, “Pollution and health: a progress update,” *Lancet Planet. Heal.*, 2022.

[2] S. Egbetokun, E. Osabuohien, T. Akinbobola, O. T. Onanuga, O. Gershon, and V. Okafor, “Environmental pollution, economic growth and institutional quality: exploring the nexus in Nigeria,” *Manag. Environ. Qual. An Int. J.*, vol. 31, no. 1, pp. 18–31, 2020.

[3] N. Egerstrom *et al.*, “Health and economic benefits of meeting WHO air quality guidelines, Western Pacific Region,” *Bull. World Health Organ.*, vol. 101, no. 2, pp. 130–139, 2023.

[4] IQAir AirVisual, “World Air Quality ReportRegion and City PM2.5 Ranking. USA,” 2019.

[5] Z. Deng *et al.*, “Mining biomarkers from routine laboratory tests in clinical records associated with air pollution health risk assessment,” *Environ. Res.*, vol. 216, p. 114639, 2023.

[6] Y. Huang, Y. Wang, T. Zhang, P. Wang, L. Huang, and Y. Guo, “Exploring health effects under specific causes of mortality based on 90 definitions of PM2. 5 and cold spell combined exposure in Shanghai, China,” *Environ. Sci. Technol.*, vol. 57, no. 6, pp. 2423–2434, 2023.

[7] P. Kumar, “A critical evaluation of air quality index models (1960--2021),” *Environ. Monit. Assess.*, vol. 194, no. 5, p. 324, 2022.

[8] R. Sen, A. K. Mandal, S. Goswami, and B. Chakraborty, “Prediction of Particulate Matter (PM2. 5) Across India Using Machine Learning Methods,” in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2*, 2023, pp. 545–556.

[9] C.-H. Wang and C.-R. Chang, “Forecasting air quality index considering socio-economic indicators and meteorological factors: A data granularity perspective,” *J. Forecast.*.

[10] S. K. Bamrah, S. Srivatsan, and K. S. Gayathri, “Region Classification for Air Quality Estimation Using Deep Learning and Machine Learning Approach,” in *Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021*, 2023, pp. 333–344.

[11] N. N. Maltare and S. Vahora, “Air Quality Index prediction using machine learning for Ahmedabad city,” *Digit. Chem. Eng.*, p. 100093, 2023.

[12] K. Ravindra *et al.*, “Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections,” *Sci. Total Environ.*, vol. 858, p. 159509, 2023.

[13] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, M. Marjani, R. Shaharudin, and M. T. Latif, “Air pollution and cardiorespiratory hospitalization, predictive modeling, and analysis using artificial intelligence techniques,” *Environ. Sci. Pollut. Res.*, vol. 28, no. 40, pp. 56759–56771, 2021.

[14] N. Ramli *et al.*, “Performance of Bayesian Model Averaging (BMA) for Short-Term Prediction of PM10 Concentration in the Peninsular Malaysia,” *Atmosphere (Basel).*, vol. 14, no. 2, p. 311, 2023.

[15] A. Tella and A.-L. Balogun, “GIS-based air quality modelling: Spatial prediction of PM10 for Selangor State, Malaysia using machine learning algorithms,” *Environ. Sci. Pollut. Res.*, pp. 1–17, 2021.

[16] N. Palanichamy, S.-C. Haw, S. Subramanian, R. Murugan, K. Govindasamy, and others, “Machine learning methods to predict particulate matter PM 2.5,” *F1000Research*, vol. 11, no. 406, p. 406, 2022.

Biography of all authors

Picture	Biography	Authorship contribution

	**Ts. Abdullah Sani Abd Rahman** obtained his first degree in Informatique majoring in Industrial Systems from the University of La Rochelle, France in 1995. He received a master's degree from Universiti Putra Malaysia in Computer Science, with specialization in Distributed Computing. Currently, he is a lecturer at the Universiti Teknologi Petronas, Malaysia and a member of the Institute of Autonomous System at the same university. His research interests are cybersecurity, data analytics and machine learning. He is also a registered Professional Technologist. He can be contacted at email: sani.arahman@utp.edu.my.	Design the research work, evaluation
	Dr. Aizal Yusrina Idris is currently a lecturer in Yanbu Industrial College, Kingdom of Saudi Arabia. She holds PhD in Education Training, in University Malaya, Malaysia. She has a Master in Information Technology (Computer Science) from the Universiti Kebangsaan Malaysia, Malaysia, and Bachelor of Information Technology (Hon.) (Industrial Computing) from the same university. Her research interest is the use of IT in teaching and learning, as well as in teacher's training, and system's evaluation methods in education. She can be contacted at idrisa@rcyci.edu.sa, aizalyusrina@gmail.com	Drafting article, camera ready
	Suhaimi Abdul Rahman is currently an instructor and Planning Specialist in Yanbu Technical Institute, Kingdom of Saudi Arabia. He has a Master in Information Technology (Computer Science) from the Universiti Kebangsaan Malaysia, Malaysia, and Bachelor of Computer Science (Hon.) from Universiti Sains Malaysia. His research interest is networking specifically in network mobility, and e-learning applications. He can be contacted at abulrahmans@rcyci.edu.sa, suhaimi.ar@gmail.com	Literature review, data collection