A PM2.5 Forewarning Algorithm Using k-Nearest Neighbors Machine Learning at Changpuek, Chiang Mai, Thailand

A Characterization of PM2.5 Levels at Changpuek, Chiang Mai, Thailand

Nopparat Pochai*
Department of Mathematics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
nopparat.po@kmitl.ac.th

Kaboon Thongtha
Department of Mathematics, Faculty of Science, Mahanakorn University of Technology, Bangkok, 10530, Thailand
kaboon.t@gmail.com

ABSTRACT

In Chiang Mai, Thailand, the air pollution issue caused by atmospheric particulate matter with a diameter of less than 2.5 μm, or PM2.5, has been identified as an ongoing crisis. PM2.5 not only has a direct impact on people's health and way of life, but it also has a negative impact on the national economy. Residents in such PM2.5-polluted locations are particularly susceptible to respiratory diseases, skin diseases, inflammatory eye diseases, and cardiovascular problems. As a result, this study is going to analyze PM2.5 data using the k-nearest neighbors machine learning algorithm as a guideline to warn people, particularly in Changpuek, Chiang Mai, Thailand, to handle the PM2.5 characterization problem.

CCS CONCEPTS

• **: Computing methodologies** → Ranking.

KEYWORDS

Chiang Mai, PM2.5, air pollution, k-nearest neighbors, machine learning algorithm

ACM Reference Format:
Nopparat Pochai and Kaboon Thongtha. 2023. A PM2.5 Forewarning Algorithm Using k-Nearest Neighbors Machine Learning at Changpuek, Chiang Mai, Thailand: A Characterization of PM2.5 Levels at Changpuek, Chiang Mai, Thailand. In *The 7th International Conference on Education and Multimedia Technology (ICEMT 2023), August 29–31, 2023, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3625704.3625749

1 INTRODUCTION

With the beginning of the modernization period, people's increased awareness of environmental and health protection has placed air pollution as one of society's primary concerns. Indeed, according to a World Health Organization report, air pollution is the most serious environmental health risk [1]. As a result, various studies have been produced and pioneered in order to decrease the threats that air pollution poses to the ecosystem.

Air quality monitoring is one of the most effective techniques to combat air pollution. Knowing the air quality allows for ideas to help mitigate its hazardous consequences. The majority of air quality monitoring devices connect the results to the air quality index (AQI). The Air Quality Index (AQI) is a method that standardizes the levels and quality of contaminated air. The higher the index, the more hazardous the air is for people, with 500 being the most dangerous [2].

Particulate matter (PM) is a suspension of solid and liquid particles in the air. These are classified as coarse, fine, and ultrafine. PM2.5 are small particles with a diameter of less than 2.5 micrometers (more than 100 times finer than a human hair) that linger in the air for a long period of time. PM2.5 is a health danger since it can travel deep into the respiratory system, reaching the lungs and the bloodstream.

Fine particle exposure may additionally harm lung function and increase diseases such as asthma and heart disease. Increases in daily PM2.5 exposure have been associated in scientific research with an increase in respiratory and cardiovascular hospital admissions, emergency department visits, and death rates.

In [3], they create a system that monitors air quality in a city using vehicle sensor networks (VSNs). In this study, the researchers recommended employing VSNs to tactically monitor air quality and construct an efficient data gathering and estimation (EDGE) system. This research emphasizes AQI as the primary criteria for assessing accuracy. They characterize the air quality in [4] by developing a model that links sensor results to AQI. Machine learning is utilized to build the model using R programming and the k-nearest neighbor (KNN) technique. In [5], they demonstrate a system that uses big data (BD) analytics to analyze CO2 levels on a logistics shipping base in Norway. The data was collected via wireless sensor networks, and the researchers employed BD as a decision support system for the health and safety of shipping personnel. [6] proposes a transportable and cost-effective technology for monitoring particulate matter (PM) air quality. In [7], they propose a technique for estimating and forecasting varying amounts of pollutants, with a focus on ozone. The scientists employed a multilabel classifier based on a machine learning technique based on Bayesian networks to assess the likelihood of pollutants exceeding a particular threshold depending on the AQI [8].

The issue of air pollution caused by atmospheric particulate matter with a diameter of less than 2.5 m, or PM2.5, has been focused on as an ongoing crisis in Chiang Mai, Thailand. PM2.5 has a significant influence not only on people's health and way of life but also

*Corresponding author.

ICEMT 2023, August 29–31, 2023, Tokyo, Japan

ACM ISBN 979-8-4007-0914-2/23/08.
https://doi.org/10.1145/3625704.3625749

Table 1: Daily PM2.5 and PM10 values on data from 1, 2, 3, and 4 days passed along 90 days since January,1, 2021 – March, 31, 2021.

Data features	Range	Unit
1-day past PM2.5	10-120	$\mu g/m^3$
2-days past PM2.5	10-120	$\mu g/m^3$
3-days past PM2.5	10-120	$\mu g/m^3$
4-days past PM2.5	10-120	$\mu g/m^3$
1-day past PM10	10-120	$\mu g/m^3$
2-days past PM10	10-120	$\mu g/m^3$
3-days past PM10	10-120	$\mu g/m^3$
4-days past PM10	10-120	$\mu g/m^3$
Warning levels	5 colors	-

Table 2: AQI Basics for Ozone and Particle Pollution [11].

Daily AQI Color	Levels of Concern	Values of Index
Green	Good	0 to 50
Yellow	Moderate	51 to 100
Orange	Unhealthy for Sensitive Groups	101 to 150
Red	Unhealthy	151 to 200
Purple	Very Unhealthy	201 to 300
Maroon	Hazardous	301 and higher

Table 3: PM2.5 warning levels.

PM2.5 ($\mu g/m^3$)	Warning levels
0-20	Blue
20-30	Green
30-40	Yellow
40-50	Orange
>50	Red

on the national economy. Residents in such PM2.5-polluted areas are predisposed to respiratory ailments, skin disorders, inflammatory eye diseases, and cardiovascular difficulties. As a consequence, this study will evaluate PM2.5 data using the k-nearest neighbors machine learning method as a guideline to warn people, particularly in Changpuek, Chiang Mai, Thailand, about how to deal with the PM2.5 characterization problem.

1.1 PM2.5 crisis in Chiang Mai, Thailand

Chiang Mai has been subjected to high PM2.5 pollution in recent months as a result of farm waste burning and forest fires in Thailand and neighboring countries. PM2.5 refers to dust particles with diameters of 2.5 micrometers or less that are readily inhaled. Long-term exposure to tiny particles has been related to a variety of chronic ailments, including acute lung and cardiac issues. According to a statement from Chiang Mai University's Maharaj Nakorn Hospital, over 12,000 individuals sought medical treatment for respiratory disorders in Chiang Mai between January and March.

2 K-NEAREST NEIGHBORS MACHINE LEARNING ALGORITHM

The k-Nearest Neighbors or KNN machine learning algorithm is a nonparametric supervised machine learning technique that uses both nominal and numerical qualities of data by picking the most frequent attribute among the KNNs or averaging the KNN values. This machine learning algorithm is one of the top ten data mining algorithms [9]. Distances are calculated to determine which of the properties from the k instances in the training data set is most comparable to a new input. The Euclidean distances for discrete data are the most widely utilized distances in the KNN method, as demonstrated below.

$$dist\,(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

Algorithm adjusting may be done based on the suitable value of k, with the value that best suits the needs of the given training set. The following algorithm demonstrates the use of KNN machine learning [10].

K Nearest Neighbors Pseudocode

1. Load the training and test data
2. Choose the value of K
3. For each point in test data:

3.1 find the Euclidean distance to all training data points

3.2 store the Euclidean distances in a list and sort it

3.3 choose the first k points

3.4 assign a class to the test point based on the majority of classes present in the chosen points

4. End

Table 4: PM2.5 and PM10 for 1-4 days past at Changpuek, Chiang Mai, Thailand since January,1, 2021 – March, 31, 2021.

Days	4-days past PM2.5	3-days past PM2.5	2-days past PM2.5	1-day past PM2.5	4-days past PM10	3-days past PM10	2-days past PM10	1-day1 past PM10
1	21	19	25	28	44	42	43	43
2	19	25	28	26	42	43	43	50
3	25	28	26	29	43	43	50	46
4	28	26	29	31	43	50	46	48
5	26	29	31	38	50	46	48	61
6	29	31	38	39	46	48	61	61
7	31	38	39	41	48	61	61	67
8	38	39	41	33	61	61	67	61
9	39	41	33	33	61	67	61	57
10	41	33	33	38	67	61	57	62
11	33	33	38	35	61	57	62	56
12	33	38	35	36	57	62	56	59
13	38	35	36	38	62	56	59	63
14	35	36	38	22	56	59	63	53
15	36	38	22	31	59	63	53	49
16	38	22	31	50	63	53	49	71
17	22	31	50	51	53	49	71	70
18	31	50	51	42	49	71	70	59
19	50	51	42	53	71	70	59	81
20	51	42	53	46	70	59	81	68
21	42	53	46	48	59	81	68	74
22	53	46	48	56	81	68	74	86
23	46	48	56	57	68	74	86	80
24	48	56	57	51	74	86	80	73
25	56	57	51	58	86	80	73	87
26	57	51	58	53	80	73	87	77
27	51	58	53	50	73	87	77	75
28	58	53	50	51	87	77	75	79
29	53	50	51	56	77	75	79	85
30	50	51	56	51	75	79	85	82
31	51	56	51	60	79	85	82	85
32	56	51	60	39	85	82	85	84
33	51	60	39	47	82	85	84	64
34	60	39	47	38	85	84	64	54
35	39	47	38	30	84	64	54	45
36	47	38	30	32	64	54	45	51
37	38	30	32	49	54	45	51	71
38	30	32	49	62	45	51	71	88
39	32	49	62	74	51	71	88	100
40	49	62	74	64	71	88	100	84
41	62	74	64	35	88	100	84	43
42	74	64	35	20	100	84	43	33
43	64	35	20	18	84	43	33	32
44	35	20	18	26	43	33	32	47
45	20	18	26	35	33	32	47	54
46	18	26	35	46	32	47	54	67
47	26	35	46	45	47	54	67	71
48	35	46	45	55	54	67	71	81
49	46	45	55	48	67	71	81	67
50	45	55	48	54	71	81	67	81
51	55	48	54	57	81	67	81	84

52	48	54	57	46	67	81	84	66
53	54	57	46	40	81	84	66	58
54	57	46	40	43	84	66	58	65
55	46	40	43	40	66	58	65	63
56	40	43	40	58	58	65	63	88
57	43	40	58	61	65	63	88	90
58	40	58	61	51	63	88	90	79
59	58	61	51	62	88	90	79	92
60	61	51	62	76	90	79	92	103
61	51	62	76	77	79	92	103	124
62	62	76	77	62	92	103	124	104
63	76	77	62	83	103	124	104	128
64	77	62	83	98	124	104	128	129
65	62	83	98	59	104	128	129	91
66	83	98	59	70	128	129	91	102
67	98	59	70	91	129	91	102	118
68	59	70	91	114	91	102	118	142
69	70	91	114	113	102	118	142	158
70	91	114	113	102	118	142	158	143
71	114	113	102	130	142	158	143	137
72	113	102	130	117	158	143	137	146
73	102	130	117	101	143	137	146	133
74	130	117	101	103	137	146	133	135
75	117	101	103	83	146	133	135	122
76	101	103	83	68	133	135	122	105
77	103	83	68	61	135	122	105	93
78	83	68	61	65	122	105	93	96
79	68	61	65	74	105	93	96	107
80	61	65	74	67	93	96	107	100
81	65	74	67	70	96	107	100	102
82	74	67	70	54	107	100	102	94
83	67	70	54	48	100	102	94	71
84	70	54	48	42	102	94	71	64
85	54	48	42	47	94	71	64	73
86	48	42	47	60	71	64	73	90
87	42	47	60	54	64	73	90	83
88	47	60	54	67	73	90	83	96
89	60	54	67	65	90	83	96	90
90	54	67	65	74	83	96	90	115

Table 5: PM2.5 and PM10 for 1-4 days past with their warning level.

Day	4-days past PM2.5	3-days pastPM2.5	2-days past PM2.5	1-day1 pastPM2.5	4-days past PM10	3-days pastPM10	2-days pastPM10	1-daypastPM10	Warning level
1	21	19	25	28	44	42	43	43	Green
2	19	25	28	26	42	43	43	50	Green
3	25	28	26	29	43	43	50	46	Green
...	...	...	...	...	...	...	...	...	...
90	54	67	65	74	83	96	90	115	Red

KNN is a good method for classification if the data set is small, aside from the fact that it is straightforward to implement. It has excellent predictive ability and can handle data sets even when the data's structure is unknown, making it a strong fit for the study approach that is being presented.

Table 6: Prediction and actual of warning level in 60 days since January,1, 2022 – March, 31, 2022.

Day	Prediction level	Actual level
1	Green	Green
2	Green	Green
3	Green	Green
...	...	...
60	Red	Red

Table 7: Precision of a 1-day forewarning prediction using 2-4 days past data features in 60 days.

k	2-days past features	3-days past features	4-days past features
1	0.9444	0.9474	0.9655
3	0.9375	0.9583	0.9629
5	0.9483	0.9642	0.9636
7	0.9483	0.9492	0.9474

2.1 Min-Max Normalization

The original data is transformed linearly in this method of data normalization. A simple method, scaling, also known as min-max scaling or min-max normalization, is scaling the range of features to scale the range in [0, 1] or [1, 1]. Depending on the data's type, the target range must be chosen. The following is the general formula for a min-max of [0, 1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

where x is an original value, x' is the normalized value.

2.2 Air quality warning dataset

In many applications, a vector generates n different pieces of information that are relevant to one particular entity or object. Quantities that can be measured or that may be calculated from the item can be included in the list of quantities. These elements are referred to as features or attributes, and such a vector can also be referred to as a feature vector.

In order to notify people, particularly in Changpuek, Chiang Mai, Thailand, about how to handle the PM2.5 characterization problem, we will assess the PM2.5 warning level using the k-nearest neighbors machine learning method. The data features that are required to be computed in this study are daily PM2.5 and PM10 values on data from 1, 2, 3, and 4 days passed along 90 days since January,1, 2021 – March, 31, 2021, as shown in Table 1.

2.3 Air Quality Index (AQI)

The short-term national ambient air quality standard for protection of public health is typically equivalent to an ambient air concentration of 100 for each pollutant. In general, AQI levels of 100 or less are considered to be good. Air quality is dangerous when AQI values are above 100, initially for some vulnerable groups of individuals, then when AQI values rise for everyone. There are six categories that make up the AQI. A varying level of health concern relates to each category. Additionally, each group has a unique color. People can immediately detect whether the air quality in their neighborhoods has reached harmful levels according to the colors as shown in Table 2[11].

Depending on the level of health concern, the PM2.5 values are classified into five warning classes, as shown in Table 3.

3 TRAINED KNN MODEL RESULTS

The creation of the model occurs next, after a thorough analysis and comprehension of the data at hand. The outcomes of fitting and training a KNN machine learning algorithm for the characterization of air quality data are shown in this part. The prediction of the one-day PM2.5 forewarn level and the actual PM2.5 level are compared in Table 3. The precision of a 1-day forewarning prediction using different data features and k, such as 2-days past, 3-days past, and 4-days past features, is shown in Table 4.

4 CONCLUSION

The research results of the PM2.5 forewarning prediction are acceptable. Using various data features and k, such as 2-days past, 3-days past, and 4-days past features, the precision of a 1-day forewarning prediction is above 90% as shown in Tables 5-7. As a result, this investigation may be utilized as a reference to build a PM2.5 warning system and inform people about how to handle the PM2.5 characterization issue, especially in Changpuek, Chiang Mai, Thailand.

ACKNOWLEDGMENTS

This paper is supported by School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand.

REFERENCES

[1] World Health Organization (WHO), "7 million premature deaths annually linked to air pollution," Mar. 2014. [Online]. Available: http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/.
[2] U.S. Environmental Protection Agency, "Air quality guide for particle pollution," US EPA, 2015.
[3] Y.C. Wang and G.W. Chen, "Efficient data gathering and estimation for metropolitan air quality monitoring by using vehicular sensor networks," IEEE Trans. Veh. Technol., vol. 66, no. 8, pp. 7234–7248, 2017.
[4] Y. Li and J. He, "Design of an intelligent indoor air quality monitoring and purification device," in 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), 2017, pp. 1147–1150.
[5] J. Molka-Danielsen, P. Engelseth, V. Olesnanikova, P. Sarafin, and R. Zalman, "Big data analytics for air quality monitoring at a logistics shipping base via autonomous wireless sensor network technologies," 2017 5th Int. Conf. Enterp. Syst., pp. 38–45, 2017.
[6] Y. Wu *et al.*, "Mobile microscopy and machine learning provide accurate and high-throughput monitoring of air quality," in 2017 IEEE Conference on Lasers and Electro-Optics, 2017.
[7] T. M. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatial-temporal data from air quality monitoring stations," in 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), 2016, pp. 58–63.
[8] U.S. Environmental Protection Agency, "Technical assistance document for the reporting of daily air quality—The air quality index (AQI)," US EPA, Dec. 2013.
[9] R. Agrawal, "k-nearest neighbor for uncertain data," Int. J. of Computer Applications, vol. 105, no. 11, pp. 13-16, 2014.
[10] Timothy M. Amado, "Air Quality Characterization Using k-Nearest Neighbors Machine Learning Algorithm via Classification and Regression Training in R", Journal of Computational Innovations and Engineering Applications 2(2), pp. 1–7, 2018.
[11] AirNow, "Air Quality Index (AQI) Basics", Sep. 2023. [Online]. Available: https://www.airnow.gov/.