Basic Health Screening by Exploiting Data Mining Techniques

Dolluck Phongphanich
Faculty of Science and Technology,
Suratthani Rajabhat University,
Suratthni, Thailand

Nattayanee Prommuang
Faculty of Science and Technology,
Suratthani Rajabhat University,
Suratthni, Thailand

Benjawan Chooprom
Faculty of Science and Technology,
Suratthani Rajabhat University,
Suratthni, Thailand

***Abstract*—This study aimed at proposing a basic health screening system based on data mining techniques in order to help related personnel on basic health screening and to facilitate citizens on self-examining health conditions. The research comprised of two steps. The first step was to create a model by using classification techniques that are Bayesian methods (Naïve Bayes, Bayesian networks, and Naïve Bayesian Updateable) and decision tree methods (C4.5, ID3, Partial Rule) to find important attributes causing the disease. In this step, the accuracy of each method was compared to the other methods to select the most efficient model as an input for the next step. The second step was to develop a basic health screening system by exploiting rules from the model developed in the first step as the second step's inputs were to classify from a citizen's health profile whether a given citizen is in a normal group, risk group or sick group. Research findings revealed two important attributes directly contributing to diabetes: Blood pressure (BP) and docetaxel (DTX). Furthermore, C4.5 algorithm provided the most accuracy with accuracy of 99.7969%, precision of 99.8%, recall of 99.8% and F-measure of 99.8%.**

***Keywords*—Bayesian methods; classification technique; data-mining; decision tree methods**

I. INTRODUCTION

According to 2015 global diabetes statistics, it revealed that more than 415 million people were diabetic patients and it is expected that by 2045, the number of diabetic patients would reach 642 million people and the trend is increasing going forward. In addition, one-eleventh of people were unexpectedly affected by diabetes, and one-seventh of birth was affected by diabetes during pregnancy, and one person will pass away from diabetes in every 6 seconds. Moreover, diabetic patients are in the risk of hypertension and other serious complications.[1] If this disease is not well addressed, mortality and morbidity rate will increase as well as increasing financial burden and the economic loss will impact to the nation.

Thailand is facing this similar situation. A report stated that diabetes is one of the 10 non-communicable diseases (NCDs) that are the main cause of death across the country. Diabetes is the outcome of risk behaviors, including over consumption of sweet, oily and salty foods, under consumption of vegetables and fruits, smoking, drinking, lack of exercises, stress and inappropriate emotional handling, thereby leading to overweight, obesity, hypertension and other circumstances. Thus, if those risk behaviors are not dealt with well, those will cause sickness, complication, disability and finally untimely death, resulting in an increase in huge healthcare expenses and economic loss.[2]

Besides Professor Chaicharn Deeroochanawong, M.D. said that diabetes is threatening Thailand, so it must be taken into account before the enormous damage will happen by applying disease screening for the risk group, increasing searches of complications in diabetic patients, performing an annual basic health screening, and focusing on early diagnosis to better protect from and control risk factors of the disease. These actions are in line with the Ministry of Public Health's policies specifying all provincial health offices as the center of carrying on activities to make citizens aware of importance of proper health behaviors, due to the fact that 2009 and 2015 statistics show that the incidence of diabetic and high blood pressure patients continuously raises from 4.7 million people in 2011 to 5.4 million people in 2015,[3] and this affects to economic and national development. Hence, he suggests that the issue must be resolved urgently and continuously. With that, the National Health Security Office provides funding supports for public health surveillance across the country by using a health survey to do an annual basic health screening inside and outside municipalities.

Nevertheless, disease screening for the risk group, increase in searches of complications in diabetic patients, running an annual basic health screening, and carrying on activities to create awareness on importance of proper health behaviors requires a lot of medical professionals. But according to Human Resources for Health Research and Development Office (HRDO)'s survey and Bureau of Policy and Strategy's health resources survey in Thailand, in 2010 there were only 26,162 physicians working for public health centers which is at the population ratio of 1 : 2,428, whereas the required

[1] Diabetes Association of Thailand under The Patronage of Her Royal Highness Princess Maha Chakri Sirindhorn, Global report on diabetes 2016 (In Thai), Retrieved Feb. 15, 2016, From the World Wide Web : http://www.dmthai.org/statistic/list

[2] Thai Health Promotion Foundation. Non-Communicable diseases 2009 (In Thai), Retrieved Feb. 15, 2016, From the World Wide Web: http://www.thaihealth.or.th/

[3] Health Statistics Development Plan No. 1 (In Thai), From the World Wide Web: http://osthailand.nic.go.th/files/ social_ sector/SDP_health291057-new6.pdf

population ratio should be 1 : 1,500 – 1,800. Aside from that, only 50.4% of total physicians are under the Ministry of Public Health while they have to take care of more 80% of people. This excludes the loss of a number of physicians due to their resignation. All of these are the issue on lack of medical professionals which lead to long waiting time for each medical appointment, incurring cost and time of travelling to meet with physicians.[4] However, most of the time for citizens is used for working; hence, many people pay less attention to medical checkup and meet with physicians in an event of emergency or when they have severe sepsis, even if there are many campaigns and encouragements for annual medical checkup derived from the ministry's policies.

Therefore, this study aimed at proposing a basic health screening system by exploiting data mining techniques in order to help related personnel on basic health screening and to facilitate citizens on self-examining health conditions. It is also compared the accuracy between Bayesian and Decision trees methods in order to select the most efficient model as an input for creating Basic Health Screening System. The results of this study can provide the knowledge of their disease risk level and knowledge for preventing the disease in a right way. Apart from that, this system helps collect statistics of diseases, correctly and quickly analyze and filter important data according to needs, because data mining is capable of analyzing, discovering, extracting relationships and finding patterns on a large amount of data efficiently. More details about capabilities of data mining will be explained in Related Work section below. A technique adopted in this study was classification which is currently well-known for data mining.

The next section, Related Works, will describe review of literature regarding to data mining techniques and patient screening. The third section called Methodology will present mythologies used in this research as well as data set and experiments. The fourth section, namely, Results and Discussion will present several discussions on experimentation. And the last section will be Conclusion and Acknowledgement.

II. Literature Review

Professor Chaicharn Deeroochanawong, M.D. analyzes diabetes situation that diabetes is negatively impacting Thailand; thus, this disease must be taken into account by increasing the diabetes screening test on the risk group, having annual screening in place to search for complications in diabetic patients, searching and resolving factors that result in poor disease control of most of diabetic patients, paying attention to early medical diagnosis and considering the ways to slow down or reduce the likelihood of complications from diabetes which is likely to be severe and costly. With this regard, data should be systematically stored for long-term usage and should cover incidence of diabetes in children and adults, while good health campaign in a light of diabetes prevention should be through step-by-step knowledge sharing and dissemination of accurate news and information to the public.[5] This is consistent to the Ministry of Public Health's policies that appoint all provincial health offices to carry on activities to raise awareness on following appropriate health behaviors to people, since 2009 and 2015 statistical data reveals that the number of diabetic and high blood pressure patients continuously increases from 4.7 million persons in 2009 to 5.4 million persons in 2015. This negatively influences economic and national advancement and that needs urgent and continuous actions. Hence, the National Health Security Office comes into play by providing funds to set up surveillance in every province. The surveillance is in a form of health survey for annual basic health screening of people living inside and outside municipalities.

From literature review, there are a number of researches adapting data mining techniques to analyze health data. To illustrate, Kittisak Sumamal [1] proposes data analysis from 1,071 records of health situation survey for citizens in BuriRam Municipality in 2012 by adopting data mining techniques. His study considers two main points: the study of relationships among diseases based on an association rule technique and study of basic health screening to classify citizens into a normal group, risk group or sick group by using a classification technique with a decision tree algorithm.

Rukthin Laoha [2] studies lung cancer risk prediction based on data mining approaches by developing a system to classify a group of patients and predict the risk of lung cancer. Classification will derive a risk score of each of factors leading to lung cancer then all risk scores will be used to analyze and predict patients with C4.5 decision tree and data of 2,215 Maha Sarakham Hospital patients between August and December 2012. These patients include 118 lung cancer patients and 2,097 non-lung cancer patients. After measuring the forecasting performance from accuracy and recall values, it was found that factors influencing the risk of lung cancer most are a heredity factor in which the risk score for people with this factor is 34.59 times of the same risk score for people without the factor, followed by smoking behavior, drinking behavior and age, respectively.

Aungkana Pijarachote [3] develops a decision support system for risk analysis of diabetes disease using data mining techniques to help analyze relationships of risk factors leading to diabetes; for instance, parents have diabetes and polyuria. The result of relationships among those risk factors from the analysis will be important information that helps medical organizations plan for diabetes prevention. The developed system consists of three parts: a data bank part for storing risk factors' data from risk group screening, data mining part that look for relationships among the risk factors, and report presentation part which is a web application.

Tapas Ranjan Baitharuaand and Subhendu Kumar Pani consider discovery of hidden patterns and relationships of medical diagnosis by learning the patterns through collected data of liver disorders to create smart medical decision support systems to aid physicians. In that paper, both researchers propose the use of J48 decision tree, Naïve Bayes, ANN,

[4] Health Statistics Development Plan No. 1 (In Thai), From the World Wide Web: http://osthailand.nic.go.th/files/ social_ sector/SDP_health291057-new6.pdf

[5] International Diabetes Federation, About Diabetes, Retrieved Feb. 15, 2016, From the World Wide Web: http:// www.idf.org/about-diabetes

ZeroR, 1BK and VFI algorithm in order to classify these diseases as well as to compare effectiveness and correction rate among them. Findings show that detection of liver diseases in an early stage is the key, as it results in improved performance of the classification models in terms of their predictive or descriptive accuracy, reduction of computation time for building models as they learn faster and enhanced understanding of those models. On top of that, the researchers present a comparative analysis of data classification accuracy using liver disorder data in various scenarios. Last but not the least, they compare predictive performance of well-known classifiers quantitatively [4].

This research similar to Pijarachote [3] and Sumamal's [1], studies in terms of data analysis to find risk factors of diabetes by using a classification technique and C4.5 algorithm like researches done by Laoha [2] and Baitharua and Pani [4]. But differentiation of this study is the adoption of several classification techniques for creating models and comparing their performance so as to find the most efficient one prior to development of the basic health screening system. Techniques used in this research include Bayesian methods (Naïve Bayes, Bayesian networks, and Naïve Bayesian Updateable) and decision tree methods (C4.5, ID3, Partial Rule).

III. METHODOLOGY

A research methodology of this study comprised of two stages. The first stage was to create a model by using a classification technique that uses a decision tree method so as to find important factors leading to incidence level of the disease. The second stage was to develop a basic health screening system by using data mining techniques to process a data set given from the first stage. Details of each stage are described below.

A. Model Building

In this model building stage, historical data was used to analyze by using classification which is one of data mining techniques. Basically, classification was done with health data of persons (general citizens or patients previous got an examination) respecting to attributes and target classes of the classification was a normal group, risk group and sick group. This model would be in a form of decision tree, which can be represented as classification rules and used as an input for development of basic health screening system in the next stage. In this research, Waikato Environment for Knowledge Analysis (WEKA) software was used to analyze data and create a model. An experiment process of this section is described below:

1) Data Selection

Data for this experiment was gathered from Bang San Health Promoting Hospital. The data was the output of 2014 and 2015 health survey of Phanom Sub-district, Phanom District, Surat Thani Province in Thailand's southern region. Specifically, the survey was conducted with 2,462 residents living in 6 villages (Mooban): Moo 1 "Ban Suratthaphirom", Moo 2 "Ban Phanom", Moo 3 "Ban Phanom Nai", Moo 4 "Ban Bang Mai Pho", Moo 6 "Ban Bang San" and Moo 11 "Ban Thung Charoen". Collected data comprised of 18 attributes: a sequence number, first name, last name, gender, weight, height, waist, docetaxel (DTX), systolic pressure, diastolic pressure, eating habit, exercise habit, stress management, smoking habit, drinking habit, maturity onset diabetes of the young (MODY), parental hypertension, and target class. All of the data was originally stored in documents which were later converted to .xlsx and .csv files.

2) Data Preprocessing

Data preprocessing is critical for data verification in terms of accuracy, completeness, missing values, noisy data, error, outliers, and inconsistency. All of these help improve data quality prior to data mining. The data preprocessing follows a process below.

a) Data cleansing

In this step, researchers performed relevance analysis or selection of attributes relevant to data mining and exclusion of duplicate or unnecessary attributes in order to reduce noisy data as follows.

Exclusion of unnecessary attributes is including a sequence number, first name and last name.

Inclusion of height and weight attributes to become Body Mass Index (BMI) with the following formula below [5].

$$BMI = (\text{Weight (kg)}) / (\text{Height (m)})^2 \quad (1)$$

Note that this BMI is applicable for Asian people including Thai which can vary according to races.

Inclusion of systolic pressure and diastolic pressure attributes to become a blood pressure attribute. The used criterion is shown in Table 1.

At the end, only 13 attributes remained, including gender, BMI, waist, DTX, BP, eating habit, exercise habit, stress management, smoking habit, drinking habit, MODY, parental hypertension, and target class.

b) Data transformation

In this step, data will be normalized to limit data distribution within a specified range or transformed to a format ready for data mining. Attributes from gathered quantitative data were either continuous or discrete; for instance, weight is continuous data while exercise and smoking habits are discrete data. To prepare data for data mining, it was transformed into a nominal scale as shown in Table 2.

TABLE I. SYSTOLIC PRESSURE AND DIASTOLIC PRESSURE ATTRIBUTES

Systolic pressure (mmHg)	Diastolic Pressure (mmHg)	Blood Pressure (BP) (mmHg)
Systolic pressure is less than 120	Diastolic pressure is less than 80	BP1is less than 120/80
Systolic pressure is between 120 – 139	Diastolic pressure is between 80 – 89	BP 2is between 120– / 139 80– 89
Systolic pressure is between 140 – 159	Diastolic pressure is between 90– 99	BP3is between 140– / 159 99 – 90
Systolic pressure is more than or equal to 160	Diastolic pressure is more than or equal to 100	BP4is more than or equal to 100/160

TABLE II. ATTRIBUTES AND MEANING

Attribute	Nominal Scale
HP_gender	Gender MALE,FEMALE
HP_SumBMI	BMI BMI1= Thin (less than 18.50), BMI2 = Normal (18.50 – 22.99), BMI3 = Plump)23.00 – 24.99(, BMI4 = Fat)25.00 – 29.99(, BMI5 = Very fat (more than or equal to 30.00)
HP_Sumhip	Waist SIZE1 =Less than 90 cm. for a man / less than 80 cm. for a woman, SIZE2 =More than or equal to 90 cm. for a man /more than or equal to 80 cm. for a woman
HP_BP	Blood pressure BP1= Less than 120/80 mmHg, BP2 = 120 – 139 / 80 – 89 mmHg, BP3 = 140 – 159 /90 – 99 mmHg, BP4 = More than or equal to 160 / 100mmHg
HP_SumDTX	DTX DTX1 =Less than 100 mg/dL, DTX2 =100 – 125 mg/dL, DTX3 = More than or equal to 126 mg/dL
HP_food	Eating habit F1=Sweet, F2= Oily, F3= Salty, F4= Normal
HP_exercise	Exercise habit EX1= Everyday, EX2=Not everyday, EX3=No exercise
HP_strain	Stress management ST1= Without management, ST2= With management, ST3= Not stressful
HP_smoking	Smoking habit YES= Smoking, NO= Not smoking
HP_drink _alcohol	Drinking habit YES= Drinking, NO= Not drinking
HP_FaHT	Parental hypertension YES= Parents have hypertension, NO= Parents do not have hypertension
HP_FaDM	MODY YES= Parents have diabetes, NO= Parents do not have diabetes
Class	Target class Normal= Normal group, Risky = Risk group, Getsick = Sick group

1) Classification

A data set from the previous step was divided into two sets to create a model: 70% of total data was a training set and 30% of total data was a testing set. The training data was then processed by using Bayesian methods (Naïve Bayes, Bayesian networks, and Naïve Bayesian Updateable) and decision tree methods (C4.5, ID3, Partial Rule). A process of this step can be summarized as shown in Fig. 1. All models were then compared to each other based on precision, recall, F-measure, and accuracy in order to implement classification rules given by the most efficient model as part of a basic health screening system development in the next step.

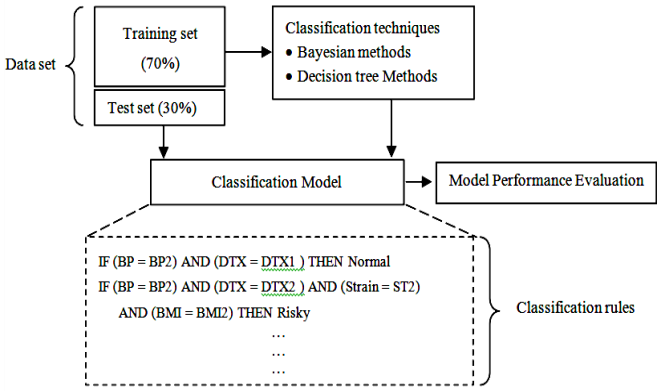

Fig. 1. A classification process.

2) Model Performance Evaluation

Model performance must be evaluated first prior to applying it for health screening. General classification model performance indicators are precision, recall, F-measure, and accuracy. In this step, the accuracy derived from a confusion matrix (see Fig. 2) of each model was compared to the same indicator from the other models to find the highest performance model. Those four indicators were calculated as follows [6].

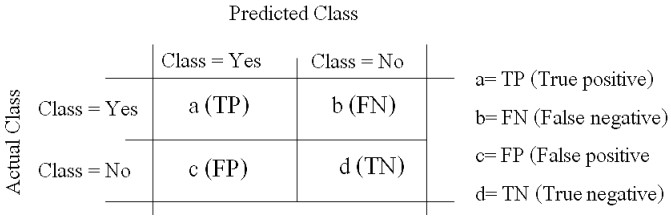

Fig. 2. A confusion matrix.

Precision: To measure precision of a particular model by considering each class [7].

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

Recall: To measure recall of a particular model by considering each class [7].

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

F-measure: To measure precision and recall at the same time for a particular model by considering each class [8].

$$\text{F-measure} = (2\times\text{Precision}\times\text{Recall}) / (\text{Precision}+ \text{Recall}) \quad (4)$$

Accuracy: To measure the model accuracy based on every class [8].

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (5)$$

B. Basic Health Screening System based on Data Mining Techniques

In this stage, classification rules from the previous stage were used to develop a basic health screening system. This system consists of screens for general users and hospitals' officers and executives. For this study, a PHP programming language was used for web application development, MySQL was used to implement a database, and 10 classification rules from a selected model were used as health screening criteria. This system is the integration of two parts: an information system and health screening system.

An information system will present useful information for users; for instance, an overall annual report of diabetic and high blood pressure patients sorted by age groups, and a report of diabetic and high blood pressure patients grouped by villages and sorted by age groups, while a health screening system is a data source of each health promotion hospital in Phanom District where officers of each health promotion hospital can register to the system to record data of its hospital.

Another part is a health screening system which can be used by officers and executives of Bang San Health Promoting Hospital as well as citizens (general users). The general users are only allowed to check their basic health by inputting their profile; for example, an eating habit, weight, exercises habit, smoking habit, etc.

For officers and executives, they can fill in patient information or searching citizens having their profile in the system by supplying a national ID, and then clicking "Screening". The system will process inputted health data and show the screening result with initial suggestions right away. Those officers and executives also want to view medical checkups for examinations by simply clicking on a national ID. They can view statistical reports like the general users do, but they can see specific reports in each area, as well as a blood pressure and diabetic level of each patient by specifying a national ID so that they can do screening.

IV. Results and Discussion

The researchers presented the experiment into two parts: the first part was the result of model building based on classification techniques and development of basic health screening system based on data mining techniques.

A. Model Building based on Classification Techniques

The accuracy comparison result from experimentation using classification techniques is shown in Table 3, which reveals that a model created from decision tree methods have higher accuracy for classification as a normal group, risk group or sick group than a model created from Bayesian methods.

TABLE III. The Comparison of Decision Tree Methods and Bayesian Methods

Techniques	Precision	Recall	F-measure	Accuracy
C4.5	99.8%	99.8%	99.8%	99.79%
Partial Rule	99.8%	99.8%	99.8%	99.79%
Induction	99.6%	99.6%	99.6%	99.51%
Bayesian Net	91.2%	91.5%	91.3%	91.47%
Naïve Bayes	91.4%	91.6%	91.4%	91.61%
Naïve Bayes Updateable	91.4%	91.6%	91.4%	91.61%

From the classification experimentation by using decision tree methods with performance comparison in terms of precision, recall, F-measure and accuracy, all three algorithms provided the similar results, and the accuracy of C4.5 algorithm was equal to that of Partial Rule algorithm; however, C4.5 algorithm had 10 rules , while Partial Rule had 9 rules. Therefore, to ensure the completeness of the system in development, the researchers asked experts to verify the accuracy of those rules. Finally, C4.5 algorithm was selected for basic health screening system development. For a chosen C4.5 algorithm, the result of classification errors of health information as a normal group, risk group and sick group is shown in Table 4.

From Table 4, the number of classification error for a normal group was 1 out of 1,381 instances, which is 0.07%. And since the number of records for training was high (1,381 instances), precision and recall was as high as 99.8% and 99.8%, respectively.

TABLE IV. A Classification Error Matrix of C4.5 Algorithm

Classification		Predicted Class		
		Normal Group	Risk Group	Sick Group
Actual Class	Normal Group	1,380	1	0
	Risk Group	3	568	1
	Sick Group	0	0	509

After verified by experts, those 10 classification rules from the model (shown in Table 5) were further adopted for development of basic health screening system. You can see that all rules were based on two attributes directly contributing to diabetes classification: Blood pressure (BP) and docetaxel (DTX).

TABLE V. BASIC HEALTH CLASSIFICATION RULES BASED ON C4.5 DECISION

Rule	Description
IF (BP=BP2) and (DTX = DTX1) Then Normal	If blood pressure is between 120 – 139 / 80 – 89 mmHg and DTX is less than 100 mg/dL, then a person belongs to a normal group.
IF (BP=BP2) and (DTX = DTX2) Then Risky	If blood pressure is between 120 – 139 / 80 – 89 mmHg and DTX is between 100 – 125 mg/dL, then a person belongs to a risk group.
IF (BP=BP2) and (DTX = DTX3) Then Getsick	If blood pressure is between 120 – 139 / 80 – 89 mmHg and DTX is more than or equal to 126 mg/dL, then a person belongs to a sick group.
IF (BP=BP1) and (DTX = DTX1) Then Normal	If blood pressure is less than 120/80 mmHg and DTX is less than 100 mg/dL, then a person belongs to a normal group.
IF (BP=BP1) and (DTX = DTX2) Then Risky	If blood pressure is less than 120/80 mmHg and DTX is between 100 – 125 mg/dL, then a person belongs to a risk group.
IF (BP=BP1) and (DTX = DTX3) Then Getsick	If blood pressure is less than 120/80 mmHg and DTX is more than or equal to 126 mg/dL, then a person belongs to a sick group.
IF (BP=BP3) and (DTX = DTX1) Then Normal	If blood pressure is between 140 – 159 /90 – 99 mmHg and DTX is less than 100 mg/dL, then a person belongs to a normal group.
IF (BP=BP3) and (DTX = DTX2) Then Risky	If blood pressure is between 140 – 159 /90 – 99 mmHg and DTX is between 100 – 125mg/dL, then a person belongs to a risk group.
IF (BP=BP3) and (DTX = DTX3) Then Getsick	If blood pressure is between 140 – 159 /90 – 99 mmHg and DTX is more than or equal to 126 mg/dL, then a person belongs to a sick group.
IF BP=BP4 Then Getsick	If blood pressure is more than or equal to160 / 100 mmHg, then a person belongs to a sick group.

B. *Development of Basic Health Screening System by using Data Mining Techniques*

The system contains screens for general users and hospitals' officers and executives as shown in Fig. 3.

Fig. 3 presents screens of the basic health screening system for general users. Fig. 3(a) is an input screen of general information, including age, gender, weight, height, waist, systolic pressure, diastolic pressure and glucose value. Fig. 3(b) is an input screen of health profile, including eating habit, exercise habit, stress management, smoking habit, drinking habit, MODY, and parental hypertension. Fig. 3(c) is a screening result screen which displays a level of sickness (normal group, risk group, sick group) and health analysis result; for instance, blood pressure is in a critical level, and calculated DTX presents a risk of diabetes. Lastly, Fig. 3(d) is an initial suggestion screen proposing helpful advices, such as additional workout, reducing dessert intakes, looking for activities to decrease stress, and keeping not drinking and smoking.

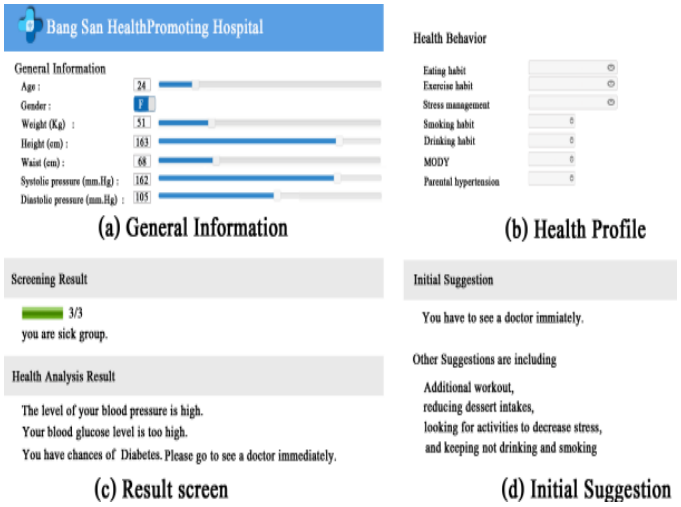

Fig. 3. (a)–(d) Screens of basic health screening system for users.

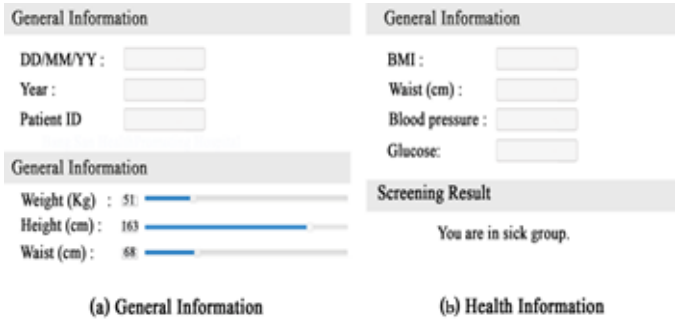

Fig. 4. (a) - (b) Patient information's input screens for officers.

Fig.4. presents screens for hospital officers. Specifically, Fig. 4(a) is an input screen of patients, including age and patient ID, while Fig. 4(b) is an input screen of health profile, including weight, height, waist, DTX, blood pressure. For a list of diabetic patients sorted by acuity levels and sub-district level patient statistics as shown in Fig. 5(a) and (b), they can be viewed by both officers and executives of a particular hospital.

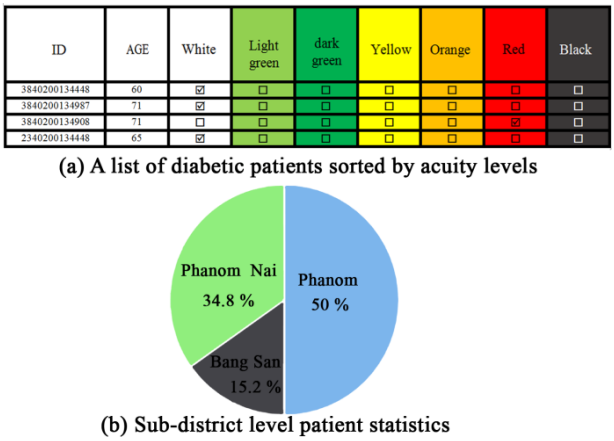

Fig. 5. (a) A list of diabetic patients sorted by acuity levels, (b) Patient statistics grouped by sub-districts.

Fig. 6. Levels of sickness in a form of color.

Fig. 5 shows a diabetic patient report sorted by acuity levels represented in a color form. For a meaning of each color, white means a normal group, light green means a risk group, dark green means a 0 level sick group, yellow means a 1st level sick group, orange means a 2nd level sick group, red means a 3rd level sick group and black means a critical sick group, such as coronary artery disease, kidney disease, and diabetic retinopathy disease. Fig. 6 summarizes the meaning of each color.

V. Conclusion

This study aimed to implement a basic health screening system based on data mining techniques to help related personnel on basic health screening and to facilitate citizens on self-examining health conditions. At first, we used Bayesian methods (Naïve Bayes, Bayesian networks, and Naïve Bayesian Updateable) and decision tree methods (C4.5, ID3, Partial Rule) to create a number of classification models and evaluated performance. Finally, a model with C4.5 algorithm implemented was selected for basic health screening system development thanks to highest accuracy. Next, we developed a basic health screening system by exploiting rules from the model developed in the previous step to classify whether a particular citizen is in a normal group, risk group or sick group. The system was successfully adopted by Bang San Health Promoting Hospital.

For limitations, since the health screening system was a pilot system, this study of basic health screening by using classification techniques only considered algorithms that provided results as classification rules; it was necessary to take the results for development of basic health screening system.

For future researches, we would like to suggest as follows:

- Accuracy of classification for some classes, such as a normal group, is very high, because a large number of instances actually belong to the normal group. Hence, a number of instances for each class in a training set should be approximately the same.
- The current classification technique is still valid even if a medical checkup input form is revised or when it is applied to another form.
- In the future, if this basic health screening system can collect health data of all citizens in a whole province and scholars would like to utilize the data for new classification by using WEKA software, they should consider analyzing data regarding to regions of instances, since people living in different regions may have different attributes.

Acknowledgements

We would like to express thanks to the Bang San Health Promoting Hospital, Thailand for giving the information in this research.

References

[1] K. Sumamal, *Basic Health Screening by Using Data Mining Techniques*, Master Thesis, Dept. Information Technology, Dhurakij Pundit University, Bangkok, Thailand, 2012.

[2] R. Laoha, *Predicting Risk Lung Cancer Patient by Data Mining Approach*, Master Thesis, Dept. Science, Khon Kaen University, Khon Kaen, Thailand, 2010.

[3] A. Pijarachote, *Decision Support System for Risk Analysis of Diabetes Disease Using Data Mining Techniques*, Master Thesis, Dept. Science, Khon Kaen University, Khon Kaen, Thailand, 2009.

[4] T. R. Baitharua, S. K. Pani, Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset, (2016) *Procedia Computer Science*, (85), pp. 862 – 870.

[5] P. Teanbun, Assessment of Nutritional Status (In Thai), Retrieved on Mar. 2, 2016, from the World Wide Web: http://www.med.cmu.ac.th/dept/nutrition/DATA/COMMON/cmunut-deptped/ped401-prasong/ped401-assessment-of-nutritional-prasong.pdf

[6] D. Phongphanich, W. Choonui, "An Internet-based Student Admission Screening System utilizing Data Mining", (2017), *International Journal of Advanced Computer Science and Applications*, vol.8, pp.207-213.

[7] L. H. Witten, E. Frank and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques, 3nd ed., Burlington, USA: Morgan Kaufmann publishers, 2011.

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann publishers, San Francisco: CA, 2006.