Forecasting Air Quality Index in Thailand Using Ensemble Method

Mr. Saksiri Lertnilkarn

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science and Information Technology

Department of Mathematics and Computer Science

FACULTY OF SCIENCE

Chulalongkorn University

Academic Year 2023

การพยากรณ์ดัชนีคุณภาพอากาศในประเทศไทยโดยใช้วิธีการผสม

นายศักดิ์สิริ เลิศนิลกาญจน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2566

Thesis Title	Forecasting Air Quality Index in Thailand Using Ensemble Method
By	Mr. Saksiri Lertnilkarn
Field of Study	Computer Science and Information Technology
Thesis Advisor	Associate Professor SUPHAKANT PHIMOLTARES, Ph.D.

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF SCIENCE
(Professor POLKIT SANGVANICH, Ph.D.)

THESIS COMMITTEE

..... Chairman
(Professor CHIDCHANOK LURSINSAP, Ph.D.)

..... Thesis Advisor
(Associate Professor SUPHAKANT PHIMOLTARES, Ph.D.)

..... Examiner
(Prem Junsawang, Ph.D.)

ศักดิ์สิริ เลิศนิลกาญจน์ : การพยากรณ์ดัชนีคุณภาพอากาศในประเทศไทยโดยใช้วิธีการผสม. (Forecasting Air Quality Index in Thailand Using Ensemble Method) อ.ที่ปรึกษาหลัก : รศ. ดร.ศุภกานต์ พิมลธเรศ

มลภาวะทางอากาศเป็นหนึ่งในปัญหาที่สำคัญที่สุดที่จำเป็นต้องแก้ไขอย่างเร่งด่วนทั่วโลก ประเทศไทยก็ต้องต่อสู้กับปัญหานี้ด้วยเช่นกันอย่างหลีกเลี่ยงไม่ได้ โดยเฉพาะอย่างยิ่งในทางตอนเหนือของประเทศไทย บริเวณนี้เผชิญกับการปนเปื้อนทางอากาศมาเป็นเวลาหลายปี ในวิทยานิพนธ์ฉบับนี้ได้นำเสนอตัวแบบบนพื้นฐานของวิธีการผสมเพื่อทำนายระดับของดัชนีคุณภาพอากาศ (เอคิวไอ) ในบริเวณนี้จากผลการลงคะแนนส่วนใหญ่ที่ได้จากขั้นตอนวิธีการจำแนกสามวิธี ได้แก่ซัพพอร์ตเวกเตอร์แมชชีน ป่าสุ่ม และเพื่อนบ้านใกล้ที่สุดเคตัว วิธีการที่นำเสนอนี้ได้ให้การเปรียบเทียบระหว่างความแม่นยำการจำแนกของการลงคะแนนและความแม่นยำของตัวแบบการจำแนกแต่ละตัวแบบจำลองนี้ใช้ชุดข้อมูลเจ็ดชุดจากสถานีตรวจวัดคุณภาพอากาศในสี่จังหวัดทางตอนเหนือของประเทศไทย ท้ายที่สุดแล้ว ตัวแบบผสมที่นำเสนอนี้ให้ค่าความแม่นยำโดยเฉลี่ยที่ 99.68% - 99.84% มากกว่าค่าความแม่นยำโดยส่วนใหญ่ที่ได้จากตัวแบบนำมาเปรียบเทียบ

สาขาวิชา	วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ	ลายมือชื่อนิสิต .....
ปีการศึกษา	2566	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6278014723 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
KEYWORD: air pollution, ensemble model, AQI

Saksiri Lertnilkarn : Forecasting Air Quality Index in Thailand Using Ensemble Method. Advisor: Assoc. Prof. SUPHAKANT PHIMOLTARES, Ph.D.

Air pollution is one of the most important problems that needs to be urgently solved around the world. Inevitably, Thailand has had to fight against it as well, particularly in the Northern Thailand. This region also has faced high air contamination for several years. In this thesis, the proposed model based on ensemble method was presented to predict the extent of air quality index (AQI) in the region from the majority vote of outputs from three classification algorithms, namely, support vector machine, random forest, and k-nearest neighbors. This proposed method made a comparison between the voted classification accuracy and the accuracies of the individual classification models. The model took advantage of seven datasets from monitoring stations in four provinces in the Northern Thailand. Eventually, the proposed ensemble model produced, on average, the accuracy rate of 99.68% - 99.84% greater than most of the accuracies of the other comparative models.

Field of Study:	Computer Science and Information Technology	Student's Signature .....
Academic Year:	2023	Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my profound gratitude towards everyone who has helped me to achieve this thesis.

This piece of work would not have been completed without suggestion, transfer of knowledge and devotion from my advisor, Associate Professor Dr. Suphakant Phimoltares. I would like to say thank you for everything he has done for me. I really appreciate to be under his supervision.

In addition, I would like to show my gratitude to all thesis committee: Professor Chidchanok Lursinsap and Dr. Prem Junsawang for their invaluable comments and suggestions.

Moreover, I would like to express my deep gratitude to Pollution Control Department of Thailand (PCD) for all the valuable datasets, which are an indispensable part for this research.

Last but not least, I would like to extend thanks to my family, especially my mother who is always by my side, and friends for all help, support and encouragements.

Saksiri Lertnilkarn

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I
INTRODUCTION

1.1 Backgrounds and Rationales

Nowadays, air pollution is a serious problem affecting many things around the world. All kinds of life are inevitably faced with this situation. When air pollutants are in the respiratory system, they can cause health problems to humans and animals, or in the worst-case scenario, put them to untimely death. Plants grown in a poorly air-polluted environment cannot grow properly or become sick with ease. In addition, buildings and structures can be gradually damaged when exposed to air pollution for a long period of time.

Thailand has also faced an unavoidable problem caused by air pollution for several years. Particularly in the northern Thailand, this chronic issue results from many causes such as burn-off activities, forest fires, or basins making air pollution concentrated in the areas. When such activities all begin, they have resulted in the very high level of air pollution in the region making it reach the high ranking of the world. Consequently, Thai people have had to suffer from the situation inevitably. Thus, the problem has become an item in Thailand's national agenda [1] that has to be solved urgently.

At present, the measurement of the seriousness of air pollution depends on air quality index (AQI). Several countries use AQI for the purpose of study and as an alarm at the unpresented situation. In general, AQI is defined differently by country. It relies on the type and number of air pollutants combined to be the AQI compositions as each country defines. For instance, in the U.S., AQI is composed of two particulate matters (PM2.5 and PM10), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ground level ozone (O_3), and carbon monoxide (CO), while AQI in India consists of all the six pollutants combined with ammonia (NH_3) and lead (Pb),

Besides, there are several countries together with Thailand finding ways to cope with the severity of air pollution. Many pieces of research have been conducted to predict the AQI level. The use of machine learning techniques has been popular as

effective tools for dealing with this problem. So far, there have been many well-known machine learning algorithms such as random forest (RF), back propagation neural network (BPNN), or deep neural network (DNN) etc. However, algorithms like DNN even can give high accuracy results, but they are computationally expensive and time-consuming.

1.2 Research Objectives

1. To classify Thai AQI into classes accurately.
2. To build a classification model taking resources less than more complicated classification model such as DNN for classifying Thai AQI.

1.3 Scope of the work

There are two issues concerned in this research:

1. AQI data covers only the data derived from Pollution Control Department (PCD), Thailand.
2. Some missing data points must be filled by approximation.

1.4 Expected Outcomes

This research aims at coming up with a classification model for classifying Thai AQI into five classes namely excellent (blue), satisfactory (green), moderate (yellow), unhealthy (orange) and very unhealthy (red) according to the pollution control department. Therefore, this methodology will be able to help the public and private sectors in Thailand to handle the air pollution problem with effectiveness and low computational cost.

CHAPTER II
LITERATURE REVIEW

In this chapter, many related pieces of research are presented to display alternative methodologies for coming up with AQI prediction models. Some works resort to the models taking several computational resources like Deep Neural Network (DNN). Many depend on methods that are not computationally expensive. There are seven relevant pieces of research provided in the following paragraphs.

First, there was a study of air quality prediction based on the principles of bayesian network [2]. The model was proposed by Ruijun Yang et al. They appied the mentioned ideas of bayesian theorem and direct acyclic graph (DAG) to calculate probability distribution and relationship among nodes. The proposed framework was composed of three steps to complete. First, 70 percent of the training data was fed on the model to build the bayesian network graph and the node probability relationship and obtain a statistical result. Second, the validation data was entered into the trained model to produce a new classification and statistical results. Last, both of the results were compared and analyzed. The experimental results, finally, were compared among other comparative classification models such as naïve bayes classifier, or support vector machine. The proposed model could outperform the other models and yield the accuracy rate up to 99.3169%.

Second, there was another air quality prediction model developed in China. The technique was proposed by Wang Zhenghua and Tian Zhihui. They introduced a hybrid method for predicting AQI by using back propagation neural network (BPNN) and genetic algorithm (GA) [3]. The proposed model made use of GA to handle some disadvantages of BPPN. BPNN could well build complex non-linear models, but it was easy to fall into local minimum error and face slow convergence speed. On the other hand, GA equipped with the global search function was used to deal with the disadvantages of BPNN. The process was started by determining the network structure and initializing the weights and thresholds of BPNN. Then, the real coding was done

according to the initial weights and thresholds, and the coded results were supplied to the GA process. Next, the optimal weights and thresholds were determined and the error between the actual and predictive results was calculated. Last, BPNN updated weights and thresholds, and the error was compared to a setting value. The process was repeated to calculate the actual and predictive error again until the error was less than the setting value. Then, the model provided the final outputs. When the prediction results were compared with the use of single BPNN, the Improved neural network provided better outcomes both with the accuracy rate of 80.44% for AQI and 82.50% for air quality grade.

Third, a work of AQI prediction was implemented under the coding setting. The authors, Lloyd H. Macatangay and Rowell M. Hernandez, used Deep Neural Network (DNN) in their study [4]. The DNN was coded with the use of Scikit-learn library. In addition, apart from doing coding to run the model, a graphic user interface (GUI) was built to make the model become user-friendly. Eventually, the proposed model needed 60 training epochs, as well as 14 hidden layers to obtain 100 percent accuracy for the predicted results.

Forth, Usha Mahalingam et al., proposed an Air Quality Index forecasting technique to which support vector machine (SVM) and back propagation neural network (BPNN) were applied [5]. The purpose was to predict the large amount of air pollutant concentration in Delhi, India. They aimed to use BPNN to reduce the error and make use of SVM to provide a bridge from linearity to non-linearity. The proposed model was divided into two stages. At the first stage, after importing the collected data, BPNN was utilized to predict the future data. Next, in the second stage, the prediction results from the first stage were fed into SVM with several kernel functions which are Linear, Quadratic, Cubic, Fine gaussian, Medium gaussian, and Coarse gaussian functions. Afterwards, SVM with different kernel functions were tested to obtain the optimal accuracy result. Finally, Medium Gaussian SVM could provide the highest accuracy of 97.3%.

Fifth, a group of researchers, Divyam Madaan et al., presented an end-to-end adaptive system called the VayuAnukulani system to make a 24-hour prediction in advance for air quality data [6]. The authors took advantage of bidirectional LSTM network with attention mechanism (BiLSTM-A) to make a concentration prediction and a classification for the levels of pollution. There were two parts in their methodology—online and offline settings. Firstly, the offline setting was composed of data collection and data preprocessing which helped perform feature extraction from past data and handle the missing records. The obtained model in this part was used for making predictions in the next part. Secondly, for the online inference, air pollution data as the streaming data were applied to the cloud server. The server updated the real-time data on the cloud storage. The machine learning model was sent to the server to train and then, be stored back to the storage and make predictions. Eventually, the prediction result of the proposed model performed better than that of the other comparative conventional techniques by 7 – 18 percent accuracies.

Sixth, a PM10 prediction model was developed by Nicolas Mejia Martinez et al. by using data in Columbia. The proposed model modified random forest by a voting method [7]. At the beginning, they designed an approach for choosing variable sets by relying on the criteria of experts, backward elimination, and forward selection. Additionally, the evaluation was done and analyzed by using three general models–random forests (RF), classification and regression trees (CART), and logistic regression (LR). The performance was compared in terms of accuracy, sensitivity, and specificity. Accuracy and specificity were considered as critical factors. As the model required the alert of high PM10 concentration levels, high sensibility became less important. In addition, the accuracy results of each model were very close to one another's results. Therefore, specificity was mainly used as the estimator, and RF with expert criteria could give the highest specificity result. After that, the authors modified RF by increasing the proportion of votes by 2% for every iteration for the first class to be selected. Finally, RF modified by voting values as the proposed model provided the most satisfactory performance of decreasing error rates.

Last, in Thailand there was a study focusing on a PM10 forecasting model. The proposed model was put forward by Chadaphim Photphanloet and Rajalida Lipikorn. They developed the prediction model with the use of genetic algorithm (GA), supervised learning neural network, and modified depth-first search algorithm (MDFS) [8]. They tried to use GA to pick only related features as GA always provided the optimal selection. In the process of GA, they selected three machine learning algorithms– Multiple linear regression (MLR), Multilayer perceptron neural network (MLP), and Support vector regression (SVR) as the fitness functions. Then, the datasets based on chosen features were sent to MLP for the first round of prediction. Next, they used MDFS to select monitoring sensors within a particular range and obtained the new data from them. The new datasets were supplemented to the existing data and the prediction process began again. The process iterated until there was no monitoring station within 25 km. The experimental results were compared among three settings namely using only three supervised learning algorithms, integrating GA with the supervised learning algorithms to make a prediction, and their proposed model, applying GA, three supervised learning models and MFDS to the prediction. Eventually, their proposed model gained the optimal results in terms of RMSE and R for predicting PM10 concentration 1 hour ahead at each monitoring station.

According to the aforementioned works, two important issues should be pondered about. Firstly, some more accurate models were liable to consume a huge number of data and computational resources. Secondly, there were several studies centering on a few hour-ahead predictions. Nevertheless, in real practice it was not proper to use them as people needed a portion of time to prepare themselves for the upcoming situation. Although some models were able to forecast the pollution levels for multiple hours ahead, they were also computationally expensive. This study aims to create a highly accurate air quality prediction model which certainly takes low computational cost and is not data intensive.

CHAPTER III
THEORETICAL BACKGROUNDS

In this study, the proposed model provided AQI prediction based on a set of selected classification algorithms. To come up with the AQI prediction model, first, the backgrounds of AQI should be understood in terms of the AQI's compositions, a formula used for calculating it, and how to perform the preprocessing stage. Then, to produce the prediction results, all 5 classification algorithms namely K-nearest neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) are introduced as the predictive tools.

3.1 AQI and Data preprocessing

3.1.1 AQI

In Thailand, the composition of AQI is composed of six kinds of pollutants– Sulfur Dioxide (SO2), Carbon Monoxide (CO), Ozone (O3), Nitrogen Dioxide (NO2), and two Particulate Matters (PM10, and PM2.5). Furthermore, the AQI is divided into 5 categories namely excellent (blue), satisfactory (green), moderate (yellow), unhealthy (orange) and very unhealthy (red). They are ranged from 0 to 201 and above.

In addition, to obtain AQI, it needs to be calculated by equation (1)

$$A = \frac{A_{max} - A_{min}}{C_{max} - C_{min}} (C - C_{min}) + A_{min} \quad (1)$$

where A : a calculated AQI corresponding to a given concentration C

C : a given concentration of an air pollutant from the measurement

C_{max} : a maximum concentration in a specific range

C_{min} : a minimum concentration in a specific range

A_{max} : a maximum AQIs corresponding to C_{max}

A_{min} : a minimum AQIs corresponding to C_{min}

Nonetheless, more than 50 percent of data from all monitoring stations do not have CO data, for the sake of consistency, this experiment chose to focus only on the data of the five pollutants instead.

3.1.2 Data preprocessing

To make the datasets become suitable for the experiment, the data preprocessing was implemented. A two-step process was performed with the original data to make it suitable for calculating air quality index and its levels on all stages of this experiment.

Firstly, to begin with the unprocessed data, for handling the values that disappeared, the next values were used to patch on them. Next, in the patched datasets for every record excluding 23 records from the beginning in every feature (PM2.5, PM10, O_3, SO_2, NO_2 and CO), the principle of 24-hour moving average was used to quantify C (the concentration of a pollutant). Then, the formula in equation (1) was used to calculate, for each feature, the AQI in the same period and its category. In addition, for all records, category features were converted into binary representation as shown in Table I before fed into all machine learning algorithms. After that, the processed data in the binary representation form were used to estimate the optimal results among simple classification algorithms.

Table 1: AQI category, color, and binary representation

AQI	Category	Daily AQI Color	Binary Representation
0 - 25	Excellent	Blue	10000
26 - 50	Satisfactory	Green	01000
51 - 100	Moderate	Yellow	00100
100-200	Unhealthy	Orange	00010
>200	Very Unhealthy	Red	00001

Secondly, when the group of classification models with the optimal results was selected, the calculation for a group of averaged air quality index on multiples of four hours, which are 4, 8, 12, 16, …, and 48-hours average bases was

performed with the processed data from the previous step. They were used to prepare for making predictions on several hours in advance. Finally, to build the ensemble model and assess the efficacy, the recently processed group of averaged air quality index was used for the last phase of the experimental process.

Additionally, according to the PCD, there is no upper bound for the last AQI level (Very Unhealthy level) in which AQI is greater than 200 and for the final concentration extent of all pollutants. Hence, in this research, the other four ranges of air quality index and concentrations of all pollutants are averaged to create the range for the upper bound of the mentioned categories. All the illustrations and details of all classifiers applied to constructing the ensemble model for this study were explained in the next section.

3.2 Classification Models

On a global scale, many classification models were used to classify AQI for several pieces of research. On one hand, some models were capable of producing results with very high accuracies and forecasting several hours ahead but consuming a number of computational resources. On the other hand, many took low computational cost, but could make a prediction simply a few hours ahead. However, in the study, the purpose was to take a balanced position between the two extremes. Five popular classification models-- K-Nearest Neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) were chosen as tools for the purpose of comparison and selection. They were used in the first stage of this study. Accuracy results of the classification algorithms under study were set as a baseline when the models were tested and used to build the proper structure of the proposed model in the second stage. The details and illustration of all classification techniques are explained below:

3.2.1 K-Nearest Neighbor (KNN)

KNN technique uses the concept of classifying datapoints based on the proximity of a targeted data point to k-nearest data points. It assumes that the similar

things lie near to each other. This method counts the number of the nearest data points according to the values of k around a targeted data point and assigns the class labels of the data points that appears most to the targeted data point. In order words, the technique relies on the notion of majority vote for providing class labels for a particular data point.

As shown in Figure 1, the graph displays data with two classes. The stars and the squares show different classes of data points. In addition, they are classified by two k values where the solid line and the dot line represents k = 3 and k = 5 respectively.

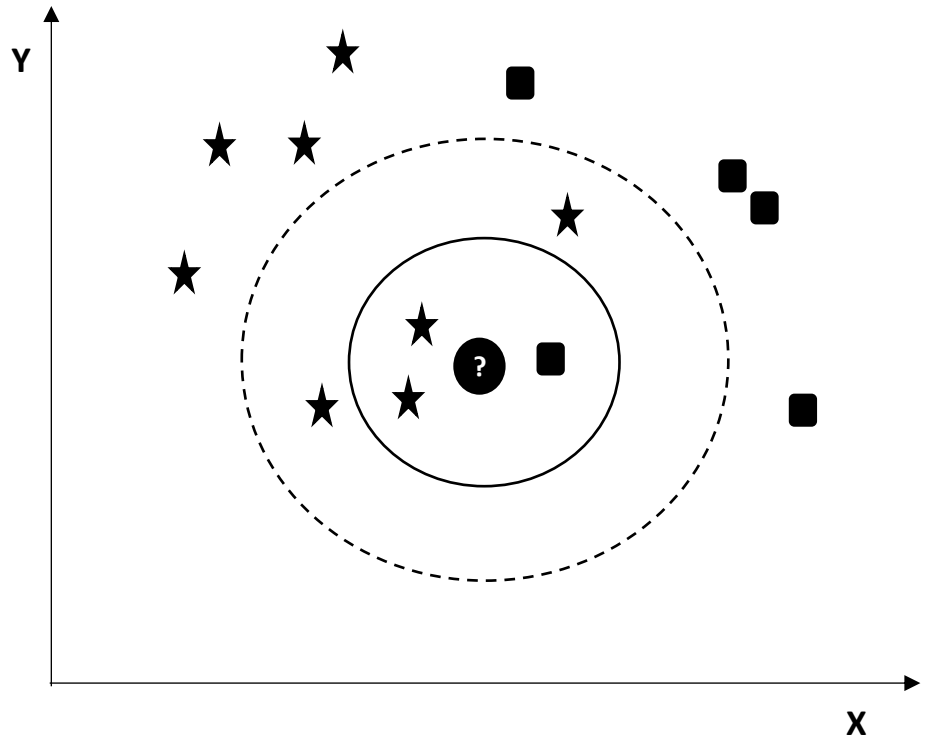


Figure 1: KNN visualization of a given data point and the data points from two classes.

Furthermore, when searching for the close proximity, KNN calculates the distances between a particular data point and the other data points. There are several

techniques used for calculating the distances such as Euclidean distance, Manhattan distance, or Minkowski distance etc.

The number of errors from the model can be reduced by running KNN algorithm many times with different k values and selecting the right k value which provides the minimum error. In this study, k value is set to 1 in every case.

3.2.2 Naïve Bayes (NB)

Naïve Bayes is a simple probabilistic classifier based on Bayes' Theorem. It is a fast and simple machine learning model. NB is often used as an approximate baseline for classification issues as it is very fast and has few parameters to be tuned making it proper to implementation on high dimensional data. NB can work under the assumption that each feature is independent.

Bayes' Theorem in this study as stated in equation (2) is used to find the conditional probability of a data record belonging to a class.

$$P(c|T) = \frac{P(T|c)P(c)}{P(T)} \quad (2)$$

where

$P(c)$: the probability of class c being true (in spite of the data).

$P(T)$: the probability of the data (in spite of the class).

$P(c|T)$: the probability of class c given the data T.

$P(T|c)$: the probability of data T given that class c is true.

In this research. It depends on Mean, Standard Deviation and Normal Distribution for coming up with likelihood of each AQI class for a data record.

3.2.3 Random Forest (RF)

To understand Random Forest, the idea of a decision tree is worth talking about first as the decision tree is the fundamental and the main structure of RF. Besides, bagging technique should be elaborated as the method for finding the final solution

among a number of decision trees. Then, RF is eventually introduced for the whole illustration.

A decision tree is a kind of supervised machine learning. It is designed for both classification and regression problems. Its main structure is a tree structure which is composed of a root node, inner nodes or decision nodes, and leaf nodes. From the root node, it is followed by several decision nodes to which a question is attached. From a series of decision nodes flow decision nodes themselves or leaf nodes that provide results or alternative answers.

In Figure 2, It shows an example of a decision tree. This tree has two decision nodes and three leaf nodes. AQI is separated by the first decision node that is acted a root of this tree with the value of 200. For the value of AQI which is greater than 200, it goes to a leaf node. Besides, for AQI that is less than or equal to 200, it moves to the second decision node. At this node, AQI is spilt by the value of 25, where it goes to either left leaf node or right leaf node.

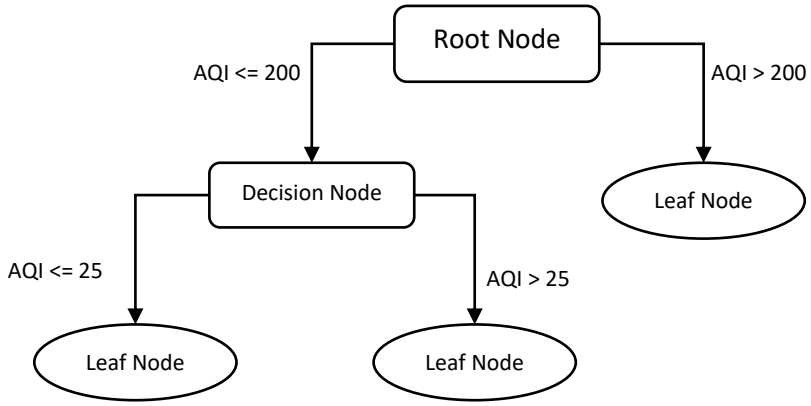


Figure 2: An Example of a decision tree

Regarding the bagging technique, it is a type of ensemble methods. This technique depends on homogeneous weak learner models. Each weak learner parallelly gives an independent prediction. The final outcome is combined from all weak learners' results by the process of averaging or majority vote. For RF, each decision tree renders

independent prediction and all of them are assessed by the majority vote process to provide the final answer. In this study, RF is set to have 100 trees.

Random Forest is a supervised learning algorithm. Its structure consists of many individual decision trees that are operated with the bagging technique. Each tree provides its independent class prediction. The class which wins the majority voting can be the class solution of the whole model.

As shown in Figure 3, the structure of random forest consists of three decision trees. Each tree is provided with the same dataset and gives its independent prediction result. All prediction results are evaluated by the majority vote to yield the final prediction result of the whole model.

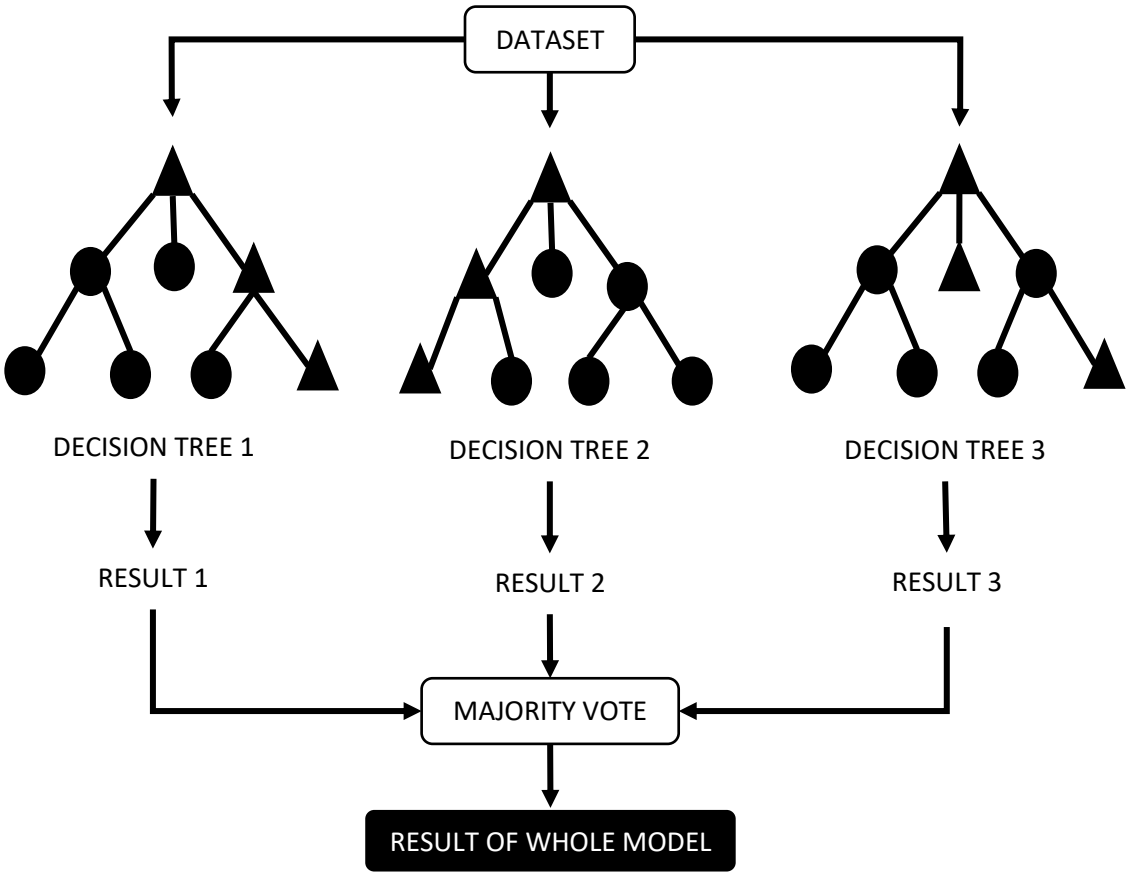


Figure 3: An illustration of Random Forest's structure

3.2.4 Support Vector Machine (SVM)

Support Vector Machine is one of the most widely used supervised learning algorithms. Its objective is to create a hyperplane that helps divide data points into classes in an N-dimensional space where N is the number of features. The optimal hyperplane should clearly maximize the distance of data points of each class so that the new data points can be easily placed into the right category.

To elaborate on hyperplanes, as shown in Figure 4, there are a decision boundary (the solid line) which help segregate data points into classes. The data points that are placed on different sides of the hyperplane can be labeled different class names. Support vectors are the data points that lie nearest to the hyperplane and affect how the hyperplane should be positioned. Deleting support vectors also changes the position of the hyperplane.

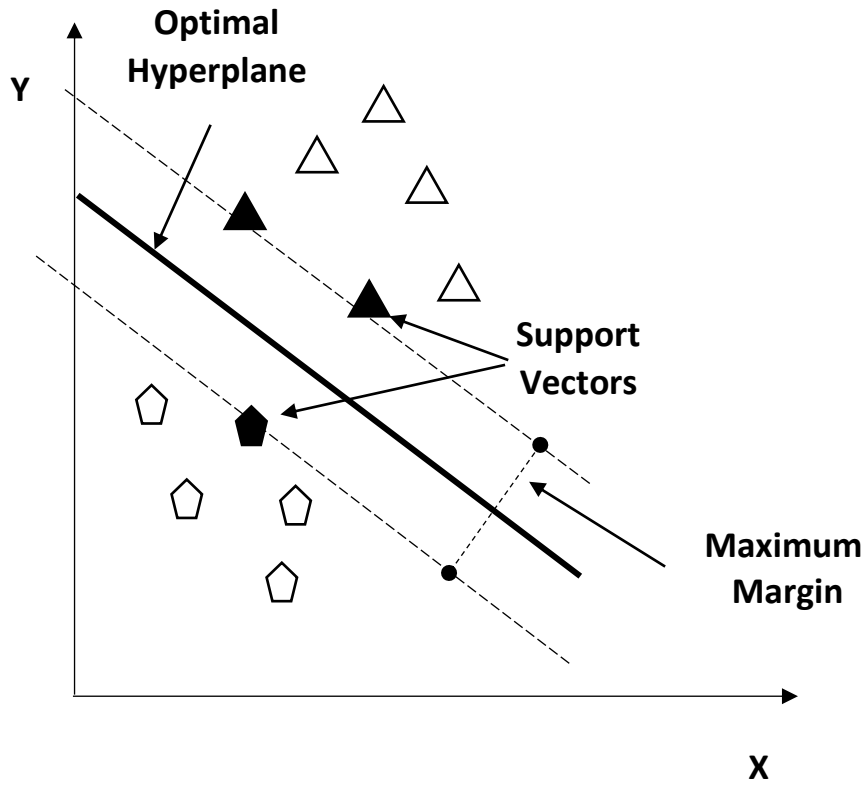


Figure 4: SVM visualization of a given data point and the data points from two classes

3.2.5 Multi-layer Perceptron (MLP)

Multi-layer Perceptron is a sort of feedforward machine learning algorithms. It is the network which its input and output form non-linear mapping. The MLP's structure comprises three main layers—one input layer, one or more hidden layers, and one output layer. Each layer contains neurons. The neuron consists of an activation function such as Sigmoid function.

To begin with the input layer, it is the first layer of neurons. Its duties are to receive the data and send it to the next layer for data processing. In the hidden layer, it can have one or more layers. The node in a hidden layer calculates weighted sum derived from the values fed from the previous layer and their corresponding weights. Then, the weighted sum is transformed by the activation function. Each linear combined result is propagated to the next layer. This propagated process is repeatedly continued from the hidden layers to the output layer. The example in Figure 5 illustrates the structure of MLP which uses only one hidden layer,

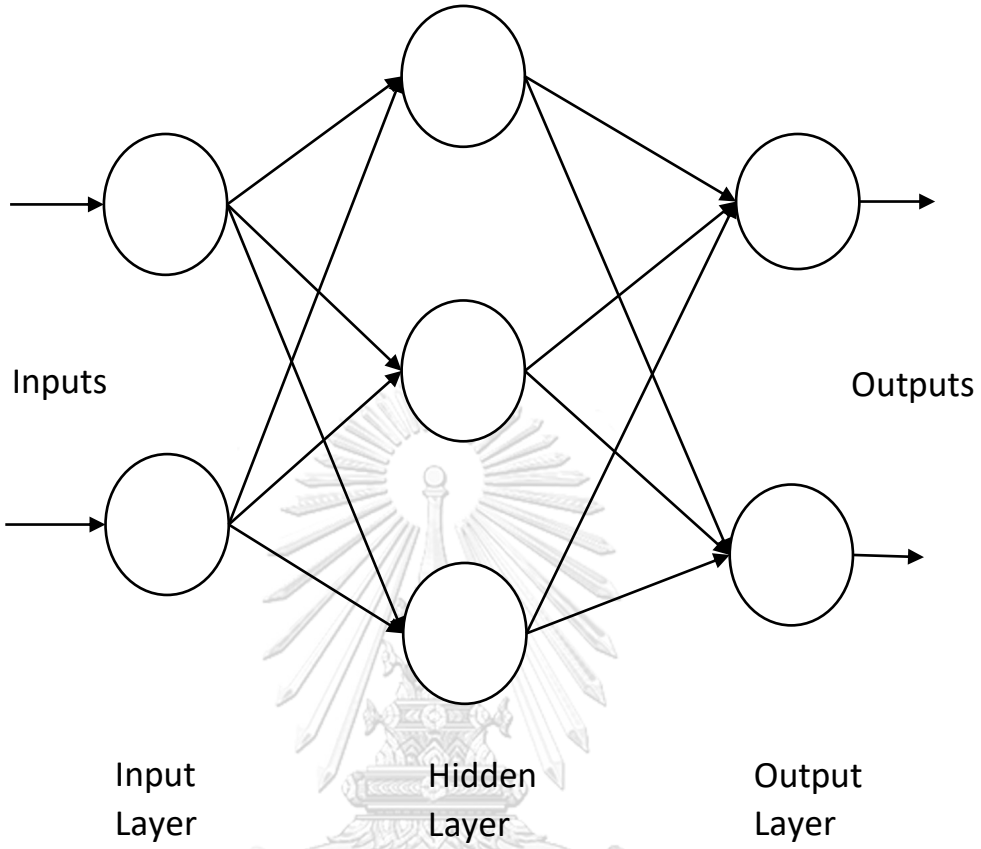


Figure 5: An example of Multi-layer Perceptron structure

This chapter gives a brief introduction of all five machine learning algorithms involved in this experiment. The details and explanation of how the ensemble model is designed and constructed are described in the chapter IV.

CHAPTER IV
PROPOSED METHODOLOGY

In this research, the proposed methodology aimed at making the prediction results become more accurate. Besides, it was designed to build the classification model taking resources less than more complicated classification models such as DNN. The general way was to find a proper combination of classification algorithms to increase the effectiveness of the prediction accuracy. Hence, an ensemble model was proposed to deal with the challenge.

The concept of the ensemble model in this research was to train many classification algorithms to make a prediction independently. Furthermore, the whole model could yield the final outcomes with the use of the majority vote of each component model. In addition, there was no the same kind of component classification models combined.

In this experiment, the proposed ensemble model consisted of three out of five tested classification models. Additionally, all classification models provided the weight equally to the final accuracy results. The same sub-models were not allowed to be implemented in this study. Moreover, all selected classification models were considered on the comparative results of the performance of each tested classification model.

There were two stages built for the proposed methodology. In the first stage, the approach was constructed on two feature types—AQI and categories. It aimed to come up with the optimal classification models to be part of the ensemble model and find the optimal structure for the ensemble model. Data used in this stage came from all seven monitoring stations.

When all five classification algorithms were trained and tested, the results showed on average that KNN, RF and SVM individually gave better accuracy results than NB and MLP did. Furthermore, after all the five comparative models were tested with the two structures of the ensemble model which were sets of five and three classification models as shown in Figure 6-7 respectively, as a whole, the combination of

KNN, RF and SVM yielded the accuracy results better than that of KNN, RF, SVM, NB and MLP combined together. Therefore, when considering the comparative result, the set of three classification algorithms namely KNN, RF and SVM was selected for the next stage of this study.

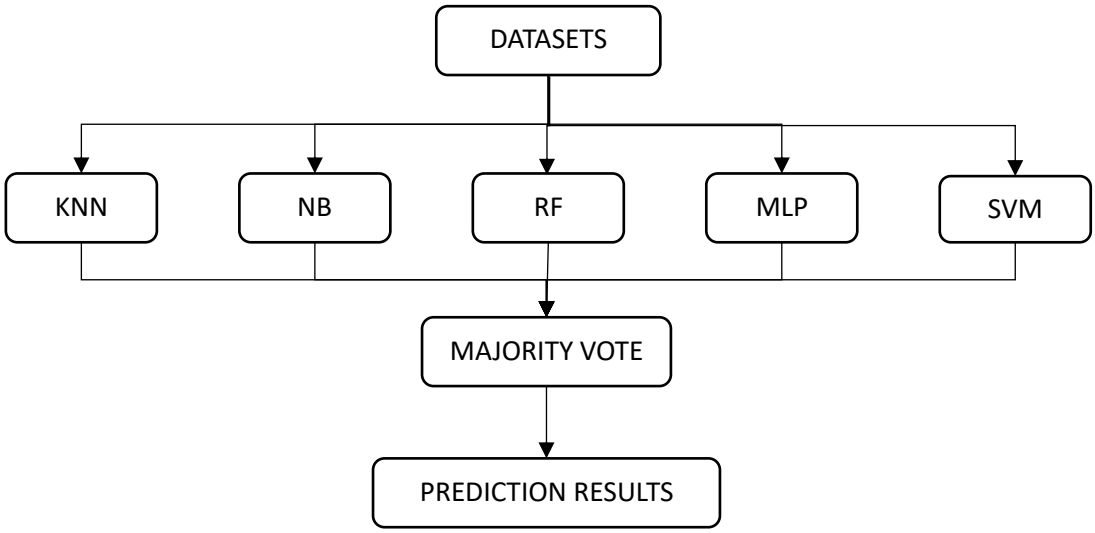


Figure 6: Structure of ensemble model based on KNN, NB, RF, MLP and SVM

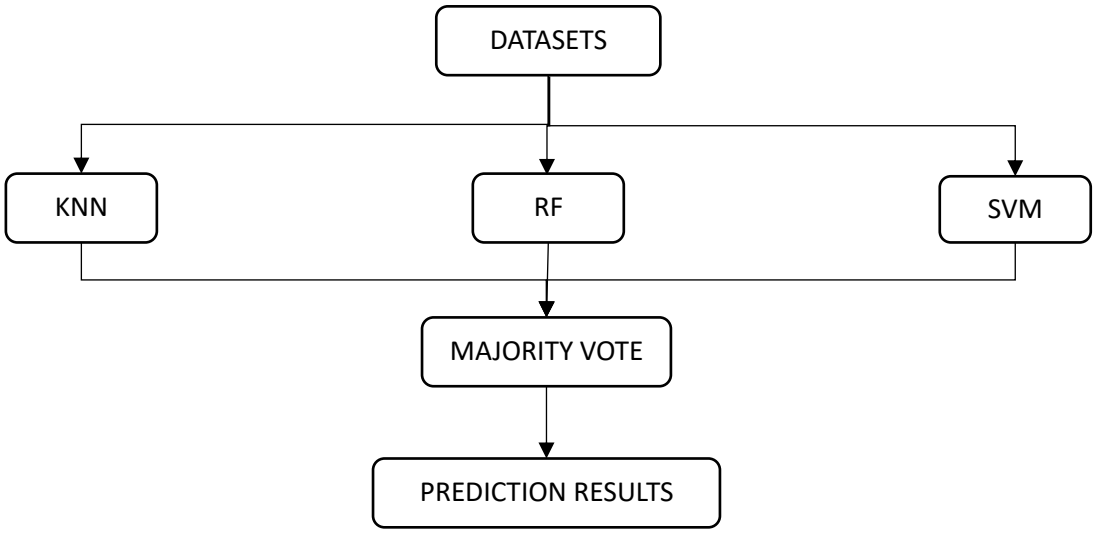


Figure 7: Structure of ensemble model based on KNN, RF and SVM

All the three chosen classification algorithms, in the second stage, were used to train the ensemble model. All processed datasets based on multiple-hours average bases, as displayed in Figure 8, were fed into all the three selected models. Eventually, each classification model provided its independent prediction result based on multiple-hours ahead prediction, as illustrated in Figure 9, and all the results were evaluated by the majority vote to forecast the final category of AQI. The details and explanation of how each classification model is set or adjusted, the experimental results, and analytical perspectives including discussion are described in chapter V.

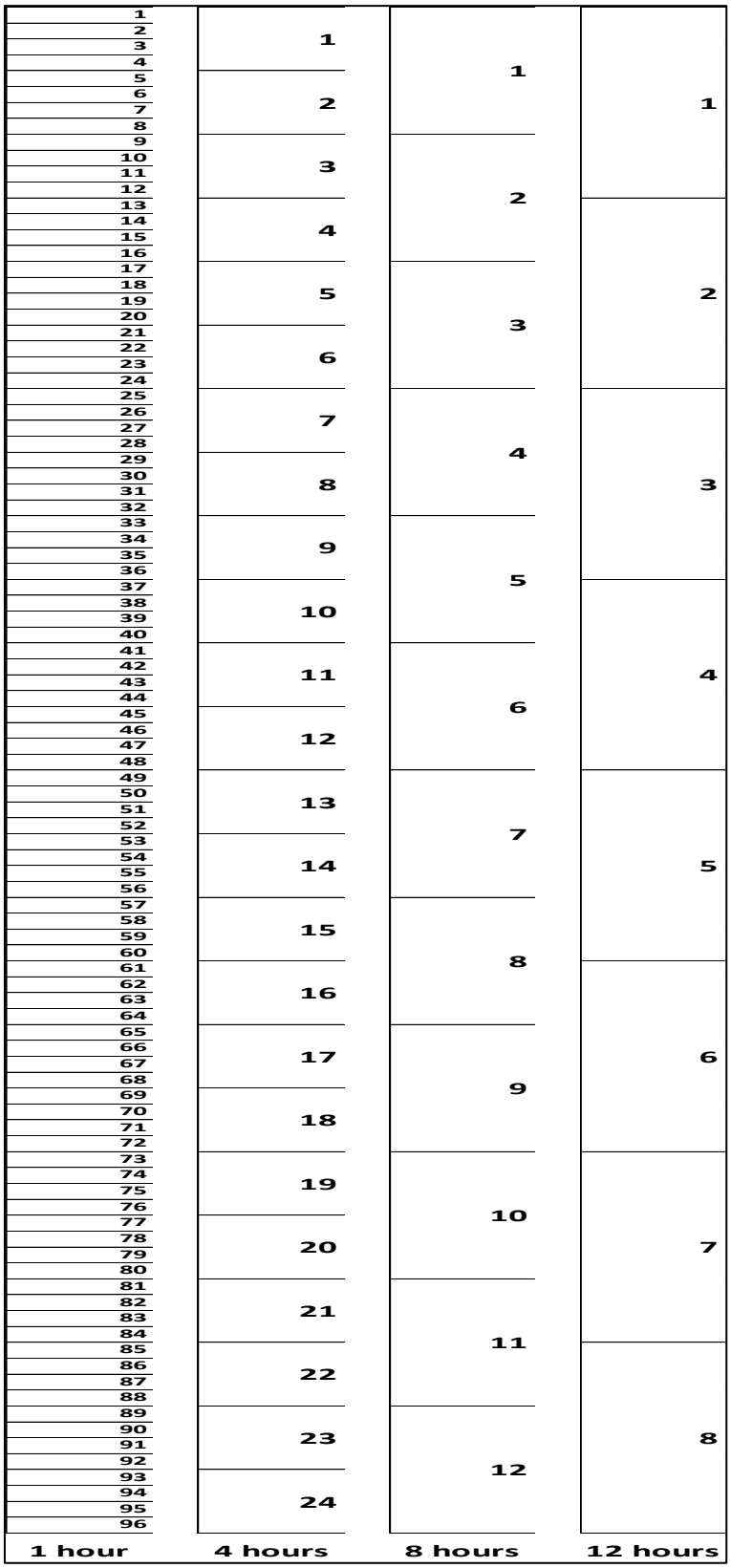


Figure 8: An illustration of multiple-hours average as inputs for 1 hour, 4 hours, 8 hours, and 12 hours

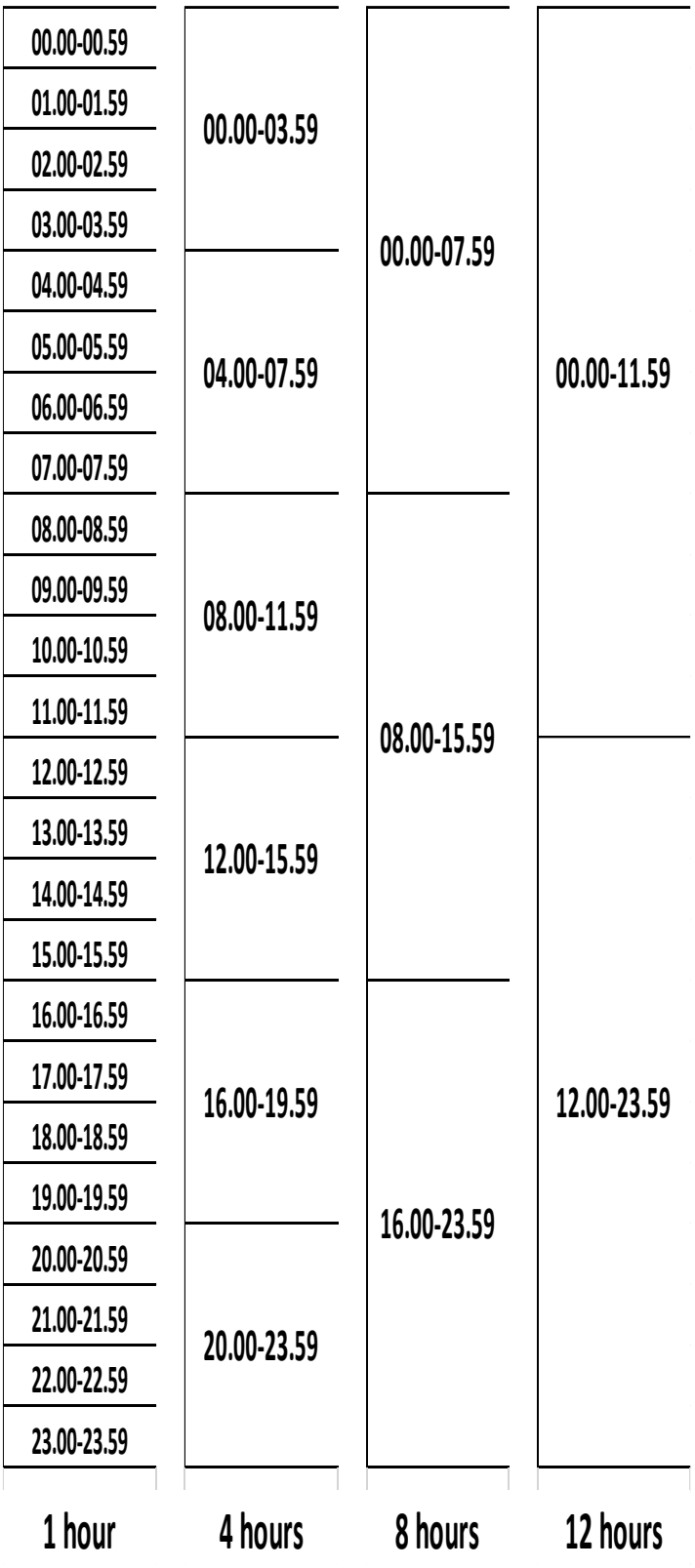
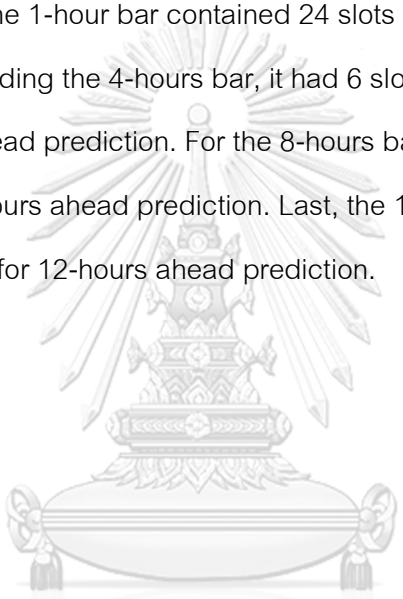


Figure 9: An illustration of multiple-hours ahead prediction as outputs in one day

As for Figure 8, the figure represented timeline of inputs in the form of multiple-hours average. The 1-hour bar showed 96 slots of time. Each slot represented 1 hour. For the 4-hours bar, there were 24 slots of time. Every slot stood for 4-hours average. In the 8-hours bar, it illustrated 12 slots of time. Each slot exhibited 8-hours average. Last, the 12-hours bar displayed 8 slots of time. All of them presented 12-hours average.

As for Figure 9, the figure showed timeline of outputs as multiple-hours ahead prediction in one day. The 1-hour bar contained 24 slots of time. Each slot meant 1 hour ahead prediction. Regarding the 4-hours bar, it had 6 slots of time. Each slot represented 4-hours ahead prediction. For the 8-hours bar, there were 3 slots of time. Each slot displayed 8-hours ahead prediction. Last, the 12-hours bar possessed 2 slots of time. Each one stood for 12-hours ahead prediction.



CHAPTER V
EXPERIMENTS

In this chapter, all components of the experiments, namely experimental setup, results, as well as analysis and discussion were described. The experimental setup part talked about information of the datasets and how to set up all classification models in this research. In the results section, experimental results from the beginning to the end of the experiments were explained. Additionally, the analysis and discussion part presented some interesting observations related to the study.

5.1 Experimental Setup

There are two important issues that are worth taking about for the experimental setup in this research, namely datasets and parameter setup. The datasets sub-section elaborates on aspects of the datasets. On the other hand, in the parameter setup sub-section, it focused on how to set and adjust all classification algorithms.

5.1.1 Datasets

All datasets used in this study were supplied by Pollution Control Department of Thailand (PCD). All the datasets were gathered from seven monitoring stations from four provinces in the northern part of Thailand namely Chiang Mai, Lampang, Chiang Rai, and Phrae. There was one station in Chiang Mai (M1), four stations in Lampang (M2, M3, M4 and M5), one station in Chiang Rai (M6), and one station in Phare (M7). The datasets comprised the information of concentration of six air pollutants which were features and used for AQI calculation. All the six features consisted of two Particulate Matters (PM10, and PM2.5), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Ozone (O3). Furthermore, all six-pollutant data, were recorded on the hourly basis from Jan 1st, 2018 at 1:00 to Feb 28th, 2021 at 24.00 and had 27720 records. All the datasets were processed, as aforementioned, according to suitability for the experiments.

As for data partitioning, all seven datasets were separated by the percentage splitting method. They were divided into 70% for the training sets and 30%

for the test sets. Besides, the class distribution of each monitoring station could be represented in Table 2-8.

Table 2: Class distribution of M1

Class	Records	Ratio
Excellent	13671	49%
Satisfactory	5931	21%
Moderate	2749	10%
Unhealthy	4014	14%
Very unhealthy	1332	5%
Total	**27697**	**100%**

Table 3: Class distribution of M2

Class	Records	Ratio
Excellent	14856	54%
Satisfactory	4784	17%
Moderate	3272	12%
Unhealthy	4001	14%
Very unhealthy	784	3%
Total	**27697**	**100%**

Table 4: Class distribution of M3

Class	Records	Ratio
Excellent	17337	63%
Satisfactory	3877	14%
Moderate	3120	11%
Unhealthy	2919	11%
Very unhealthy	444	2%
Total	**27697**	**100%**

Table 5: Class distribution of M4

Class	Records	Ratio
Excellent	18353	66%
Satisfactory	2999	11%
Moderate	2299	8%
Unhealthy	3260	12%
Very unhealthy	786	3%
Total	**27697**	**100%**

Table 6: Class distribution of M5

Class	Records	Ratio
Excellent	15169	55%
Satisfactory	4194	15%
Moderate	3144	11%
Unhealthy	4663	17%
Very unhealthy	527	2%
Total	**27697**	**100%**

Table 7: Class distribution of M6

Class	Records	Ratio
Excellent	17220	62%
Satisfactory	4648	17%
Moderate	2374	9%
Unhealthy	2195	8%
Very unhealthy	1260	5%
Total	**27697**	**100%**

Table 8: Class distribution of M7

Class	Records	Ratio
Excellent	15284	55%
Satisfactory	4858	18%
Moderate	3600	13%
Unhealthy	3245	12%
Very unhealthy	710	3%
Total	**27697**	**100%**

Regarding the ratios of Tables 2-8, the classes possessing the top three highest percentage took up 84 percent or more for each table namely Excellent (49%), Satisfactory (21%) and Unhealthy (14%) for M1, Excellent (54%), Satisfactory (17%) and Unhealthy (14%) for M2, Excellent (63%), Satisfactory (14%) and Moderate equal to Unhealthy (11%) for M3, Excellent (66%), Unhealthy (12%) and Satisfactory (11%) for M4, Excellent (55%), Unhealthy (17%) and Satisfactory (15%) for M5, Excellent (62%), Satisfactory (17%) and Moderate (9%) for M6, as well as Excellent (55%), Satisfactory (18%) and Moderate (13%) for M7.

All the datasets of seven monitoring stations from the period that was mentioned above were applied to testing the proposed model and all five comparative

classification models. This can be practical for explicating data sufficiency and creating an efficient model.

5.1.2 Tool Setup

The experiment was made on a platform called WEKA (Waikato Environment for Knowledge Analysis), a data mining tool developed by The University of Waikato on Intel(R) Core (TM) i5-8250U CPU @ processor base frequency of 1.60GHz; configurable TDP-up base frequency of 1.80 GHz.

In addition, to achieve the optimal results of the proposed model in this research. Some parameter adjustments for classification algorithms needed to be made before the model training. For KNN, k-value was established at 1. RF was based on 100 decision trees. The activation function for MLP was Sigmoid function. Additionally, An adjustment on some parameters of SVM helped make it become more competent on the comparison stage. After fine-tuning, the optimal key parameters were set to RBF Kernel, gamma = 5 and c = 150. These parameter adjustments were prepared for making SVM become superior to NB and MLP. After that, SVM with the adjusted parameters was beneficial to the ensemble model to come up with the final results.

5.2 Results

At the beginning of the experiments, the datasets from all seven monitoring stations were applied to training and testing all five simple classification algorithms to predict AQI category at the next hour. The experimental results of this phase were shown in Table 9.

Table 9: Accuracy results for comparison of single classification models

Station	Accuracy (%)				
	Classification Models				
	SVM	MLP	KNN	RF	NB
M1	99.398	97.112	99.976	99.904	97.617
M2	99.711	96.739	99.904	99.904	99.013
M3	99.567	97.015	99.976	99.976	99.374
M4	99.699	97.870	99.964	99.964	98.809
M5	99.795	97.713	99.928	99.952	98.484
M6	99.410	97.124	99.964	99.964	97.280
M7	99.591	97.039	99.976	99.976	98.688
Average	**99.596**	**97.230**	**99.955**	**99.948**	**98.466**

Based on the average values, the accuracy results of SVM, MLP, KNN, RF, and NB were at 99.596%, 97230%, 99.955%, 99.948%, and 98.466% respectively. It showed that KNN, RF, and SVM as a single model delivered greater performance than NB and MLP did. As a result, this study was continued to search for the optimal structure for the proposed ensemble model under two scenarios. The first scenario was based on three classification models namely KNN, RF and SVM. For the second scenario, the structure depended on all the five classification algorithms combined. The accuracy results of the models under both scenarios were compared and evaluated with each other.

In Table 10, under the experiment of two mentioned scenarios, on only the next-hour basis and on average, the combination of three classification models yielded superior performance to that of all the five classification algorithms with the accuracies at 99.955% and 99.715% respectively. Therefore, as the combination of KNN, RF, and SVM provided better performance, it would be used as the structure of the final ensemble model in the next phase.

Table 10: Accuracy results from the combination of three and five classification algorithms

Station	Accuracy (%)	
	Five classification models	Three classification models
M1	99.434	99.976
M2	99.711	99.904
M3	99.976	99.976
M4	99.880	99.964
M5	99.832	99.928
M6	99.519	99.964
M7	99.651	99.976
Average	**99.715**	**99.955**

To construct the final proposed model, all three selected classifiers, in the last phase, were fused together. The processed data sets from each monitoring station were prepared on the mentioned multiples of four-hour bases and fed onto all selected classifiers to train and test the ensemble model.

Next, after the model was trained and tested, the last outcomes of the ensemble model were evaluated. The majority vote principle was applied to the final outcome of all individual classification models. In Figures 10 – 16, The curves of accuracy regarding multiple-hours average for every station were given. The X axis represented the hour average on multiples of four-hours bases from 1 hour to 48 hours average.

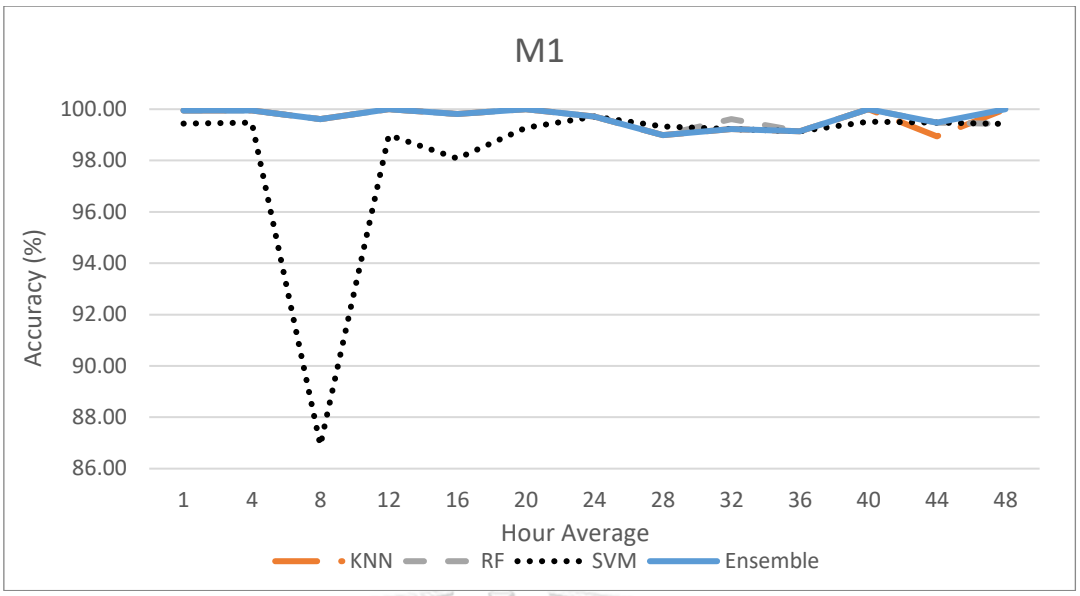


Figure 10: Accuracy curve with regard to hour average for M1

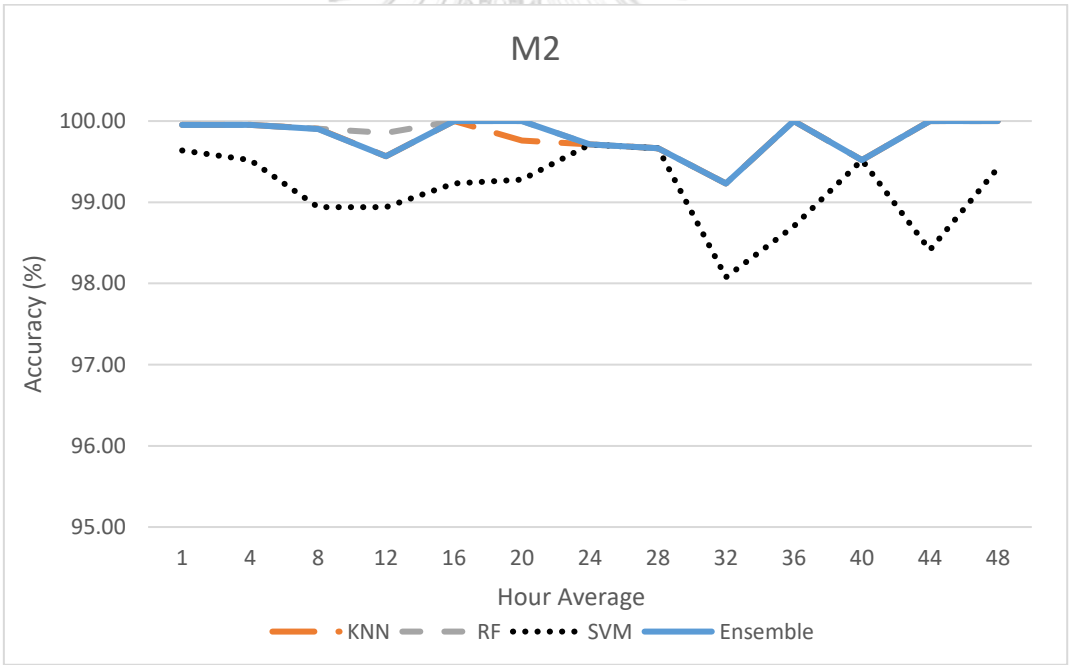


Figure 11: Accuracy curve with regard to hour average for M2

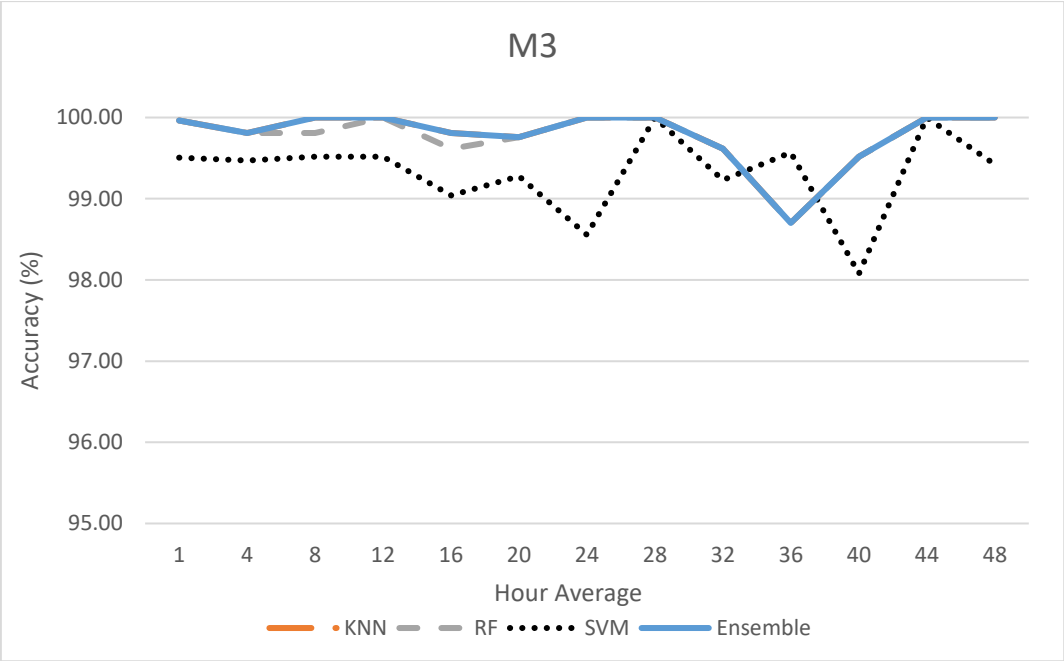


Figure 12: Accuracy curve with regard to hour average for M3

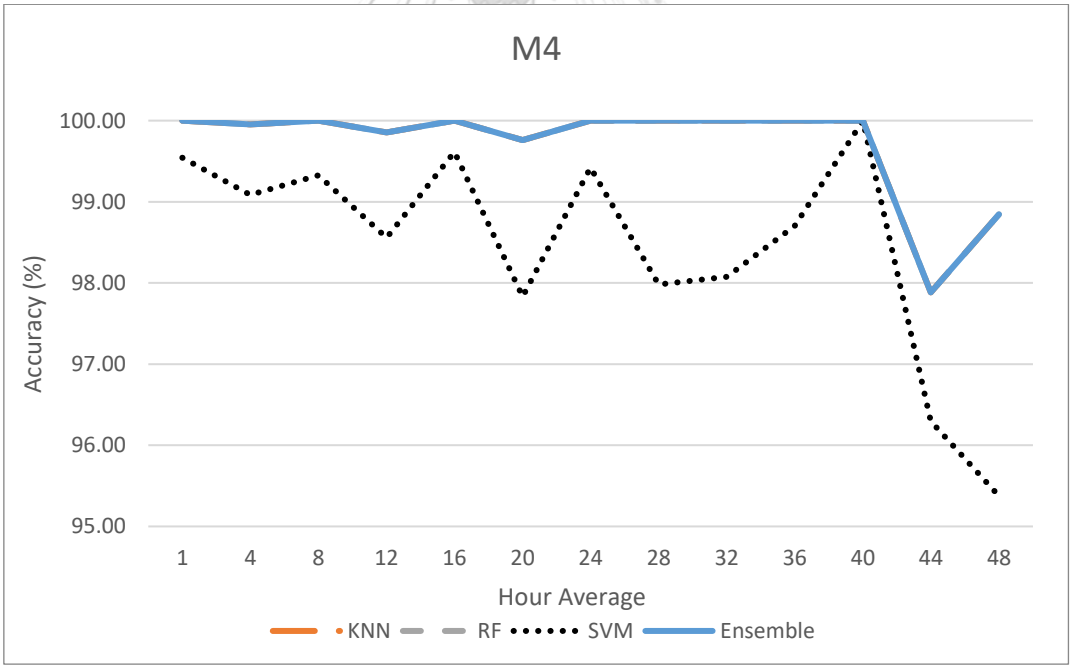


Figure 13: Accuracy curve with regard to hour average for M4

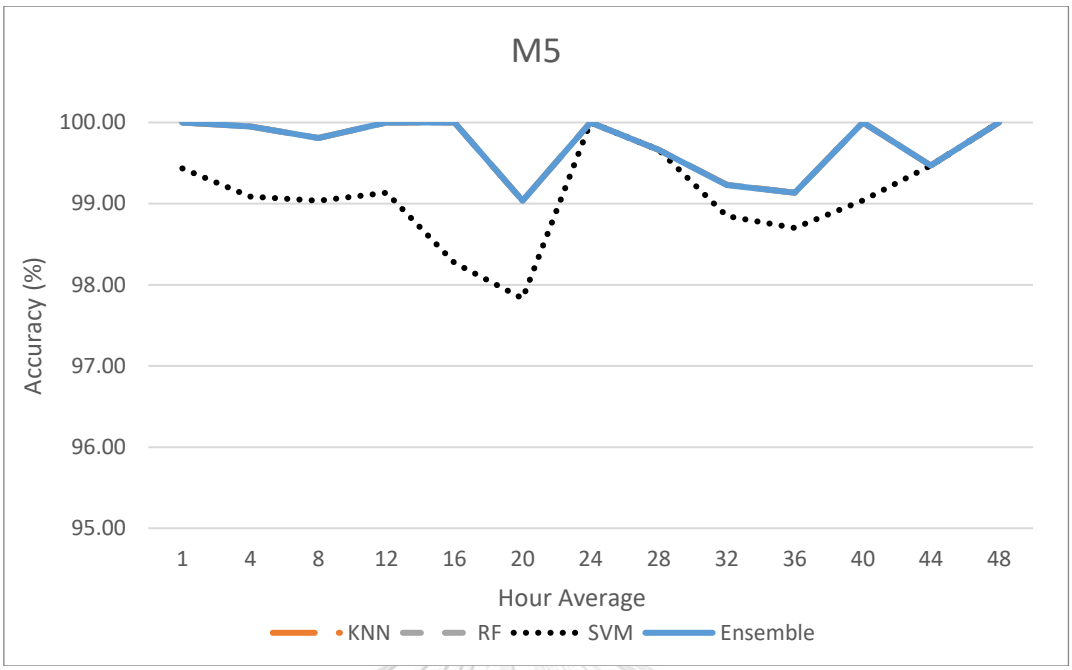


Figure 14: Accuracy curve with regard to hour average for M5

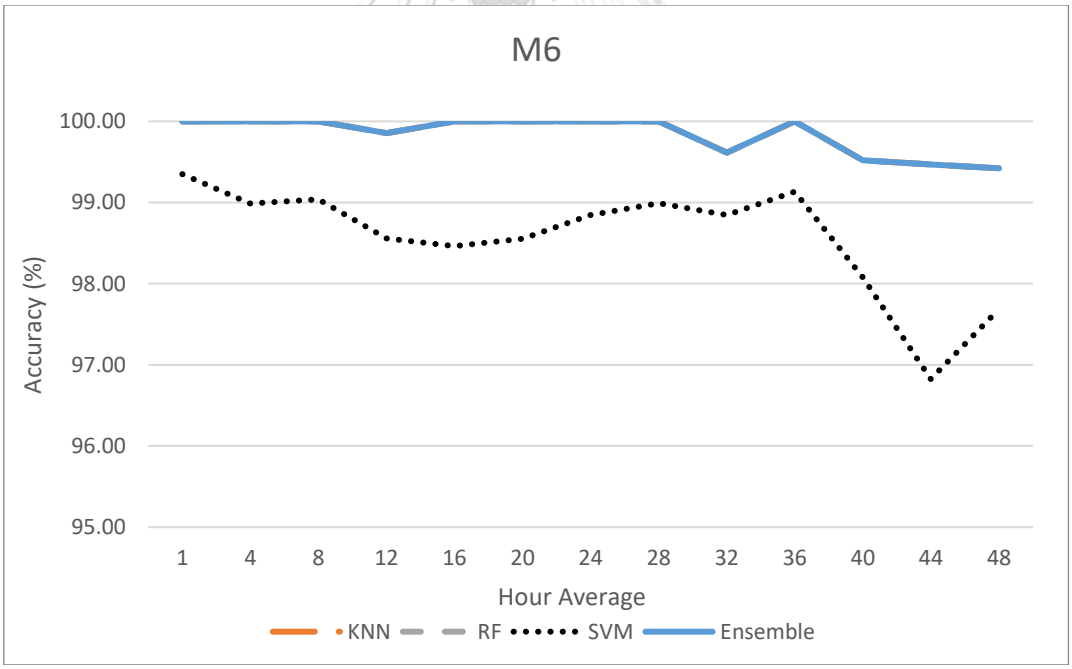


Figure 15: Accuracy curve with regard to hour average for M6

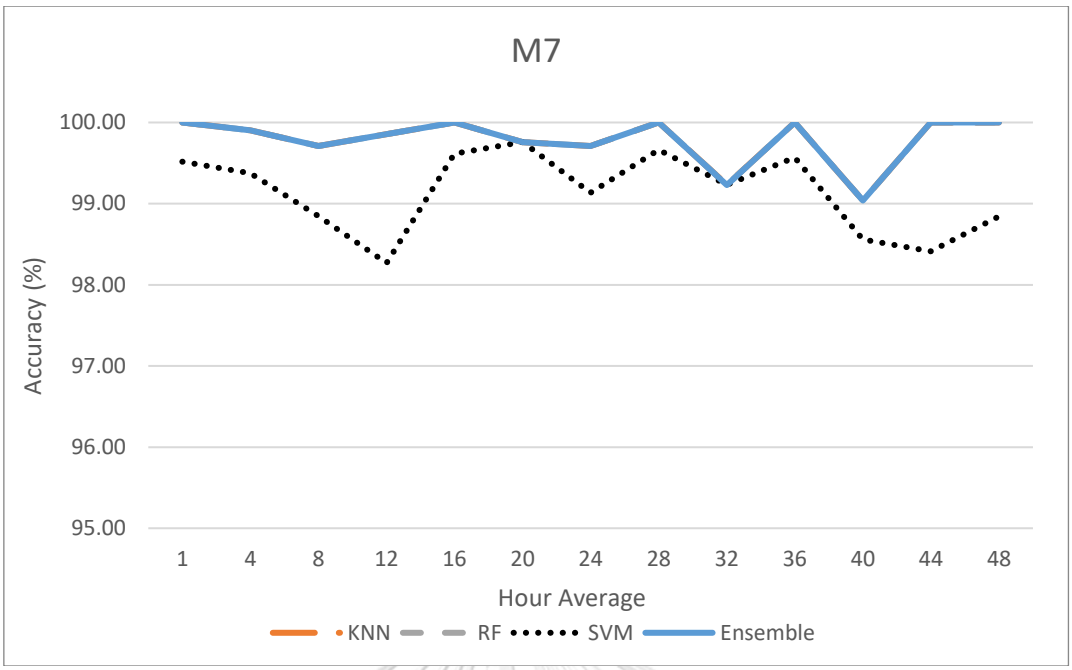


Figure 16: Accuracy curve with regard to hour average for M7

When considering Figures 10 – 16, datasets of M4, M5, M6, and M7 could produce the accuracy rates better than or equal to those of the other comparative models for all period of multiple-hours average.

Based on three chosen classification models, as shown in Table 11, the ensemble model could produce accuracy results on average for all periods better than or equal to those of the other compared existing models from 6 out of 7 datasets (excluding M2). The overall performance was superior to those of the other compared existing models. The average accuracy rates for M1, M3, M4, M5, M6, M7 datasets were at 99.682%, 99.783%, 99.715%, 99.715%, 99.837%, and 99.785%, respectively.

Table 11: Average accuracy results of component models compared to ensemble model

Station	Average accuracy (%)			
	Single classification models			Ensemble model
	SVM	KNN	RF	
M1	98.306	99.641	99.667	99.682
M2	99.158	99.789	99.830	99.808
M3	99.322	99.783	99.753	99.783
M4	98.447	99.715	99.715	99.715
M5	99.116	99.715	99.715	99.715
M6	98.566	99.837	99.837	99.837
M7	99.138	99.785	99.785	99.785

Tables 12-18 presented the details of the accuracy results of the purposed ensemble model and all three component models—KNN, RF, and SVM on multiple-hours averages for all datasets from seven monitoring stations.

Table 12: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M1

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M1	1	99.434	99.952	99.952	99.952
	4	99.470	99.952	99.952	99.952
	8	86.911	99.615	99.615	99.615
	12	98.990	100.000	100.000	100.000
	16	98.077	99.808	99.808	99.808
	20	99.277	100.000	100.000	100.000
	24	99.711	99.711	99.711	99.711
	28	99.327	98.990	98.990	98.990
	32	99.231	99.231	99.615	99.231
	36	99.134	99.134	99.134	99.134
	40	99.519	100.000	100.000	100.000
	44	99.471	98.942	99.471	99.471
	48	99.422	100.000	99.422	100.000
	Average	98.306	99.641	99.667	99.682

Table 13: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M2

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M2	1	99.639	99.952	99.952	99.952
	4	99.519	99.952	99.952	99.952
	8	98.941	99.904	99.904	99.904
	12	98.941	99.567	99.856	99.567
	16	99.231	100.000	100.000	100.000
	20	99.277	99.759	100.000	100.000
	24	99.711	99.711	99.711	99.711
	28	99.663	99.663	99.663	99.663
	32	98.077	99.231	99.231	99.231
	36	98.701	100.000	100.000	100.000
	40	99.519	99.519	99.519	99.519
	44	98.413	100.000	100.000	100.000
	48	99.422	100.000	100.000	100.000
	Average	**99.158**	**99.789**	**99.830**	**99.808**

Table 14: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M3

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M3	1	99.507	99.964	99.964	99.964
	4	99.470	99.807	99.807	99.807
	8	99.519	100.000	99.808	100.000
	12	99.519	100.000	100.000	100.000
	16	99.039	99.808	99.615	99.808
	20	99.277	99.759	99.759	99.759
	24	98.555	100.000	100.000	100.000
	28	100.000	100.000	100.000	100.000
	32	99.231	99.615	99.615	99.615
	36	99.567	98.701	98.701	98.701
	40	98.077	99.519	99.519	99.519
	44	100.000	100.000	100.000	100.000
	48	99.422	100.000	100.000	100.000
	Average	**99.322**	**99.783**	**99.753**	**99.783**

Table 15: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M4

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M4	1	99.543	100.000	100.000	100.000
	4	99.085	99.952	99.952	99.952
	8	99.326	100.000	100.000	100.000
	12	98.557	99.856	99.856	99.856
	16	99.615	100.000	100.000	100.000
	20	97.831	99.759	99.759	99.759
	24	99.422	100.000	100.000	100.000
	28	97.980	100.000	100.000	100.000
	32	98.077	100.000	100.000	100.000
	36	98.701	100.000	100.000	100.000
	40	100.000	100.000	100.000	100.000
	44	96.296	97.884	97.884	97.884
	48	95.376	98.844	98.844	98.844
	Average	**98.447**	**99.715**	**99.715**	**99.715**

Table 16: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M5

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M5	1	99.434	100.000	100.000	100.000
	4	99.085	99.952	99.952	99.952
	8	99.038	99.808	99.808	99.808
	12	99.134	100.000	100.000	100.000
	16	98.269	100.000	100.000	100.000
	20	97.831	99.036	99.036	99.036
	24	100.000	100.000	100.000	100.000
	28	99.663	99.663	99.663	99.663
	32	98.846	99.231	99.231	99.231
	36	98.701	99.134	99.134	99.134
	40	99.039	100.000	100.000	100.000
	44	99.471	99.471	99.471	99.471
	48	100.000	100.000	100.000	100.000
	Average	**99.116**	**99.715**	**99.715**	**99.715**

Table 17: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M6

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M6	1	99.350	100.000	100.000	100.000
	4	98.989	100.000	100.000	100.000
	8	99.038	100.000	100.000	100.000
	12	98.557	99.856	99.856	99.856
	16	98.462	100.000	100.000	100.000
	20	98.554	100.000	100.000	100.000
	24	98.844	100.000	100.000	100.000
	28	98.990	100.000	100.000	100.000
	32	98.846	99.615	99.615	99.615
	36	99.134	100.000	100.000	100.000
	40	98.077	99.519	99.519	99.519
	44	96.825	99.471	99.471	99.471
	48	97.688	99.422	99.422	99.422
	Average	**98.566**	**99.837**	**99.837**	**99.837**

Table 18: Average accuracy results of component models compared to ensemble model on multiple-hours averages for M7

Station	Hour Average	Accuracy (%)			
		SVM	KNN	RF	Ensemble model
M7	1	99.519	100.000	100.000	100.000
	4	99.374	99.904	99.904	99.904
	8	98.845	99.711	99.711	99.711
	12	98.268	99.856	99.856	99.856
	16	99.615	100.000	100.000	100.000
	20	99.759	99.759	99.759	99.759
	24	99.133	99.711	99.711	99.711
	28	99.663	100.000	100.000	100.000
	32	99.231	99.231	99.231	99.231
	36	99.567	100.000	100.000	100.000
	40	98.558	99.039	99.039	99.038
	44	98.413	100.000	100.000	100.000
	48	98.844	100.000	100.000	100.000
	Average	**99.138**	**99.785**	**99.785**	**99.785**

As for M6, as illustrated in Figure 15, the proposed ensemble model provided the greatest average accuracy result on average at 99.837%. This is because the model could achieve 100% accuracy results for eight targeted periods of time while the poorest accuracy result of this monitoring station was at 48-hour average as shown in Table 17.

Furthermore, all sets of data showed the proclivity of the model on multiple-hours ahead predictions which were still trustworthy. The accuracy results from all seven sets

of data displayed the effective performance of producing at least 97.884% of accuracy rate for all targeted periods of time as presented in Tables 12-18.

In addition, Tables 19-25 displayed the confusion matrices of RF. They were created to explain the summary where the machine learning model was confused when it made predictions. The columns explained the predicted class. On the contrary, the rows were used for representing the actual class. This study provided the model with inputs for 8,309 instances for each monitoring station.

Table 19: Confusion matrix of RF for M1

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	4169	0	0	0	0
	Satisfactory	0	1799	0	0	0
	Moderate	0	0	798	0	0
	Unhealthy	0	0	3	1174	1
	Very unhealthy	0	0	0	0	365

Table 20: Confusion matrix of RF for M2

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	4535	0	0	0	0
	Satisfactory	0	1412	0	0	0
	Moderate	0	2	980	0	0
	Unhealthy	0	0	2	1151	0
	Very unhealthy	0	0	0	0	227

Table 21: Confusion matrix of RF for M3

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	5285	0	0	0	0
	Satisfactory	0	1107	0	0	0
	Moderate	0	2	942	0	0
	Unhealthy	0	0	1	847	0
	Very unhealthy	0	0	0	0	125

Table 22: Confusion matrix of RF for M4

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	5592	0	0	0	0
	Satisfactory	0	847	0	0	0
	Moderate	0	0	682	0	0
	Unhealthy	0	0	0	974	0
	Very unhealthy	0	0	0	0	214

Table 23: Confusion matrix of RF for M5

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	4646	0	0	0	0
	Satisfactory	0	1246	0	0	0
	Moderate	0	0	923	0	0
	Unhealthy	0	0	0	1350	0
	Very unhealthy	0	0	0	0	144

Table 24: Confusion matrix of RF for M6

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	5241	0	0	0	0
	Satisfactory	0	1370	0	0	0
	Moderate	0	0	702	0	0
	Unhealthy	0	0	0	634	0
	Very unhealthy	0	0	0	0	362

Table 25: Confusion matrix of RF for M7

		Predicted				
		Excellent	Satisfactory	Moderate	Unhealthy	Very unhealthy
Actual	Excellent	4661	0	0	0	0
	Satisfactory	0	1459	0	0	0
	Moderate	0	0	1051	0	0
	Unhealthy	0	0	0	929	0
	Very unhealthy	0	0	0	0	209

The classified results suggested that in M4-M7 datasets as displayed in Table 22 - 25, there was no wrong prediction in all classes. However, for M1 in Table 19 there were four wrong predictions on unhealthy class. For M2 in Table 20, four wrong predictions occurred in class of moderate (2 instances) and unhealthy (2 instances). For M3 in Table 21, there were three wrong predictions in moderate class (2 instances) and unhealthy class (1 instances).

5.3 Analysis and Discussion

With respect to the speed of training and testing the proposed model, the operation on the proposed model did not take too much time to finish. Even though the technique of the ensemble model needed the cooperation of many classification models. All the single models were able to be trained and tested independently and parallelly. All classification algorithms could separately receive input and provide output without being involved with one another. Additionally, the final prediction originated simply on the majority vote of the output of each single model. Therefore, the proposed ensemble technique was not time-consuming at all for the entire experimental process.

In addition, on the aspects of all classification algorithms including the proposed model, overall, the prediction results went unstable between 86.91% and 100% accuracy results from 1-hour average to 48-hour average. It was possible that seasonal patterns and trends of air pollution in Thailand [9] were the cause of this phenomenon, and they also brought about the volatility to all the classifiers.

Moreover, the performance of the proposed model, as illustrated in Table 11, produced the lowest prediction result at M1. It gave the accuracy rate at 99.682%. It might be due to the reasons of the geological location of M1 and some uncertain activities occurring in the area. Both causes could have an unpredictable impact on the seasonal pollution distribution. Regarding the first cause, Chiangmai province which was the location of M1 had the geological location as a basin, a low-lying area among mountains. This kind of area caused stagnation to the air flow and trapped air pollution within the area. This impact inevitably led to the high amount of accumulated air pollution in the area. As regards the second reason, at M1 there were some activities which were terribly unpredictable happening in the region. Such situations as burn-off activities, large forest fires, haze from neighboring countries, or El Nino, an unusual weather condition [10] were all very difficult to control. Thus, these situations might be the cause of fluctuations in the amount and change of air pollution in the region.

CHAPTER VI
SUMMARY AND SUGGESTIONS

From the beginning to the end, for the entire experimental process done in this thesis, as a whole, three chosen classification algorithms namely KNN, RF, and SVM made a great impact on forecasting AQI for multiple hours ahead. On average, for all seven targeted datasets, they produced accuracy results ranging from 98.306% to 99.837%. When considering the proposed ensemble model's performance, it can provide the accuracy results ranging from 99.682% to 99.837%. The proposed model is capable of presenting the overall results greater than those of the comparative single models. Hence, we suggest that the proposed model should be used to formulate strategies for dealing with the problem of air pollution. Besides, it is also useful for warning people to know about the level of its seriousness beforehand.

In the future, a more complicated model could be employed to obtain better predicted results for AQI. The future model would avoid depending on any deep learning model to conserve the same principle of taking low computational cost and spending not much time as well as data. The new development should be beneficial in improving, on average, the accuracies of the prediction model, making it become more precise and training the model faster. Moreover, as this experiment used binary representation as inputs and outputs, for future work, it is possible to design the method under the consideration with the characteristic of utilizing binary representation.

REFERENCES

1. 2019; Available from: https://resolution.soc.go.th/?prep_id=99333474.
2. Yang, R., F. Yan, and N. Zhao. *Urban air quality based on Bayesian network*. in *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*. 2017.
3. Zhenghua, W. and T. Zhihui. *Prediction of Air Quality Index Based on Improved Neural Network*. in *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*. 2017.
4. Macatangay, L.H. and R.M. Hernandez. *A Deep Learning-Based Prediction and Simulator of Harmful Air Pollutants: A Case from the Philippines*. in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*. 2020.
5. Mahalingam, U., et al. *A Machine Learning Model for Air Quality Prediction for Smart Cities*. in *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*. 2019.
6. Madaan, D., et al. *VayuAnukulani: Adaptive Memory Networks for Air Pollution Forecasting*. in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2019.
7. Martínez, N.M., et al. *Machine Learning Techniques for PM10 Levels Forecast in Bogotá*. in *2018 ICAI Workshops (ICAIW)*. 2018.
8. Photphanloet, C. and R. Lipikorn, *PM10 concentration forecast using modified depth-first search and supervised learning neural network*. Science of The Total Environment, 2020. **727**: p. 138507.
9. Sukkhum, S., et al., *Seasonal Patterns and Trends of Air Pollution in the Upper Northern Thailand from 2004 to 2018*. Aerosol and Air Quality Research, 2022. **22**(5): p. 210318.
10. PostReporters. *Flights diverted as Chiang mai haze reaches critical level*. 2019; Available from: https://www.bangkokpost.com/thailand/general/1653828.

VITA

NAME	Saksiri Lertnilkarn
DATE OF BIRTH	29 October 1983
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	B.Sc. in Mathematics, Faculty of Science, Mahidol University, 2005. B.A. in Business and Managerial Economics, Faculty of Economics, Chulalongkorn University, 2009.
HOME ADDRESS	41 Soi Bangbon 1 Soi 11 Yak 8 Bangbon 1,Rd Khwaeng Khlong Bang Phran, Bangbon District, Bangkok, 10150