The Efficiency of Measure Particulate Matter Device using Classification Techniques and Association Rules Discovery for Factor Analysis Influencing Particulate Matter

Pongpat Singsri[1]* and Sahapat Chalachai[2]

[1]Information and Communication Technology, Science and Technology, Rajamangala University of Technology Tawan-ok, Chonburi, 20110, Thailand

[2]Agricultural Engineering and Technology, Agriculture and Natural Resources, Rajamangala University of Technology Tawan-ok, Chonburi, 20110, Thailand

* Corresponding author. E-mail address: pongpat_si@rmutto.ac.th

Received: 10 August 2020; Revised: 10 November 2020; Accepted: 17 November 2020

Abstract

The objectives of the research were to develop low-cost Measure Particulate Matter Device ($PM_{2.5}$) with real-time display and analyze data using classification techniques and association rules discovery. The developed device has a microcontroller with a $PM_{2.5}$ dust detector using infrared light, temperature and humidity sensor, and raindrop sensor. The device is being installed at the weather station in Laem Chabang Municipal Stadium (32T), Chonburi, Thailand. The weather station belongs to the Air Quality and Noise Management Bureau, Pollution Control Department, Ministry of Natural Resources and Environment. The data is collected every 20 seconds for 50 days, received a total of 196,023 records. The collected data is divided into 2 data sets; the first data set is the data that device and weather station measures at the same time and the second data set are the average hourly data of that device and weather station measure. The two sets of data compared to measure the efficiency of the device. The data was analyzed by data mining using Classification Techniques with Decision Tree algorithm and Association Rule with the Apriori algorithm. The results of this research showed that the device has the accuracy in measuring at ± 0.5 (± 5.77%) with hourly data set and has the accuracy in measuring at ± 1.0 (± 8.46%) with the hourly average data set. Decision Tree algorithm has accuracy for the forecast at 62.36%. Apriori algorithm gave the highest confidence in Association Rules at 78% with $PM_{2.5}$ between 20.00 - 24.99 μg/m^3 relation to a temperature between 30.00° - 34.99° C

Keywords: Particulate Matter 2.5, microcontroller, Decision Tree Algorithm, Apriori Algorithm

Introduction

The world around experiencing more air pollution such as The US Embassy in Beijing found that fine particulate matter has particulate matter that has a diameter of fewer than 2.5 microns (PM2.5) measured 568 micrograms per cubic meter (μg/m3) on January 15, 2015 and the Weather Station in the east of Beijing measured the maximum of $PM_{2.5}$ at 631 μg/m^3 (Dailynews, 2015). In Thailand, Chiang Mai Province Offices for Natural Resources and Environment (2019) found that the air quality is at a level that exceeds the standard value and measured $PM_{2.5}$ at 183 μg/m^{3}on April 5, 2019. Currently, there is still a high level of $PM_{2.5}$. The monitoring of air pollution from AirVisual (2020) on February 25, 2020, found Dhaka, Bangladesh, ranked no.1 with the highest amount of $PM_{2.5}$ at 196 μg/m^{3f,} followed by Delhi, India, with $PM_{2.5}$ value at 179.5 μg/m^{3}and Kolkata, India, with $PM_{2.5}$value at 128 μg/m^3. In Bangkok and the metropolitan region of Thailand on January 11, 2020, the air pollution levels by Real-time Air Quality Index (2020) found $PM_{2.5}$ value at 180 μg/m.^{3} The good air quality level, as determined by the World Health Organization, is 25 μg/m^{3,} and Thailand has set a good air quality level at 50 μg/m^3.

To measure $PM_{2.5,}$ mostly required the devices to import from foreign countries and cost up to 200,000 - 10,000,000 baht (depending on the measurement technique and measurement resolution) and the currently

developed equipment are three times cheaper than imported from abroad but still considered to be a heavy and costly (Intra, 2009).

Figure 1 Real-time measuring and sampling of $PM_{2.5}$ and PM_{10}

Kiatwattanacharoen and Jayasvasti has developed a real-time particle detector (PM_{10}) or Real-Time Air Particulate Matters Monitoring, as shown in Figure 2. It is priced at around 100,000 baht cheaper than foreign countries, which cost about 400,000 baht. However, still considered to be expensive, not easy to carry, and cannot manage by internet. (2013).

(a) (b)

Figure 2 Real-Time Measuring and Sampling of PM_{10}. (a) Research Institute for Health Science (2012) (b) Manager Online (2013)

From the research articles and developed real-time measuring of $PM_{2.5}$ and PM_{10} devices in Thailand and microcontroller technology, the researcher has the idea to develop a prototype particulate matter device with a low cost and high efficiency. The measured data were analyzed for data analysis by Classification Techniques (Decision Trees) and Association Rules (Apriori algorithm) to be used to analyze the trend of particles matter in the air at ground level, using Data Mining analysis. Data Mining is a method of finding all knowledge, relationships, and patterns that are hidden in a large amount of data. Data mining automates the exploration and analysis of data in meaningful and legal ways, with these relationships demonstrating useful knowledge. The core of data mining is to search for knowledge in an extensive database. Once knowledge has been acquired, it is used to make decisions. (Songsiri, Rakthanmanon, & Waiyamai, 2001)

The decision tree has been used for grouping data and has theories to create decision trees as usually in the form of rules in "IF (conditions) Then (results)." Decision Trees, including "node" replace trees, branch represents the test conditions, and use leaves instead of the result or answer class. Which in selecting which attribute is the node in the tree will use the "Information Gain." The feature with the highest gain is chosen as the root node. The gain is calculated from the Entropy of all data. (Rattanapoka, 2013). The use of classification techniques, one of the effective techniques for data mining in order to analyze the probability of occurrence, factors, and the relationship of Factor Analysis Influencing Student Retire using Classification Technique case study of Rajamangala University of Technology Tawan-ok, Bangphra Campus (Singsri, Phiromlap, Saramas, & Koodsamrong, 2018), This algorithm still able to Quail Gender Identification Considering from the External Factors of Quail Eggs (Singsri, Suksawatchon, & Suksawatchon, 2013) (Suksawatchon, Suksawatchon, & Singsri, 2014) (Suksawatchon & Singsri, 2014), and can using the Decision Trees method for dust detection from MODIS satellite image (Ataei, Mohammadzadeh, & Abkar, 2015).

Methods and Materials

1. Research tools

This research uses data collected from the developed prototype and various applications as follows.

1. Arduino IDE application is text editor for coding with C language, used for creating and saving the commands into microcontroller memory and used for running device by command and save the results.

2. Microsoft Excel application is a program used to gather information and sort data in a format used in the processing and use of the pivot function to summarize data. And then, use the results to calculate the accuracy and error by using the equation to find the accuracy of the measured results when compared with the data of the measurement station in terms of percentages as shown in Equation (1) and (2) respectively.

$$accuracy = average(Ax{:}Ay) \quad (1)$$

where Ax is the difference between data from the prototype and data from the weather station of the first data set and Ay is the difference between data from the prototype and data from the weather station of the last data set.

$$error = ((Cx-Bx)/Bx) * 100 \quad (2)$$

where Bx is the data that the weather station can measured and Cx is the data that the prototype can measured.

3. RapidMiner Studio Trial 9.3.001 application, used for data analysis by classification technique with Decision Tree method (C4.5 Algorithm) and Association Rules with Apriori algorithm.

2. Data collection

This research use data collected from the prototype and data from the Pollution Control Department at Laem Chabang Municipal Stadium Weather Station (32T). The data collection has the following steps.

1. After the researcher developed the commands set, the particulate matters measuring device, and other sensors devices were installed at Laem Chabang Municipal Stadium Weather Station (32T). Because it's an air

monitoring station that can measure $PM_{2.5}$ located near a large industrial estate and also an area with large construction, this will allow the device to measure a variety of dust values than installed in a place with the wind that is calm and does not dust.

2. The device installed near the detector of the station and set the time of the device the same as that of the weather station to measure.

3. Collected data every 20 seconds, and the data saved into SD-Card. The researchers collected data that was measured every seven days during the data collection period.

4. When the first round of data collection is completed, the researcher used the data obtained to analyze the data and modify commands to increase the accuracy of the measurement of the device.

5. Install the device and store the data again at the same location.

6. Evaluated the efficiency of measuring $PM_{2.5}$ of the device with Laem Chabang Municipal Stadium Weather Station (32T).

7. The data obtained was analyzed by data mining.

The first set of data was collected from June 28 to July 28, 2018, which received a total of 126,056 data sets. The second set of data was collected from August 6 to August 18, 2018, and August 22 to August 27, 2018, received a total of 69,967 data sets. The data collected for 50 days and has 196,023 data sets.

Divided the data into two groups, with the first group being hourly data, and the second group is average hourly data, as shown in Table 1.

Table 1 The dataset used for processing.

Group	Number (data sets)
1st Group	1,093
2nd Group	1,093
Daily average data	50

3. Data analysis and statistics

This research uses two types of analysis as follows.

1. Analysis to measure the performance of prototype device used two methods of comparing values were to 1) compare data that the prototype device obtained at the same time as the weather station can measure and obtained, and 2) compare the average hourly data that the prototype device obtained at the same time as the weather station can measure and obtain. Then, each data sets were calculated to be accurate from the statistical correlations of data from the prototype device, compared with the data from the Laem Chabang Municipal Stadium Weather Station (32T) that has the same measurement unit is "Micrograms per cubic meter" ($\mu g/m^3$) with errors.

2. Data analysis using RapidMiner Studio Trial 9.3.001 application, by using the classification technique with Decision Tree method (C4.5 Algorithm) in order to make decision rules for the forecasting, used Association Rules with Apriori algorithm in order to find the relationship of data, and use 10-fold cross-validation to evaluate the model. After that, it is considered in the development of the command set to increase the efficiency of the device. The factors for analysis as shown in Table 1.

Table 2 Data collection

Data type	Type of value
Date	dd/mm/YYYY
Time	hh:mm:ss
Temperature	Actual value (Decision Tree)
	Acronym (Apriori)
Humidity	Actual value (Decision Tree)
	Acronym (Apriori)
Raining	Acronym
$PM_{2.5}$	Acronym

From Table 2, The researcher was an acronym to represent the data as shown in Table 2, because the data format is not suitable for analysis. The factors were an influence to measure $PM_{2.5}$ by high temperature, low humidity and no rain relate to low particulate matter because hot air help particulate matter to rise higher, and low temperature, high humidity and raining relate to high particulate matter because high pressure forces the particulate matter to float lower.

In Table 3, The researcher was divided particulate matter factor into 7 groups, temperature factor into 6 groups, humidity factor into 5 groups, and rain into 2 groups by average number and type of data.

Table 3 Data representation

Value	Acronym
Particulate matter 0.00 - 4.99 $\mu g/m^3$	Adust
Particulate matter 5.00 - 9.99 $\mu g/m^3$	Bdust
Particulate matter 10.00 - 14.99 $\mu g/m^3$	Cdust
Particulate matter 15.00 - 19.99 $\mu g/m^3$	Ddust
Particulate matter 20.00 - 24.99 $\mu g/m^3$	Edust
Particulate matter 25.00 - 29.99 $\mu g/m^3$	Fdust
Particulate matter 30.00 $\mu g/m^3$ or more	Gdust
Temperature 25.00° C - 29.99° C	Atemp
Temperature 30.00° C - 34.99 ° C	Btemp
Temperature 35.00 ° C - 39.99 ° C	Ctemp
Temperature 40.00 ° C - 44.99 ° C	Dtemp
Temperature 45.00 ° C - 49.99 ° C	Etemp
Temperature 50.00 ° C or more	Ftemp
Humidity 50.00% - 59.99%	Ahumi
Humidity 60.00% - 69.99%	Bhumi
Humidity 70.00% - 79.99%	Chumi
Humidity 80.00% - 89.99%	Dhumi
Humidity 90.00% - 100.00%	Ehumi
Raining	YES
No rain	NO

4. Experimentation and data collection places

The development of the prototype device, data processing, and the conclusion of the experiment completed at Embedded and Automation Systems Laboratory, Information and Communication Technology Program,

Faculty of Science and Technology, Rajamangala University of Technology Tawan-ok, 43 Moo.6, Bang Phra sub-district, Sriracha district, Chonburi, Thailand. The data is collected at Laem Chabang Municipal Stadium Weather Station (32T) Sriracha district, Chonburi, Thailand.

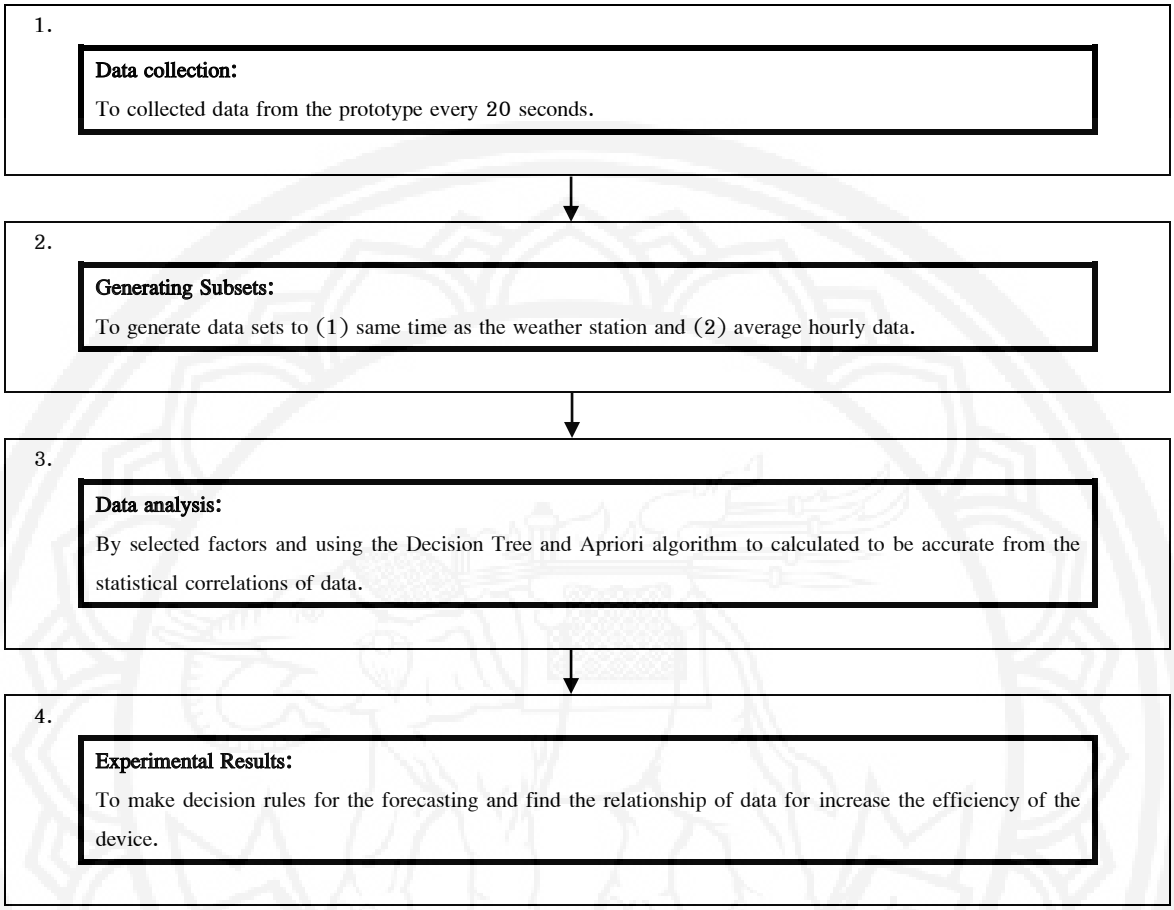

Figure 3 The proposed method framework

Results and discussion

1. Development prototype

The researchers have developed a prototype device for measuring $PM_{2.5}$ by using various materials and equipment as follows,

1. Microcontroller, in this research, we used Arduino Uno R3, as shown in Figure 4(a), for reading the value from sensors, showing results on-screen (LCD 20x4), and saving data into SD-Card.

2. Datalog and Real-Time Clock (RTC) is a circuit that has an SD-Card storage unit and has a timer circuit in itself, as shown in Figure 4(b).

3. Digital Humidity and Temperature sensor (DHT22) is a sensor device that can measure temperature and humidity in the air in itself and give an actual value by a digital signal, as shown in Figure 4(c).

4. Dust sensor (GP2Y10 PM2.5 module) is a sensor device to particulate matter measure. Work by using infrared firing and then measuring the reflection of light from particulate matters, as shown in Figure 4(d).

5. Raindrop sensor is a device that uses the principle of water droplets falling onto the contacts of the circuit board to complete the cycle and has a signal, as shown in Figure 4(e).

6. LCD Display is LCD can display 20 characters in one line and can display four lines on the device, as shown in Figure 4(f).

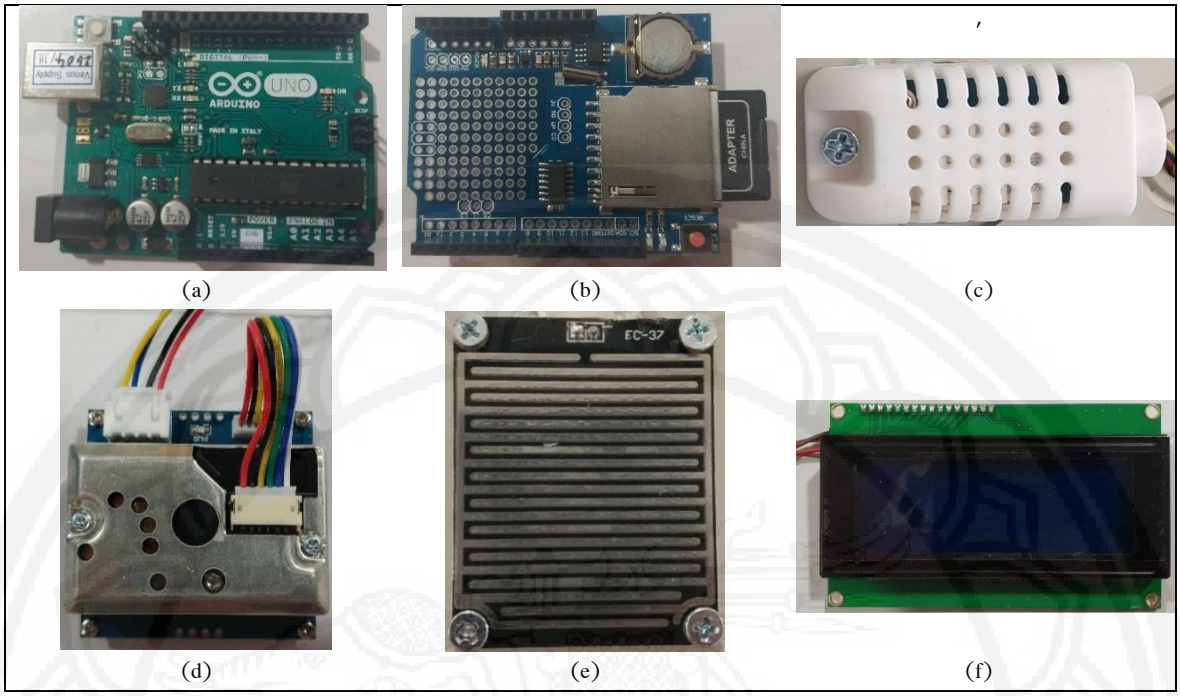

Figure 4 Device and sensors

The devices and components fitted into the plastic box, and commands were set up into a microcontroller to control the device. When the prototype development has been completed, it is installed to measure and collect data, as shown in Figure 5. The device collected data every 20 seconds.

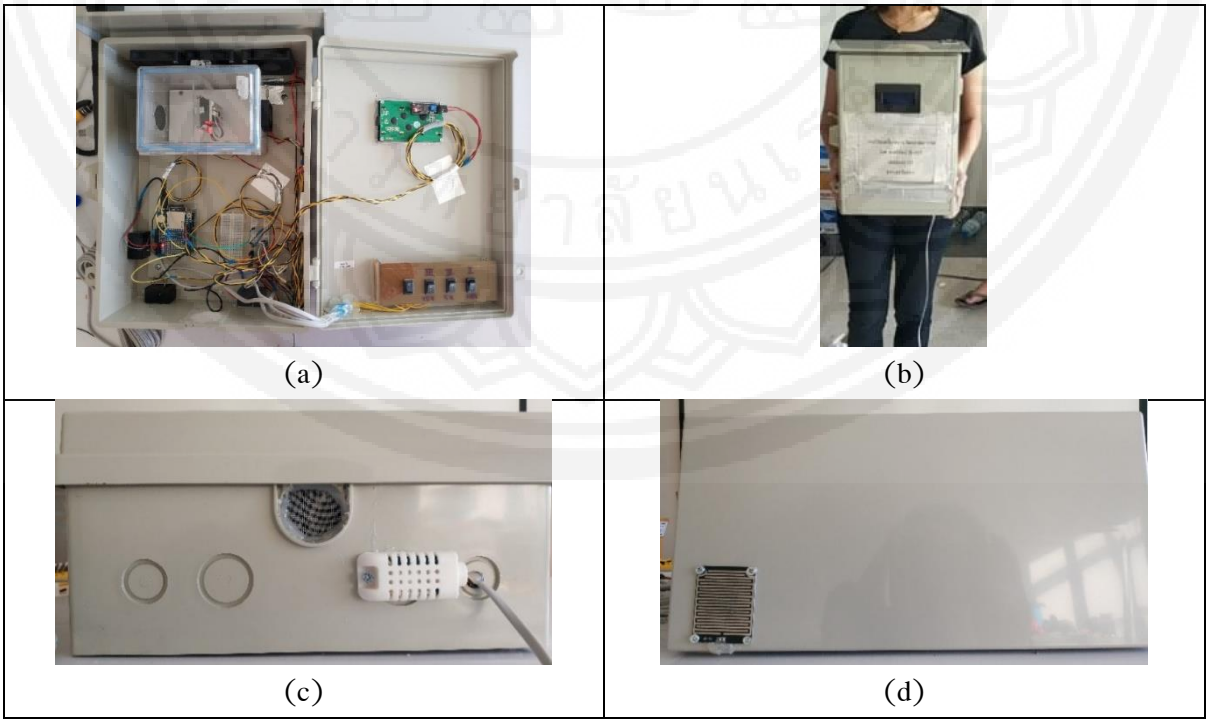

Figure 5 Prototype device for measuring $PM_{2.5}$ (a = inside, b = overall, c = below, d = on-top)

The device is being installed at Laem Chabang Municipal Stadium Weather Station (32T) near the air suction pipe of the station, as shown in Figure 6.

Figure 6 Installed point

Table 4 Accuracy and error of prototype

Data	Accuracy	Error (%)
First data collection (hourly data)	±0.5	6.72
First data collection (hourly average data)	±1.5	9.13
Second data collection (hourly data)	±0.5	4.82
Second data collection (hourly average data)	±1.0	7.79
Summary (hourly data)	±0.5	5.77
Summary (hourly average data)	±1.0	8.46

From Table 4, The performance of prototype device has highest accuracy and lowest error at ± 0.5 (± 5.77%) with hourly data set and ± 1.0 (± 8.46%) with the hourly average data set.

2. Results from forecasting and modeling processing

Before data processing, the researcher selected the factors that had the highest effect on Decision Trees analysis using RapidMiner Studio Trial 9.3.001 application, as shown in Figure 7. The result has three good factors that affect the results, such as time, temperature, and humidity.

Before data processing, the researcher selected the factors that had the highest effect for the Apriori algorithm with RapidMiner Studio Trial 9.3.001 application, as shown in Figure 8. The result has four good factors that affect the results, such as date, time, temperature, and humidity.

From the analysis of all three factors using Classification techniques with Decision Trees method to make decisions for the forecasting and using 10-fold cross-validation to evaluate the model. Analysis of all five factors using Association Rules with the Apriori algorithm to find the relationship of data. The RapidMiner Studio Trial 9.3.001 application used for processing and finding the factors that affect the change of $PM_{2.5,}$ as shown in Table 5.

Figure 7 selected the factors for Decision Trees method

Figure 8 selected the factors for Apriori algorithm

Table 5 Data analysis results

Techniques	Accuracy (%)
Decision Trees with minimum number object at 25	62.36
Apriori algorithm	78

Table 5 gives decision rules from decision trees with minimum number object at 25 and has the highest accuracy at 62.36%. Apriori algorithm has the highest confidence at 0.78 from the relationship of Edust => Btemp that means particulate matter 20.00 – 24.99 μg/m^3 has a relation with temperature 30.00°C – 34.99°C at 78%.

Discussions

Comparison of data from the real-time measuring of $PM_{2.5}$ and PM_{10} with data from air quality monitoring station, Pollution Control Department found that the information obtained is likely to be similar (Supasri, Intra, Jomjunyong, & Sampattagul, 2018). Same as this prototype, but this prototype has a much lower cost in a similar performance

The accuracy of decision tree method in this research has higher than Dust Detection from MODIS Satellite Image by the Normalized Difference Dust Index (NDDI) Indicators and Ln (b1) are used for dust detection over bright surface, have the accuracy at 58%. And dark surfaces have the accuracy at 53% is achieved using the NDDI and BTD (BT20-BT31) (Ataei et al., 2015).

Conclusion and Suggestions

The results of this research showed that the device has the accuracy in $PM_{2.5}$ measuring at ± 0.5 (± 5.77%) with first data set and has the accuracy in $PM_{2.5}$ measuring at ± 1.0 (± 8.46%) with the second data set.

Selected the factors that had the highest effect on Decision Tree algorithm as time, temperature, and humidity. That has accuracy for the forecast at 62.36%. Selected the factors that had the highest effect on Apriori algorithm as date, time, temperature, humidity, and dust. That gave the highest confidence in Association Rules at 78% with $PM_{2.5}$ between 20.00 – 24.99 μg/m^3 relation to a temperature between 30.00° – 34.99° C. This result showed the hot air help particulate matter to rise higher and make low particulate matter at the ground.

The GP2Y10 PM2.5 module is a sensor device to detect particulate matter by using infrared light. This sensor has a potential measurement error because of the light from the sun. Nowadays, there is a laser light measuring device to detect particulate matter, can be measured more accurately than using infrared light, and can measure up to $PM_{0.1}$.

In the future work, that can make the real-time alarm and report system in the 6-D Digital Secure City System (Bang Phra Municipality Application), Bang Phra Sub-district Municipality, Chonburi, Thailand.

Acknowledgments

Financial support for this research was provided by the National Research Council of Thailand, project no. 543884.

Thank you, Mr.Wissanu Wingpad, Air Quality Division staff, Bureau of Air and Noise Quality Management, Pollution Control Department, Ministry of Natural Resources and Environment. That helps in requesting access to the area and advise on air quality monitoring stations at Laem Chabang Municipal Stadium Weather Station (32T)

References

AirVisual. (2020). *Air Quality Index (AQI).* Retrieved from https://www.iqair.com

Ataei, S. H., Mohammadzadeh, A., & Abkar, A. A. (2015). Using Decision Tree Method for Dust Detection from MODIS Satellite Image. *Journal of Geomatics Science and Technology, 4*(4), 151-160.

Chiang Mai Province Offices for Natural Resources and Environment. (2019). *The air quality is at a level that exceeds the standard value.* Retrieved from https://www.chiangmai.mnre.go.th/th/news/more/1380

Dailynews. (2015). *Air pollution in in Beijing soared 20 times.* Retrieved from https://www.dailynews.co.th/foreign/294174

Intra, P. (2009). *Real-Time Measuring and Sampling of $PM_{2.5}$ and PM_{10}. Research promotion and development section, Research management mission, National Research council of Thailand.* Retrieved from https://www.priv.nrct.go.th/shopping/home/show_product.php?research_id=517

Manager Online. (2013). *Installed Measure Particulate Matter Device (Real Time) at Lampang Child Center.* Retrieved from manager.co.th/Local/ViewNews.aspx?NewsID=9560000043806

Rattanapoka, C. (2013). *Teaching documents 030523111 Introduction to Artificial Intelligence.* Bangkok: King Mongkut's University of Technology North Bangkok.

Real-time Air Quality Index. (2020). *NRCT, BKK Air Pollution.* Retrieved from https://www.aqicn.org/city/thailand/bkk/nrct/

Research Institute for Health Science. (2012). *Research Institute for Health Science are successful invention of the small dust particle analyzer to stimulate the resolution of forest fire smog from the local level, General Division, Office of the University, Chiang Mai University.* Retrieved from http://www.prcmu.cmu.ac.th/perin_detail.php?perin_id=327

Singsri, P., Phiromlap, S., Saramas, S., & Koodsamrong, R. (2018). Factor Analysis Influencing Student Retire Using Classification Technique Case Study Rajamangala University of Technology Tawan-ok, *LRU Research Conference 2018* (pp. 363-371). Bangphra Campus.

Singsri, P., Suksawatchon, J., & Suksawatchon, U. (2013). Apply of Classification Techniques for Factor Analysis Influencing Quail Gender Identification Considering from the External Factors of Quail Eggs. *National Conference on Computing and Information Technology*, *9,* 240 - 247.

Songsiri, C., Rakthanmanon, T., & Waiyamai, K. (2001). Applying a data mining technique to help students in selecting their majors. *Kasetsart University Annual Conference: Engineering*, *39,* 43-50.

Suksawatchon, J., Suksawatchon, U., & Singsri, P. (2014). Feature Selection and Efficiency Comparison of Classification Techniques for Quail Gender Forecasting from the External Factors of Quail Eggs. *National Conference on Computing and Information Technology*, *10,* 515 - 521.

Suksawatchon, U., & Singsri, P. (2014). Factor Analysis Influencing Quail Gender Classification Considering from the External Factors of Quail Eggs Using Classification Techniques. *International Joint Conference on Computer Science and Software Engineering*, *11,* 297-301.

Supasri, T., Intra, P., Jomjunyong, S., & Sampattagul, S. (2018). Evaluation of Particulate Matter Concentration by Using a Wireless Sensor System for Continuous Monitoring of Particulate Air Pollution in Northern of Thailand. *Journal of Innovative Technology Research*, *2*(1), 69-83.