AIR QUALITY CLASSIFICATION IN THAILAND

KATTARIYA KUJAROENTAVON

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(INFORMATION TECHNOLOGY MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2015

Thesis
entitled

AIR QUALITY CLASSIFICATION IN THAILAND

…………………………………….............
Miss Kattariya Kujaroentavon
Candidate

....
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Major advisor

....
Asst. Prof. Adisorn Leelasantitham,
Ph.D. (Electrical Engineering)
Co-advisor

....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Co-advisor

…………………………………….............
Prof. Banchong Mahaisavariya,
M.D., Dip. (Thai Board of Orthopedics)
Dean
Faculty of Graduate Studies,
Mahidol University

……………………………………………
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Program Director
Master of Science Program in Information Technology Management
Faculty of Engineering
Mahidol University

Thesis
entitled
AIR QUALITY CLASSIFICATION IN THAILAND

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Information Technology Management)
on
January 2, 2015

……………………………………...........
Miss Kattariya Kujaroentavon
Candidate

....
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Chair

……………………………………......
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Member

……………………………………………
Asst. Prof. Adisorn Leelasantitham,
Ph.D. (Electrical Engineering)
Member

……………………………………......
Asst. Prof. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Member

……………………………………………
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Member

……………………………………......
Prof. Banchong Mahaisavariya,
M.D., Dip. (Thai Board of Orthopedics)
Dean
Faculty of Graduate Studies,
Mahidol University

……………………………………………
Lect. Worawit Israngkul,
M.S. (Technical Management))
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

The author wishes to express his profound gratitude and to respectfully dedicate this work to his parent and family members for their endless encouragements, love and sacrifices. The author is most grateful to his advisor, Asst. Prof. Supaporn Kiattisin, for his valuable supervision, supports and encouragements throughout the study.

Grateful acknowledgements for the Co-advisor Asst. Prof. Adisorn Leelasantitham and Dr. Sotarat Thammaboosadee. In addition, grateful acknowledgements are made to Asst. Prof. Waranyu Wongseree members of thesis committee, for their valuable suggestions and comments.

Kattariya Kujaroentavon

AIR QUALITY CLASSIFICATION IN THAILAND

KATTARIYA KUJAROENTAVON 5638538 EGTI/M

M.Sc. (INFORMATION TECHNOLOGY MANAGEMENT)

THESIS ADVISORY COMMITTEE: SUPAPORN KIATTISIN, Ph.D., ADISORN LEELASANTITHAM, Ph.D., SOTARAT THAMMABOOSADEE, Ph.D.

ABSTRACT

This thesis purposes air quality classification based on six variables of the air quality index (AQI) in Thailand i.e. O_3, NO_2, CO, SO_2, PM_{10} and levels of health concerns. The classification results are compared using JRip, Multi-layer Perceptron and C4.5 decision tree. The results show that averaging the accuracies of the classifications used by the C4.5, JRip, Multi-layer Perceptron produce approximate values of 90.98, 90.36 and 88.18, respectively, which in terms of the overview in Thailand is 88.29 Therefore, this study suggests that the topography and climate are factors affecting the differences in the rules in the C4.5 decision tree and the levels of the air quality index.

KEY WORDS: AIR QUALITY INDEX / CLASSIFICATION / C4.5 DECISION TREE

67 pages

การจำแนกคุณภาพอากาศในประเทศไทย

AIR QUALITY CLASSIFICATION IN THAILAND

แคททริยา คู่เจริญถาวร 5638538 EGTI/M

วท.ม. (การจัดการเทคโนโลยีสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : สุภาภรณ์ เกียรติสิน, Ph.D., อดิศร ลีลาสันติธรรม, Ph.D., โฆษฑ์รัตต ธรรมบุษดี, Ph.D.

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอวิธีการจำแนกคุณภาพอากาศ 6 ตัวแปรตามดัชนีคุณภาพอากาศในประเทศไทย คือ ก๊าซโอโซน, ก๊าซไนโตรเจนไดออกไซด์, ก๊าซคาร์บอนมอนอกไซด์, ก๊าซซัลเฟอร์ไดออกไซด์, ฝุ่นละอองขนาดเล็กกว่า 10 ไมครอน และระดับผลกระทบที่ส่งผลต่อสุขภาพ ผลของการจัดหมวดหมู่จะถูกนำมาเปรียบเทียบโดยอัลกอริทึม JRip, Multi-layer Perceptron และ C4.5 decision tree ผลการศึกษาพบว่าค่าเฉลี่ยความถูกต้องของการจำแนกประเภทที่ใช้โดย C4.5, JRip, Multi-layer Perceptron ค่าเฉลี่ยอยู่ที่ 90.98, 90.36 และ 88.18 ตามลำดับ ซึ่งในแง่ของภาพรวมทั้งประเทศค่าเฉลี่ยอยู่ที่ 88.29 ดังนั้นการศึกษานี้แสดงให้เห็นว่าสภาพภูมิประเทศ และสภาพภูมิอากาศเป็นปัจจัยที่มีผลต่อความแตกต่างของกฎอัลกอริทึม C4.5 decision tree และระดับดัชนีคุณภาพอากาศ

67 หน้า

CONTENTS

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

1.1 Background

At the present, the air pollution is one of the most concerned problems in Thailand. It is closely related with and mostly generated from the industrialization, transportation and construction sectors affecting to the climate because of the damage severity. In this case, it bases on the categories and concentrations of air pollutants including the duration of exposure to air pollutants and the environmentally degrading effect of the urban physical development that directly causes the effects. Air pollution leads to the lower level of air quality, so it promotes the greater risk on health. Especially the human living in the downtown that people get the bad atmosphere and much dust into their lungs. From statistics of respirator's patients in 2007, 242,405 patients are up to 305,929 in 2008 and the patients increased from 363,744 in 2009 to 365,372 in 2010. Moreover, it is up to 381,184 in 2011. As of this, it can be seen that the statistic results of respirators patients are increased in every years in Fig. 1.1 [1].

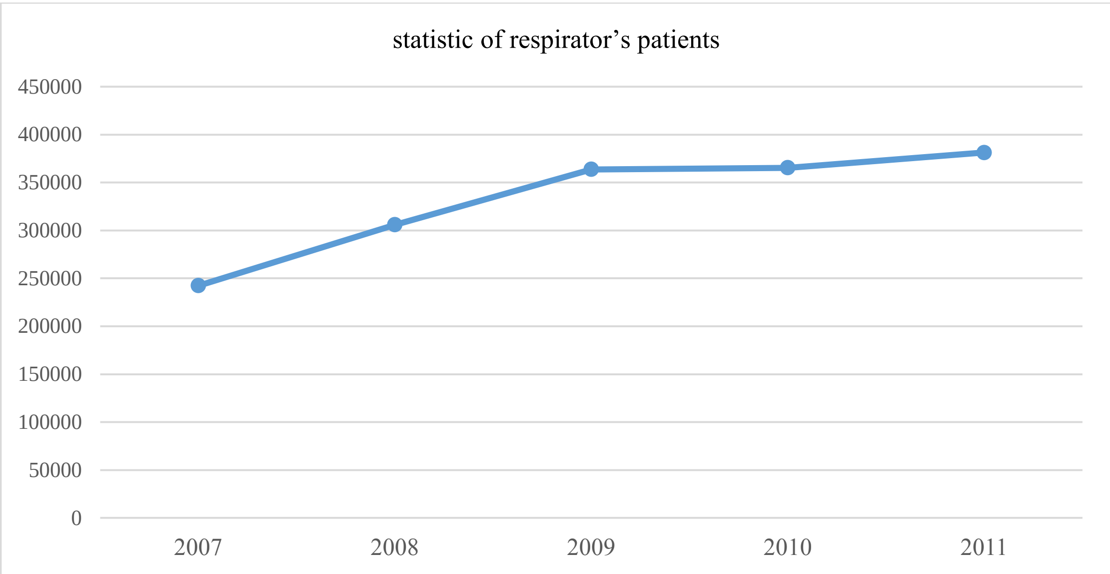

Figure 1.1 statistic numbers of respirators patients.

This thesis has used the air quality index or AQI for the air quality assessment and management in Thailand now. However, the air quality index in Thailand can be divided to 6 levels that contain good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy and Hazardous. The air pollution levels are calculated from particulate matter ten micron (PM10), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Nitrogen Dioxide (NO2) and Ozone (O3) factors from 67 monitoring stations in 29 provinces of Thailand as the following Fig. 1.2.

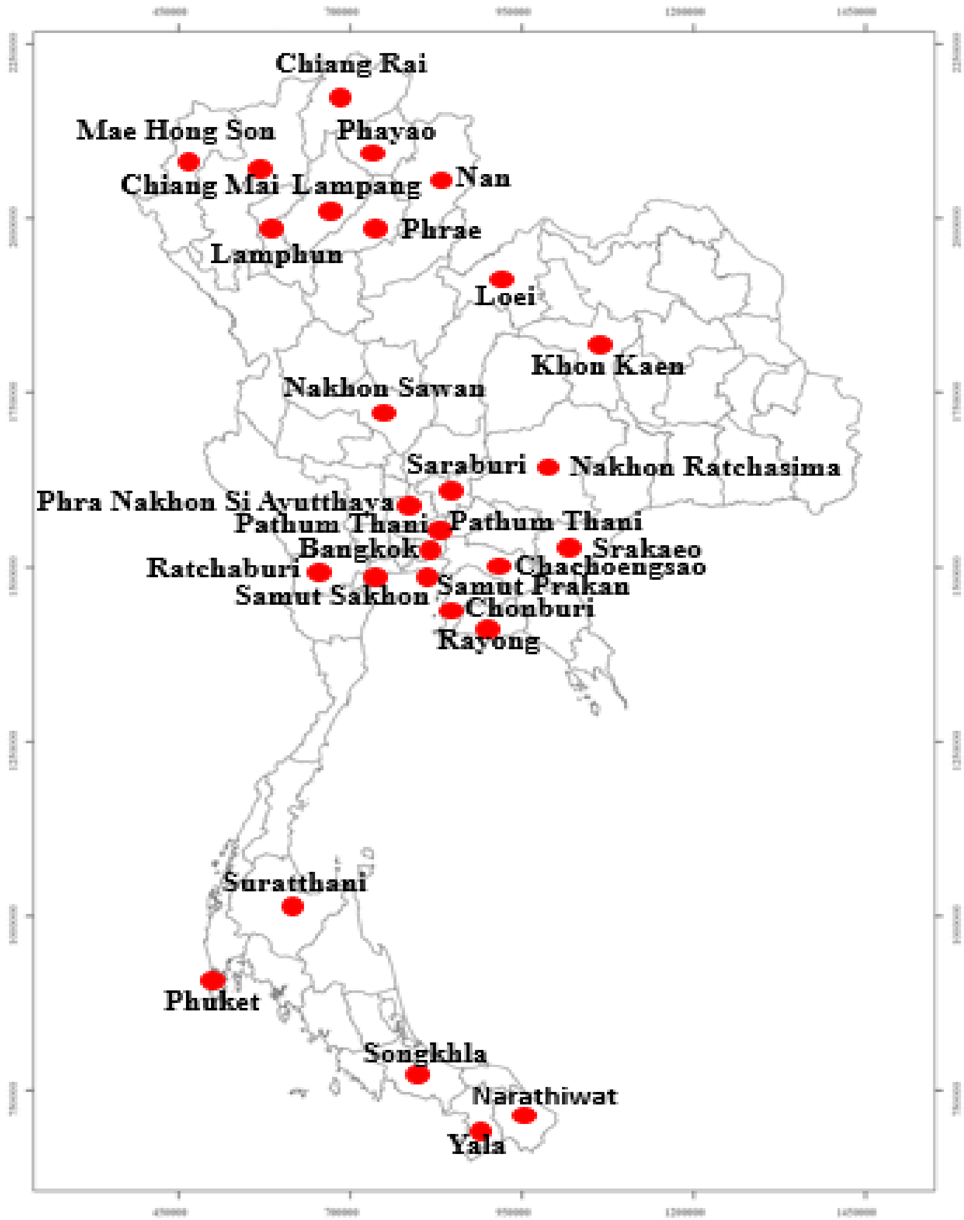

Figure 1.2 Map of monitoring stations in Thailand.

This thesis proposes the method of data mining for comparing and analyzing the different classification of 5 regions in Thailand (Northern, Central, Eastern, North-Eastern and southern) and overview of Thailand. In this case, the researcher selects 3 techniques consisting of the JRip, Multi-layer Perceptron and C4.5 decision tree that the classification will be used to help classifying the air quality index by not calculating from pollutant concentration.

1.2 Objectives

1.2.1 To create the model and analyze the air quality index in Thailand.

1.2.2 To compare the results of the JRip, Multi-layer Perceptron and C4.5 decision tree.

1.2.3 To compare the results of air quality index between region and overview of Thailand.

1.3 Scope of Work

This study uses the pollutant concentration's information of 62 monitoring stations within 29 provinces from the pollution control department (PCD) over a period of three years (January 2011 to December 2013).

1.4 Expected Results

1.4.1 To classify and predict the air quality index levels.

1.4.2 To compare the ways that are proper for the air quality index's classification between JRip, Multi-layer Perceptron and C4.5 decision tree.

1.4.3 To suggest the interested parties to study and research on the air quality.

CHAPTER II
LITERATURE REVIEW AND RELATED THEORIES

This research is related to the development and analysis of the air quality index by using the classification as the essential technique of data mining. Consequently, this thesis bases on the basic knowledge, theory and research as following:

2.1 Literature Review

Nowadays, the research about developing the air quality index by using a classification is the essential technique of data mining. For example,

In 2000, David Nerini, et.al. presented the results of daily forecasts for the dissolved oxygen rates in a lagoon, the 'Etang de Berre'. The prediction model is displayed in term of a binary decision tree. For the purpose of a transfer procedure, it is to improve the prediction error of the tree model. Results are obtained on the 'Etang de Berre' data set allow to describe and precise the effects of the environmental variables on the dissolved oxygen dynamics. The transfer procedure is applied after the tree building process gives the prediction accuracy about 17% [2].

In 2006, Ioannis, N. et.al. presented the air quality forecast that it was one of the core elements of Air Quality Management and Information Systems. Such systems are usually set up in order to serve early warning and information provision for public in Athens, Greece. However, this paper performs a comparative study between various air quality by using the forecasting methods and tools which describes the comparison work performed between several statistical methods and classification algorithms. In this case, it is based on the basis of performance. The results are compared by using IBk - K-nearest neighbors' classifier and ADTree-Alternative Decision Trees. For the results, they are showed that the average accuracies of the classifications used by the IBk - K-nearest neighbors' classifier and ADTree-Alternative Decision Trees are between 59.68% and 85.38%, respectively. The classification algorithms seem to have an advantage when comparing with the statistical one, achieving better performance concerning air quality management-related decisions taken on the basis of threshold values used [3].

In 2006, Nahun loya, et.al presented the models based on decision trees and neural network models for predicting the ozone levels by working with a data set of the Atmospheric Monitoring System of Mexico City (SIMAT), including the measurements hour by hour, during 2010 - 2011. As of this, the data comes from three meteorological stations: Pedregal, Tlalnepantla and Xalostoc in Mexico City. The data set contains 8 parameters: four chemical variables and four meteorological variables. Depending on our results, it’s possible to predict ozone levels by using these parameters, with an accuracy of 94.4% [4].

In 2008, Kasparova Milova, et.al presented the air quality model by using a decision trees in the Czech Republic locality. In this case, it focuses on daily observations of air polluting substances concentrations in the Pardubice region. After data collection, data description, and data preprocessing, we works on the creation of classification models and the analysis of the achieved results. As modeling algorithms, we select C5.0 algorithm, boosting, and CHAID method [5].

In 2011, Mohammad Hossein Sowlata, et.al. presented how to develop a novel, fuzzy-based air quality index (FAQI1) to handle the limitations. The index is developed by presenting the study, based on fuzzy logic that is considered as one of the most common computational methods of artificial intelligence. In addition to criteria air pollutants (i.e. CO, SO_2, PM_{10}, O_3, NO_2), benzene, toluene, ethylbenzene, xylene, and 1, 3-butadiene are also taken into account in the index proposed because of their considerable health effects. The different weighting factors are then assigned to each pollutant according to the priority. Trapezoidal membership functions are employed for classifications and the final index consists of 72 inference rules. To assess the performance of the index, a case study is carried out employing air quality data at five different sampling stations in Tehran, Iran, during January 2008 to December 2009, results of which are then compared to the results obtained from USEPA air quality index (AQI) [6].

In 2011, Minyue Zhao presented the decision tree for classification of air pollution index that the study area is in China, deals with the norms of the API, including density of total suspended particulate, density of SO_2, density of NO_2 and etc. For showing the graphical analysis, it demonstrates a tree shape of the classification of the API and a map of the spatial distribution of the target attribute's categories which illustrate the practicability of spatial decision tree [7].

In 2012, Hone-Jay Chu, et.al presented the identification that controls the factors of ground-level ozone levels over southwestern Taiwan by using a decision tree to obtain quantitative insight into spatial distributions of precursor compound emissions and the effects of meteorological conditions on ozone levels. As of this, they are useful for refining the monitoring plans and developing the management strategies [8].

According to this passage, researcher is introduced the rules of separated air quality classification which influences for healthy to support the decision about the separated air quality classification by combining the information about concentration of pollutants. This thesis introduces the rule of separated air quality classification by considering the level of healthy concern and using 3 algorithm techniques that contain JRip, Multi-layer Perceptron and C4.5 decision tree. The classification technique can refer to their results in order to analyze the factors causing the pollution problem. However, this chapter presents the literature review basic knowledge about the air pollution system, air quality index and technique for the following research.

2.2 Related Theories

2.2.1 Air pollution system

Air pollution is the impure air status that is higher than normal status for a long time. As of this, it will cause the danger to human, animals, plants and properties. In addition, it can be occurred in the nature e.g. dust in gale, volcano eruption, forest fires and natural gas. However, the air pollution occurring in the nature influence less to human because the source is far, so the pollution quantity transferring to the environment is low. For the human activities causing the air pollution, it contains the exhaust of motor in the factory, agro industry and the evaporation caused from garbage and waste. However, the air pollution system is occurred by 3 important parts as following:

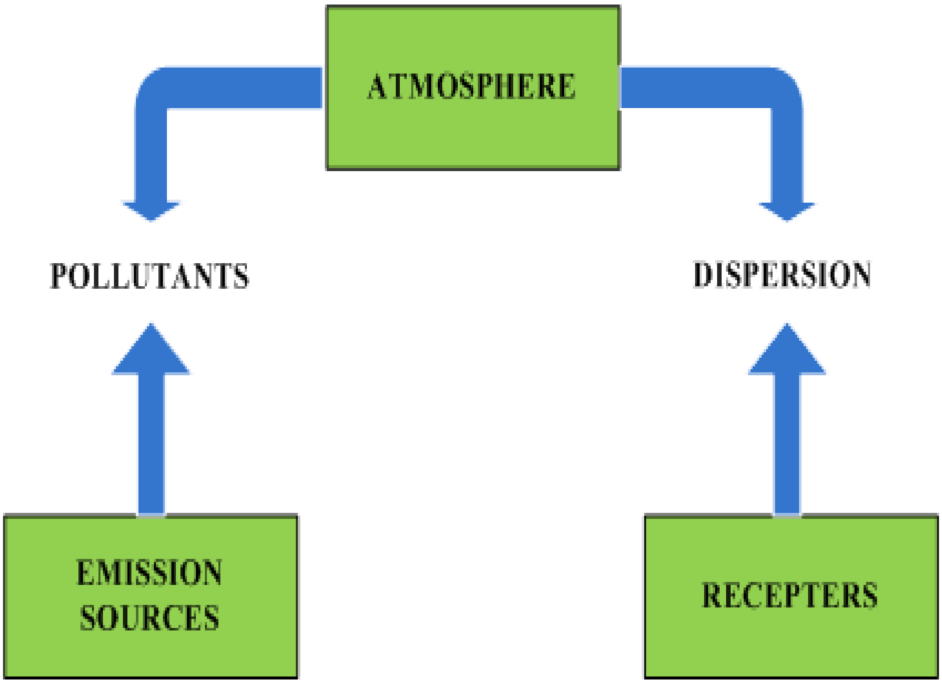

Figure 2.1 Air pollution system.

2.2.1.1 Source

Source is the origin of air pollution and omits that pollution into the atmosphere. However, the kinds and quantities of the air pollution depend on the type of source and the air pollution controlling standard. For example, the huge industrial factory, the traffic, construction and incineration.

2.2.1.2 Atmosphere

Atmosphere is supporter about the air pollution from source, so the atmosphere is the main factor to show about dissemination of air pollution and especially the air pollution form is the transmitter to spread the pollution from the source to receiver. However, the necessary factors of the atmosphere contain the direction of the wind, the velocity and the temperature.

2.2.1.3 Receptor

Receptor is the surface to contact to the air pollution, so it causes the damage and danger. However, the severity of the effect depends on the type and quantities of the air pollution, the duration and the sensitivity of the receiver. In this case, the important affected people are the human, trees, water and community.

2.2.2 Source of air pollution

The source of air pollution can be divided into 2 categories as following:

2.2.2.1 Natural Source

The natural source is one of the original sources to spread the pollution into the atmosphere in term of the natural process and without any human actions. The examples of natural disaster are volcano, gale and forest fire that all of this disasters cause the dirty air containing dust, smut, cinders and various types of gas. As of this, it causes the air pollution widely in Thailand. Especially, the important problems are the storm and the forest fire sometimes. However, the forest fire is the big problem in Thailand and it is found that it widely occurs in a huge area in summer. In addition to lose the forest, it leads to the unclean air; dust and smut are spread over until causing the looking obstacles, including CO2, SO2 and oxide of Nitrogen are gathered a lot in the atmosphere.

2.2.2.2 Man-made sources

Man-made sources are the sources causing from the human activities, it causes the air pollution to spread to the atmosphere. However, the man-made sources can be divided into 2 types:

- **Mobile Source**

The examples of mobile source are transportation in all land, water and air by using the various vehicles such as cars, motorcycles, trains, motorboats and planes that they have the fuel combustion and then leave all pollution and gas into the air. If the fuel combustion is incomplete, it will cause the prison gases that most of these prison gases are left by the intake. However, the prison gases will be released into the air in high quantities and concentration, if it is in the traffic jam and crowded areas.

So the air qualities around those areas are bad until leading to the health effect for people who live there, including to the property damage, buildings. The example gases that are released from cars are CO, CO2, oxide of sulfur, oxide of nitrogen and Hydrocarbon. However, all of these are often released from diesel engine.

- **Stationary Sources**

Combustible pollution occurs from the various fuel combustion such as incineration, forest fire, coal and etc. All of these contain the smoke, smut, and gases. However, the quantities of smuts and gases depend on the quantities and qualities of fuel and how to burn. If the fuel combustion is incomplete, it will cause the higher gases, smut and smoke than complete fuel combustion.

2.2.3 Type of Pollution

Quality of atmosphere in general are 6 kinds including

2.2.3.1 Particle Matter (PM)

The particulate matter (Particle pollution) is solid or liquid atom that has diameter about 0.001 micron (1 micron = 0.000001 m) about dust micro atom to scale of coarse sand. The small particulate size is 500 micron that is the size of coarse sand. Atom is suspensions in the atmosphere during a few seconds to month depending on the size. In addition, atom can do interaction to other substances depending on the atom size and chemical reaction in atoms because the chemical compound can erode the metal or break the plants and bring about the healthy also. The air pollution standards of air quality index refer to two size of the particle matter as following:

- **Particle Matter Ten Micrometers (PM_{10})**

The coarse particulate matter or PM10 particles is the fraction of particulates in air that diameters are less than 10 micrometers (<10 μm). It primarily comes from river beds, agriculture dust, road dust, construction sites, mining operations, and similar activities.

- **Particle Matter Twenty-Five Micrometers (PM_{25})**

The diameter of fine particulate matter or PM25 is less than 2.5 micrometers which has smaller than particle matter ten micrometers. In this case, the fine particulate matter is a product of combustion, primarily caused by fuel burn such as power plants, vehicles, wood burning stoves, and wild land fires. The diameters of these particles are less than 2.5 micrometers that are small enough to potentially pose significant health risks of people.

2.2.3.2 Carbon monoxide (CO)

Carbon monoxide or CO is a colorless and odorless gas in the atmosphere that will remain longer 2 to 4 months caused by the incomplete burning of materials. In this case, they contain carbon and transport fuels in the most of activities human of the primarily.

2.2.3.3 Sulfur oxides (SO_2)

In the atmosphere, Sulfur oxides are mostly found in the form of sulfur dioxide (SO2) that is the colorless, non-flammable and non-explosive gas. So they may cause taste, if there is high volume. When the sulfur dioxide needs a long time to convert to sulfur, potash, sulfuric acid and sulfate salts. The reaction of catalytic or chemical exposure (Photochemical Reaction) in the air of sulfur dioxide comes from the sulfur combustion that appears in the fuels from petroleum and coal. However, sulfur dioxide is the pollutants originating mainly from the industrial and diesel of engine.

2.2.3.4 Nitrogen oxides (NO_2)

Nitrogen dioxide is same high reacting gas called Oxide's nitrogen that originates the combustion in the high temperature and it is the main substance in this group. It causes the air pollution Nitrogen dioxide that can react in spray to become Nitric eroding the metal. In addition, it can react to the light, so it falls down and can be visible in the atmosphere. For Nitrogen dioxide, it will be drained from vehicle and industrial factory.

2.2.3.5 Ozone (O_3)

Ozone is one photochemical oxidant collection type caused by a chemical reaction of Ozone; it is Photochemical Oxidation that occurs between Hydrocarbon and Nitrogen's oxide by using light to increase the reaction. Another Photochemical including Aldehyde Ketone and Peroxyacetyl Nitrate (PAN) causing

Photochemical Smog like foggy in atmosphere. Consequently, the high levels of ozone are generally observed during hot, still sunny, summertime weather. But general Ozone is irritated, irrita respiratory and reduced the lung function.

2.2.3.6 Volatile organic compounds (VOC)

VOC is only the compounds of hydrogen and carbon, while VOC may contain other elements produced by incomplete combustion of hydrocarbon fuels and by evaporation sometimes. Therefore, the main attribute is evaporation in the normal temperature and normal pressure that carbon atom and hydrogen are main factors. However, the various compounds as Oxygen, Fluoride, Chloride, Bromide, Sulfur and Nitrogen and separate VOCs below Molecule structure can be divided into 2 groups as Table 2.1.

Table 2.1 Molecule structure of VOCs [9].

VOCs	Example of VOCs
Non-halogenated Hydrocarbon	- **Aliphatic Hydrocarbons** such as Fuel oil, Industrial Sovent, Propane, 1,3 – Butadiene, Gasoline, Hexane. - **Alcohol, Aldehyde, Ketone** such as Ethyl Alcohol, Methyl Alcohol, Formaldehyde. - **Aromatic Hydrocarbons** such as Toluene, Xylene, Benzene, Naphthalene, Styrene, Phenol.
Halogenated Hydrocarbon	- 1,1,1,2-Terachloroethane - 1,1,1-Trichloroethane - 1,1,2,2- Tetracholoroethane - 1,1,2 – Tetracholoroethane - 1,1 - Dichloroethane - 1,1 – Dichloroethylene -- 1,2,2 – Trifluoroethane (Freon 113) - Bromoform - Bromomethane - Carbon tetrachloride

Table 2.1 Molecule structure of VOCs [9]. (Cont.)

VOCs	Example of VOCs
Halogenated Hydrocarbon	- Chloroform - Methylene chloride - Vinyl chloride - Vinyl tricholoride - Vinylidene chloride - 1,1,1,2-Terachloroethane - 1,1,1-Trichloroethane - 1,1,2,2- Tetracholoroethane - 1,1,2 – Tetracholoroethane - 1,1 - Dichloroethane - 1,1 – Dichloroethylene

2.2.4 Air Quality Index

The first air quality index naming the “Pollutant Standard Index” (PSI) was developed and introduced by United States Environmental Protection Agency, it take into five majors (criteria) consideration for the air pollutants, namely, CO, SO_2, PM_{10}, O_3, and NO_2. In 1999, the index was further completed and replaced by the Air Quality Index or AQI. However, the index is mostly used for the air quality assessment and management [10].

In generally, air quality report is appraised with intensity that is toxic to the air quantity. When comparing to the air quantity standard whether it is over the limitation or not. Normally, people know if the intensity is over than the standard, it is not dangerous for health. On the other hand, they don’t know the limitation to be the dangerous and how to do? Therefore, the air quality in air quality index system (Pollution Control Department 2004) is calculated to compare with air quality standard that the air pollutants are 5 kinds included Ozone (O_3) average 1 hour, Nitrogen dioxide (NO_2) average 1 hour, Carbon monoxide (CO) average 8 hour, Sulfur dioxide (SO_2) average 24 hours and micro dust less than 10 micron (PM_{10}) average 24 hours. However, the air quality index is calculated for that day only. In each level of healthy concern, it contains many levels as the Tables 2.2 - 2.3.

Table 2.2 The levels air quality based on health impacts [11].

Air Quality Index	Protect of Health
Good	No health impacts are expected when air quality is in this range.
Moderate	Unusually sensitive people should consider limiting prolonged outdoor exertion.
Unhealthy for Sensitive Groups	The following groups should limit prolonged outdoor exertion - People with lung disease, such as asthma - Children and older adults - People who are active outdoors
Very Unhealthy	The following groups should avoid prolonged outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit prolonged outdoor exertion.
Hazardous	The following groups should avoid all outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit outdoor exertion.

Table 2.3 The air quality index for level of health concern [12].

Air Quality Index (AQI)	Levels of Healthy Concern
0 to 50	Good
51 to 100	Moderate
101 to 200	Unhealthy for Sensitive Groups
201 to 300	Very Unhealthy
More than 300	Hazardous

From Table 2.3, the air quality index is divided by using the specific color to each AQI level that the first level (good) is blue the second (moderate) is green, the third (unhealthy) is yellow, the forth (very unhealthy) is orange and the fifth level (hazard) is red. However, AQI standard does not exceed to 100. To calculate the air quality index in daily, it will be done with the intensity of the air pollutants as follow:

$$I_i = \frac{I_{ij}+1- I_{ij}}{X_{ij}+1- X_{ij}} \left(X_i - X_{ij}\right) + I_{ij}. \tag{2.1}$$

Where X_i = The pollutant concentration from the measurement results.

X_{ij} = The pollutant concentration is the minimum of the range, with the X_i values.

X_{ij} + 1 = The pollutant concentration to the maximum of the range with the X_i values.

I_i = The sub-index of air quality

I_{ij} = Air Quality sub-index is the minimum value of a range of values that I_i.

I_{ij} +1 = Air Quality sub-index is the maximum value of a range of values that I_i.

AQI = Air quality index

2.2.5 C4.5 Decision Tree

Decision tree algorithm is the main key to classify that the decision tree is in the form of flowchart and the structure contains root, node shows attribute, branches and leafs show group or class defined which decision tree learning is the learning in term of decision tree to show the difference between class or group and predict which class of the information following Figure 2.2.

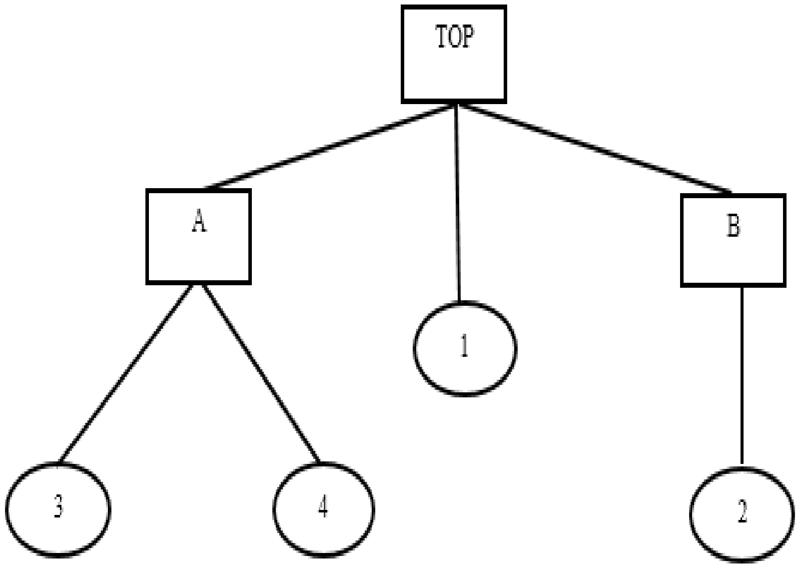

Figure 2.2 Model Decision Tree

In step C4.5 decision tree, it is the step ID3 extension developed by Ross Quinlan that is used for planting to use as one of the decision factors. In the data classification, the gain and data prediction (Entropy) are used the same as ID3, but adding from ID3 step as below [13]:

1) Be able to use both continuous and discrete data. For the continuous data, step C4.5 will create the threshold and classify into 2 parts that are the more and less group and equal with the starting point.

2) Be able to use with training data by marking with ‘?’ and exclude that value from the entropy calculation.

3) Be able to use with the abnormal value and damage.

4) Be able to apply the pruning tree with the decision.

For the model that is used for the class classification, it uses the concept of plants by selecting the most important attribute to be the root node. In this case, it uses the highest gain ratio as the root node and the next node to use for calculating the gain ratio needs to find the split information and entropy before.

- **Entropy equation**

$$Entropy\ (s) = \sum_{i=1}^{e} - P_1\ log_2\ P_1 \quad (2.2)$$

By S is attributed to be measured.

P_1 is the ratio of members equal to the number of member groups.

- **Information Gain equation**

$$GAIN\ (S, A) = Entropy\ (S) - \sum_{Value(A)} \frac{|S_v|}{|S|}\ Entropy\ (S_v). \quad (2.3)$$

By A is Attribute A.

S_v is Subset of attribute Valuable V.

S is Members of samples.

- **Split Information equation**

$$Split\ Information\ (S, A) = \sum_{i=0}^{n} \frac{|S_1|}{|S|} Log = \frac{|S_1|}{|S|}. \quad (2.4)$$

- **Gain Ratio equation**

$$GAIN\ RATIO\ (S, A) = \frac{GAIN\ (sSA)}{Split\ Information\ (S,A)}. \quad (2.5)$$

2.2.6 JRip

Ripper rule (Cohen, 1995) Forming contains 2 phrases that are the first phrase - determining the initial rule and the second one – identify the post-process rule optimization. In this case, the training data can be divided to “growing set” and “pruning set” that the algorithm creates the connection with greedy rule. However, RIPPER tries to find the best value for growing and pruning the data. Whenever it is finished, it will get the same sample that covers the training set. Then it will be deleted and the remaining training data will be divided again after learning in order to solve the problems from the wrong classification. However, this action will be done until satisfying the results. [14].

2.2.7 Multi-layer perceptron

The most common neural network model is the multi-layer perceptron (MLP). This type of neural network is known as a supervised learning not the answers is right or wrong [15]. The aim of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown a graphical representation of a multi-layer perceptron is show below:

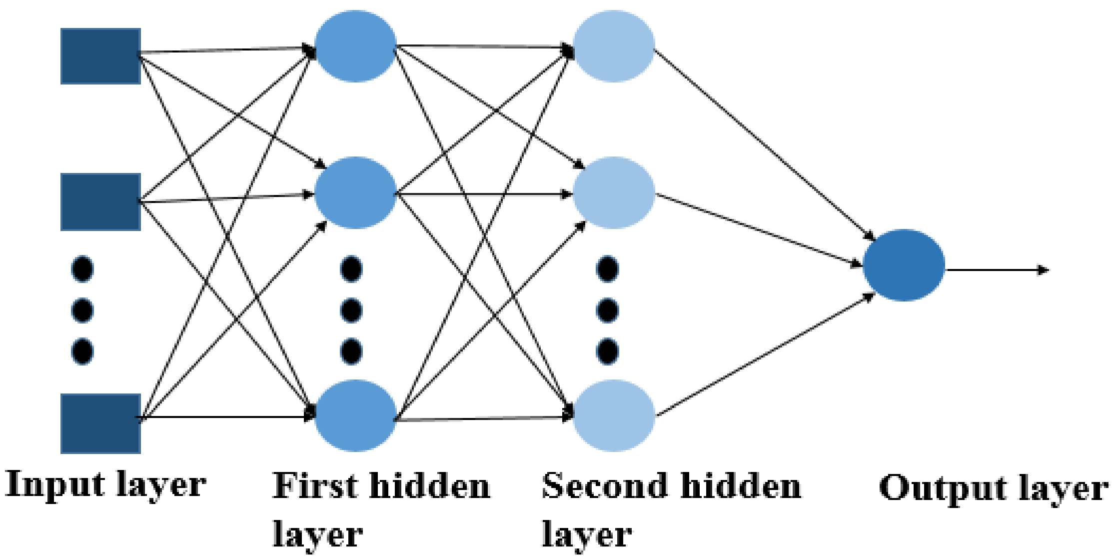

Figure 2.3 The multi-layer perceptron.

The multi-layer perceptron learn using an algorithm called backpropagation. With the input data is repeatedly presented to the neural network. With each presentation the output of the neural network is compared to the desired output and an error is computed. This error is then back propagated to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output. Multi-layer perceptron is the neural network containing many layers that each layer comprises of node like neurons. It is the line weight connecting between node of each layer (W Matrix), bias-vextor (b) and the output vector (a). In this case, m is the layer index being at the top when p is the input vector. The output calculation of the neural network at M layer can be written as the below equation:

$$a^{m+1} = f^{m+1}(W^{m+1} a^m + b^{m+1}), \tag{2.6}$$

where $m = 0, 2, \dots, M-1,$

$$a^0 = p, \tag{2.7}$$

$$a = a^m .$$

2.2.8 K-fold Cross-validation

K-fold cross-validation technique (Ron, 1995) is the method of efficiency measurement for the model prediction. For the basic of this technique, it is the sampling by starting with the data division calling fold and testing some parts of data by predicting the model information. In case of precision sampling by *k* groups, the data can be divided to *k* groups equally and then calculated the precision value for *k* times. In each round, it needs to create the classification model by using the learning data for *k*-1 and 1 testing data (not the learning data) [16].

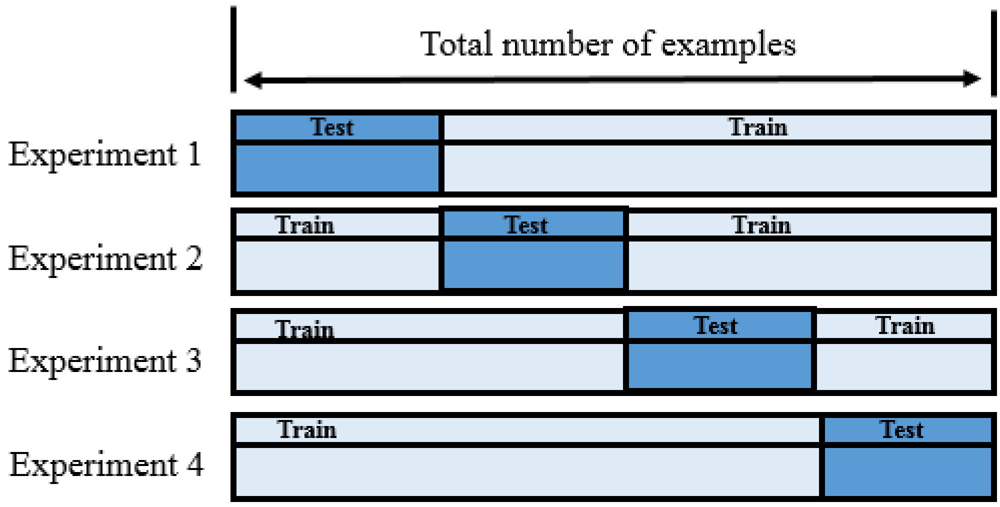

Figure 2.4 10-fold cross validation.

According to Fig. 2.4, the first data set is used as the testing data and the second data set to tenth data set are used as the learning data in the first working round, giving the result of a classification model. The second round uses the second data set as the testing data, but the first and third to tenth data sets are used as the learning data. After that, the result is one classification model also. However, this process will be repeated until the tenth round that the tenth data set is used as the testing data, but the first to ninth data sets are used as the learning data and it finally gets the other one classification set.

CHAPTER III
PROPOSED METHODS

This chapter presents the research methodology including data that are used to perform the classification of air quality criterions.

3.1 Data used in the Study

The data is used in this research, taken from 67 monitoring station of 29 provinces in Thailand of the pollution concentration development by using total air pollutants 5 kind CO, SO_2, PM_{10}, O_3, and NO_2 of Thailand.

3.2 Research Tools

In this thesis, we use Waikato Environment for Knowledge Analysis (WEKA) Version 3.6.12© [17] in order to obtain the simulation result.

3.3 Steps of Research Methodology

Aim of this thesis is used data mining to create model by using classification technique. This thesis was separated step for used data mining in 5 steps

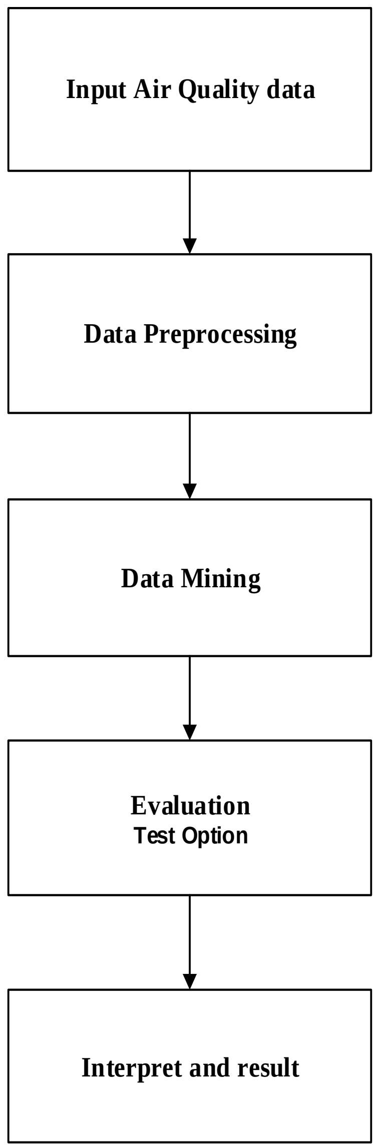

Figure 3.1 Steps of research.

3.2.1 Input Air Quality Data

For this research, the data is collected to use in the next step that the collected raw data of the pollutant factors are focused on 67 monitoring station of 29 provinces in Thailand. However, they can be shown as Table 3.1.

Table 3.1 Data of the pollution concentration in Thailand.

Date	SO_2	NO_2	CO (1hr)	CO (8hr)	Ozone	PM_{10}	AQI
31-Dec-11	0	14	0.8	0.8	17	49.4	56
30-Dec-11	1	9	0.7	0.7	27	49	56
29-Dec-11	1	8	0.6	0.5	32	60.9	63
28-Dec-11	1	12	0.1	0.1	29	68.6	68
…………	1	13	0.2	0.2	25	48.5	88
…………	1	11	0.2	0.2	21	36.5	54
31-Dec-13	0	12	0.3	0.3	17	37.5	49

3.2.2 Data Pre-Processing

It takes a long time for this method because this model uses the data mining depending on the data quality. So if the data or some parts of data are wrong, the proceeded result will be false also. In this research, before importing the system into data mining, the following steps need to be made:

3.2.2.1 Data Cleaning

The process of data cleaning e.g. checking the data with a null data and outliers. In this step, it is very important for the result of data mining. If data is not clean, it may cause the wrong result or no data consistency.

- **Data Integration**

In this procedure, it is to integrate data from multiple sources to the same data set in order to provide access to the data mining.

- **Data Selection**

After data collection, the researchers choose the concentration of the interested pollutants that is the most important step. The attribute data mining selection must be consistent with the attribution.

3.2.2.2 Data Transformation

In this process, the researchers choose the data mining technology to use for classification, the decision tree model, JRip and Multi-layer perceptron are required to convert the data according to prescribed techniques, including converting the files to the selected program.

3.2.3 Data Mining

For the proceeded data in this research, the researchers use the classification technique containing decision tree model, JRip and Multi-layer perceptron to create a decision rules to divide the air quality criteria.

3.2.4 Pattern Evaluation

When the model or results, then the evaluation process patterns from data mining or measure the effectiveness of the model to gauge reliability of the model in this research a model of multi-model analysis. So have evaluated each model for a good part of the impairment and should be used to model the selected test option to verify the accuracy of the training data using tests option with 10- Fold cross-validation in this research.

3.2.5 Interpret and Result

Understanding the decision to divide the air quality by converting the results to make them understand easier. The model is included the analyzed results and summarized the air quality classification rules from this research.

CHAPTER IV
RESULTS AND DISCUSSION

This chapter presents the classification of air quality index data divided into two part. The first part presents classification of the overview in Thailand (29 provinces). And, the second part present classification divided data information in Thailand into 5 region following Table 4.1. By compared used algorithm C4.5 decision tree, JRip and Multi-layer Perceptron. The experiment using is t-test which is used 10-fold cross-validation.

4.1 Data Information of monitoring station in Thailand

The air quality from 62 monitoring station of 29 provinces in Thailand as following Table. 4.1.

Table 4.1 The monitoring stations of Thailand.

Regions	Provinces	Number of Stations
Northern	Chiang Mai	2
	Lampang	4
	Nakhon Sawan	1
	Chiang Rai	2
	Mae Hong Son	1
	Nan	1
	Lamphun	1
	Phrae	1
	Phayao	1
Northeastern	Khon Kaen	1
	Nakhon Ratchasima	1
	Loei	1

Table 4.1 The monitoring stations of Thailand. (Cont.)

Regions	Provinces	Number of Stations
Central	Bangkok	17
	Samut Prakan	5
	Pathum Thani	1
	Samut Sakhon	2
	Nonthaburi	2
	Phra Nakhon Si Ayutthaya	1
	Saraburi	2
	Ratchaburi	1
Eastern	Rayong	4
	Chonburi	3
	Chachoengsao	1
	Srakaeo	1
Southern	Suratthani	1
	Phuket	1
	Songkhla	1
	Narathiwat	1
	Yala	1
Total		62

4.2 The Classification Results

The classification results divided into 7 datasets are the overview in Thailand, Northern, Northeastern, Central, Eastern, Western and Southern.

4.2.1 Overview in Thailand

There are 9,179 data that are the input of WEKA program as shown in Fig. 4.1 and the details of data in each criteria can be performed as Table 4.2.

Table 4.2 Input data of the overview in Thailand to WEKA program.

Class	Description	Number of Data
Good	Good	4231
Moder	Moderate	4572
Sensitive	Unhealthy for Sensitive Groups	371
Unhealthy	Unhealthy	5
Total		**9,179**

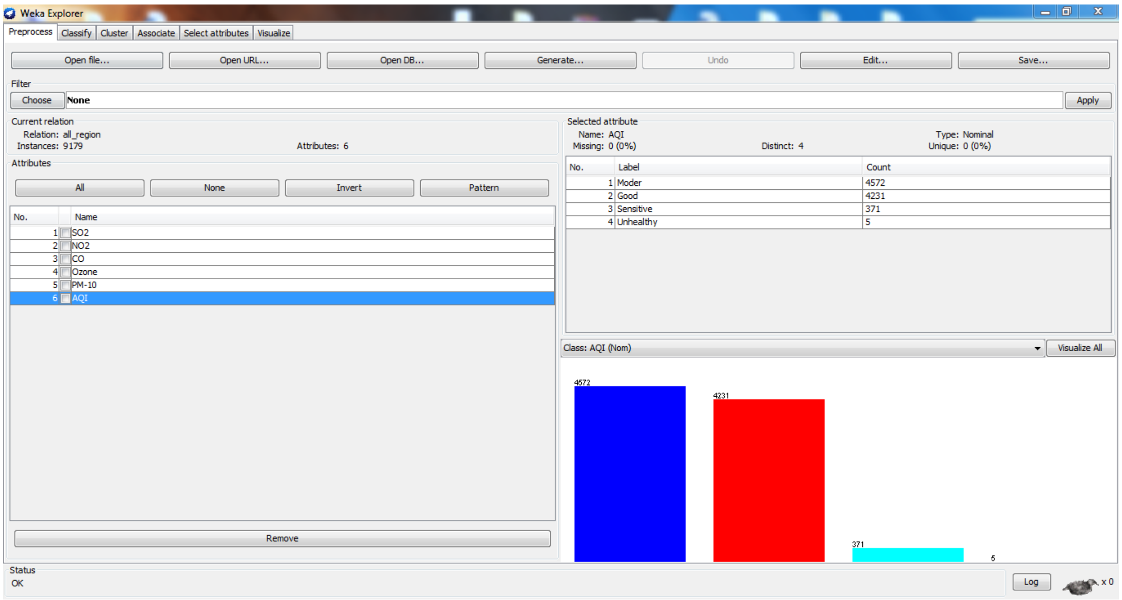

Figure 4.1 Data Information for WEKA Program.

Table 4.3 The classification summarization from the overview in Thailand data set.

Algorithm	Accuracy (%)	Incorrectly (%)	Kappa (1)
C4.5 Decision tree	88.29	11.71	0.78
Multilayer Perceptron	85.82	14.18	0.73
JRip	88.19	11.81	0.78

Table 4.3 shows that the classifications' results used by C4.5, JRip and Multi-layer Perceptron are about 88.29, 85.82 and 88.19. As of this, it can be found that the best efficiency classification algorithm is C4.5 decision tree.

4.2.2 Northern

There are 1,034 data that are the input of WEKA program as shown in Fig. 4.2 and the details of data in each criteria can be performed as Table 4.4.

Table 4.4 Input data of the northern to WEKA program.

Class	Description	Number of Data
Good	Good	341
Moder	Moderate	658
Sensitive	Unhealthy for Sensitive Groups	35
Total		**1,034**

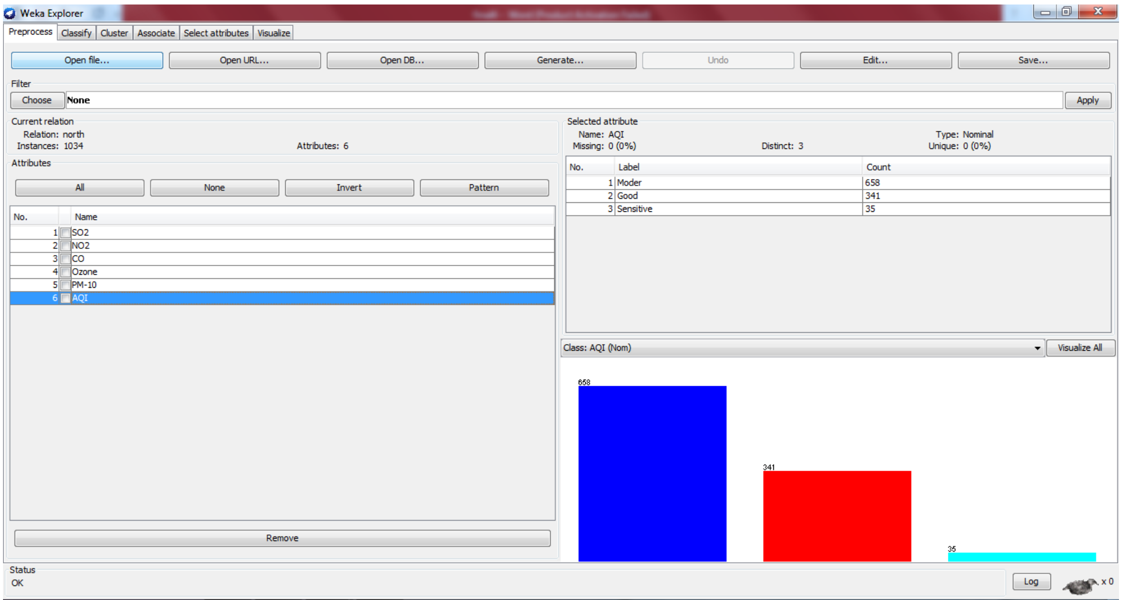

Figure 4.2 Data Information for WEKA Program.

Table 4.5 The classification summarization from the northern data set.

Algorithm	Accuracy (%)	Incorrectly (%)	Kappa (1)
C4.5 Decision tree	88.30	11.70	0.76
Multilayer Perceptron	86.36	13.64	0.70
JRip	88.68	11.32	0.76

Table 4.5 shows that the classifications' results used by C4.5, JRip and Multi-layer Perceptron are about 88.30, 86.36 and 88.68. As of this, it can be found that the best efficiency classification algorithm is C4.5 decision tree.

4.2.3 northeastern

There are 1,026 data that are the input of WEKA program as shown in Fig. 4.3 and the details of data in each criteria can be performed as Table 4.6.

Table 4.6 Input data of the northeastern to WEKA program.

Class	Description	Number of Data
Good	Good	560
Moder	Moderate	465
Sensitive	Unhealthy for Sensitive Groups	1
Total		**1,026**

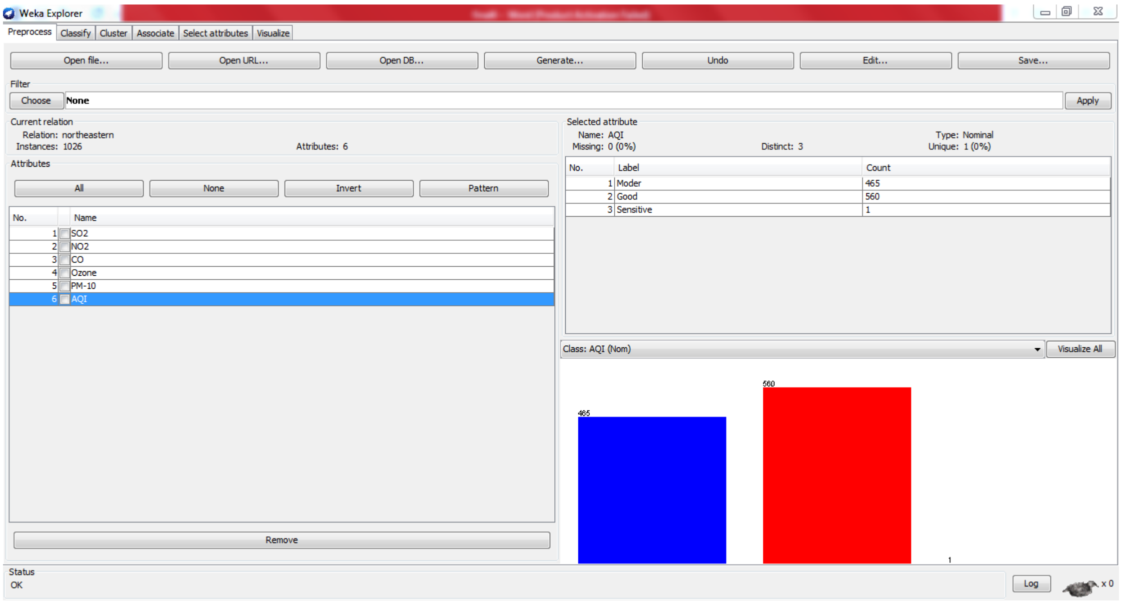

Figure 4.3 Data Information for WEKA Program.

Table 4.7 The classification summarization from the northeastern data set.

Algorithm	Accuracy (%)	Incorrectly (%)	Kappa (1)
C4.5 Decision tree	97.78	2.22	0.95
Multilayer Perceptron	96.98	3.02	0.94
JRip	97.66	2.34	0.93

Table 4.7 shows that the classifications' results used by C4.5, JRip and Multi-layer Perceptron are about 97.78, 96.98 and 97.66. As of this, it can be found that the best efficiency classification algorithm is C4.5 decision tree.

4.2.4 Central

There are 3,388 data that are the input of WEKA program as shown in Fig. 4.4 and the details of data in each criteria can be performed as Table 4.8.

Table 4.8 Input data of the central to WEKA program.

Class	Description	Number of Data
Good	Good	1,153
Moder	Moderate	1,963
Sensitive	Unhealthy for Sensitive Groups	262
Unhealthy	Unhealthy	5
Total		**3,388**

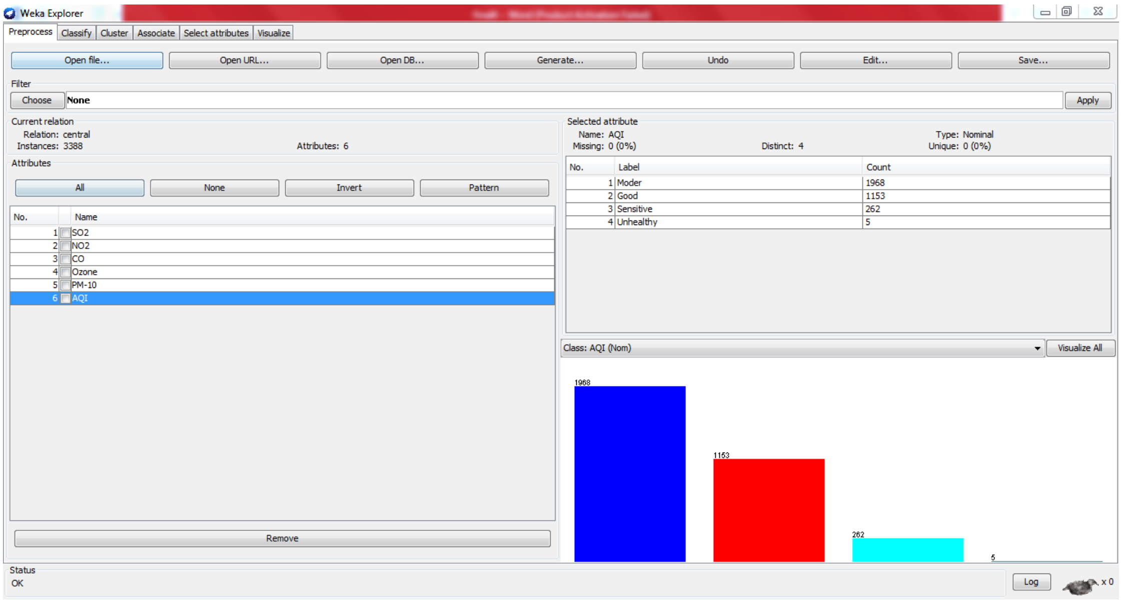

Figure 4.4 Data Information for WEKA Program.

Table 4.9 The classification summarization from the central data set.

Algorithm	Accuracy (%)	Incorrectly (%)	Kappa (1)
C4.5 Decision tree	88.68	11.32	0.78
Multilayer Perceptron	86.92	13.08	0.75
JRip	88.08	11.92	0.76

Table 4.9 shows that the classifications' results used by C4.5, JRip and Multi-layer Perceptron are about 88.68, 86.92 and 88.08. As of this, it can be found that the best efficiency classification algorithm is C4.5 decision tree.

4.2.5 Eastern

There are 1,749 data that are the input of WEKA program as shown in Fig. 4.5 and the details of data in each criteria can be performed as Table 4.10.

Table 4.10 Input data of eastern to WEKA program.

Class	Description	Number of Data
Good	Good	947
Moder	Moderate	60
Sensitive	Unhealthy for Sensitive Groups	742
Total		**1,749**

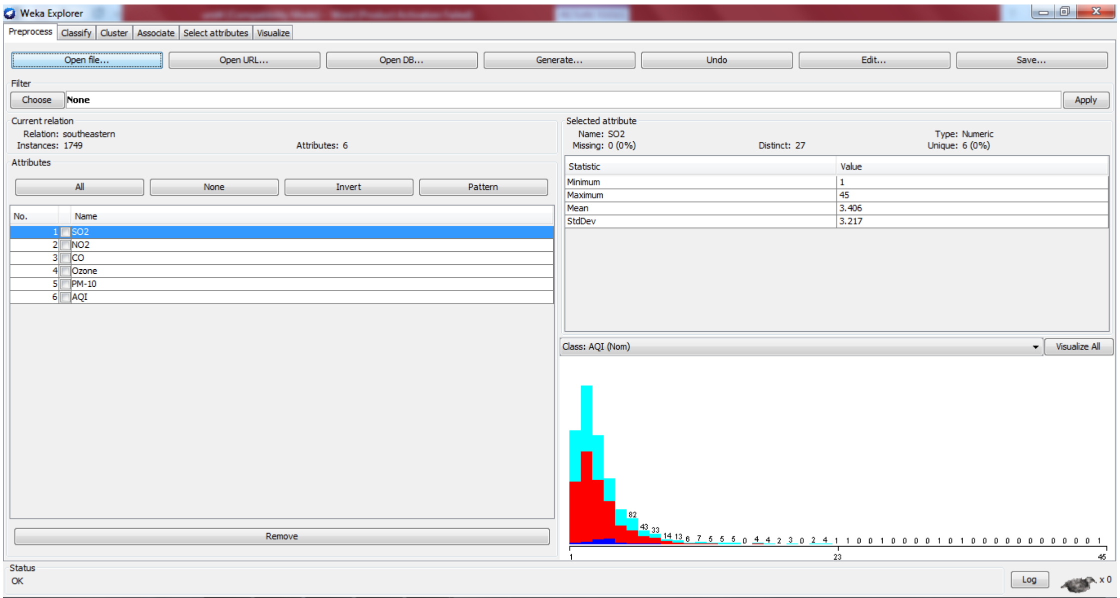

Figure 4.5 Data Information for WEKA Program.

Table 4.11 The classification summarization from the eastern data set.

Algorithm	Accuracy (%)	Incorrectly (%)	Kappa (1)
C4.5 Decision tree	83.02	16.98	0.67
Multilayer Perceptron	81.93	18.07	0.65
JRip	82.16	17.84	0.66

Table 4.11 shows that the classifications' results used by C4.5, JRip and Multi-layer Perceptron are about 83.02, 81.93 and 82.16. As of this, it can be found that the best efficiency classification algorithm is C4.5 decision tree.

4.2.6 Southern

There are 1,353 data that are the input of WEKA program as shown in Figure 4.6 and the details of data in each criteria can be performed as Table 4.12.

Table 4.12 Input data of southern to WEKA program.

Class	Description	Number of Data
Good	Good	1,095
Moder	Moderate	258
Total		**1,353**

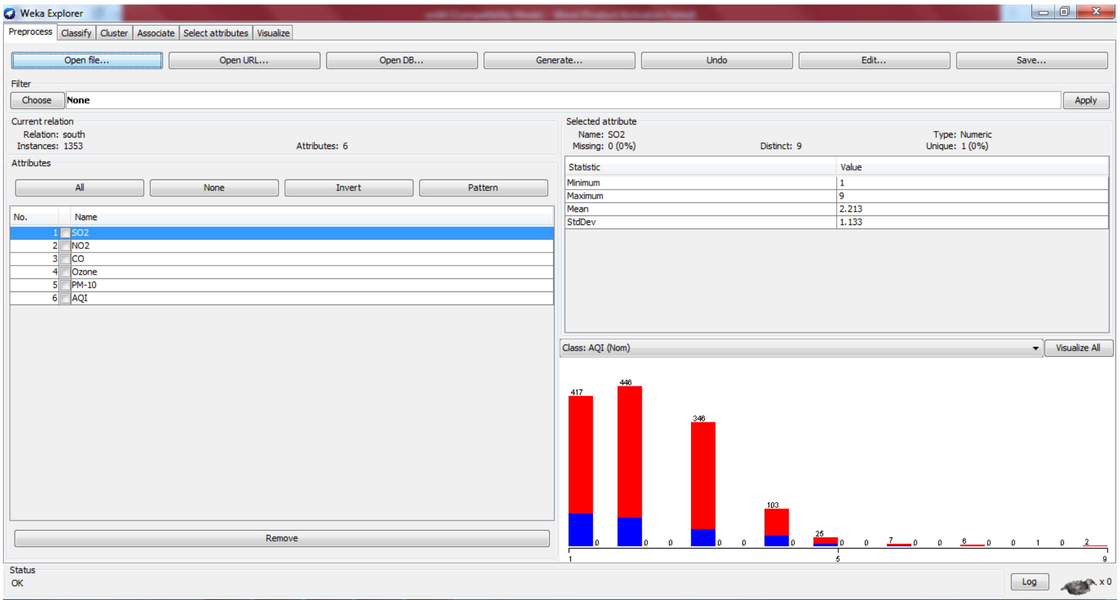

Figure 4.6 Data Information for WEKA Program.

Table 4.13 The classification summarization from the southern data set.

Algorithm	Accuracy (%)	Incorrectly (%)	Kappa (1)
C4.5 Decision tree	97.78	2.22	0.92
Multilayer Perceptron	97.19	2.80	0.91
JRip	97.56	2.44	0.92

Table 4.13 shows that the classifications' results used by C4.5, JRip and Multi-layer Perceptron are about 97.78, 97.19 and 97.56. As of this, it can be found that the best efficiency classification algorithm is C4.5 decision tree.

4.3 Discussion

From experiments methodology based on C4.5 decision tree, JRip, Multi-layer Perceptron to assess air quality is proposed Table 4.14 shows how to identify the most effective of the air quality is C4.5 decision tree algorithms in Thailand with accuracy 88.29, northern accuracy 88.68, northeastern accuracy 97.78, central accuracy 88.68, eastern accuracy 83.02, western accuracy 92.37 and southern accuracy 97.78.Which Figs. 4.7 - 4.13 show that the different of rules.

Table 4.14 The C4.5 decission tree summarization with 10-Fold Validation.

Data	**Algorithm**		
	C4.5 Decision tree (%)	**Multi-layer Perceptron (%)**	**JRip (%)**
Overview in Thailand	88.29	85.82	88.19
Northern	88.68	86.36	88.30
Northeastern	97.78	96.98	97.66
Central	88.68	86.92	88.08
Eastern	83.02	81.93	82.16
Western	92.37	91.73	92.05
Southern	97.78	97.19	97.56

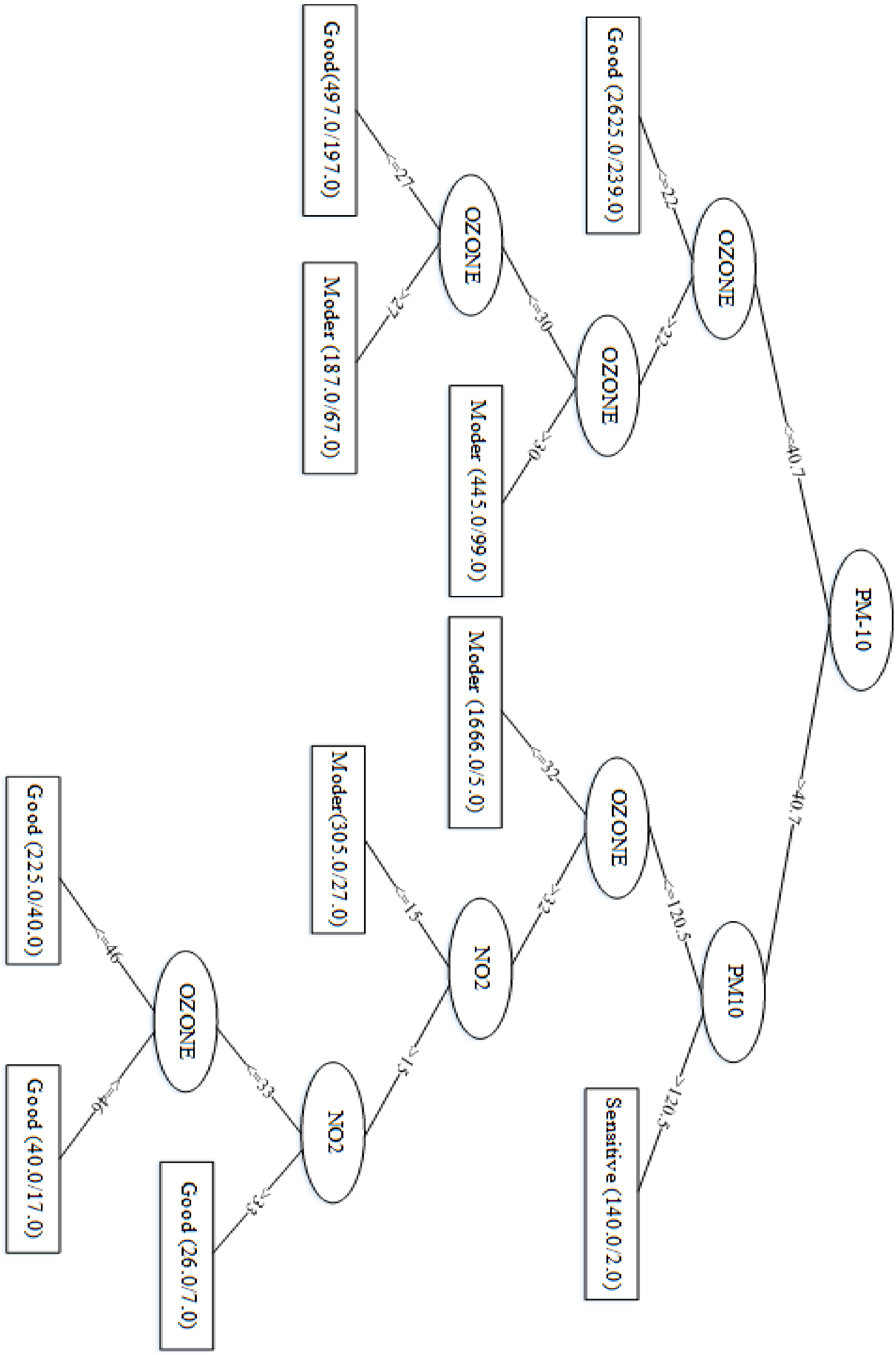

Figure 4.7 The result classification of the air quality index in thailand with the C4.5 decision tree.

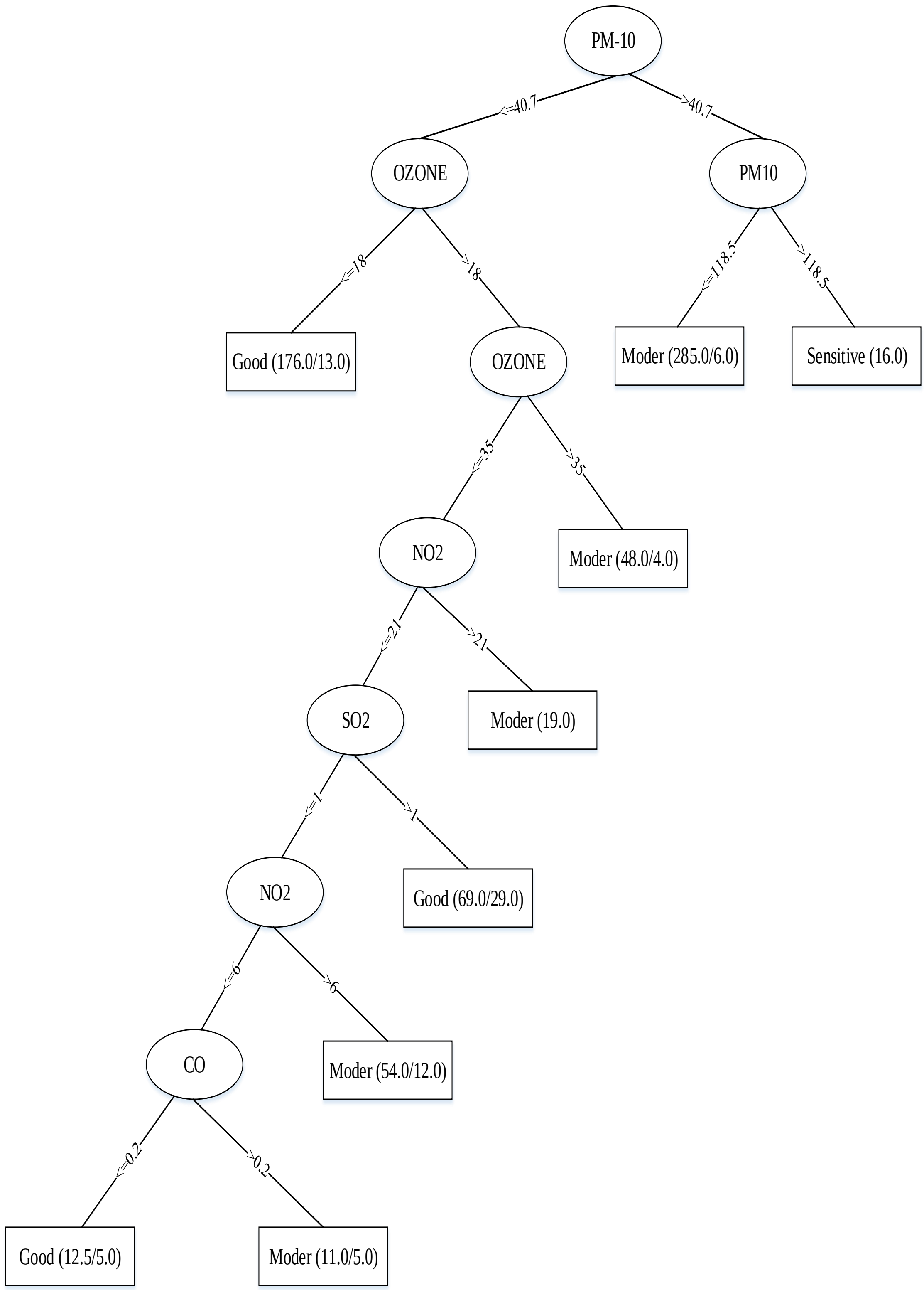

Figure 4.8 The result classification of the air quality index in northern with the C4.5 decision tree.

Figure 4.9 The result classification of the air quality index in northeastern with the C4.5 decision tree.

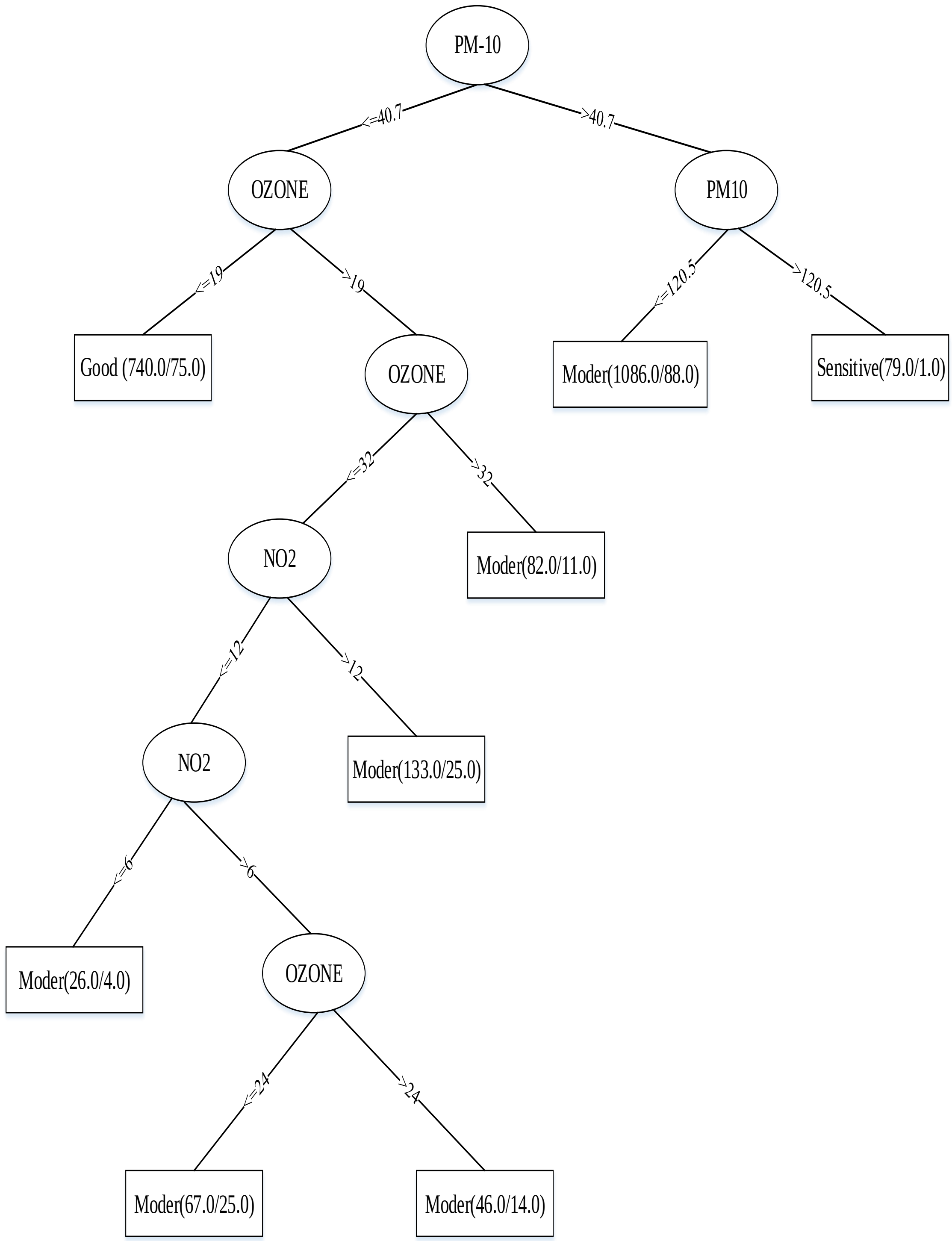

Figure 4.10 The result classification of the air quality index in central with the C4.5 decision tree.

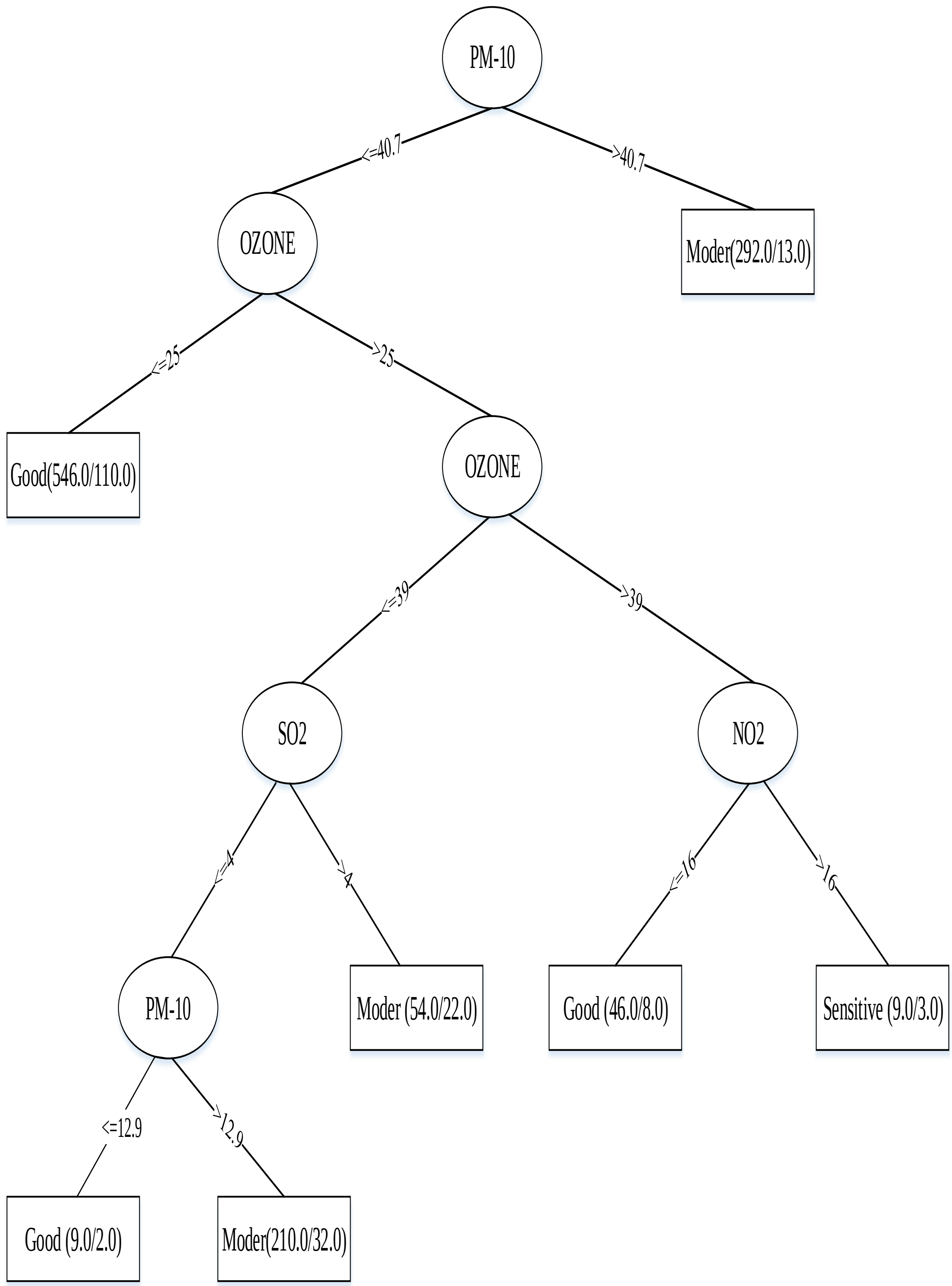

Figure 4.11 The result classification of the air quality index in eastern with the C4.5 decision tree.

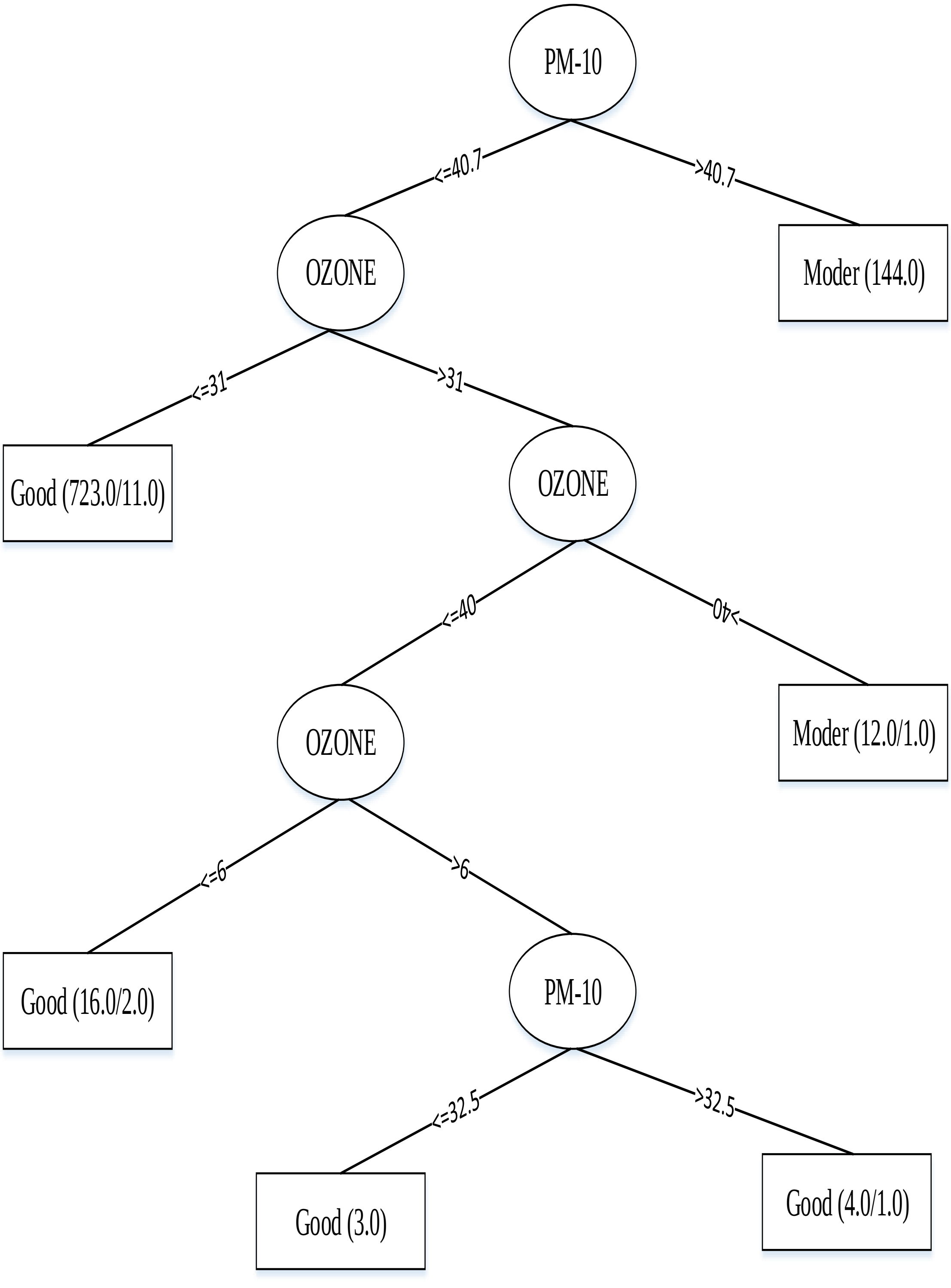

Figure 4.12 The result classification of the air quality index in southern with the C4.5 decision tree.

CHAPTER V
CONCLUSION

This chapter will describe the conclusion of the air quality classification in Thailand and the recommendation for further study as follow:

5.1 Conclusion

This study present the use of data mining for air quality classification. The results are shown that the C4.5 decision tree algorithm has the most effective and the suitable for classification of the air quality criteria as shown in Figs 4.7 - 4.14. When analyzed tree models are found that the topography and climate are factors affecting the differences in the rules in the C4.5 decision tree in each regions, as follow:

The result classification of overview in Thailand

```
PM-10 <= 40.7
|  Ozone <= 22: Good (2625.0/239.0)
|  Ozone > 22
|  |  Ozone <= 30
|  |  |  Ozone <= 27: Good (497.0/197.0)
|  |  |  Ozone > 27: Moder (187.0/67.0)
|  |  Ozone > 30: Moder (445.0/99.0)
PM-10 > 40.7
|  PM-10 <= 120.5
|  |  Ozone <= 32: Moder (1666.0/5.0)
|  |  Ozone > 32
|  |  |  NO2 <= 15: Moder (305.0/27.0)
|  |  |  NO2 > 15
```

| | | | NO2 <= 33
| | | | | Ozone <= 46: Moder (225.0/40.0)
| | | | | Ozone > 46: Sensitive (40.0/17.0)
| | | | NO2 > 33: Sensitive (26.0/7.0)
| PM-10 > 120.5: Sensitive (104.0/2.0)

The result of classification of northeastern

PM-10 <= 40.6
| Ozone <= 27: Good (369.0/4.0)
| Ozone > 27
| | PM-10 <= 29.4: Good (7.0/1.0)
| | PM-10 > 29.4: Moder (12.0/3.0)
PM-10 > 40.6: Moder (296.0)

The rules of C4.5 decision tree showed that for the different in rules of overview in Thailand *IF PM-10 <= 40.7 AND Ozone <= 22 THEN Good* and northeastern *IF PM-10 <= 40.6 AND Ozone <= 27 THEN Good.*

5.2 Future Work

For classification of air quality that influence to healthy population in the future. It can use in difference places. In air quality thesis that influenced to healthy including concentration of pollutants in each station of Thailand to fix problems in each points more. That shows about factors were relative or change result of concentration of pollutants in each kind concentration of pollutants may be change away. This thesis, researcher use 5 variants are Ozone, NO_2, CO, SO_2 and PM_{10} which others variants with air quality. And can use to analysis in same away.

REFERENCES

[1] Komchadluek (2013, May 28) People's problems Air pollution [Online]. URL http://www.komchadluek.net

[2] Nerini, David, Jean Pierre Durbec, and Claude Manté. "Analysis of oxygen rate time series in a strongly polluted lagoon using a regression tree method."Ecological modelling 133.1 (2000): 95-105.

[3] Athanasiadis, Ioannis N., Kostas D. Karatzas, and Pericles A. Mitkas. "Classification techniques for air quality forecasting." Proceeding of the 5th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, Riva del Garda, Italy. 2006.

[4] Loya, Nahun, et al. "Forecast of air quality based on ozone by decision trees and neural networks." Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2013. 97-106.

[5] Miloslova, K., and K. Jiri. "Air quality modelling by decision trees in the Czech Republic locality." 8th WSEAS International Conference on Applied Informatics and Communications (AIC ‘08) Rhodes, Greece. 2008.

[6] Sowlat, Mohammad Hossein, et al. "A novel, fuzzy-based air quality index (FAQI) for air quality assessment." Atmospheric Environment 45.12 (2011): 2050-2059.

[7] Zhao, Minyue, and Xiang Li. "An application of spatial decision tree for classification of air pollution index." Geoinformatics, 2011 19th International Conference on. IEEE, 2011.

[8] Chu, Hone-Jay, et al. "Identifying controlling factors of ground-level ozone levels over southwestern Taiwan using a decision tree." Atmospheric Environment 60 (2012): 142-152.

[9] Pollution Control Department .Air Quality [Online]. URL http://www.pcd.go.th

[10] airnow .Air Quality Index (AQI) - A Guide to Air Quality and Your Health [Online]. URL http://airnow.gov.

[11] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

[12] United States Environmental Protection Agency, July 1999, “Guideline for Reportng of Daily Air Quality - Air Quality Index (AQI),” 40 CFR Part 58, Appendix G.

[13] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[14] Cohen, William W. "Fast effective rule induction." Proceedings of the twelfth international conference on machine learning. 1995.

[15] Rajput, Anil, et al. "J48 and JRIP rules for e-governance data." International Journal of Computer Science and Security (IJCSS) 5.2 (2011): 201.

[16] Pal, Sankar K., and Sushmita Mitra. "Multilayer perceptron, fuzzy sets, and classification." IEEE Transactions on Neural Networks 3.5 (1992): 683-697.

[17] Machine Learning Group at the University of Waikato (2013) Data Mining Software in Java [Online]. URL http://www.cs.waikato.ac.nz/ml/weka

APPENDICES

APPENDIX A
EXPERIMENTAL OUTPUT

Overview in Thailand

C4.5 decision tree output

Test mode: 10-fold cross-validation

Number of Leaves: 10
Size of the tree: 19

PM-10 <= 40.7
| Ozone <= 22: Good (2625.0/239.0)
| Ozone > 22
| | Ozone <= 30
| | | Ozone <= 27: Good (497.0/197.0)
| | | Ozone > 27: Moder (187.0/67.0)
| | Ozone > 30: Moder (445.0/99.0)
PM-10 > 40.7
| PM-10 <= 120.5
| | Ozone <= 32: Moder (1666.0/5.0)
| | Ozone > 32
| | | NO2 <= 15: Moder (305.0/27.0)
| | | NO2 > 15
| | | | NO2 <= 33
| | | | | Ozone <= 46: Moder (225.0/40.0)
| | | | | Ozone > 46: Sensitive (40.0/17.0)
| | | | NO2 > 33: Sensitive (26.0/7.0)
| PM-10 > 120.5: Sensitive (104.0/2.0)

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	8104	88.2885 %
Incorrectly Classified Instances	1075	11.7115 %
Kappa statistic	0.7773	
Mean absolute error	0.0869	
Root mean squared error	0.2117	
Relative absolute error	32.3048 %	
Root relative squared error	57.7264 %	
Total Number of Instances	9179	

JRIP rules output

Number of Rules : 20

(PM-10 >= 251.5) => AQI=Unhealthy (3.0/1.0)
(PM-10 >= 120.3) and (PM-10 >= 121.2) => AQI=Sensitive (145.0/0.0)
(Ozone >= 34) and (Ozone >= 48) and (PM-10 >= 70.5) and (NO2 >= 16) => AQI=Sensitive (35.0/10.0)
(Ozone >= 33) and (Ozone >= 48) and (NO2 >= 18) => AQI=Sensitive (23.0/9.0)
(Ozone >= 33) and (PM-10 >= 57.7) and (Ozone >= 50) and (Ozone >= 62) and (NO2 >= 6) => AQI=Sensitive (17.0/3.0)
(Ozone >= 33) and (NO2 >= 27) and (SO2 >= 6) and (NO2 >= 35) => AQI=Sensitive (21.0/4.0)
(Ozone >= 33) and (NO2 >= 16) and (Ozone >= 41) and (NO2 >= 27) and (SO2 >= 5) => AQI=Sensitive (9.0/2.0)
(PM-10 <= 40.7) and (Ozone <= 21) and (Ozone <= 15) => AQI=Good (2358.0/53.0)
(PM-10 <= 40.7) and (Ozone <= 23) and (Ozone <= 18) => AQI=Good (713.0/97.0)
(PM-10 <= 40.7) and (Ozone <= 26) and (NO2 <= 7) => AQI=Good (645.0/134.0)
(PM-10 <= 40.7) and (Ozone <= 27) and (Ozone <= 22) and (NO2 <= 10) => AQI=Good (203.0/48.0)

(PM-10 <= 40.6) and (Ozone <= 27) and (NO2 <= 17) and (PM-10 <= 34.6) and (Ozone <= 24) and (NO2 <= 14) => AQI=Good (207.0/56.0)
(PM-10 <= 40.7) and (Ozone <= 27) and (NO2 <= 9) => AQI=Good (123.0/53.0)
(PM-10 <= 40.7) and (Ozone <= 29) and (Ozone <= 20) and (SO2 <= 3) => AQI=Good (70.0/21.0)
(PM-10 <= 40.7) and (Ozone <= 27) and (NO2 <= 10) => AQI=Good (20.0/6.0)
(PM-10 <= 40.7) and (Ozone <= 29) and (PM-10 <= 21.8) and (PM-10 >= 18.4) => AQI=Good (49.0/18.0)
(PM-10 <= 40.7) and (Ozone <= 27) and (PM-10 <= 36.2) and (PM-10 >= 28.8) and (Ozone <= 22) => AQI=Good (43.0/15.0)
(PM-10 <= 40.7) and (Ozone <= 32) and (Ozone <= 27) and (CO >= 0.8) and (SO2 <= 3) => AQI=Good (48.0/16.0)
(PM-10 <= 40.5) and (Ozone <= 33) and (NO2 <= 9) and (SO2 >= 5) => AQI=Good (30.0/12.0)
=> AQI=Moder (4417.0/398.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	8095	88.1904 %
Incorrectly Classified Instances	1084	11.8096 %
Kappa statistic	0.7805	
Mean absolute error	0.093	
Root mean squared error	0.2193	
Relative absolute error	34.5885 %	
Root relative squared error	59.8169 %	
Total Number of Instances	9179	

Multilayer Perceptron Stratified cross-validation

Correctly Classified Instances	2945	86.9244 %
Incorrectly Classified Instances	443	13.0756 %
Kappa statistic	0.7522	

Mean absolute error	0.0963
Root mean squared error	0.2204
Relative absolute error	35.5883 %
Root relative squared error	59.9459 %
Total Number of Instances	3388

Northern

C4.5 decision tree output

Test mode: 10-fold cross-validation

Number of Leaves: 9

Size of the tree: 17

```
PM-10 <= 40.7
|  Ozone <= 18: Good (176.0/13.0)
|  Ozone > 18
|  |  Ozone <= 35
|  |  |  NO2 <= 21
|  |  |  |  SO2 <= 1
|  |  |  |  |  NO2 <= 6
|  |  |  |  |  |  CO <= 0.2: Good (12.0/5.0)
|  |  |  |  |  |  CO > 0.2: Moder (11.0/5.0)
|  |  |  |  |  NO2 > 6: Moder (54.0/12.0)
|  |  |  |  SO2 > 1: Good (69.0/29.0)
|  |  |  NO2 > 21: Moder (19.0)
|  |  Ozone > 35: Moder (48.0/1.0)
PM-10 > 40.7
|  PM-10 <= 118.5: Moder (285.0/6.0)
|  PM-10 > 118.5: Sensitive (16.0)
```

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	917	88.6847 %
Incorrectly Classified Instances	117	11.3153 %
Kappa statistic	0.7551	
Mean absolute error	0.1001	
Root mean squared error	0.2403	
Relative absolute error	30.9077 %	
Root relative squared error	59.743 %	
Total Number of Instances	1034	

JRIP rules output

Number of Rules : 7

(PM-10 >= 121.2) => AQI=Sensitive (24.0/0.0)

(Ozone >= 59) => AQI=Sensitive (5.0/2.0)

(PM-10 <= 40.6) and (Ozone <= 18) => AQI=Good (260.0/16.0)

(PM-10 <= 40.6) and (Ozone <= 35) and (PM-10 <= 14.9) and (SO2 <= 7) => AQI=Good (23.0/2.0)

(PM-10 <= 40.6) and (Ozone <= 24) and (SO2 >= 3) => AQI=Good (25.0/4.0)

(PM-10 <= 40.5) and (Ozone <= 32) and (NO2 <= 20) and (NO2 >= 16) and (PM-10 >= 24.6) => AQI=Good (12.0/1.0)

=> AQI=Moder (685.0/52.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	913	88.2979 %
Incorrectly Classified Instances	121	11.7021 %
Kappa statistic	0.7627	
Mean absolute error	0.1077	
Root mean squared error	0.2564	
Relative absolute error	33.2586 %	

Root relative squared error	63.7522 %
Total Number of Instances	1034

Multilayer Perceptron Stratified cross-validation

Correctly Classified Instances	893	86.3636 %
Incorrectly Classified Instances	141	13.6364 %
Kappa statistic	0.7005	
Mean absolute error	0.1096	
Root mean squared error	0.2436	
Relative absolute error	33.8576 %	
Root relative squared error	60.5711 %	
Total Number of Instances	1034	

Central

C4.5 decision tree output

Test mode: 10-fold cross-validation

Number of Leaves: 8

Size of the tree: 15

```
PM-10 <= 40.7
|  Ozone <= 19: Good (740.0/75.0)
|  Ozone > 19
|  |  Ozone <= 32
|  |  |  NO2 <= 12
|  |  |  |  NO2 <= 6: Good (26.0/4.0)
|  |  |  |  NO2 > 6
|  |  |  |  |  Ozone <= 24: Good (67.0/25.0)
```

| | | | | | Ozone > 24: Moder (46.0/14.0)
| | | NO2 > 12: Moder (133.0/25.0)
| | Ozone > 32: Moder (82.0/11.0)
PM-10 > 40.7
| PM-10 <= 120.5: Moder (1086.0/88.0)
| PM-10 > 120.5: Sensitive (79.0/1.0)

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	2984	88.0756 %
Incorrectly Classified Instances	404	11.9244 %
Kappa statistic	0.776	
Mean absolute error	0.0888	
Root mean squared error	0.2171	
Relative absolute error	32.8359 %	
Root relative squared error	59.036 %	
Total Number of Instances	3388	

JRIP rules output

Number of Rules : 7

(PM-10 >= 121.2) => AQI=Sensitive (24.0/0.0)
(Ozone >= 59) => AQI=Sensitive (5.0/2.0)
(PM-10 <= 40.6) and (Ozone <= 18) => AQI=Good (260.0/16.0)
(PM-10 <= 40.6) and (Ozone <= 35) and (PM-10 <= 14.9) and (SO2 <= 7) => AQI=Good (23.0/2.0)
(PM-10 <= 40.6) and (Ozone <= 24) and (SO2 >= 3) => AQI=Good (25.0/4.0)
(PM-10 <= 40.5) and (Ozone <= 32) and (NO2 <= 20) and (NO2 >= 16) and (PM-10 >= 24.6) => AQI=Good (12.0/1.0)
=> AQI=Moder (685.0/52.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	917	88.6847 %
Incorrectly Classified Instances	117	11.3153 %
Kappa statistic	0.7627	
Mean absolute error	0.1077	
Root mean squared error	0.2564	
Relative absolute error	33.2586 %	
Root relative squared error	63.7522 %	
Total Number of Instances	1034	

Multilayer Perceptron Stratified cross-validation

Correctly Classified Instances	893	86.3636 %
Incorrectly Classified Instances	141	13.6364 %
Kappa statistic	0.7005	
Mean absolute error	0.1096	
Root mean squared error	0.2436	
Relative absolute error	33.8576 %	
Root relative squared error	60.5711 %	
Total Number of Instances	1034	

Northeastern

C4.5 decision tree output

Test mode: 10-fold cross-validation

Number of Leaves: 4

Size of the tree: 7

PM-10 <= 40.6
| Ozone <= 27: Good (369.0/4.0)
| Ozone > 27
| | PM-10 <= 29.4: Good (7.0/1.0)
| | PM-10 > 29.4: Moder (12.0/3.0)
PM-10 > 40.6: Moder (296.0)

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	1002	97.6608 %
Incorrectly Classified Instances	24	2.3392 %
Kappa statistic	0.9528	
Mean absolute error	0.0231	
Root mean squared error	0.1118	
Relative absolute error	6.9753 %	
Root relative squared error	27.4713 %	
Total Number of Instances	1026	

JRIP rules output

Number of Rules : 3

(PM-10 >= 40.6) => AQI=Moder (442.0/1.0)
(Ozone >= 28) and (PM-10 >= 29.5) => AQI=Moder (20.0/6.0)
=> AQI=Good (564.0/10.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	998	97.271 %
Incorrectly Classified Instances	28	2.729 %
Kappa statistic	0.9449	

Mean absolute error	0.0268
Root mean squared error	0.1269
Relative absolute error	8.0976 %
Root relative squared error	31.1926 %
Total Number of Instances	1026

Multilayer Perceptron Stratified cross-validation

Correctly Classified Instances	995	96.9786 %
Incorrectly Classified Instances	31	3.0214 %
Kappa statistic	0.939	
Mean absolute error	0.0264	
Root mean squared error	0.1276	
Relative absolute error	7.9583 %	
Root relative squared error	31.3475 %	
Total Number of Instances	1026	

Eastern

C4.5 decision tree output

Test mode: 10-fold cross-validation

Number of Leaves: 7

Size of the tree: 13

PM-10 <= 40.7

| Ozone <= 25: Good (546.0/110.0)

| Ozone > 25

| | Ozone <= 39

| | | SO2 <= 4

| | | | | PM-10 <= 12.9: Good (9.0/2.0)
| | | | PM-10 > 12.9: Moder (210.0/32.0)
| | | SO2 > 4: Moder (54.0/22.0)
| | Ozone > 39
| | | NO2 <= 16: Moder (46.0/8.0)
| | | NO2 > 16: Sensitive (9.0/3.0)
PM-10 > 40.7: Moder (292.0/23.0)

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	1452	83.0189 %
Incorrectly Classified Instances	297	16.9811 %
Kappa statistic	0.6462	
Mean absolute error	0.1744	
Root mean squared error	0.3037	
Relative absolute error	49.7402 %	
Root relative squared error	72.5585 %	
Total Number of Instances	1749	

JRIP rules output

Number of Rules : 9

(Ozone >= 34) and (Ozone >= 52) and (NO2 >= 8) => AQI=Sensitive (34.0/10.0)
(Ozone >= 34) and (NO2 >= 18) and (PM-10 <= 38.5) and (Ozone >= 41) => AQI=Sensitive (9.0/3.0)
(Ozone <= 24) and (PM-10 <= 40.5) and (Ozone <= 18) and (Ozone <= 12) => AQI=Good (232.0/2.0)
(PM-10 <= 23.9) and (Ozone <= 22) and (NO2 <= 9) => AQI=Good (194.0/20.0)
(PM-10 <= 39.9) and (Ozone <= 24) and (Ozone <= 18) and (PM-10 >= 26.7) => AQI=Good (66.0/5.0)

(PM-10 <= 33.2) and (Ozone <= 27) and (Ozone <= 20) and (NO2 <= 13) => AQI=Good (89.0/21.0)

(PM-10 <= 30.7) and (Ozone <= 27) and (PM-10 <= 19.3) and (NO2 <= 8) => AQI=Good (62.0/15.0)

(PM-10 <= 39.8) and (Ozone <= 26) and (SO2 >= 4) and (PM-10 >= 29.8) => AQI=Good (24.0/7.0)

=> AQI=Moder (1039.0/175.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	1437	82.1612 %
Incorrectly Classified Instances	312	17.8388 %
Kappa statistic	0.6557	
Mean absolute error	0.1801	
Root mean squared error	0.3117	
Relative absolute error	51.3575 %	
Root relative squared error	74.4675 %	
Total Number of Instances	1749	

Multilayer Perceptron Stratified cross-validation

Correctly Classified Instances	1433	81.9325 %
Incorrectly Classified Instances	316	18.0675 %
Kappa statistic	0.6712	
Mean absolute error	0.1592	
Root mean squared error	0.289	
Relative absolute error	45.403 %	
Root relative squared error	69.0311 %	
Total Number of Instances	1749	

Southern

C4.5 decision tree output

Test mode: 10-fold cross-validation

Number of Leaves: 6

Size of the tree: 11

```
PM-10 <= 40.7
|  Ozone <= 31: Good (723.0/11.0)
|  Ozone > 31
|  |  Ozone <= 40
|  |  |  NO2 <= 6: Good (16.0/2.0)
|  |  |  NO2 > 6
|  |  |  |  PM-10 <= 32.5: Moder (3.0)
|  |  |  |  PM-10 > 32.5: Good (4.0/1.0)
|  |  Ozone > 40: Moder (12.0/1.0)
PM-10 > 40.7: Moder (144.0)
```

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	1323	97.7827 %
Incorrectly Classified Instances	30	2.2173 %
Kappa statistic	0.9174	
Mean absolute error	0.0401	
Root mean squared error	0.1512	
Relative absolute error	12.9916 %	
Root relative squared error	38.4949 %	
Total Number of Instances	1353	

JRIP rules output

Number of Rules : 3

(PM-10 >= 40.8) => AQI=Moder (215.0/0.0)
(Ozone >= 32) and (Ozone >= 39) => AQI=Moder (21.0/3.0)
=> AQI=Good (1117.0/25.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	1320	97.561 %
Incorrectly Classified Instances	33	2.439 %
Kappa statistic	0.9255	
Mean absolute error	0.0407	
Root mean squared error	0.1471	
Relative absolute error	13.1688 %	
Root relative squared error	37.4402 %	
Total Number of Instances	1353	

Multilayer Perceptron Stratified cross-validation

Correctly Classified Instances	1353	97.1914 %
Incorrectly Classified Instances	39.3299 %	2.8086 %
Kappa statistic	14.1436 %	
Mean absolute error	0.1545	
Root mean squared error	0.0437	
Relative absolute error	0.9059	
Root relative squared error	38	
Total Number of Instances	1315	

APPENDIX B
ATTRIBUTES USED IN THE EXPERIMENTS

Attributes used in the experiments

No.	Code	Station
1.	2t	Bansomdejchaopraya Rajabhat University, Bangkok
2.	05t	Thai Meteorological Department Bangna,Bangkok
3.	10t	National Housing Authority Klongchan, Bangkok
4.	11t	National Housing Authority Huaykwang, Bangkok
5.	12t	Nonsi Witthaya School, Bangkok
6.	13t	EGAT, Nonthaburi
7.	14t	Highway District, Samut Sakhon
8.	15t	Mathayomwatsing School, Bangkok
9.	17t	Residence for Dept. of Primary Industries and Mines, Samut Prakan
10.	18t	City Hall, Samut Prakan
11.	19t	National Housing Authority Bangplee, Samut Prakan
12.	20t	Bangkok University Rangsit Campus, Pathum Thani
13.	21t	Ayutthaya Witthayalai School, Ayutthaya
14.	22t	Sukhothai Thammathirat Open University, Nonthaburi
15.	24t	Na Phralan Police Station Saraburi
16.	26t	Regional Environmental Office 8, Ratchaburi
17.	27t	Samut Sakhon Wittayalai School, Samut Sakhon
18.	28t	Pluak Daeng Public Health Office, Rayong
19.	29t	Health Promotion Hospital Maptaput, Rayong
20.	30t	Agricultural Office, Rayong

No.	Code	Station
21.	31t	Field Crop Research Center, Rayong
22.	33t	Health Promotion Hospital Ban Khao Hin,Chonburi
23.	35t	City Hall, Chiangmai
24.	36t	Yupparaj Wittayalai School, Chiangmai
25.	37t	Lampang Meteorological Station
26.	38t	Health Promotion Hospital Sob Pad, Lampang
27.	39t	Health Promotion Hospital Ta See, Lampang
28.	40t	Provincial Waterworks Authority Mae Moh, Lampang
29.	41t	Nakhonsawan Irrigation , Nakhon Sawan
30.	42t	Regional Environmental Office 14, Surat Thani
31.	43t	Municipal Health Center 1, Phuket
32.	44t	Hat Yai Municipality, Songkhla
33.	46t	Hydro Division, Water Resources Office Region 4, Khonkaen
34.	47t	Municipal Waste Water Pumping Station, Nakhon Ratchasima
35.	52t	Thonburi Power Sub-Station, Bangkok
36.	53t	Chokchai Police Station, Bangkok
37.	54t	National Housing Authority Dindaeng, Bangkok
28.	57t	Natural Resources and Environment Office, Chiangrai
29.	58t	Natural Resources and Environment Office, Mae Hongson
40.	60t	Municipality Office, Tungsadao, Chachoengsao
41.	61t	Bodindecha (Sing Singhaseni) School, Bangkok
42.	62t	City Hall, Narathiwat
43.	63t	White Elephant Park, Yala
44.	67t	Municipality Office, Nan
45.	68t	Provincial Administrative Stadium, Lamphun
46.	69t	Phrae Meteorological Station
47.	70t	Knowledge Park, Phayao
48.	71t	Sriaranyothai Kindergarten, Aranyaprathed, Sa Kaeo

No.	Code	Station
49.	72t	Provincial Health Office, Loei
50.	73t	Maesai Health Office, Chiangrai
51.	74t	Rayong Government Complex
52.	m2	Tak (Mobile 2)
53.	m4	Mobile 4
54.	a03	Ratburana Post Office, Bangkok
55.	a07	Chandrakasem Rajabhat University, Bangkok
56.	a08	Prabadang Rehabiltation Center, Samut Prakan
57.	a16	South Bangkok Power Plant, Samut Prakan
58.	a25	Khao Noi Fire Station, Saraburi
59.	a32	Laem Chabang Municipal Stadium, Chonburi
60.	a34	General Education Office, Chonburi
62.	a48	Ministry of Science and Technology, Bangkok
63.	a49	Department of Land Transport, Bangkok
64.	a50	Chulalongkorn Hospital, Bangkok
65.	a59	Public Relations Department, Bangkok

Table 1 How to collect the data.

Variables	Method	Height (M)	Range
Carbon Monoxide	Non-Dispersive Infrared Detection	3	0 - 50 ppm
Nitric Oxide	Chemiluminescence	3	0 - 500 ppb
Oxides of Nitrogen	Chemiluminescence	3	0 - 500 ppb
Nitrogen Dioxide	Chemiluminescence	3	0 - 500 ppb
Sulphur Dioxide	UV Fluorescence	3	0 - 500 ppb
Ozone	UV Absorption Photometry	3	0 - 500 ppb

APPENDIX C

Air Quality Classification in Thailand Based on Decision Tree

Kattariya Kujaroentavon, Supaporn Kiattisin, Adisorn Leelasantitham and Sotarat Thammaboosadee
Information Technology Management Program Faculty of Engineering, Mahidol University
25/25 Phutthamonthon 4Rd., Salaya Nakhon Pathom 73170, Thailand
E-mail: kat_aretee@hotmail.com, supaporn.kit@mahidol.ac.th, adisorn.lee@mahidol.ac.th, zotarat@gmail.com

***Abstract*—The paper presents a model for management classifier air quality by algorithm of decision tree using air quality index in Thailand including a pollutant's concentration e.g. O_3, NO_2, CO, SO_2, PM_{10} and levels of healthy concern. The purpose of this research is to establish rules of separated air quality classification by levels of healthy concern. The results of this study are correctly classified into instances of training set of 96.80% and testing set of 91.07%. The ROC curve shows that the training set data and testing set data are similar to such results. The algorithm of decision tree can use to become rules of separated air quality classification by levels of healthy concern.**

Keywords—air quality, Model, Classification, Levels of Healthy Concern, Decision Tree, air quality, Model, Classification, Levels of Healthy Concern, Decision Tree

I. INTRODUCTION

Air Pollution is main problem of people will met for affect health and respiratory. Almost this problem happened in downtown. People smell bad atmosphere and many dust into lungs. From statistic of respirator's patients. In 2007, patients 242,405 up to became 305,929 in 2008. In 2009, patients 363,744 up to became 365,372 in 2010. Finally In 2011 up to 381,184 following Fig. 1.

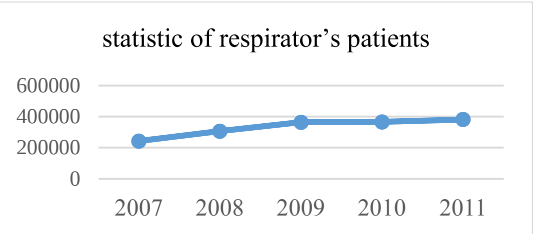

Fig. 1. Show statistic's patients

The results of statistic numbers of respirators patients were up very years. Although respirators patients someone they relative with air pollution from traffic problem which happened by directly and indirectly following Table 1. All of them from a pollutant's concentration of dusts less than 10 micron nitrogen dioxide) NO2) carbon dioxide)CO) sulfur dioxide) SO2 (and Ozone left out to atmosphere effected to health of the people with directly [1].

Table 2 Show levels Air quality health impacts [2]

Air Quality Index	**Protect Your Health**
Good	No health impacts are expected when air quality is in this range.
Moderate	Unusually sensitive people should consider limiting prolonged outdoor exertion.
Unhealthy for Sensitive Groups	The following groups should limit prolonged outdoor exertion - People with lung disease, such as asthma - Children and older adults - People who are active outdoors
Unhealthy	The following groups should avoid prolonged outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit prolonged outdoor exertion.
Very Unhealthy	The following groups should avoid all outdoor exertion: - People with lung disease, such as asthma - Children and older adults - People who are active outdoors Everyone else should limit outdoor exertion.

The first air quality index, name the “Pollutant Standard Index” (PSI), was developed and introduced by United States Environmental Protection Agency, taking into consideration five major (criteria) air pollutants, namely, CO, SO_2, PM_{10}, O_3, and NO_2. In 1999, the index was further completed and replaced by the Air Quality Index or AQI. The most widely used index for air quality assessment and management. $PM_{2.5}$ and 8-hr average ozone [3].

Nowadays the paper about develop air quality index by used a classification is an essential technique of data mining. Such as used fuzzy inference system to separated air quality classification by used pollutant's concentration by added concentration of pollutants benzene, toluene, ethyl benzene, xylene, and 1, 3 -butadiene standards for air quality classification [4]. Used neural network Model by classification technique to forecast air quality for reduce pollution problem which population can prepare with population effect before [5] and use classification technique to make model Decision Tree. To assignment results of concentration of pollutants which influenced for healthy of population [6]. Used a decision tree to forecast daily dissolved oxygen rates in a lagoon along the French Mediterranean sea coast [7]. Including used a decision tree identifying controlling factors of ground-level ozone levels over southwestern Taiwan [8].

From this passage. Researcher was introduced rules of separated air quality classification which influenced for healthy. To support decision for separated air quality classification. By combined the information about concentration of pollutants. This paper introduced rule of separated air quality classification by level of healthy concern and used decision tree which technique of classification and can use the results of them to analysis factors is caused to happened the pollution problem more standard with directly.

II. METHODOLOGY

Aim of this paper is use data mining to create model by using decision tree with classifier technique. This paper separate step for use data mining in 5 steps with the following Fig. 2.

A. Input Air Quality Data

First, combined information about factors which influenced for levels of air Quality such as PM_{10}, PM_{25}, So_2)1 hour(, So_2
2) 4 hour (etc. Collect data about concentration of pollutants in each kinds in Thailand for 2012-2013.

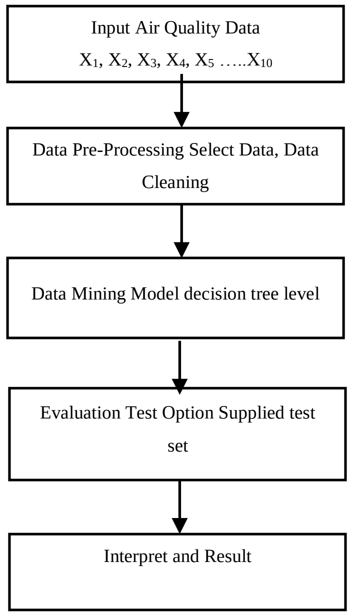

Fig. 2. Data mining in 5 steps

B. Data Pre-Processing

Using the data to Pre-processing before process chose interested Attribute and repeated data out. Included missing value data noisy data and inconsistent data. After clean the data, adapted data for using in data mining step.

Form concentration of pollution. The result of air quality’s pollution control department. The pollution in 2012-2013, Thailand has concentration of pollutants separated level which influenced for healthy in 4 levels are good, moderate, unhealthy for sensitive groups and unhealthy. From the standard of air quality classification in Thailand have 6 levels. In Table 2. By air quality index from 0-100 is an air quality in normal atmosphere. If air quality index more 100 is show that concentration of pollutants has over standard.

In this paper used a concentration of pollutants have kinds including Ozone NO_2 CO SO_2 PM_{10}. For created rule to separated levels of Healthy Concern Ozone classification in Thailand.

Table 3 Levels of Healthy Concern [9]

Air Quality Index)AQI (Values	**Levels of Healthy Concern**
0 to 50	Good
51 – 100	Moderate
101 – 150	Unhealthy for Sensitive Groups
151 – 200	Unhealthy
201 – 300	Very Unhealthy
301 to 500	Hazardous

Table 4 The Concerntration of pollutants [10]

Attribute	**Attribute Name**	**Average)hour(**	**Descriptions**
1	SO_2	24	Sulfur dioxide
2	NO_2	1	Nitrogen dioxide
3	CO	8	Carbon dioxide
4	Pm_{10}	24	Dust less than 10 micron
5	Ozone	1	Ozone Average
6	Level	-	Levels of Healthy Concern

III. DATA MINING

A decision Tree is decision method. It consists of a root, nodes, branches and leafs) terminals (which the results will happened when the situation started, it shows in decision form and divided in each ways to decision. The Following Fig. 3.

Decision tree model started to separate air quality classification of decision from "root" calculated information gain to used attribute in each nodes of tree attribute. Anyone has most the information gain result or less Entropy result will be attribute of node. And remaining data will calculate information gain again. Using the following formula

Entropy equation

$$\text{Entropy (s)} = \sum_{i=1}^{e} -p_1 \log_2 p_1 \quad (1)$$

By S is attribute to be measured.

P_1 is ratio of members in groups to the number all members of sample

Information Gain

$$\text{GAIN (S, A)} = \text{Entropy (S)} - \sum_{\text{value (A)}} \frac{|S_V|}{|S|} Entropy(S_V) \quad (2)$$

By A is attribute A

S_v is members of attribute V valuable

S is number of samples

Split Information

$$\text{Split Information (S, A)} = - \sum_{i=1}^{n} \frac{|S_1|}{|S|} \log_2 \frac{|S_1|}{|S|} \quad (3)$$

Gain Ratio

$$\text{GAIN RATIO(S, A)} = \frac{\text{Gain}(S, A)}{\text{Split Information}(S, A)} \quad (4)$$

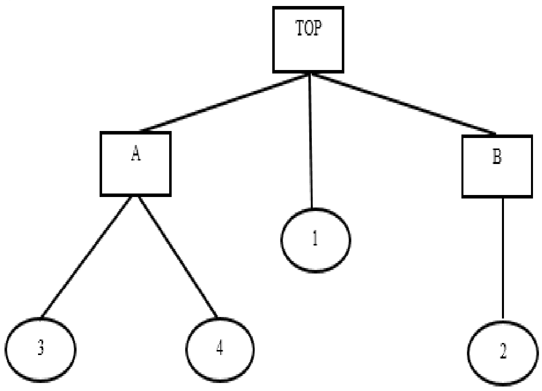

Fig. 4. Decision tree model

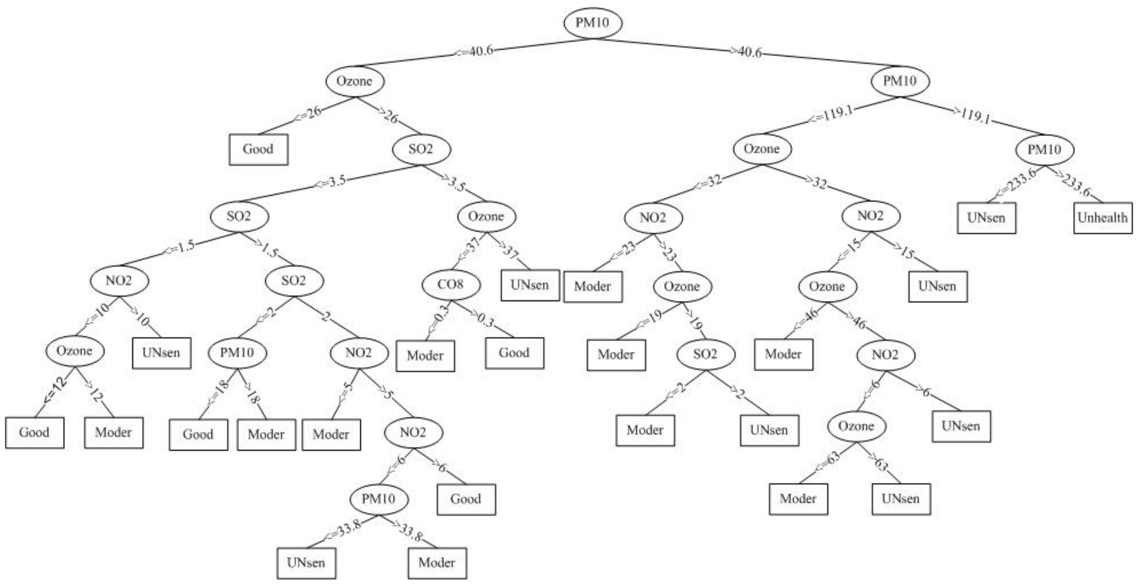

Fig. 3. Shows the tree from the classification of the Air Quality Index to Levels of Health Concern with the decision tree.

A. *Evaluation, Interpret and Result*

Evaluation interpret and result are data which processed by using attribute the following table 3. In data mining, research chose to use decision tree

technique for create air quality classification model. For separate levels of air quality which influence for healthy. Using examined data in test option supplied test set. Divide in a sets first set is training set 70% and test set 30% for testing model's quality. Last step is compare efficient model in ROC curve form compare results between ROC curve form on training set and test

IV. RESULT AND DISCUSSION

In classification, separate data in levels which influence for healthy include Good, Moderate, Unhealthy for sensitive groups and Unhealthy by using algorithm decision tree following Fig. 4. Used evaluation test option supplied test set which divided air quality data in 2 sets are 70% for training set and 30% for test set.

Result of correctly classified instances 's training set can predict data with correctly 96.8% has Incorrectly Classified Instances 3.55% and result correctly classified instances of test set is 91.07% incorrectly classified instances 8.93% following Table 4.

Table 5 show result of correctly classified Training set data and Test set data

Data	**correctly Instances**
Training Set	96.8%
Test Set	91.07%

From result of training set and test set can create receiver aerator characteristic or ROC curve to make relative graph between true positive rate with false positive rate by cut – off point Following Fig. 5.

Compared efficient for process result's algorithm between training set data and test set data. ROC curve result and cut point of Training set is X (0.79), Y (0.997) and cut point of test set X (0.121), Y (0.999) that show training set data and test set data have algorithm nearby results.

This paper used concentration of pollutants in Thailand based, include concentration of pollutants in air 5 kinds are Ozone, NO_2, CO, SO_2 and PM_{10} in each provinces. To create model with decision tree for use rules to separate air quality which levels to influence healthy.

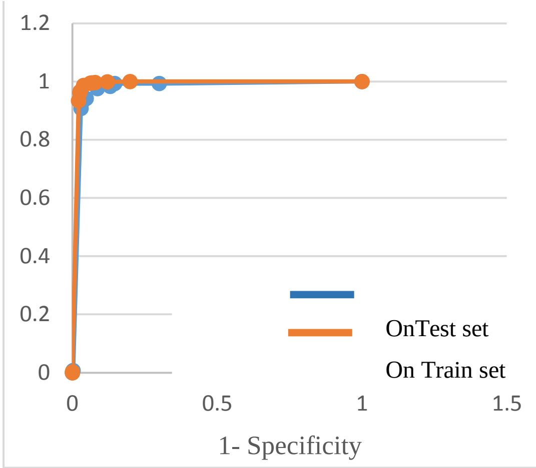

Fig. 5. Shows result of ROC curve on Trainnig set and Test Set

From Fig. 6, use rules of decision tree with classification technique processed to show that levels of healthy concern in each provinces in Thailand. The concentration of pollutants in 5 kinds are Ozone, NO_2, CO, SO_2 and PM_{10} to show that levels of healthy concern in colors each that provinces. They have 6 levels in air quality in Thailand based. First level good is green, Second level moderate is yellow, third level unhealthy for sensitive groups is orange, Forth level unhealthy is red, Fifth level very unhealthy is purple and Sixth level hazardous is maroon [11] which the colors will change with input data in each areas Following Fig. 7.

V. CONCUSSION

For classification of air quality that influence to healthy population in the future. It can use in difference places. In Air quality paper that influenced to healthy including concentration of pollutants in each station of Thailand to fix problems in each points more. That shows about factors were relative or change result of concentration of pollutants in each kind concentration of pollutants may be change away. This paper, researcher use 5 variants are Ozone, NO2, CO, SO2 and PM10 which others variants with air quality. And can use to analysis in same ways.

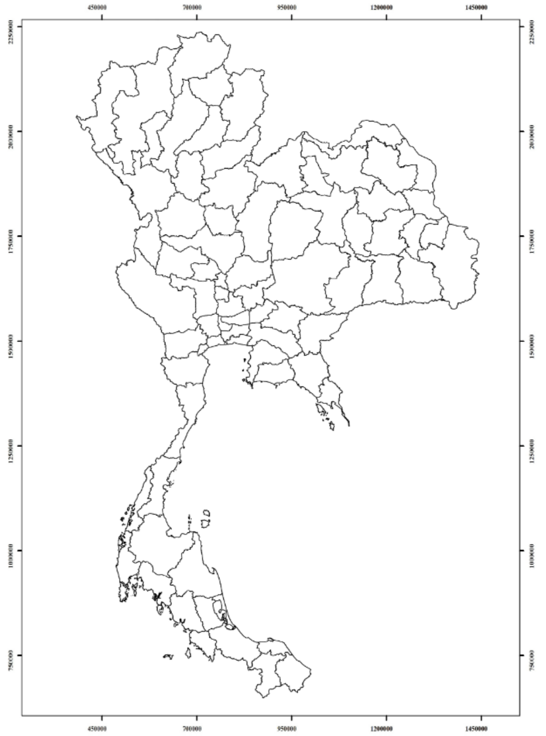

Fig. 6. The map of Thailand

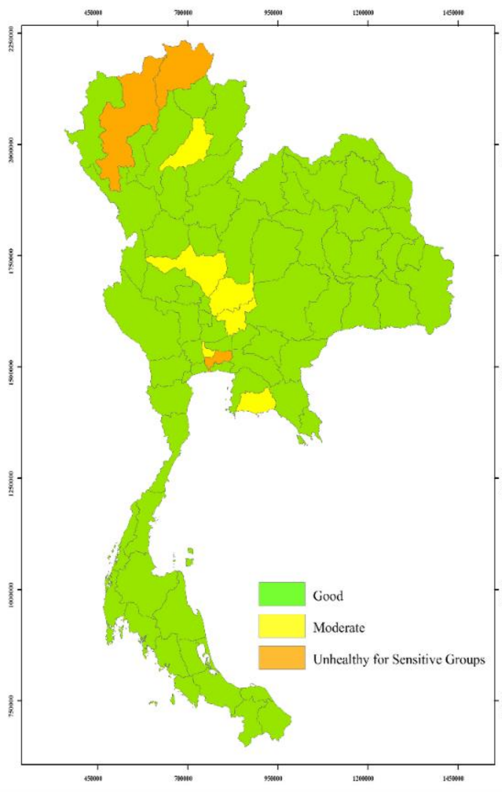

Fig. 7. Shows color level of Healthy Concern on map of Thailand

References

[1] Ioannis N. Athanasiadis, Kostas D. Karatzas, Pericles A. Mitkas3, “Classification techniques for air quality forecasting,” *Conference*, pp.1-7, August. 2006.

[2] komchadluek) .2013, May 28(. People's problems! Air pollution [Online]. URL http://www.komchadluek.net

[3] airnow . Air Quality Index (AQI) - A Guide to Air Quality and Your Health [Online]. URL http://airnow.gov

[4] Pollution Control Department . Air Quality [Online]. URL http://www.pcd.go.th

[5] Mohammad Hossein Sowlat, Hamed Gharibi, Masud Yunesian, Maryam Tayefeh Mahmoudi, Saeedeh Lotfi. “A novel, fuzzy-based air quality index (FAQI) for air quality assessment,” Volume 45, Issue 12, pp. 2050–2059, Apr. 2011.

[6] KAŠPAROVÁ MILOSLAVA, KŘUPKA JIŘÍ, “Air Quality Modelling by Decision Trees in the Czech Republic Locality,” *Conference*, pp. 1-6, August 2008.

[7] Hand, David J, “Measuring classifier performance: A coherent alternative to the area under the ROC curve " Machine Learning, Volume 77, pp 103-123, Jun 2009.

[8] M. Amarnath, V. Sugumaran, Hemantha Kumar, “Exploiting sound signals for fault diagnosis of bearings using decision tree,” *Journal*, vol. 46, pp. 1250–1256, Apr. 2013.

[9] Mevlut Ture, Fusun Tokatlib, Imran Kurtc “Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients,” *Journal*, vol. 36, pp. 2017-2016, Mar. 2009.

[10] Hone-Jay Chu, Chuan-Yao Lin, Churn-Jung Liau, Yi-Ming Kuo, “Identifying controlling factors of ground-level ozone levels over southwestern Taiwan using a decision tree,” *Journal*, vol. 60, pp. 142-152, Dec. 2012.

[11] D. Nerini, J.P. Durbec, C. Mante, “Analysis of oxygen rate time series in a strongly polluted lagoon using a regression tree method,” *Journal*, pp. 95–105, 2000

BIOGRAPHY

NAME	Miss Kattariya Kujaroentavon
DATE OF BIRTH	27 April 1991
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Prince of Songkla University, 2009-2012
	Bachelor of Sciences Program
	(Information Technology Business)
	Mahidol University, 2013-2015
	Master of Science (Information Technology Management)
HOME ADDRESS	227/91, Makamtier Sub-district, Muang District Suratthani 84000
	Tel 081-477-5519
	Email: kat_aretee@hotmail.com
PUBLICATION / PRESENTATION	Air Quality Classification in Thailand Based on Decision Tree, The proceeding of the 2014 7th Biomedical Engineering International Conference (BMEiCON), pp. 1-5.