Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 69 (2015) 132 – 142

Procedia Computer Science

7th International Conference on Advances in Information Technology

Comparison of Classifiers for the Risk of Diabetes Prediction

Nongyao Nai-arun[a,*], Rungruttikarn Moungmai[a]

[a]*Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Thailand.*

Abstract

This paper applied a use of algorithms to classify the risk of diabetes mellitus. Four well known classification models that are Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes were first examined. Then, Bagging and Boosting techniques were investigated for improving the robustness of such models. Additionally, Random Forest was not ignored to evaluate in the study. Findings suggest that the best performance of disease risk classification is Random Forest algorithm. Therefore, its model was used to create a web application for predicting a class of the diabetes risk.

Peer-review under responsibility of the Organizing Committee of IAIT2015

Keywords: diabetes; random forest; logistic regression; artificial neural networks; decision tree; naive bayes; bagging; boosting.

1. Introduction

Diabetes mellitus (DM) is a chronic non-communicable disease. The disease has been closely followed by World Health Organization (WHO) and International Diabetes Federation (IDF) since worldwide number of diabetes increase continuously. It was found that there were 387 million people with diabetes in year 2014 and have a tendency to be 592 million patients in the next 20 years[1]. IDF also found that almost half of diabetes in South East Asia is undiagnosed. According to these amounts, the disease should be controlled and properly maintained for efficient and sustainable prevention.

The annual report 2013 of Department of Disease Control, Ministry of Public Health, reports that diabetes is the top three of chronic non-communicable diseases in Thailand[2,3]. The statistics shows that 1 of 13 adults Thais had

* Corresponding author. Tel.: +6-681-643-8016; fax: +6-656-882-531
E-mail address: nongyao25@hotmail.com, r.moungmai@gmail.com

Peer-review under responsibility of the Organizing Committee of IAIT2015
doi:10.1016/j.procs.2015.10.014

diabetes and the total number of people with diabetes is not less than 3 million. In the future, there will be more than 7 million people are at risk of diabetes[4]. The report also indicates that the number of diabetes is likely to be increased every year. Consider diabetes death rate, there are about 12 dead with diabetes in every 100 thousand people. This can be seen that the rate is a small number however this amount is only the dead with diabetes. In fact, diabetes is an important cause of other diseases such as stroke and heart diseases which are the top three of chronic non-communicable diseases and have high death rates[3]. It also leads to the destruction of cells in the body such as nerves, blood vessels, heart, eyes and kidneys[4,5].

Nowadays, the situation of diabetes in Thailand has been concerned due to 1 of 3 diabetes patients is undiagnosed and unaware. It was also found that the age of patients trends to decrease. Moreover, the number of female is more likely than male with diabetes and the patients are obese people more than the non-obese[5].

As mentioned above, a study of disease classification is considered since it holds great potential for improving human health and personal treatment. In this paper, four popular classifiers for disease risk prediction are studied. These algorithms consists Decision Tree, Artificial Neural Network, Logistic Regression and Naïve Bayes. After that Bagging and Boosting techniques are combined with those algorithms to improve the robustness of each model. At the end, Random Forest algorithm is applied.

The objective of this study is to predict the risk of diabetes for everyone without the need of blood test or going to a hospital. The study also aims to encourage and promote good health of people. In addition, the diabetes prediction will be created as a simple diagnosis application and will be published by a website. However, this application is only an initial diagnosis. People who found that they are in the diabetes risk group should go to see a doctor for formal diagnosis to prevent themselves from serious diabetes.

2. Material and Method

The format of this study is as shown in Fig. 1.

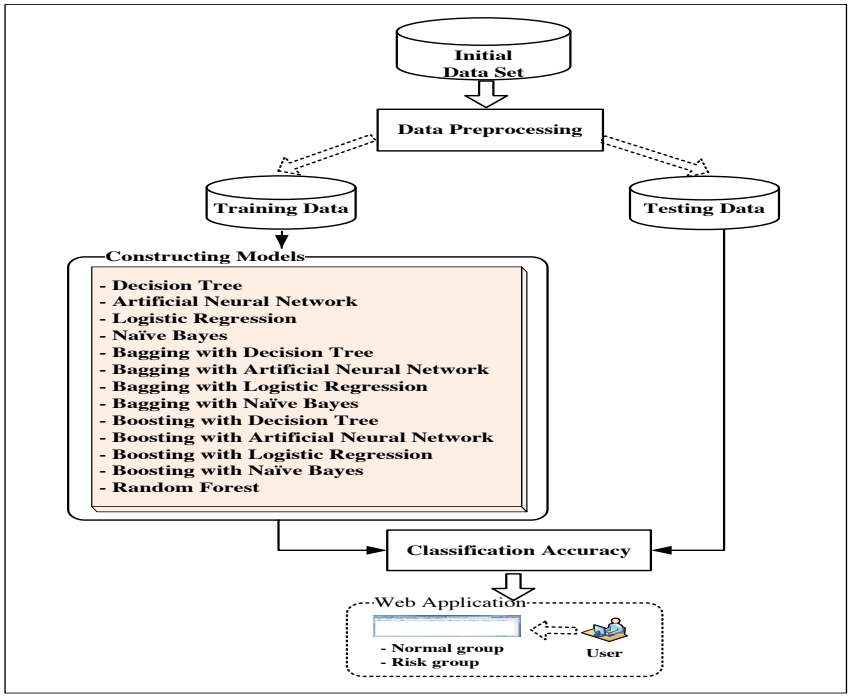

Fig. 1. The process of conceptual framework.

Fig. 1 presents four stages of the process of conceptual framework in the study. The process starts with data manipulation. Next, thirteen models will be investigated for finding a prediction model. Then, accuracy of each model will be calculated and compared for seeking the best model. The study ends up with creating a web application.

2.1. Data preprocessing

An initial data set was collected from 26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital during 2012 – 2013. Each person filled a screening form that will be used to identify diabetes risk group in the study. In this process, the data set was manipulated as follow;

Firstly, data where were collected from each PCU was integrated together.

Secondly, some variables needed a transformation such as age and BMI. For instance, in the screening form, there are no age and BMI information. Hence, age had to be calculated from date of birth and BMI was calculated from weight (kg) divided by height (meter) squared.

Table 1. Input and output variables.

No	Variables	Description	Value
1	BMI	Body Mass Index (kg/m^2)	numeric
2	AGE	Age (year)	numeric
3	WEIGHT	Weight (kg)	numeric
4	WAIST_CM	Height (cm)	numeric
5	BPH	Systolic blood pressure (mmHg)	numeric
6	BPL	Diastolic blood pressure (mmHg)	numeric
7	DM_FAMILY	History of Diabetes in family	1: Have
			2: No have
			9: Unknown
8	HT_FAMILY	History of Hypertension in family	1: Have
			2: No have
			9: Unknown
9	ALCOHOL	Alcohol drinking	1: No smoke
			2: Rarely
			3: Occasionally
			4: Often
			9: Unknown
10	SMOKE	Smoking behaviour	1: No smoke
			2: Rarely
			3: Occasionally
			4: Often
			9: Unknown
11	Sex	Sex	1: Female
			2: Male
12	**CLASS**	**1: Normal group**	
		2: Diabetes risk group	

Thirdly, input and output variables were defined by consulting a medical person and, then, selecting from general information in the screening form. These variables were considered based on correlation and causes of the diabetes.

Hence, there are eleven input variables and a dichotomous output variable as shown in Table 1. The table displays all variables used in the models, their description and values of each variable. Note that variable one to eleven are input variables where the first six variables are numeric and the remains are categorical and the last variable is the dichotomous, called class. The class variable is the fasting blood sugar (FBS) which is divided into two groups, normal and risk. The normal group is a people who have FBS less than 100 mg/dl whereas a people who have FBS between 100–125 mg/dl will be put in the risk group[2]. In this paper, people who have FBS more than 125 mg/dl are not included due to they are classified into a diabetes group that will be separated in another database and will be treated as patients.

Lastly, all the missing values were taken off from this study.

Therefore, the final data set used in the study consists of 30,122 people who can be divided into two groups. One is normal group, 19,145 people, and another is diabetes risk group, 10,977 people. These data will be used in the next stage.

2.2. Constructing models

In this stage, performances of four well know algorithms which are Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes were first considered. Decision Tree algorithm[6,7] is a simple well known approach that predicts a disease risk class based on several input variables and a use of decision tree. The tree consists three types of nodes, a root node, child node and leaf node. The algorithm starts with defining a root node from the most relationship between each input and output variables. Next, a child node is selected by calculating Information Gain (IG) which is given by

$$IG(parent, child) = Entropy(parent) - \left[p(c_1) \times Entropy(c_1) + p(c_2) \times Entropy(c_2) ... \right] \quad (1)$$

where Entropy(c_i) = – p(c_i) log p(c_i) and p(c_i) is a probability of child node i. Then a node that has the highest IG will be a parent for the next generation. This process is repeated until it gets a leaf node and completed decision tree. An example which is some parts of the model in this study is as shown in Fig. 2.

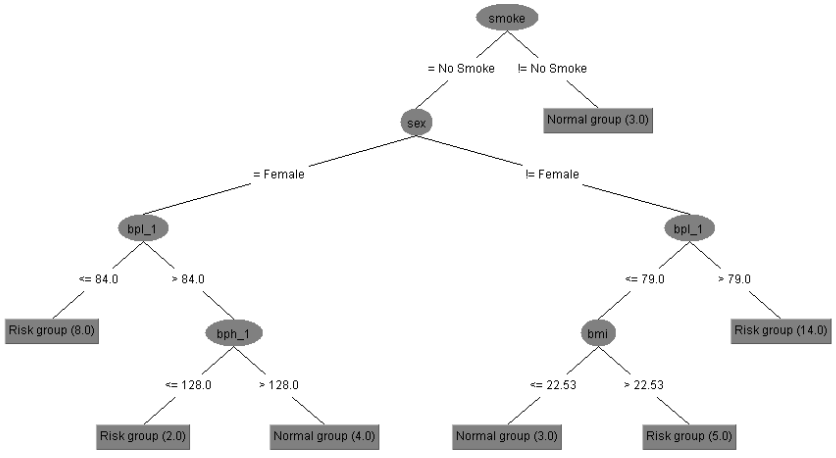

Fig. 2. Example of Decision Tree Model.

Artificial Neural Networks algorithm (ANNs)[8,9,10] called Neural Networks, was developed by copying an idea of the function of nervous systems in particular the brain. This algorithm can be used to evaluate a function of a large number of input variables. It is also suitable for both categorical and numeric output variables[7]. The structure of this algorithm is as shown in Fig. 3. The figure displays a process of Artificial Neural Networks algorithm in the study.

Firstly, a number of hidden layers are defined from sum of number of input and output variables divided by two. For example, there are twenty-three input variables and a dichotomous output variable. Hence the number of hidden layers is equal to twelve. Note that a categorical variable is counted by the number of its levels. Next, values of these layers are calculated from sigmoid function which is given by

$$f(x) = \frac{1}{1+e^{x}} \tag{2}$$

where x is sum of product of input values and numeric weights. Then, output layers are calculated in the same spirit of the hidden layers. Finally, a class will be predicted by selecting the highest value from output layers.

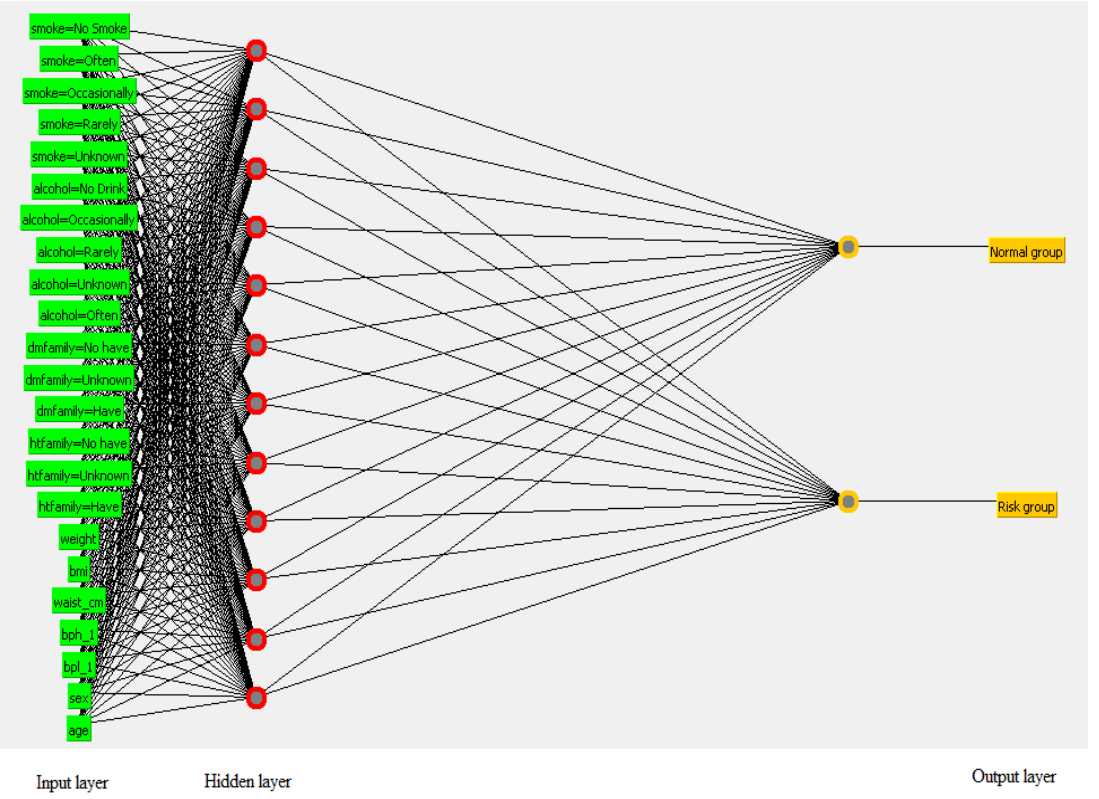

Fig. 3. Artificial Neural Network Model.

Logistic regression algorithm, sometimes called logit model, is a common model for dichotomous output variables[11] and was extended for disease classification prediction[12]. Suppose that there are p input variables where their values are indicated by $x_1, x_2, ..., x_p$. Let z be a probability that an event will occur and 1-z be a probability that the event will not occur. The logistic regression model is given by

$$\log\left(\frac{z}{1-z}\right) = \log it(z) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3}$$

or can be written by

$$z = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}} \tag{4}$$

where 0 is the intercept and β_1, β_1, ..., β_p are the regression coefficients.

Naive Bayes algorithm[13] is a simple classifier that based on the Bayes' Theorem. Let $class_i$ be diabetes risk group i and V be input variables that are used in a model and under the assumption of all variables are independent. To predict a class of diabetes risk, a model of Naive Bayes can be defined by

$$P(class_i \mid V) = \frac{P(V \mid class_i) \times P(class_i)}{P(V)} \tag{5}$$

where $P(class_i|V)$ is a posterior probability of a training data set with variable V that will be $class_i$. $P(V|class_i)$ is a likelihood of a training data set of $class_i$ and variable V where V is equal to $V_1 \cap V_2 \cap ... \cap V_M$. $P(class_i)$ is a probability of diabetes risk group i. The above model can be written as

$$P(class_i \mid V) = \frac{P(V_1 \mid class_i) \times P(V_2 \mid class_i) \times ... \times P(V_M \mid class_i) \times P(class_i)}{P(V)} \tag{6}$$

Hence, the prediction class will be $class_i$ when it gives the highest value of $P(class_i|V)$.

To modify such models accuracy, Bagging and Boosting are combined with those models. These approaches have been shown that they can improve classification accuracy[14]. Bagging algorithm, sometimes called bootstrap aggregation, was developed by Leo Breiman[15] and was introduced to avoid over fitting and reduce variance of the predicting model[16]. Suppose that a data set consists of N data. The algorithm starts with random generating a training data set. Then, its model where gives a prediction class is constructed. After this procedure is repeated several times where each time generates with data replacement, the final output prediction will be presented by a majority vote of those model predictions.

Boosting algorithm was developed by Schapire[17]. and was suggested that it can reduce an error of weak classifier due to the procedure of repeated construction of such classifier[14,17]. It was also developed for binary classification models by Freund and Schapire, called Adaboost[16]. This algorithm procedure begins with applying a weight to all observations in a training data set. Next, its classifier, for example Decision Tree, Artificial Neural Networks, Logistic Regression and Naïve Bayes, is modeled. After that these steps are repeated several times and their predicting classes are combined. Therefore, the last prediction output is from a majority vote of the combination[18,19].

In addition, Random Forest where was developed from trees algorithm and Bagging algorithm is modelled. Breiman[20] who developed the algorithm found that it can potentially improve classification accuracy. It is also work well with a data set with large number of input variables[21,22,23,24]. The algorithm is started by creating a combination of trees which each will vote for a class as shown in Fig. 4. The figure presents how to model the Random Forest. Suppose that there are N data and M input variables in a data set where the real data used in this paper compose of 30,122 data and 11 input variables. Let k be the number of sampling groups, n_i and m_i be number of data and variables in group i where i is equal to 1, 2, ... and k. Each sampling group is as followed;

1. n_i data where n_i is not greater than N are selected randomly from N.
2. m_i variables where m_i is not greater M are selected randomly from M.
3. A tree is grown and gives a prediction class.

After Step 1 to 3 was repeated for k times, these trees become a forest. Then the classification will be selected by a majority vote of all trees in the forest. Note that all data have to be returned to the data set before selecting a new sampling group. Therefore, there are thirteen models that will be evaluated in this process as shown in Fig. 1.

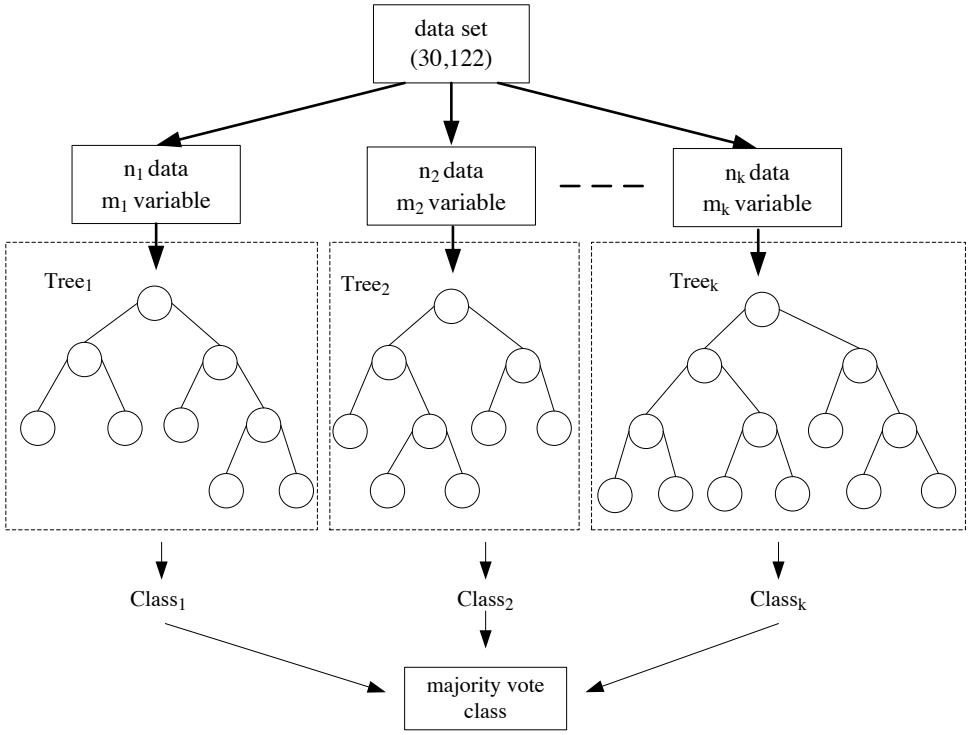

Fig. 4. Random Forest Model.

2.3. Classification accuracy

To assess the performances of those models, classification accuracy of each model will be calculated. After the model was constructed, it will be tested by using 10-folds cross validation for avoiding model over fitting[7,8]. Let TP, FP, TN and FN be the number of true positives, false positives, true negative and false negatives respectively. Therefore accuracy is defined by

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7}$$

However, this might not be enough for evaluating the robustness of each model therefore ROC curve will be considered. ROC curve describes a performance of an algorithm without the consideration of class distribution[7]. The curve is creating by plotting the Sensitivity which is given by

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\% \tag{8}$$

and 100 – Specificity which is

$$Specificity = \frac{TN}{(FP + TN)} \times 100\%. \quad (9)$$

2.4. Web application

The best algorithm where gives the highest accuracy or ROC Curve value is selected for creating a web application of diabetes risk prediction. The application shows prediction result that identifies two types of diabetes risk classes, normal group and diabetes risk group. The normal group means a person who does not have diabetes while another group means they might have diabetes or trend to have diabetes in the future. The application also displays accuracy of prediction (%) that presents the percentages of how much the model forecasting accurate. In this process, Programming PHP and Database MySQL are applied.

3. Experimental Results

In this section, the classification accuracy of thirteen models in the previous section is presented to assess the performance of each model as shown in Table 2.

Table 2. Comparison of classification accuracy of 13 models.

Models	Accuracy (%)
Decision Tree (DT)	85.090
Artificial Neural Network (ANN)	84.532
Logistic Regression (LR)	82.308
Naïve Bayes (NB)	81.010
Bagging with Decision Tree (BG+DT)	85.333
Bagging with Artificial Neural Network (BG+ANN)	85.324
Bagging with Logistic Regression ((BG+LR)	82.318
Bagging with Naïve Bayes ((BG+NB)	80.960
Boosting with Decision Tree (BT+DT)	84.098
Boosting with Artificial Neural Network (BT+ANN)	84.815
Boosting with Logistic Regression (BT+LR)	82.312
Boosting with Naïve Bayes (BT+NB)	81.019
Random Forest (RF)	**85.558**

Table 2 displays the results of comparison of the classification accuracy of thirteen models. The top five accuracy are Random Forest, Bagging with Decision Tree, Bagging with Artificial Neural Network, Decision Tree and Boosting with Decision Tree models which are 85.558%, 85.333%, 85.324%, 85.090% and 84.815% respectively. It can be seen that most of them are based on Decision Tree algorithms. Hence, Decision Tree model works well with this data set. The least accuracy is from the model of Bagging with Naïve Bayes, 80.960%. The results also suggest that Bagging and Boosting techniques improve the accuracy of Decision Tree, Artificial Neural Network and Logistic Regression models. The accuracy of Naïve Bayes model is only improved by Boosting technique. On the other hand, the accuracy of Bagging with Naïve Bayes model, 80.960%, is less than the accuracy of Naïve Bayes only, 81.010%, but not by much. However, these accuracies are not greater than the Random Forest accuracy. Note that Random Forest was developed from the combination of Trees and Bagging algorithms therefore the accuracy of Bagging with Decision Tree model, 85.333%, is closely to the accuracy of Random Forest model, 85.558%. In order to be confirmed the accuracy of prediction, the use of ROC Curve was applied as shown in Fig. 5.

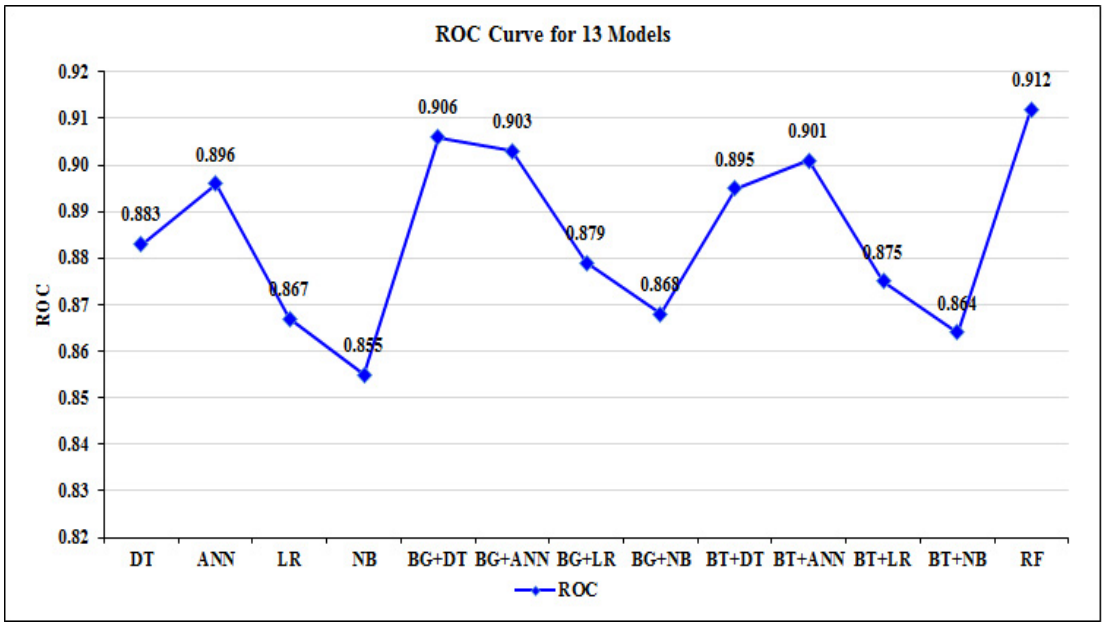

Fig. 5. Comparison of ROC Curve values from 13 models.

Fig. 5 clearly presents that the values of ROC Curve is greater than the accuracy but not by much. The top five ROC Curve values are Random Forest, Bagging with Decision Tree, Bagging with Artificial Neural Network, Boosting with Artificial Neural Network and Artificial Neural Network models which are 0.912, 0.906, 0.903, 0.901 and 0.896 respectively. While the least ROC Curve value is Naïve Bayes model, 0.855. Notice that, overall, ROC Curve valued of Artificial Neural Network is greater than value of Decision Tree whereas accuracy of Decision Tree is greater than accuracy of Artificial Neural Network. However, the best ROC curve value 91.2% is still from the same model as the highest accuracy, Random Forest model. In the same spirit of accuracy, the figure shows that Bagging and Boosting methods improve the ROC Curve values of Decision Tree, Artificial Neural Network, Logistic Regression and Naïve Bayes models.

After the best model, Random Forest, was selected, a web application was designed for forecasting the risk of diabetes by using such model. The application, includes two pages that are input and output screens as shown in Fig. 6 (a) and Fig. 6 (b) respectively. A user has to fill his/her general information in the input screen which are Name, Smoking behaviour, Alcohol drinking, History of diabetes in family, History of hypertension in family, weight(kg), body mass index(kg/m^2), waist(cm), systolic blood pressure, diastolic blood pressure, sex and age before clicking Random Forest bottom. In this stage, the program takes a while for evaluating. After that, the prediction will show the diabetes class which is normal group or risk group and an accuracy percentage of the prediction as shown in Fig. 6 (b).

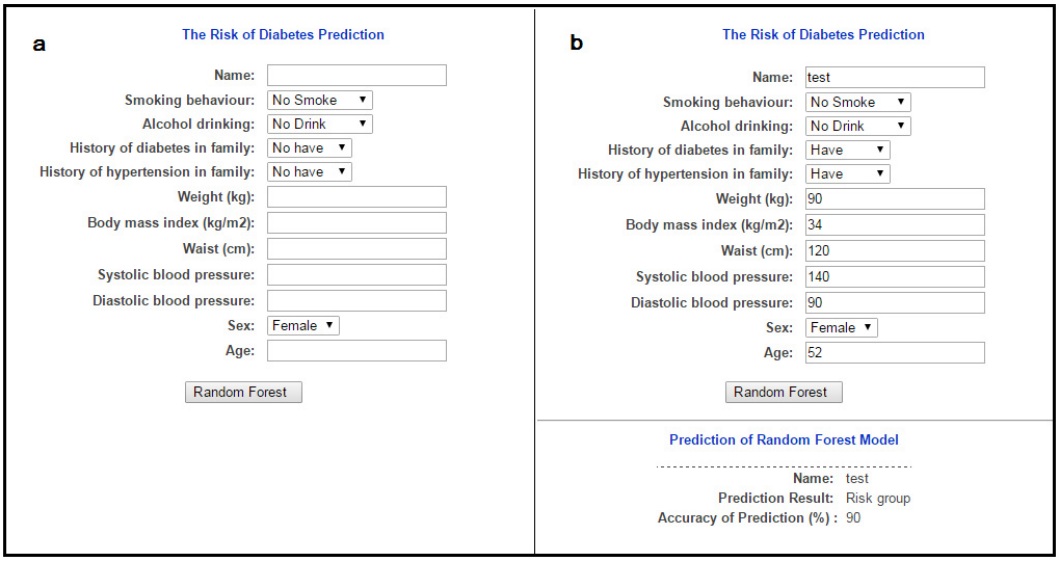

Fig. 6. (a) Web application input screen; (b) Web application output screen.

4. Conclusion

In this work, we proposed a web application by using a use of disease classifiers and a real data set. The data used in this creation are general information of 30,122 people who were collected from 26 Primary Care Units in Sawanpracharak Regional Hospital during 2012 – 2013. Before creating the web application, thirteen classification models were evaluated for seeking a predicting model. These models consist Decision Tree, Neural Network, Logistic Regression, Naïve Bayes and Random Forest algorithms including combination of Bagging and Boosting techniques except Random Forest algorithm. To investigate the robustness of each model, accuracy and ROC Curve were calculated and compared with others. The results reveal that Random Forest was ranked first in both accuracy and ROC Curve. This might be because of variable selection. In the process of Random Forest, data were not only chosen randomly but also input variables were random selected by considering important variables. Hence, this causes accuracy values increase. Therefore this algorithm was selected to model the diabetes risk prediction and used for creating the application.

Acknowledgements

The data set evaluated in this study was collected from 26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital

References

1. International Diabetes Federation. Retrieve 3 July 2015, from http://www.idf.org/diabetesatlas/update-2014.
2. Ministry of Public Health. *Surveillance control and prevention system of DM and HT in Thailand: Policy to action*. Thailand; 2013.
3. Ministry of Public Health. *Annual report 2013*, Bureau of policy and strategy. Thailand; 2013.
4. Thai encyclopedia for youth. Retrieve 9 June 2015, from http://kanchanapisek.or.th/kp6/sub/book/book.php?book=35&chap=8&page=t35-8-infodetail01.html.
5. World Health Organization. Retrieve 18 June 2015, from http://www.who.int/diabetes/en/.
6. Quinlan JR. Induction of decision tree. *Machine Learning*, 1; 1986. p. 81-89.
7. Han J, Kanber M. Pei J. *Data Mining: Concepts and Techniques*, 3rd ed. USA: Morgan Kaufman; 2012.

8. Witten IH, Frank E. Data mining: *Practical machine learning tools and techniques*. 2nd ed. USA: Morgan Kaufmann; 2005.
9. Wang C, Li L, Wang L, Ping Z, Flory MT, Wang G, Xi Y, Li W. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach. *Journal of Diabetes Research and Clinical Practice*, 100; 2013. p. 111-118.
10. Temurtas H, Yumusak N, Temurtas F. A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Application*, 36; 2009. p. 610-615.
11. Agresti A. *An Introduction to Categorical Data Analysis*. 2nd ed. New York: Wiley; 1996.
12. Tabaei B, Herman W. A Multivariate logistic regression equation to screen for diabetes. *Diabetes Care*, 25; 2002. p. 1999–2003.
13. Han J, Kamber M, Pei J, *Data Mining Concepts and Techniques*, 3rd ed. USA: Morgan Kaufman; 2012.
14. Dietterich TG. An Experimental Comparison of Three Methods for Construction Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40; 2000. p. 139-157.
15. Breiman L. Bagging Predictors. *Machine Learning*, 26(2); 1996. p. 123-140.
16. Nai-arun N, Sittidech P. Ensemble Learning Model for Diabetes Classification. *Journal of Advanced Materials Research*, Vols.931-932; 2014. p. 1427-1431.
17. Schapire RT. *The Boosting Approach to Machine Learning, An Overview. Nonlinear Estimation and Classification*. New York: Springer; 2003.
18. Liang G, Zhang C. Empirical Study of Bagging Predictors on Medical Data, *Proceedings of the 9th Australasian Data Mining Conference*, 121; 2011. p. 31-40.
19. Yang P, Yang YH, Zhou BB, Zomaya AY. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4); 2010. p. 296-308.
20. Breiman L. Random Forests. *Machine Learning*, 45(1); 2001. p. 5–32.
21. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision tree. *Journal of Computer Science*, 9(5); 2012. p. 272-278.
22. Sittidech P, Nai-arun N. Random Forest Analysis on Diabetes Complication Data. *Proceeding of the IASTED International Conference*; 2014. p. 315-320.
23. Kellie J, Archer, Ryan VK. Empirical characterization of random forest variable importance measures. *Journal of Computational Statistics & Data Analysis*, 52; 2008. p. 2249-2260.
24. Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and Results of new tests, *Journal of Pattern Recognition*, 44; 2011. p. 330-349.