COMBINATION OF ENVIRONMENTAL AND ECONOMIC FACTORS IN CORONARY HEART DISEASE PREVALENCE IDENTIFICATION

Chatnarong Fukprapi [1], Sotarat Thammaboosadee [1,*], and Jean Paolo Lacap [2]
[1]Information Technology Management Division, Faculty of Engineering, Mahidol University, Thailand
[2] School of Business and Accountancy, Holy Angel University, Philippines

ABSTRACT

World Health Organization (WHO) reports that coronary heart disease (CHD), one of the non-communicable diseases (NCD), is the leading cause of death around the world. The main risk factors are mostly medical factors such as hypertension, diabetes and physical inactivity. This research proposes new additional factors including economic and environmental factors to create a predictive model of coronary heart disease in global aspect using data mining process. The based medical risk factors and new blended variables were reviewed from WHO report and some reliable related research. The historical data were collected from public health organization reports. The classification techniques used to predict for the prevalence of coronary disease were experimented by several techniques. The finding of this research showed that the decision tree algorithm provided the best classification model, and gradient boosted tree algorithm provided the best regression model. The most important factor of the decision tree model was an average income per household. The result of this research can present a risk of CHD on visualization to support the management of medical resources.

Keywords: data mining, coronary heart disease, predictive modeling

1. INTRODUCTION

The number of Non-communicable diseases (NCDs) patients seriously gradually increases every years. NCDs also known as chronic diseases, tend to be of long duration and are the result of a combination of genetic, physiological, environmental and behaviours factors. The typical kind of NCDs are cardiovascular diseases (e.g. heart attacks and stroke), cancers, chronic respiratory diseases, and diabetes. World health organization (WHO) reported that there were 70% of all deaths worldwide.

Manuscript received on Sep 20, 2018; revised on Nov 16, 2018.
*Corresponding author Email: sotarat.tha@mahidol.ac.th Information Technology Management Division, Faculty of Engineering, Mahidol University, Thailand

Moreover, in low and middle-income countries, 82% of premature deaths (died before reaching 70 years of age) were also caused by NCDs [1]. NCDs are also one of the serious concerns in Thailand which has the highest mortality rate comparing with other diseases [2]. In 2015, 71% (354,000 people) of deaths were caused by NCDs [3].

Bureau of Non-communicable Disease [4] reported that the mortality rates caused by NCDs including cardiovascular diseases (CVDs), ischemic heart disease, stroke, hypertension, and diabetes have been rapidly increasing, especially the cardiovascular diseases Figure 1 presents the mortality rates of NCDs including Cardiovascular Diseases (I00 - I99), Ischemic heart disease (I20 - I25), Stroke (I60 - I69), Hypertension(I10 - I15), and Diabetes (E10 – E14) from 2008 to 2014 [4].

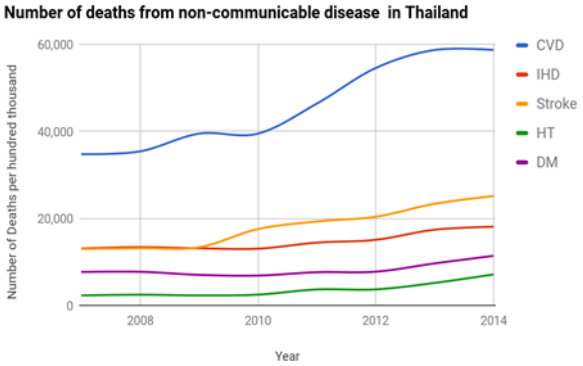

Figure 1. The Number of deaths from NCD in Thailand.

CVDs are the diseases that associated with heart and blood vessels. According to the Figure 1, the CVDs is the highest mortality rate number comparing with others diseases. Moreover, the WHO reports that Coronary Heart Disease (CHD) is the most violent CVDs which can have a directly negative effect on human living [5]. Therefore, the prevention of CHD is strongly necessary to establish for increasing the quality of life of the citizen [6].

Concerning to the predictive modeling, most of existing research employed the medical information to predict coronary heart disease, whereas the finding from relevant research found that the additional non-medical factors such as economic and environmental factors [7][8]. They are also related to occur of CHD. Therefore, they should be considered to be involved for modeling to increase the performance of model.

According to the described motivation, this paper aims to propose a CHD risk predictive model based on

the opened provincial data of medical, economic, and environmental factors integrating with medical statistic data.

2. RELATED RESEARCH

There were some published research related to data science application in cardiovascular-family disease with factors from various fields. Ghosh [7] studied the association between death rates of cardiovascular and air pollution in the Southern part of California. The carbon dioxide emissions and PM2.5 were selected to find the relationship with the number of deaths by CHD in individual area. The results of this study showed that the burden of near-roadway air pollution (NRAP) had high possibility to increase the prevalence of CHD. Thus, this research suggested that reducing greenhouse gases can consequently reduce the negative effects of air pollution on human health. This research strongly confirms that air pollution affects increasing of the prevalence of CHD. Therefore, it suggests that air environment factors should be included in modeling as one of risk factors to predict the trend of coronary heart disease in the future.

According to the study proposed by Munzel [8], this research studied health effects which were caused by noise from road site. This research reviewed various related research domains [9] [10] in both direct and indirect effects then the finding found that the noise affects to low life quality due to insufficient of sleep time. Moreover, in the field of epidemiology, the noise was proved that it has a relationship with increasing the number of CHD patients. The advantage of this research is a description in detail of several ways of noise effect on human health, and identify other risk factors which lead to the cause of CHD.

Kim, J. Lee and Y. Lee [11] proposed the classification data mining model for prediction the occurrence possibility of CHD by using risk factors consists of age, sex, cholesterol, systolic, diastolic, smoking, and diabetes. Those factors were gathered from the Korean Nation Health and Nutrition Examination Survey VI with total 748 cases. The hybrid techniques of fuzzy logic and decision tree are integratively constructed the model. Interestingly, this kind of model had higher effective than Artificial Neural Network (ANN), Logistic Regression (LR), Support Vector Machine (SVM), and traditional C5.0 Decision Tree. However, from the future model development plan, the effective of the model still needs to be improved because this results of this study have a low accuracy (maximum is 0.6951). This study in the appropriate model for learning how to adapt the Hybrid technique for build the model.

Vijayashree and Narayanaiyengar [12] studied the prediction of the CHD prevalence by using data mining process for proving the importance of prediction before occurring of the disease. There collected and selected the risk factors from seven existing works [13] [14] [15] [16] [17] [18] [19] which are age, angina, blood cholesterol levels, diabetes, diet, genes, hypertension, obesity, physical inactivity, smoking, and work stress. Moreover, all of those research indicated in detail of what is the essential indicators for prediction. This in-depth review is imperative for factors selection for CHD predictive model building.

Additionally, Abdulah [20] presented the data mining model for prediction of number of CHD patients by using age, sex, trestbps, chest pain type and biochemical factors (serum cholesterol;TC, fasting blood sugar; FBS, thalach, exang, old peak, restecg, slope, the number of vessels colored by fluoroscopy and thal) as the risk factors for analysis. This study compares two methods which are Random Forest Classifier and Decision trees. The results show that random forest classifier has higher accuracy and precision. The advantage of this study is to study from big data. These can get more variety of risk factors. However, it still limited to the method of analysis which uses only algorithm; tree model type. These gave high accuracy (63.33). This study should apply other methods such as neural network or support vector machine to analyst the data. It might be increasing the accuracy of the experiment.

3. RESEARCH METHODOLOGY

This research focuses on the development of coronary heart disease risk predictive model. The model was developed based on the provincial data of risk factors of coronary heart disease which is provided from the public organization in Thailand. The data mining technique is chosen as a tool for data analysis. A schematic diagram in Figure 2 describes the overall methodology framework of this research.

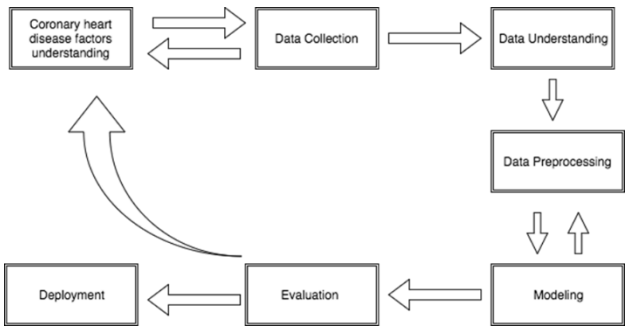

Figure 2. Research procedure.

3.1 Coronary heart disease factors understanding

For understanding, many reports and research were reviewed for identifying the possible risk factors. For simplification, the sources were divided into three main sources; WHO, Bureau of Non-Communicable Disease of Thailand, and related publications. The risk health factors were gathered from the first two sources. The economic and environmental factors were selected from the related researches. The process is shown in Figure 3. Therefore, the scoped of risk factors are hypertension, tobacco use, diabetes, physical inactivity, unhealthy diet, cholesterol,

overweight, drinking alcohol, air pollution, household income.

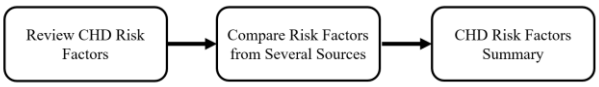

Figure 3. Risk factor understanding process.

3.2 Data collection

The numbers of patients under CHD risk factors category were selected to represent the situation and possibility of CHD in Thailand. The available data were collected from four sources which are Health Data Center (HDC), Government Open data Thailand, National Statistical Organization Thailand (NSO), and Ministry of Energy. The data collection process is followed as Figure 4.

Figure 4. Data collection process.

3.2.1 Results of frequency of engagement

This system is developed by Ministry of Public health to collect data from all hospital in Thailand. This system provides the CHD data and some of CHD risk factors. The benefit of this system is that the data is primary instead of estimation. However, disadvantage of this system is about data lacking because of unreported data of some hospitals.

The number of patients with coronary heart disease, hypertension, diabetes and the prevalence of obesity were collected from this source during 2013 to 2015.

3.2.2 Data from Government Open Data Thailand

This system was developed by government organization for the purpose of easily access to government data. The data were analyzed and submitted by organizations which has a directed responsibility on the data such as the Ministry of Public Health, the Ministry of Energy, and the National Statistical Office. Thus, the data from this system come from various sources. This system also provides data via open data platforms. The advantage of this system is easy-to-access and well-categorizing. Disadvantage of this system is that most of data conducted by survey method which may lead to bias. The CHD risk factor data that were selected from this system is the household income data and debt.

3.2.3 Data from National Statistic Organization (NSO)

NSO is an organization under the ministry of digital economy and society. The responsibilities of NSO are statics operation to support effective planning and decision support. In addition, the government policies was monitored and evaluated by this organization. NSO also conducts other necessary database which are produced through efficient processes and expertise for all sectors.

The risk factors data of CHD including alcohol, tobacco, physical activity, and nutrition. These data were surveyed only 3 year in province level. This system also provided number of patient with hypertension, number of patient with diabetes and number of patient with CHD In 2001 to 2012.

3.2.4 Data from Thailand Energy Ministry

The Ministry of Energy conducts public website for providing the energy data. The purpose is to create powerful energy database which can be used to prepare provincial energy strategic plan and increase knowledge and understanding of the relevant sectors for energy database usage. The website has opened the information include energy infrastructure, power consumption, petroleum reserves, greenhouse gas emissions and energy projects. These data were reported in 2001 to 2015. Since the finding of research found that greenhouse gas emissions is one of coronary heart disease risk factors, it will be included for consideration.

3.3 Data understanding

Since the data has been collected from many sources, the analyzed pattern of these data is necessary to identify a number and characteristic of the attributes in order to perform data integration. The data understanding process started from describing data, data exploration, validating the data quality, as shown in Figure 5.

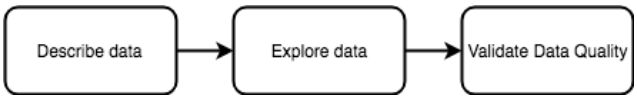

Figure 5. Data understanding process.

3.4 Data preprocessing

The results of data understanding show that some data were reported every year while data from other sources were collected biannually which cause the missing value in yearly manner. In the first step, all data was transformed to the same pattern then integrate them together using province name and the year of data. After data were integrated, the records that have missing attribute was removed, and the outlier detection was performed to detect abnormal data. The discretization method was performed to suppurate the prevalence of CHD. Finally, they were normalized using z-transformation to reduce the variance of data. All data preprocessing process presented in Figure 6.

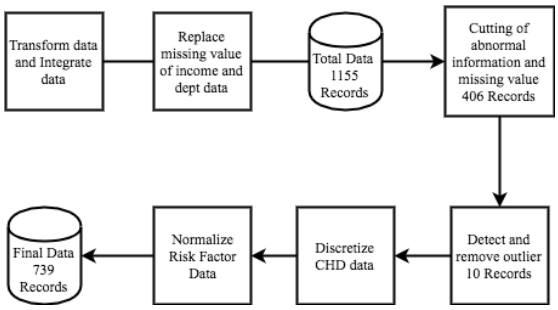

Figure 6. Data preprocessing process.

3.4.1 Data transformation and Integration

Since the data was collected form different sort, they have different formats. They have to be transformed into the homogeneous structure. The first process of transformation is joining of the province name and year into single dataset. The consequence of data integration is shown Table 1.

Table 1: The joined data

Attribute	Description
Province	Province name
Year	Fiscal year
CHD	number of patients with coronary heart disease
Obesity	number of population who have BMI ≥ 25 kg/m^2
Hight_Blood	number of patients with hypertension
Diabetes	number of patients with diabetes
Alcohol	Proportion of Alcohol consuming
Tobacco	Proportion of Tobacco consuming
Income	Average Monthly Income Per Household
loan	Average Debt Per Household
CO_2	concentration of carbon dioxide
N_2O	concentration of nitrous oxide
CH_4	concentration of methane

3.4.2 Replacing of missing value of income and debt data

According to both income and debt data, they were surveyed biannually. The lacking data was replaced by the interpolation method, the calculation of an average value between next year and previous year.

3.4.3 Removal of abnormal data and missing value

The finding of data exploration with Figure 7 found that the abnormal data in 2007 and 2008 might be occurred by the error of data input process due to unfound of the news or phenomena which are causes of GHGs increasing. The Figure 7 presents CH_4, CO_2, and N_2O which increase abnormally. To avoid error in prediction, the data in 2007 to 2008 were removed. Moreover, the data of Bangkok and Bung Kan province were not included in this studied because they were not reported in this data source. Apart from record dimension and focusing on attribute projection, the alcohol, tobacco, physical activity, nutrition, obesity, and poverty were removed because these data were reported less than 5 year. It reduce the number of training data that mean it may reduce the performance of model. Table 2 shows the removed attributes with their removal criteria.

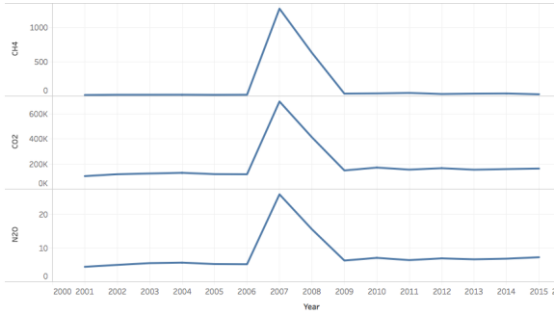

Figure 7. Abnormal data during 2007 to 2008.

Table 2: The removed attributes

Attribute	Reason for removing
Alcohol	The data were collected only 2 years.
Tobacco	The data were collected only 3 years.
Obesity	The data can integate with the others only 3 years.
Physical activity	The data were not reported in province level.
Nutrition	The data were not reported in province level.
Diabetes	The data were collected only 2 years.

3.4.4 Outlier detection

In this part, the CHD data were suppurated by frequency. The researcher experimented two method of separated CHD data. The first method is suppurating CHD data to 3 groups: [0, 1108.5], [1108.6, 2021.0, and [2021.1, infinity). The number of record of data per group was shown in Figure 8. The second method is suppurating CHD data to 2 groups: [0, 1521.5] and [1521.6, infinity). The number of record of data per group was shown in Figure 9.This system was developed by government organization for the purpose of easily access to government data. The data were analyzed and submitted by organizations which has a directed responsibility on the data such as the Ministry of Public Health, the Ministry of Energy, and the National Statistical Office. Thus, the data from this system come from various sources. This system also provides data via open data platforms. The advantage of this system is easy-to-access and well-categorizing. Disadvantage of this system is that most of data conducted by survey method which may lead to bias. The CHD risk factor data that were selected from this system is the household income data and debt.

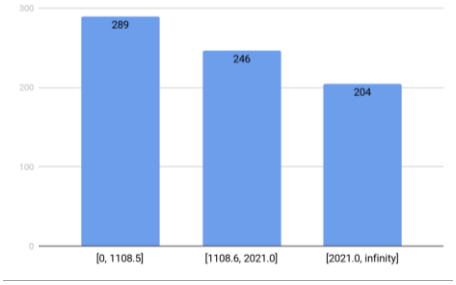

Figure 8. The number of records per group in method 1.

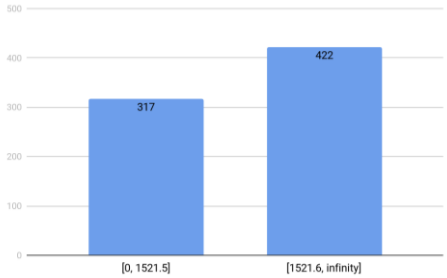

Figure 9. The number of records per group in method 2.

3.4.5 Data Normalization

The risk factor data that were used in this research is medical factors, economic factors, and environment factors that mean they have different unit. This research chose z-transformation [22] technique to fit in a specific range. The range was rescaled based on arithmetic mean and standard deviation which will transformed into -3.00 to +3.00 approximately.

3.4.6 Feature Selection

Feature selection is a technique to select the significant attribute from massive and high-dimension dataset. This research applied feature selection technique including Sequential Forward Feature Selection (SFFS), Sequential Backward Feature Selection (SBFS), and Principal Components Analysis (PCA) to select the significant risk factor of CHD.

Sequential Forward Feature Selection (SFFS) [23] starts from empty attribute and selects one attribute then adds it in the testing process. The testing process, it is creating the model from the attributes which were selected. In this research use linear regression to model the testing model. Each step of adding new attributes, only one attribute which has the highest performance was selected. Forward selection will stop adding new attributes when performance is not increased.

Sequential Backward Feature Selection (SBFS) [23] starts from all of the candidate attributes and removes one attribute. The testing process of backward selection like the testing process of forward selection. Each step of removing attributes, the attribute which makes the model has the highest performance, when it was removed will be removed.

Principal component analysis (PCA) [24] is decreasing the dimension of data by creating new attribute from candidate attribute. The lack of interpretation of new generation is the disadvantage of PCA.

3.5 Model building

The model building process was developed by using data mining classification technique. In the classification technique, there are two types of algorithm: classification and regression. This research compares algorithm by measuring the performance of model. The classification algorithms are compared between decision tree and support vector machine. The regression algorithms are comparative experimented among linear regression, neural network, polynomial regression, support vector machine, and gradient boosted trees. The appropriate parameters of each algorithm were found by experimentation. The Figure 10 presents process of model building.

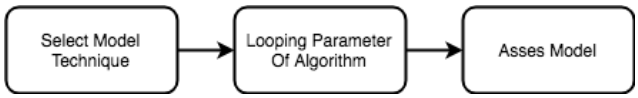

Figure 10. Model building process.

3.5.1 Decision tree

C4.5 Decision tree algorithm [25] creates a decision tree from the sample data using the concept of information entropy which selects the best attribute to create a node for making a decision. The gain value are used as the criteria for choosing the root node. The node which has highest score of gain will be selected to be root node.

3.5.2 Linear regression

Linear regression [26] is a statistical technique that used for numerical prediction. The model of linear regression is the equation that determine the strength of the relationship between dependent variable. Figure 11 is an example of linear regression which shows the relationship of independent variable and dependent variable. The result is easily applied but it is usually not fit with the high complexity data.

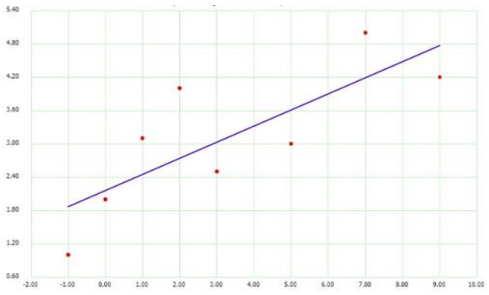

Figure 11. Example of linear regression model.

3.5.3 Polynomial regression

Polynomial regression [27] is a technique for numerical prediction likes linear regression but its equation is nonlinear. The higher degree of polynomial

produces more critical points. Figure 12 shows an example of third degree polynomial regression. The advantages of this method is the model may fit with the high complex data more than linear regression whereas applying polynomial regression with small data, the model may over fitting.

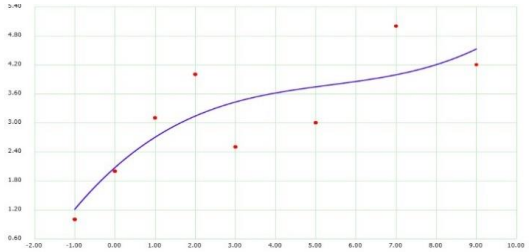

Figure 12. Example of polynomial regression model.

3.5.4 Support Vector Machine

Support Vector Machine (SVM) [28] is usually used for identification the pattern of data by set all collected data into the feature space and create a straight line for the segmentation of the data; called Hyperplane. Many straight lines are created to separate the data set. Thus, calculation for finding the value of Margin is important part of this method. The Margin value is the sum of distance between the straight lines which is using for separate data set (separate line) to the straight line which is the most closest to the data and parallel with the separate line. The straight line which selected for study have to have highest Margin value when they are compared with others. Figure 13 demonstrates the SVM model.

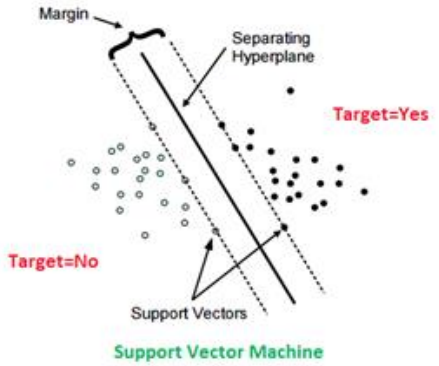

Figure 13. Support Vector Machine.

3.5.5 Neural network

Neural network (NN) [30] is a simulation of human brain function for supervised learning problem. The first step is creating a node similar with neurons, and each of node has an equation to calculate the input data. The simple calculation is usually applied in this method which is sigmoid function. The sigmoid function is similar to nerve cells of human which is used for remembering. The nodes are connected with each node to be a network. The line that connects the node to other nodes they have their specific weighting values. The weighting value is used as a factors to multiply with the value which are sent from the node. The nodes in the neural network are divided into three types. There are nodes for receiving input data (input nodes); the number of input nodes related to attribute data, nodes for delivering the output data (output nodes) and hidden nodes (located between the input nodes and output nodes) for increasing effective of learning [31]. Figure 14 presents the three layers Neural Network with one or more hidden layer and Figure 15 shows the neural network structure including bias and activation function for analyze input data and calculate the result.

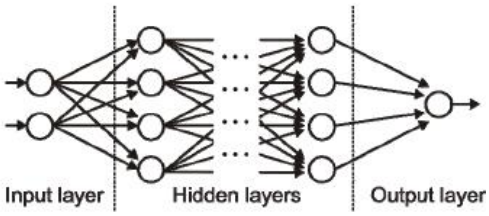

Figure 14. Neural Network Structure.

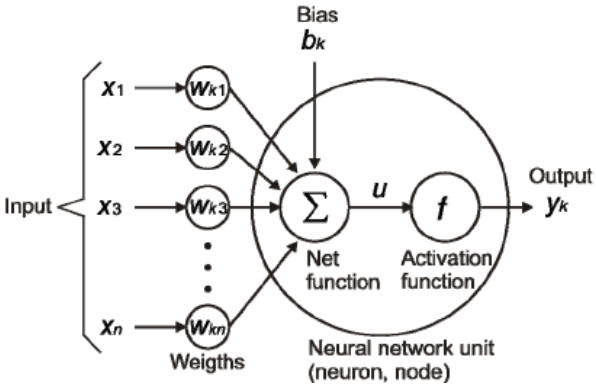

Figure 15. Neural network structure with activation function.

3.5.6 Gradient boosted tree

Gradient Boosted Tree (GBT) [32] creates multiples decision tree for making more powerful decision. The concept of gradient boosted tree is building the series of decision trees by the pieces of tree attempts to correct the mistake of previous tree, as demonstrated in Figure 16.

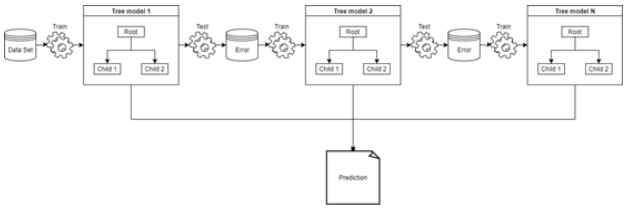

Figure 16. Gradient boosted tree.

3.6 Evaluation

The evaluation model used in this research is k-fold cross-validation [33]. This technique separate one data set to k data set. One of all data set is testing set and the rest

are training set. The selected of testing set will be cycle until all set was selected to be testing set. In testing process measured accuracy and relative error to identify the best model that has the lowest relative error or highest accuracy. This research use k equal to three due to the limited of training data.

3.7 Deployment

The visualization will be used to present the results from predictive model and also it can analyze prevalence and forecast future trend of CHD.

3.8 Setting up of experiment

According to all described processes, the algorithms and methods are then experimented for hyper-parameter tuning in order to find the best combination of data pre-processing and classification and regression algorithms and their parameter set. The list of experimental setup is shown in Figure 17.

Feature selection methods
- SFFS
- SBFS
- PCA
- Do nothing

Algorithms
- NN
  - Hidden Layer: [1,2]
  - Hidden node: [2, 20]
  - Learning rates: [0.1, 0.3]
  - Momentum: [0.1, 0.3]
  - Decay: [0,1]
- SVM
  - Nu: [0.1, 0.5]
  - Gamma: [0, 100]
  - Kernel type: Sigmoid
- Linear regression
  - Max iterations: [1, 100]
- Polynomial regression
  - Max degree: [2, 5]
  - Replication [1, 5]
- GBT
  - Maximal depth: [5, 10]
  - Learning rates: [0.025, 1]

Figure 17. The setting up of experiment.

4. RESULTS AND DISCUSSION

This section presents the experimental results in performance of model which based on linear regression, polynomial regression, ANN, SVM, GBT, and decision tree. Preliminarily, Table 3 and 4 show the correlation test of all attributes.

Table 3: The correlation test part 1

Arrtibutes	CHD rate	Obesity	CO_2	N_2O	CH_4	Income
CHD rate	1	0.270	-0.004	0.004	-0.018	-0.279
Obesity	0.270	1	0.366	0.388	0.325	0.003
CO_2	-0.004	0.366	1	0.998	0.966	0.139
N_2O	0.004	0.388	0.998	1	0.958	0.147
CH_4	-0.018	0.325	0.966	0.958	1	0.029
Income	-0.279	0.003	0.139	0.147	0.029	1
debt	-0.025	0.116	0.045	0.054	-0.009	0.611
HT rate	-0.150	-0.161	-0.033	-0.030	-0.041	0.547
DM rate	-0.117	0.066	-0.038	-0.035	-0.042	0.538
Tabaco	0.301		-0.004	-0.004	0.019	-0.404
Alcohol	0.237		-0.241	-0.240	-0.226	-0.367

Table 4: The correlation test part 2

Arrtibutes	Debt	HT rate	DM rate	Tabaco	alcohol
CHD rate	-0.025	-0.150	-0.117	0.301	0.237
Obesity	0.116	-0.161	0.066		
CO_2	0.045	-0.033	-0.038	-0.004	-0.241
N_2O	0.054	-0.030	-0.035	-0.004	-0.240
CH_4	-0.009	-0.041	-0.042	0.019	-0.226
Income	0.611	0.547	0.538	-0.404	-0.367
Debt	1	0.406	0.418	-0.034	0.184
HT rate	0.406	1	0.949	-0.291	-0.145
DM rate	0.418	0.949	1	-0.233	-0.084
Tabaco	-0.034	-0.291	-0.291	1	0.130
alcohol	0.184	-0.145	-0.145	-0.130	1

From Table 3 and 4, the CHD data has weak uphill linear relationship with all factors. The HT rate has strong uphill linear relationship with DM rate. Income has moderate uphill relationship with debt. Interestingly, it also has the same direction of relationship with HT rate and DM rate, which are prior proved medical factors. The environmental factors has strong uphill linear relationship among themselves.

4.1 The model performances

The first step of comparative performance experiment is to compare the feature selection techniques which are forward selection (SFFS), backward selection (SBFS), PCA, and no feature selection approach. The second step is comparing discretizing of CHD data between 2-class and 3-class fashion. The last step is comparing the performance of predictive algorithms which are Linear Regression (LR), Polynomial Regression (PR), Neural

Network with 1 hidden layer (NN1), Neural Network with two hidden layers (NN2), Gradient Boosted Tree (GBT), and Support Vector Machine (SVM). The relative error is used as an indicator of numerical regression model as shown in Table 5 while the accuracy is applied for nominal classification model, as shown in Table 6.

Table 5: The relative error of regression model

Algorithm	Feature selection	Relative error	
		2-class	3-class
LR	SFFS	0.3983	0.3384
	SBFS	0.3721	0.3559
	PCA	0.3969	0.3430
	Do nothing	0.3722	0.3514
PR	SFFS	0.3839	0.3400
	SBFS	0.4108	0.4428
	PCA	0.3797	0.3310
	Do nothing	0.5143	0.5356
NN1	SFFS	0.4023	0.3337
	SBFS	0.3673	0.3452
	PCA	0.3976	0.3395
	Do nothing	0.3750	0.3435
NN2	SFFS	0.4039	0.3138
	SBFS	0.3799	0.3307
	PCA	0.4007	0.3148
	Do nothing	0.3789	0.3398
GBT	SFFS	0.3664	0.3134
	SBFS	0.2445	0.2854
	PCA	0.2817	0.3086
	Do nothing	**0.2110**	0.2797
SVM	SFFS	0.5003	0.2294
	SBFS	0.5021	0.2280
	PCA	0.4999	**0.2267**
	Do nothing	0.5023	0.2281

Table 6: The accuracy of classification model

Algorithm	Feature selection	Accuracy	
		2-class	3-class
DT	SFFS	**0.851**	0.705
	SBFS	0.825	0.696
	PCA	0.596	0.397
	Do nothing	0.841	0.688
SVM	SFFS	0.668	0.411
	SBFS	0.661	0.422
	PCA	0.606	0.412
	Do nothing	0.669	0.444

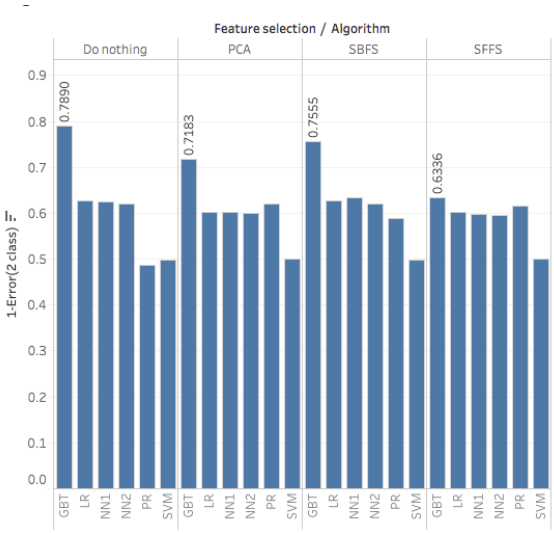

Figure 18. The performance of regression model with two-class discretization.

Figure 18 shows the performance of regression model measured by the relative error. These models predict the prevalence of CHD in 2-class style; high and low. The highest performance model is GBT without feature selection method that means all factors relate with the CHD prevalence.

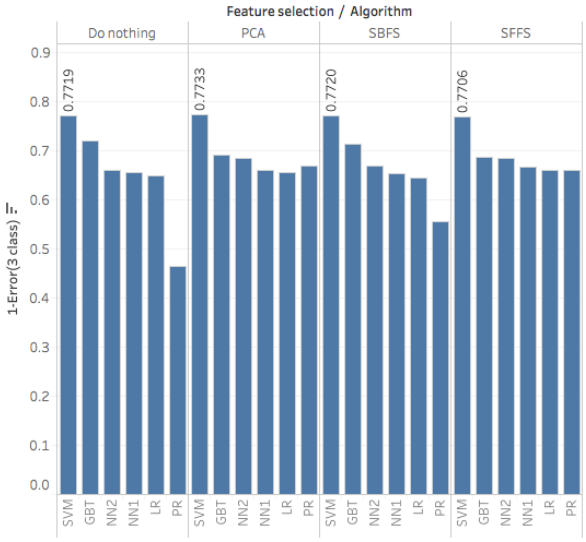

Figure 19. The performance of regression model with three-class discretization.

Figure 19 shows the performance of regression model measured by the relative error. These model has predict the prevalence of coronary heart disease in 3-class style; high, medium, and low. As can be seen the best model which has high performance is SVM and the best feature selection of this model is PCA. This experiment shows the SVM algorithm has high performance when the data has more complex but the SVM model is difficult to interpret.

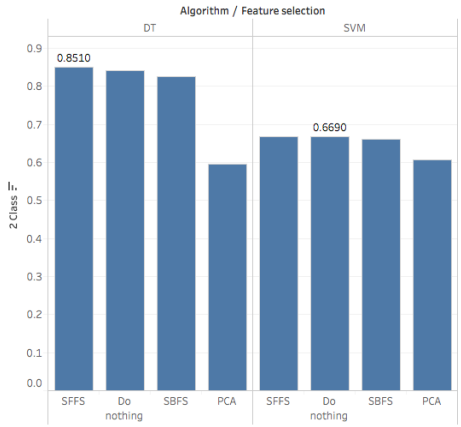

Figure 20. The comparing performance of regression model between algorithm and feature selection in 2-class discretization.

Figure 20 shows the performance of classification model by measure the accuracy in two-class style; high and low. The best model which has the highest performance is Decision Tree and the best feature selection of this model is SFFS. This model is easy to interpret and high accuracy.

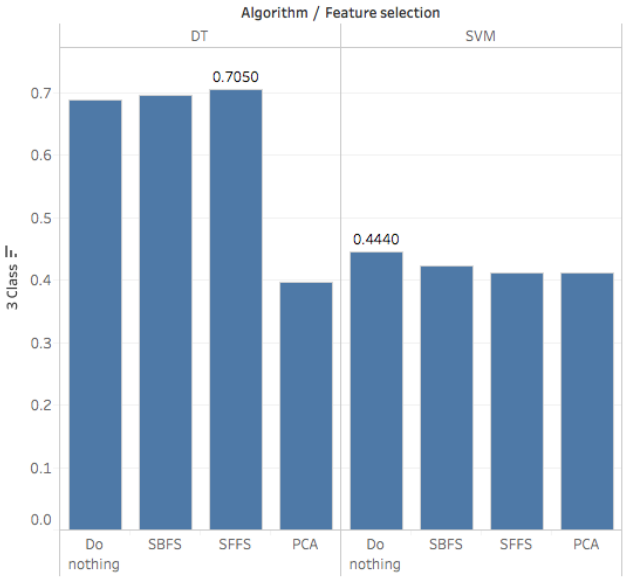

Figure 21. The performance of classification model.

Figure 21 shows the performance of classification model by measure the accuracy. These model has predict the prevalence of coronary heart disease in 3-class; high medium and low. The best model and the best feature selection are the same with 2-class classification model but the accuracy is decreased, because the 3-class discretization is more complexity than 2-class style.

4.2 The best parameters of each algorithm

The optimum parameter values for all regression and classification algorithms are presented in Table 7 and Table 8 respectively.

Table 7: The best parameters of regression model

Algorithm	2-class structure	3-class structure
PR	max degree = 2 replication = 4	max degree = 5 replication = 1
NN1	training cycle = 2,500 learning rate = 0.3 momentun = 0.15 hidden nodes = 16	training cycle = 2,500 learning rate = 0.3 momentun = 0.25 hidden nodes = 18

Table 8: The best parameters of classification model

Algorithm	2-class structure	3-class structure
NN2	training cycle = 2,500 learning rate = 0.25 momentun = 0.25 hidden nodes = 18	training cycle = 2,500 learning rate = 0.1 momentun = 0.25 hidden nodes = 18
GBT	max depth = 9 learning rate = 0.025	max depth = 9 learning rate = 0.025

4.3 The deployed model

The deployed models consists of 2 types: nominal classification and numerical regression. The greatest regression model has been produced by GBT algorithm. There are 20 trees in the GBT model. The highest accuracy classification model has been produced by the DT algorithm as shown in Figure 22. The attributes description is shown in Table 9.

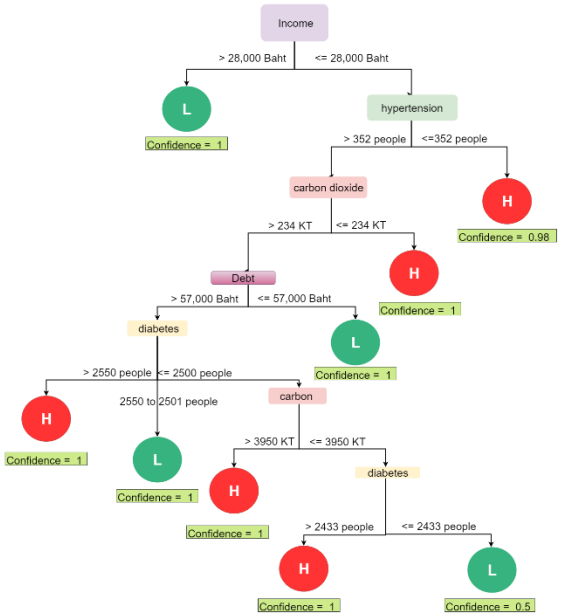

Figure 22. The decision tree model of CHD prediction.

Table 9: Attributes description of decision tree

Attribute	Description
Income	Provincial income
Hypertension	Number of patients with hypertension per 100,000 population
Carbon	Provincial carbon dioxide concentration
Debt	Provincial average debt per household
Diabetes	Number of patients with diabetes per 100,000 population

4.4 Data visualization by Business Intelligence

The results shows that the Gradient Boosted Trees is the best model for predict CHD prevalence. Approaching to more representative performance information, the data visualization is applied to present the predictive data and the prevalence of CHD. The Figure 23 and 24 presents the actual data and predictive data of CHD. Figure 23 shows the comparative provincial predictive results represented in color shade while Figure 24 shows the trend and difference of actual and predicted values.

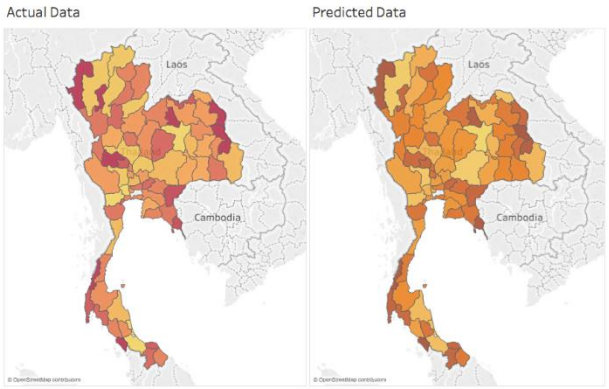

Figure 23. The actual prevalence and predictive prevalence per province.

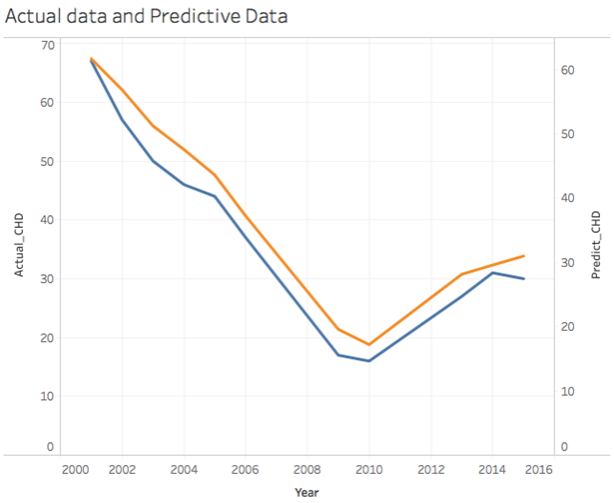

Figure 24. The actual prevalence and predictive prevalence per year.

5. CONCLUSION

The evidence from several publications attempts to define the prediction model of CHD using the medical factor. In addition, the environmental factor and economic are reported about the relation with CHD. This research attempt to collect the provincial data of CHD risk factors focus on medical, environmental and economic factors. Based on available data in Thailand, the risk factor data that are used in this research including number of patients with coronary heart disease, number of population who have BMI ≥ 25 kg/m2, number of patients with hypertension, number of patients with diabetes, average monthly Income per Household, average Debt Per Household, concentration of carbon dioxide, concentration of nitrous oxide, concentration of methane. The best classification model was created using decision tree and the best regression model was created using gradient boosted trees. The gradient boosted trees was presented in appendices. In addition, the researchers has analyze the model and found that the average income per household are the first priority which used to identify the prevalence of coronary heart disease. In deployment

part, the result of CHD prediction was presented using business intelligence to support the decision of medical service management [34].

The main limitation is the years of reported data are the difference that made the number training data are reduced when all data are integrated together. The regression algorithms are performed to create the prediction prevalence of CHD. The gradient boosted trees is chosen to create a model because it is the lowest error in prediction of CHD prevalence.

The future work will focus on improving the accuracy of 3-class classification model to increase perspective of prevalence prediction. The regression model of this research is hard to apply, it should be improved by transforming to new indicator. According to the finding of this research report that the economic factors and environmental factors related with the prevalence of CHD, we will collect more data about the economic factors and environmental factors to use in the next research.

REFERENCES

[1] "NCDs | Major NCDs and their risk factors," *World Health Organization*, 07-Apr-2016. [Online]. Available: http://www.who.int/ncds/introduction/en/. [Accessed: 02-Jan-2017].

[2] S. Srivanichakorn, "Morbidity and mortality situation of non-communicable diseases (diabetes type 2 and cardiovascular diseases) in Thailand during 2010-2014," *Disease Control Journal*, vol. 43, no. 4, pp. 379–390, Dec. 2017.

[3] "Noncommunicable Diseases Progress Monitor 2015," World Health Organization, 13-Nov-2015. [Online]. Available: http://www.who.int/nmh/publications/ncd-progress-monitor-2015/en/. [Accessed: 15-Jan-2017].

[4] "Annual Report 2015," *Annual Report 2015*, Jan-2016. [Online]. Available: http://www.thaincd.com/document/file/download/paper-manual/Annual-report-2015.pdf. [Accessed: 15-Jan-2017].

[5] "Cardiovascular diseases (CVDs)," *World Health Organization*, 17-May-2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 19-Oct-2017].

[6] K. Liu, M. L. Daviglus, C. M. Loria, L. A. Colangelo, B. Spring, A. C. Moller, and D. M. Lloyd-Jones, "Healthy Lifestyle Through Young Adulthood and the Presence of Low Cardiovascular Disease Risk Profile in Middle Age: The Coronary Artery Risk Development in (Young) Adults (CARDIA) Study," *Circulation*, vol. 125, no. 8, pp. 996–1004, 2012.

[7] R. Ghosh, F. Lurmann, L. Perez, B. Penfold, S. Brandt, J. Wilson, M. Milet, N. Künzli, and R. Mcconnell, "Near-Roadway Air Pollution and Coronary Heart Disease: Burden of Disease and Potential Impact of a Greenhouse Gas Reduction Strategy in Southern California," *Environmental Health Perspectives*, vol. 124, no. 2, pp. 193–200, 2016.

[8] T. Munzel, T. Gori, W. Babisch, and M. Basner, "Cardiovascular effects of environmental noise exposure," *European Heart Journal*, vol. 35, no. 13, pp. 829–836, Sep. 2014.

[9] W. Babisch, "Updated exposure-response relationship between road traffic noise and coronary heart diseases: A meta-analysis" *Noise and Health*, vol. 16, no. 68, p. 1-9, 2014.

[10] W. Babisch, "The Noise/Stress Concept, Risk Assessment and Research Needs" *Noise Health*, vol. 4, no. 16, p. 1-11, 2002.

[11] J. Kim, J. Lee, and Y. Lee, "Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree," *Healthcare Informatics Research*, vol. 21, no. 3, p. 167, 2015.

[12] J. Vijayashree and N. S. Narayanaiyengar, "Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 4, pp. 139–148, 2016.

[13] G. Purusothaman and P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease," *Indian Journal of Science and Technology*, vol. 8, no. 12, 2015.

[14] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," *2010 5th International Conference on Computer Science & Education*, 2010.

[15] C. S.dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, 2012.

[16] T. J. Peter and K. Somasundaram, "An empirical study on prediction of heart disease using classification data mining techniques" *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012)*, Nagapattinam, Tamil Nadu, pp. 514-518, 2012.

[17] S. Joshi and M. K. Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques" *Computational Intelligence in Data Mining - Volume 2 Smart Innovation, Systems and Technologies*, pp. 503–511, Nov. 2014.

[18] A. Makwana and J. Patel, "Decision Support System for Heart Disease Prediction using Data Mining Techniques," *International Journal of Computer Applications*, vol. 117, no. 22, pp. 1–5, 2015.

[19] S. Ratnakar, K. Rajeswari, and R. Jacob, "Prediction of Heart Disease using Genetic Algorithm for Selection of Optimal Reduced set of Attributes" *International Journal of Advanced Computational Engineering and Networking*, vol. 1, no. 2, pp. 51–55, 2013.

[20] Abdullah, A. Sheik, and R. Rajalaxmi. "A data mining model for predicting the coronary heart disease using random forest classifier." *International Conference in Recent Trends in Computational Methods, Communication and Controls*, pp. 22-25, 2012.

[21] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, Jan. 2000.

[22] E. R. Kanasewich, *Time sequence analysis in geophysics*, 3rd ed. Alberta: The University of Alberta Press, 1981.

[23] D. R. Martin, *Multiple regression analysis*. David Martin Associates, 1983.

[24] H. Hotelling, *Analysis of a Complex of Statistical Variables into Principal Components*. Baltimore: Warwick & York, 1933.

[25] J. R. Quinlan, *C4.5: Programming for Machine Learning*, Morgan Kauffmann 38, 1993.

[26] K. Hope, *Methods of multivariate analysis: with handbook of multivariate methods programmed in Atlas Autocode*. New York: Gordon and Breach, 1969.

[27] Y. Chang, C. Hsieh, K. Chang, M. Ringgaard, and C. Lin, "Training and Testing Low-Degree Polynomial Data Mappings via Linear SVM", *Journal of Machine Learning Research*, vol. 11, pp.1471–1490, 2010.

[28] M. Jändel, "A Neural Support Vector Machine," *Neural Networks*, vol. 23, no. 5, pp. 607–613, 2010.
[29] C. S. Sindhu, T. H. Sai, C. Swathi, and S. K. Babu, "Predictive Analytics Using Support Vector Machine", *International Journal for Modern Trends in Science and Technology*, vol. 3, no. 2, pp.19–23, 2017.
[30] F. Rosenblatt, "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms," *Brain Theory*, pp. 245–248, 1961.
[31] G. Palm, "Warren McCulloch and Walter Pitts: A Logical Calculus of the Ideas Immanent in Nervous Activity," *Brain Theory*, pp. 229–230, 1986.
[32] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[33] J. S. U. Hjorth, *Computer intensive statistical methods: validation model selection and bootstrap*. Boca Raton, Fla: Chapman & Hall / CRC, 1999.
[34] T. Mayakul and S. Darakorn Na Ayutthaya, "A Digital Prescription Refill System Based On Healthcare Standard In Thailand", *International Journal of Applied Biomedical Engineering*, vol. 11, no. 1, pp.28-35, 2018.

Chatnarong Fukprapi received his B.Eng. in Computer Engineering from Kasetsart University and M.Sc. in Information Technology Management from Mahidol University, Thailand, in 2011 and 2018, respectively. His research interests include Data Science and health informatics. healthcare domain.

Asst. Prof. Dr. Sotarat Thammaboosadee received the B.Eng., M.Sc. degrees in Computer Engineering and Technology of Information System Management from Mahidol University, Thailand, in 2003 and 2005, respectively. He also got a Ph.D. in Information Technology from King Mongkut's University Technology Thonburi, Thailand in 2013. He is now an assistant professor in Information Technology at Technology Information System Management Division, Faculty of Engineering, Mahidol University, Thailand. His research interests include Data Governance, Data Mining, Health Informatics, and Human Resource Analytics.

Dr. Jean Paolo G. Lacap is the Vice President for Administration and Quality Assurance of the City College of Angeles, Angeles City, Philippines. He obtained his Doctor in Business Management at the Philippine Women's University, Master in Business Administration at Angeles University Foundation, Bachelor of Arts Major in Economics at the University of the Philippines. He is one of the founding board members of Pampanga Research Educators Organization and the Vice President – Luzon of PCDEB.