Pertanika J. Sci. & Technol. 31 (1): 123 - 135 (2023)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Prediction of Daily Air Pollutants Concentration and Air Pollutant Index Using Machine Learning Approach

Nurul A'isyah Mustakim[1], Ahmad Zia Ul-Saufie[1]*, Wan Nur Shaziayani[2], Norazian Mohamad Noor[3] and Sofianita Mutalib[1]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Selangor, Malaysia

[2]Department of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Pulau Pinang, Permatang Pauh Campus, 13500 UiTM, Permatang Pauh, Penang, Malaysia

[3]Faculty of Civil Engineering Technology, Universiti Malaysia Perlis, Kompleks Pengajian Jejawi 3, 02600 UniMAP, Arau, Perlis, Malaysia

ABSTRACT

The major air pollutants in Malaysia that contribute to air pollution are carbon monoxide, sulfur dioxide, nitrogen dioxide, ozone, and particulate matter. Predicting the air pollutants concentration can help the government to monitor air quality and provide awareness to the public. Therefore, this study aims to overcome the problem by predicting the air pollutants concentration for the next day. This study focuses on an industrial, the Petaling Jaya monitoring station in Selangor. The data is obtained from the Department of Environment, which contains the dataset from 2004 to 2018. Subsequently, this study is conducted to construct predictive modeling that can predict the air pollutants concentrations for the next day using a tree-based approach. From the comparison of the three models, a random forest is a best-proposed model. The results of PM_{10} concentration prediction for the random forest is the best performance which is shown by RMSE (15.7611–19.0153), NAE (0.6508–0.8216), and R^2 (0.346–0.5911). For SO_2, the RMSE was 0.0016–0.0017, the NAE was 0.7056–0.8052, and the R^2 was 0.3219–0.4676. The RMSE (0.0062–0.0075), the NAE (0.7892–0.9591), and the R^2 (0.0814–0.3609) for NO_2. The RMSE (0.3438–0.3975), NAE (0.7387–0.9015), and R^2 (0.2005–0.4399) for CO were all within acceptable limits. For O_3, the RMSE

ARTICLE INFO

Article history:
Received: 26 February 2022
Accepted: 05 July 2022
Published: 19 August 2022

DOI: https://doi.org/10.47836/pjst.31.1.08

E-mail addresses:
nurulaisyahmustakim@gmail.com (Nurul A'isyah Mustakim)
ahmadzia101@uitm.edu.my (Ahmad Zia Ul-Saufie)
shaziayani@uitm.edu.my (Wan Nur Shaziayani)
norazian@unimap.edu.my (Norazian Mohamad Noor)
sofi@fskm.uitm.edu.my (Sofianita Mutalib)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

was 0.0051–0.0057, the NAE was 0.8386–0.9263, and the R^2 was 0.1379–0.2953. The API calculation results indicate that PM_{10} is a significant pollutant in representing the API.

Keywords: data mining, decision tree, gradient boosted trees, Modeling, PM_{10}, random forest

INTRODUCTION

The Malaysian Department of Environment (DOE) plays a vital role by providing real-time Air Pollutant Index (API) readings on the Air Pollutant Index Management System (APIMS) website. It can help people to know the actual situation of air quality status. However, people can only find out the current and past API readings. People need to cancel plans for outdoor activities immediately if the air quality status is unhealthy or worse. It would be better if people could find the air quality status for the next few days, like the weather forecast (Ul-Saufie et al., 2012). Therefore, this study will predict the API for the next day.

Furthermore, there are many studies on the prediction of air pollutants concentration. However, most studies only focus on predicting one pollutant or a few pollutants only such as (Hamid et al., 2017; Shaadan et al., 2019; Alias et al., 2021; Shaziayani et al., 2021), and there is limited study on predicting all pollutants concentration in a study. Thus, this study will predict the concentration of all major air pollutants: ozone, carbon monoxide, nitrogen dioxide, sulfur dioxide, and particulate matter less than 10 micrometers (PM_{10}). Besides, there is limited study on predicting API since most researchers only predict the concentration of air pollutants. Therefore, this study has continued to predict API after predicting air pollution concentrations.

Additionally, most researchers use artificial neural networks (ANN) to predict the concentration of air pollutants. The researchers also always predict only one or a few air pollutants in a study, as shown in Table 1. In Malaysia, there are limited studies on the prediction of air pollution concentrations using a tree-based approach and limited studies to predict all major air pollution concentrations. Furthermore, the tree-based approach is suitable to be applied to air pollution data, given that the data is not normally distributed, which can be handled using this method. However, a haze event will result in extreme air pollutants concentration distribution values. The extreme values will affect the normality of the distribution. Therefore, a method with a non-normality assumption is needed to predict the concentration.

Therefore, decision trees, random forests, and gradient boosted trees were used in this study to predict the air pollution concentrations for the next day. Thus, tree-based modeling is chosen to predict the air pollutants concentration because the models are not sensitive and affected by extreme values or outliers (Hu et al., 2018). The study of predicting air pollutants concentration using tree-based models is also limited, given that most researchers focus on regression-based models and time series.

Table 1
The summary of methods used in previous studies

Authors	MLR	ANN	DT	RF	GBT
Thomas & Jacko (2007)	√	√			
Cai et al. (2009)		√			
Moustris et al. (2010)		√			
Arhami et al. (2013)		√			
Rahman et al. (2013)		√			
Sekar et al. (2016)		√	√		
Moazami et al. (2016)		√			
Masih (2019)			√	√	
Qadeer & Jeon (2019)					√
Watson et al. (2019)	√	√		√	√
Alpan & Sekeroglu (2020)			√	√	
Shams et al. (2020)	√	√			
Lu et al. (2021)				√	
Shaziayani et al. (2021)					√

*Abbreviations of the methods: MLR: Multiple Linear Regression, ANN: Artificial Neural Networks, DT: Decision Tree, RF: Random Forest, GBT: Gradient Boosted Trees.

MATERIALS AND METHODS

The secondary data is obtained from the Department of Environment (DOE), which contains the data from 2004 to 2019 for the Petaling Jaya air monitoring station (CA0016). Petaling Jaya is an industrial area, and the station is located at Sri Petaling Primary School (N03° 06.612, E101° 42.274'). The number of observations for the dataset is 5763. In addition, the data for each variable is the daily average data to predict the next day's air pollutant concentration. Therefore, each variable in the dataset is very important to understand. It is to get an overview of the data that will potentially be useful for moving on to steps in the data analysis process. At this stage, the data would be evaluated in every way, such as quality, accuracy, and the representative of the data. The variables included in the study for predicting the air pollutants concentration in Petaling Jaya are carbon monoxide (CO), particulate matter 10 microns or less in diameter (PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), temperature, and relative humidity. Dedovic et al. (2016) mentioned that temperature is an important meteorological parameter in the formation of secondary PM_{10}. The relative humidity is also a strong predictor of PM_{10} concentration and was used to replace the rainfall parameter because DOE does not measure the data. At the same time, $PM_{2.5}$ is not considered in the prediction because $PM_{2.5}$ monitoring in Malaysia started in mid-2017, while the report on $PM_{2.5}$ was only available in 2018. Table 2 shows the summary of sample size for the dataset. The numbers of missing values for all variables are below 0.6%.

Table 2
Summary of dataset sample size

Variables	Number of Samples	Non-Missing Values	Missing Values	Percentage of Missing Values
O_3	5398	5372	26	0.48%
CO	5398	5371	27	0.50%
NO_2	5398	5369	29	0.54%
SO_2	5398	5366	32	0.59%
PM_{10}	5398	5397	1	0.02%
Temperature	5398	5397	1	0.02%
Relative Humidity	5398	5397	1	0.02%

Model Development

Three models were used in predicting the air pollution concentration. The models are Decision Tree (DT), Random Forest (RF), and Gradient Boosted Trees (GBT). Besides, in the process of model development, model prediction, and model evaluation, RapidMiner Studio was used to predict the air pollution concentration.

Decision Tree (DT) is a popular machine learning algorithm that can solve the problem by transforming the data into a tree representation. DT algorithms can be used to solve both regression and classification problems. In addition, DT can work well with both nominal and numeric data types. The structure of DT consists of nodes, edges or branches and leaf nodes. The nodes represent a test for the values on a certain attribute. Each edge or branch represents a test's outcome and connects to the next node or leaf. Each leaf node (or terminal node) holds a class label that predicts the outcome. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. DT algorithm splits data into subsets based on an attribute value. The process continues for each consequent subset until the target is found.

Random forest (RF) determines variable importance by randomly permuting (shuffling) a given variable. In this way, the variable should have no relationship with the response. RFs are one of the methods for assembling a collection (or forest) of decision trees with the bagging technique. It trains many trees concurrently and uses the majority judgment developed in DTs as the RF model's final decision. The difference in accuracy in the random forest using the original data and the random forest predictions using the shuffled variable is then calculated. Next, a single variable importance measure is computed as the average of these differences across every tree in the forest (Breiman et al., 2002). Finally, a single variable importance measure is computed as the average of these differences across every tree in the forest.

Boosted regression tree models are developed by integrating two algorithms; Decision trees are used as the main methods for classifying the datasets through a supervised method,

and then boosting is used to aggregate their outputs to obtain the total prediction. Common boosting algorithms applied are Adaboost and gradient boosting. The gradient boosting trees (GBT) method is slightly different from Adaboost. Instead of using the weighted average of individual outputs as the final outputs, GBT uses a loss function to minimize loss (as an optimization function) and converge upon a final output value. Moreover, gradient boosting uses short, less-complex decision trees instead of decision stumps. A larger number of gradient boosting iterations reduces training set errors. However, raising the number of gradients boosting iterations too high will increase the overfitting. So, monitoring the error of prediction from a distinct validation data set can help choose the optimal value for the number of gradients boosting iterations.

The year 2004 to 2018 will be divided into two parts which are 70% of the data for model training and another 30% for model testing, as suggested by Arabameri et al. (2019). Then the results of each model were compared to find the best-proposed model for predicting each air pollutant. Finally, the best-proposed model was applied to the dataset for the year 2019 to calculate the air pollutant index.

Table 3 shows the general model of DT, RF and GBT in predicting air pollutants concentration. The predicted value is represented by δ while the observed value is represented by D. Each method (DT, RF, and GBT) is used to predict the next day's concentrations for each pollutant with CO, PM_{10}, NO_2, SO_2, O_3, T, and RH of the current day as predictors.

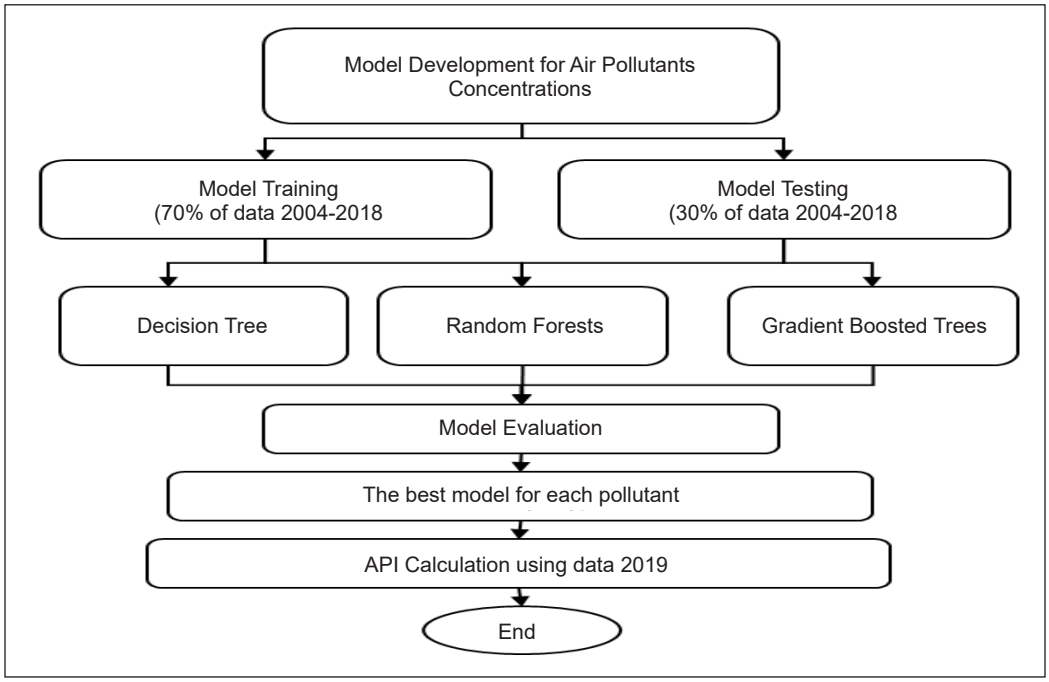

Figure 1. Model prediction workflow

Table 3
General model for the next day's prediction

Prediction	Model
Next day (δ+1)	CO (δ +1) ~ DT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	CO (δ +1) ~ RF [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	CO (δ +1) ~ GBT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
Next day (δ+1)	PM_{10} (δ +1) ~ DT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	PM_{10} (δ +1) ~ RF [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	PM_{10} (δ +1) ~ GBT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
Next day (δ+1)	NO_2 (δ +1) ~ DT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	NO_2 (δ +1) ~ RF [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	NO_2 (δ +1) ~ GBT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
Next day (δ+1)	SO_2 (δ +1) ~ DT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	SO_2 (δ +1) ~ RF [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	SO_2 (δ +1) ~ GBT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
Next day (δ+1)	O_3 (δ +1) ~ DT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	O_3 (δ +1) ~ RF [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]
	O_3 (δ +1) ~ GBT [$CO_{(D)}$, $PM_{10\,(D)}$, $NO_{2\,(D)}$, $SO_{2\,(D)}$, $O_{3\,(D)}$, $T_{(D)}$, $RH_{(D)}$]

Model Evaluation

The performance of the models is evaluated using statistical comparisons, which were root mean square error (RMSE), normalized absolute error (NAE), and squared correlation (R^2). The R^2 is used to assess the model's accuracy; values closer to 1 imply more accuracy. While to quantify a model's error, the RMSE and NAE are used; a value close to 0 indicates a minimal error. The models that provide the best prediction values were selected as the best-proposed models for predicting the concentration of air pollutants. Below is the Equations 1-3 for each indicator:

$$NAE = \frac{\sum_{i=1}^{n}|P_i - O_i|}{\sum_{i=1}^{n}|\bar{O} - O_i|} \quad (1)$$

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{\frac{1}{2}} \quad (2)$$

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}[(P_i - \bar{P})(O_i - \bar{O})]}{\sigma_P \sigma_O}\right]^2 \quad (3)$$

where,

$\boldsymbol{N}$ = Total number of sample size;

O_i = Observed values of i-th day
P_i = Predicted values of i-th day
$\overline{P}$ = Mean of the predicted values of one set of daily monitoring records
$\overline{O}$ = Mean of the observed values of one set of daily monitoring records
S_P = Standard deviation of the predicted values of one set of daily monitoring records
S_Q = Standard deviation of the observed values of one set of daily monitoring records

The results of RMSE, NAE, and R^2 for the three models were compared to determine the best model. First, the lowest value of RMSE and NAE was ranked as 1, the second lowest value was ranked as 2, and the highest was ranked as 3. For R^2, the method for ranking is the opposite of RMSE and NAE. The highest value of R^2 was ranked as 1, the second highest value was ranked as 2, and the lowest was ranked as 3. Then, the ranked values for RMSE, NAE, and R^2 were summed up to find the lowest total ranking values among the three models, which indicate the best (Shaziayani et al., 2021).

API Calculation

The first step to calculating the air pollution index (API) is the sub-index for each pollutant. The predicted values of air pollutant concentration were used to calculate the sub-index. Table 4 shows the formulas used by the Department of Environment (DOE) in calculating the sub-index. Let X represent predicted concentration and Y represent the index.

Then, the highest sub-index among the air pollutants was chosen as the API for the predicted day. Figure 2 shows the determination of API calculation. After calculating the sub-index for all pollutants, the sub-index results were compared, and the highest was chosen as a maximum index, API.

Table 4
Equations of sub-index calculation

Air Pollutants	Values of Predicted Concentration (X)	Formula of Sub-Index (Y)
Carbon Monoxide (CO)	X < 9 ppm	$Y = X \times 11.11111$
	9 < X < 15 ppm	$Y = 100 + [(X - 9) \times 16.66667]$
	15 < X < 30 ppm	$Y = 200 + [(X - 15) \times 6.66667]$
	X > 30 ppm	$Y = 300 + [(X - 30) \times 10]$
Ozone (O_3)	X < 0.2 ppm	$Y = X \times 1000$
	0.2 < X < 0.4 ppm	$Y = 200 + [(X - 0.2) \times 500]$
	X > 0.4 ppm	$Y = 300 + [(X - 0.4) \times 1000]$
Nitrogen Dioxide (NO_2)	X < 0.17 ppm	$Y = X \times 588.23529$
	0.17 < X < 0.6 ppm	$Y = 100 + [(X - 0.17) \times 232.56]$
	0.6 < X < 1.2 ppm	$Y = 200 + [(X - 0.6) \times 166.667]$
	X > 1.2 ppm	$Y = 300 + [(X - 1.2) \times 250]$

Table 4 *(continue)*

Air Pollutants	Values of Predicted Concentration (X)	Formula of Sub-Index (Y)
Sulfur Dioxide (SO_2)	X < 0.04 ppm	$Y = X \times 2500$
	0.04 < X < 0.3 ppm	$Y = 100 + [(X - 0.04) \times 384.61]$
	0.3 < X < 0.6 ppm	$Y = 200 + [(X - 0.3) \times 333.333]$
	X > 0.6 ppm	$Y = 300 + [(X - 0.6) \times 500]$
Particulate Matter (PM_{10})	X < 50 μg/m³	$Y = X$
	50 < X < 150 μg/m³	$Y = 50 + [(X - 50) \times 0.5]$
	150 < X < 350 μg/m³	$Y = 100 + [(X - 150) \times 0.5]$
	350 < X < 420 μg/m³	$Y = 200 + [(X - 350) \times 1.4286]$
	420 < X < 500 μg/m³	$Y = 300 + [(X - 420) \times 1.25]$
	X > 500 μg/m³	$Y = 400 + (X - 500)$

Source. Department of Environment (1997)

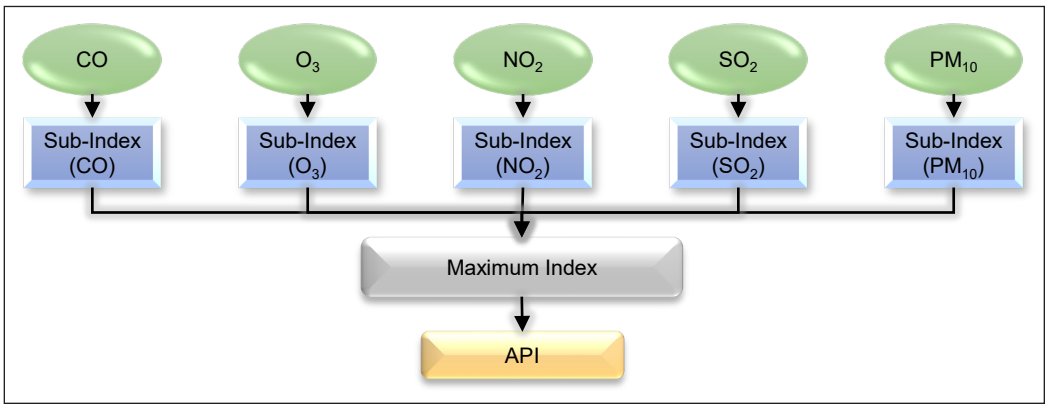

Figure 2. Diagram of API calculation
Source. Department of Environment (2017)

RESULTS AND DISCUSSION

The central tendency and dispersion measures are shown in Table 5 from 2004 to 2018. This research focuses on five major air pollutants at Petaling Jaya station, namely O_3, CO, NO_2, SO_2, and PM_{10}. From this study, PM_{10} is the one with the most unstable data as it has many outliers compared to others. O_3, CO, NO_2, and SO_2 also have outliers but not as much as PM_{10}. Besides, PM_{10} has the highest standard deviation, which means that PM_{10} concentrations are far from the mean of the set and spread over a wider range. Meanwhile, the standard deviation for O_3, CO, NO_2, and SO_2 are all less than one, which means the values tend to be close to the set mean.

The highest concentration of PM_{10} was recorded in Petaling Jaya on August 11, 2005. According to Shaharuddin and Noorazuan (2006), massive fires caused by agricultural activities in Kampar, Pelalawaan, Indragiri Hulu, and Bengkalis provinces, as well as

peat forest fires in Rokan Hulu and Rokan Hilir, were the main sources of PM_{10} emissions entering the Malaysian atmosphere from 10–12 August 2005.

Lastly, it reveals that all the variables are not normally distributed because the skewness is not equal to zero. Furthermore, besides the skewness results, it indicates that O_3, CO, SO_2, and PM_{10} are skewed right because the values are positive, while NO_2, temperature, and relative humidity are skewed left because the values are negative. Thus, the non-normal distribution data reinforces the reason it is necessary to use the tree-based approach because the approach does not require assumptions of normality (Elith et al., 2008).

Table 5
Summary of dataset statistic

Variables	Mean	Median	Standard Deviation	Skewness	Maximum
O_3	0.0142	0.0130	0.0063	0.6269	0.0420
CO	1.3201	1.2660	0.4664	1.1080	6.7730
NO_2	0.0285	0.0280	0.0079	-0.0313	0.0610
SO_2	0.0041	0.0040	0.0021	1.0787	0.0240
PM_{10}	48.1261	43.5830	24.9671	4.9070	482.2080
Temperature	27.0160	28.0279	5.3528	-4.1983	33.1370
Relative Humidity	71.6447	73.0000	9.4837	-2.2168	94.2071

Prediction Model

The results of RMSE, NAE, and R^2 in predicting air pollution concentrations using DT, RF, and GBT are shown in Table 6. First, the best-proposed model is chosen by choosing the lowest RMSE, lowest NAE, and highest R^2. Next, the RMSE, NAE, and R^2 were ranked to make it easier. The best performance was ranked 1, followed by 2 and 3. After that, the ranks were summed up to choose the best-proposed model. It is revealed that the random forest is the best-proposed model for predicting PM_{10} concentration because it has the lowest RMSE, NAE, and second highest R^2 in predicting the next day's air pollution concentration.

Next, the results also show that random forest is the best-proposed model for predicting SO_2, NO_2, O_3, and CO concentrations for the next day since it has the lowest error and highest accuracy, with a total of rank values of 4 for NO_2 and 3 for SO_2, O_3, and CO. These results are similar to Alpan and Sekeroglu (2020), where the RF model achieved the best results in predicting air pollution concentration compared to the other two models in their study.

Air Pollutant Index

The best-proposed model, the random forest, is applied to the dataset for the year 2019 to predict the air pollution concentration. As shown in Table 7, the predicted air pollution concentration is used to calculate the sub-index, and the maximum sub-index is chosen as

Table 6
Model evaluation of PM_{10} concentration prediction

Air pollutants	Model	Model Evaluation			Rank			
		RMSE	NAE	R^2	RMSE	NAE	R^2	Sum
PM_{10}	Decision tree	17.4991	0.7098	0.4790	2	2	3	7
	Random forest	**15.7611**	**0.6508**	0.5764	1	1	2	**4**
	Gradient boosted tree	19.2405	0.7895	**0.5911**	3	3	1	7
SO_2	Decision tree	0.0017	0.7642	0.3685	2	2	3	7
	Random forest	**0.0016**	**0.7056**	**0.4676**	1	1	1	**3**
	Gradient boosted tree	0.0018	0.8350	0.4620	3	3	2	8
NO_2	Decision tree	0.0072	0.9161	0.2507	3	3	3	9
	Random forest	**0.0062**	**0.7892**	0.3587	1	1	2	**4**
	Gradient boosted tree	0.0067	0.8610	**0.3609**	2	2	1	5
O_3	Decision tree	0.0056	0.9072	0.2161	3	3	3	9
	Random forest	**0.0051**	**0.8386**	**0.2953**	1	1	1	**3**
	Gradient boosted tree	0.0055	0.8963	0.2877	2	2	2	6
CO	Decision tree	0.4108	0.8567	0.2896	3	3	3	9
	Random forest	**0.3438**	**0.7387**	**0.4399**	1	1	1	**3**
	Gradient boosted tree	0.3901	0.8434	0.4243	2	2	2	6

Table 7
API calculation of next day

ID	Date	Sub-Index (PM_{10})	Sub-Index (SO_2)	Sub-Index (NO_2)	Sub-Index (O_3)	Sub-Index (CO)	API (Max Sub-Index)	Pollutant of Max Sub-Index
1	2019.01.01	31.6825	3.4266	13.5366	15.2291	13.1300	31.6825	PM_{10}
2	2019.01.02	30.7458	3.4560	13.6141	11.1102	13.9577	30.7458	PM_{10}
3	2019.01.03	35.2744	3.6125	13.9373	16.4702	12.3811	35.2744	PM_{10}
4	2019.01.04	35.9763	3.9619	13.8152	15.1340	13.4098	35.9763	PM_{10}
5	2019.01.05	38.3268	4.3707	14.9920	16.6031	13.2073	38.3268	PM_{10}
6	2019.01.06	35.7925	3.4983	13.3914	16.5355	13.0631	35.7925	PM_{10}
7	2019.01.07	31.0234	4.2816	12.3576	10.2682	13.0592	31.0234	PM_{10}
8	2019.01.08	35.9258	4.3158	15.2095	14.1415	14.8572	35.9258	PM_{10}
9	2019.01.09	35.6401	3.9545	13.4431	14.4477	14.8709	35.6401	PM_{10}
10	2019.01.10	31.2531	3.9623	12.6887	12.0756	13.1288	31.2531	PM_{10}

the next day's air pollution index (API). In this section, only 10 observations are shown from 365 days in 2019 as an example of results in calculating and selecting the API index. For instance, on January 1, 2019, the highest value for the sub-index was PM_{10} concentration. Therefore, PM_{10} was selected as the API for this date. Overall, the results showed that all APIs selected PM_{10} because the index of PM_{10} is the highest among the air pollutants.

CONCLUSION

The main finding showed that the random forest was chosen as the best-proposed model for predicting the concentrations of O_3, CO, NO_2, SO_2, and PM_{10} for next-day prediction. The random forest has the best performance and is compared using RMSE, NAE, and R^2. Then, the best random forest model is applied to the dataset for the year 2019. The predicted air pollutants concentrations for 2019 are used to calculate the sub-index. Then, the maximum sub-index was chosen as API. From the results of API calculation, all data of the maximum sub-index represent by PM_{10}. It indicates that PM_{10} is a significant pollutant in calculating the API.

The main contribution of this study is that, besides using this tree-based model to predict air pollution concentration, this study also converted air pollution prediction concentration to sub-index API. This procedure will help the local authorities to make predictions on air pollutants in Malaysia and as a tool for the early warning system.

ACKNOWLEDGEMENT

The Ministry of Science, Technology & Innovation (MOSTI) under Technology Development Fund 1(TDF04211363) funded this research. Thanks to the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, for their support and the Department of Environment Malaysia for providing air quality monitoring data.

REFERENCES

Alias, S. N., Hamid, N. Z. A., Saleh, S. H. M., & Bidin, B. (2021). Predicting carbon monoxide time series between different settlements area in Malaysia through chaotic approach. *Journal of Science and Mathematics Letters, 9*, 45-54. https://doi.org/https://doi.org/10.37134/jsml.vol9.sp.6.2021

Alpan, K., & Sekeroglu, B. (2020). Prediction of pollutant concentrations by meteorological data using machine learning algorithms. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *44*(4/W3), 21-27. https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-21-2020

Arabameri, A., Pradhan, B., & Lombardo, L. (2019). Comparative assessment using boosted regression trees, binary logistic regression, frequency ratio and numerical risk factor for gully erosion susceptibility modelling. *Catena*, *183*(6), Article 104223. https://doi.org/10.1016/j.catena.2019.104223

Arhami, M., Kamali, N., & Rajabi, M. M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environmental Science and Pollution Research*, *20*(7), 4777-4789. https://doi.org/10.1007/s11356-012-1451-6

Breiman, L., Culter, A., Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2*, 18-22.

Cai, M., Yin, Y., & Xie, M. (2009). Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transportation Research Part D: Transport and Environment*, *14*(1), 32-41. https://doi.org/10.1016/j.trd.2008.10.004

Dedovic, M. M., Avdakovic, S., Turkovic, I., Dautbasic, N., & Konjic, T. (2016). Forecasting PM10 concentrations using neural networks and system for improving air quality. In *2016 xi international symposium on telecommunications (bihtel)* (pp. 1-6). IEEE Publishing. https://doi.org/10.1109/BIHTEL.2016.7775721

Department of Environment. (1997). *A guide to air pollution index in Malaysia (API)*. Ministry of Science, Technology and the Environment. https://aqicn.org/images/aqi-scales/malaysia-api-guide.pdf

Department of Environment. (2017). *AIR pollutant index (API) calculation*. Ministry of Environment and Water. http://apims.doe.gov.my/public_v2/pdf/API_Calculation.pdf

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Hamid, H. A., Japeri, A. Z. U. S. M., & Ahmat, H. (2017). Characteristic and prediction of carbon monoxide concentration using time series analysis in selected urban area in Malaysia. In *MATEC Web of Conferences* (Vol. 103, p. 05001). EDP Sciences.. https://doi.org/10.1051/matecconf/201710305001

Hu, Y., Scavia, D., & Kerkez, B. (2018). Are all data useful? Inferring causality to predict flows across sewer and drainage systems using directed information and boosted regression trees. *Water Research, 145,* 697-706.

Lu, J., Zhang, Y., Chen, M., Wang, L., Zhao, S., Pu, X., & Chen, X. (2021). Estimation of monthly 1 km resolution PM2.5 concentrations using a random forest model over "2 + 26" cities, China. *Urban Climate*, *35*, Article 100734. https://doi.org/10.1016/j.uclim.2020.100734

Masih, A. (2019). Application of random forest algorithm to predict the atmospheric concentration of NO_2. In *2019 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)* (pp. 252-255). AIP Publishing LLC. https://doi.org/10.1063/1.5137947

Moazami, S., Noori, R., Amiri, B. J., Yeganeh, B., Partani, S., & Safavi, S. (2016). Reliable prediction of carbon monoxide using developed support vector machine. *Atmospheric Pollution Research*, *7*(3), 412-418. https://doi.org/10.1016/j.apr.2015.10.022

Moustris, K. P., Ziomas, I. C., & Paliatsos, A. G. (2010). 3-day-ahead forecasting of regional pollution index for the pollutants NO_2, CO, SO_2, and O_3 using artificial neural networks in athens, Greece. *Water, Air, and Soil Pollution*, *209*(1-4), 29-43. https://doi.org/10.1007/s11270-009-0179-5

Qadeer, K., & Jeon, M. (2019). Prediction of PM10 concentration in South Korea using gradient tree boosting models. In *PervasiveHealth: Pervasive Computing Technologies for Healthcare* (pp. 1-6). ACM Publishing. https://doi.org/10.1145/3387168.3387234

Rahman, N. H. A., Lee, M. H., Latif, M. T., & Suhartono, S. (2013). Forecasting of air pollution index with artificial neural network. *Jurnal Teknologi, 63*(2), 59-64. https://doi.org/10.11113/jt.v63.1913

Sekar, C., Gurjar, B. R., Ojha, C. S. P., & Goyal, M. K. (2016). Potential assessment of neural network and decision tree algorithms for forecasting ambient PM2.5 and CO concentrations: Case study. *Journal of Hazardous, Toxic, and Radioactive Waste*, *20*(4), 1-9. https://doi.org/10.1061/(asce)hz.2153-5515.0000276

Shaadan, N., Rusdi, M. S., Azmi, N. N. S. N. M., Talib, S. F., & Azmi, W. A. W. (2019). Time series model for Carbon Monoxide (CO) at several industrial sites in Peninsular Malaysia. *Malaysian Journal of Computing (MJoC), 4*(1), 246-260.

Shaharuddin, A., & Noorazuan, M. H. (2006). Kebakaran hutan dan isu pencemaran udara di Malaysia: Kes jerebu pada Ogos 2005 [Forest fires and air pollution issues in Malaysia: The case of haze on August 2005]. *UKM Journal Article Repository, 1*(1), 1-19.

Shams, S. R., Jahani, A., Moeinaddini, M., & Khorasani, N. (2020). Air carbon monoxide forecasting using an artificial neural network in comparison with multiple regression. *Modeling Earth Systems and Environment*, *6*(3), 1467-1475. https://doi.org/10.1007/s40808-020-00762-5

Shaziayani, W. N., Ul-Saufie, A. Z., Ahmat, H., & Al-Jumeily, D. (2021). Coupling of quantile regression into boosted regression trees (BRT) technique in forecasting emission model of PM10 concentration. *Air Quality, Atmosphere and Health, 14*, 1647-1663. https://doi.org/10.1007/s11869-021-01045-3

Thomas, S., & Jacko, R. B. (2007). Model for forecasting expressway fine particulate matter and carbon monoxide concentration: Application of regression and neural network models. *Journal of the Air and Waste Management Association*, *57*(4), 480-488. https://doi.org/10.3155/1047-3289.57.4.480

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, A., & Hamid, H. A. (2012). Performance of multiple linear regression model for long-term PM10 concentration prediction based on gaseous and meteorological parameters. *Journal of Applied Sciences, 12*(14), 1488-1494.

Watson, G. L., Telesca, D., Reid, C. E., Pfister, G. G., & Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution*, *254,* Article 112792. https://doi.org/10.1016/j.envpol.2019.06.088