Lecture Notes in Networks and Systems 251

Phayung Meesad
Sunantha Sodsee
Watchareewan Jitsakul
Sakchai Tangwannawit *Editors*

Recent Advances in Information and Communication Technology 2021

Proceedings of the 17th International Conference on Computing and Information Technology (IC2IT 2021)

Springer

Lecture Notes in Networks and Systems

Volume 251

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/15179

Contents

Herb Suggestion Model for Treat Disease by Symptoms Using Decision Tree Techniques

Chumpol Mokarat(✉) and Vipa Thananant

Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-Ok, Bangkok 10400, Thailand
{chumpol_mo,vipa_th}@rmutto.ac.th

Abstract. This research paper presents designing and developing model to support a decision making for choosing the traditional herbal medicine to relieve the symptoms in 8 groups of diseases or symptoms and used 38 attributes for filtering the diseases to generate a model. The attributes information related to the diagnosis of disease or symptoms are collected from the knowledge of the herbal medicine books and online medias for use as a learning and testing dataset of 230 records by using the J48 algorithm to create the rule-based from the decision tree model. After that, the model performance is evaluated with k-fold cross validation method and find a suitable accuracy model in the Weka software. From the experimental results, it is found that the data accuracy is 94.40%, a prediction in conjunction with the precision is 0.950, recall is 0.944, and f-measure is 0.938. It is effective to recommend herbs for the symptomatic treatment in alleviating the initial symptoms.

Keywords: Symptoms analysis · Herb suggestion · Decision tree · J48 algorithm

1 Introduction

Nowadays, Thai traditional medicine and alternative medicine are considered as another channel for people to turn their attention to as a health care alternative. They are used in conjunction treatment guidelines with modern medicine. In addition, during the past 20–30 years, although the synthetic chemical drugs were effectively, but also they contained chemical poisoning and they could cause side effects. This is the reason for the social trend towards in the natural products and this trend has grown in popularity [1]. This is because modern medicine only may not be able to cover all health problems. And it is found that there is a problem with the lack of demand for public health services to meet the needs of the people. Therefore, there is a need for an alternative to health care with herbs to assist in treating the initial symptoms.

It is the information that can be used as public health information, especially in the field of Thai traditional medicine. It can be used to study, understand and present as a channel for herbal recommendations for treatment according to the disease group or initial symptoms. It helps users to access the information more conveniently through various applications. Including health care facilities can also be used to provide services

P. Meesad et al. (Eds.): IC2IT 2021, LNNS 251, pp. 34–43, 2021.
https://doi.org/10.1007/978-3-030-79757-7_4

to clients. The data mining [2] technology is a part of Knowledge Discovery in Databases (KDD) [2]. The data mining is a definition to describe the process of modifying large databases for querying areas of interest, including previously unknown data formats. It is also a suitable choice for analysis, design and build decision model of factors that cause diseases or symptoms. And it can be further developed into an application that can suggest herbs for symptomatic treatment with decision tree [2] technique. This technique is a predictive model which can be used to represent both classifiers and regression models. Conducting research with decision tree refers to a hierarchical model of decision and its sequential outcome. It is presented with algorithms such as C4.5 [2] and J48 [2]. Several researches applied decision tree techniques such as Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor [3], Expert system for rice disease diagnosis [4], and Decision support systems for longan leaf disease diagnosis with decision tree technical [5].

Therefore, the researchers use traditional herb information to analyze, design and develop into a model to be used as a prototype to support application development that can suggest herbs for symptomatic treatment with decision tree technique. The dataset consists of the disease or symptom group data, the training data and test data. It is developed with the J48 algorithm to generate rules and evaluate model performance with k-fold cross validation and find the accuracy model with Weka software [6]. This paper is organized as follows. The introduction is described in Sect. 1. Section 2 reviews the related works and backgrounds. Section 3 describes the research approaches. Section 4 describes the research results. The conclusion is in Sect. 5 and suggestion the future research directions is in Sect. 6.

2 Backgrounds and Related Works

2.1 Herb

Herbs [7] can grow naturally and have meaning for human life, especially in the health dimension both promoting health and treating disease. Most herbs for primary health care are medicinal plants. The use of herbs as medicine must consider the nature of each herb, herbs environment for growing, season and period of harvesting. These are important factors in determining the quality of herbs. Each herb contains many substances such as carbohydrates, fats, essential oils, etc. Herbs for primary health work can be used to treat diseases or illness, such as Gastritis (Turmeric and Banana), constipation (Cassia alata), etc.

2.2 Data Mining

Data mining [2] is a definition to describe the process of modifying large databases for querying areas of interest, including previously unknown data formats. Nowadays, accessing a variety of information is very important and essential in building the data mining. Various tools and techniques can automatically analyze large amounts of data, such as RapidMiner [8], Weka [6], etc. Most of the data mining techniques are based on inductive learning [9]. It is an accuracy model by making forecasts from a sufficient number of training datasets under fundamental assumptions of the inductive learning. The trained model is applicable to future unseen examples.

2.3 Decision Tree

Decision Tree [2] is a predictive model which can be used to represent both classifiers and regression models. Conducting research with decision tree refers to a hierarchical model of decision and its sequential outcome. The decision makers employ decision trees to identify that can best achieve their goals. A classification tree is used to classify an object or an instance. An insurer uses a predefined set of classes such as risky/non-risk and their attributes values such as age or gender. The classification tree is often applied in areas such as finance, marketing, engineering and medicine. The classification tree can also be used as a survey or search technique. The decision tree classifier that is used to facilitate the underwriting process of bank mortgage applications is shown in Fig. 1.

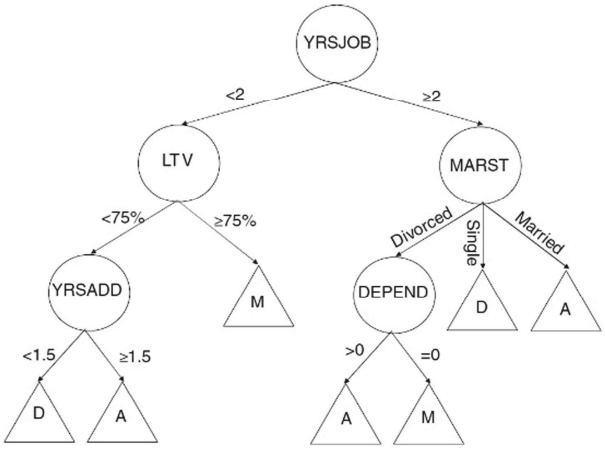

Fig. 1. Decision tree classifier is used to facilitate the underwriting process of mortgage applications of a certain bank [2]

2.4 J48 Algorithm

J48 Algorithm [2] is an open source Java implementation of the C4.5 algorithms in the Weka data mining tool. The J48 algorithm is a reuse of the C4.5 algorithm therefore it performs similar operations to C4.5. The C4.5 algorithm is enhanced from the ID3 algorithm developed by Ross Quinlan [10]. ID3 uses the gain ratio as a criterion for separating pathways. The split path ends when the number of split instances falls below the specified threshold. Error-based pruning is performed after the growing phase. C4.5 can handle numeric attributes.

2.5 K-Fold Cross Validation

k-Fold cross validation [11] is an alternative to Leave-one-out cross-validation (LOOCV). k-Fold Cross Validation randomly divides the set of observations into k groups (folds), which are equal in size. The first fold is treated as a validation set and the first fold is the same size as the remaining k-1 folds. The mean squared error, MSE, is then computed on the observations in the remaining k-1 folds. This procedure is repeated

k times and each time, a different group of observations is treated as a validation set. After finishing k times, the MSE results of each time are a set of mean squared error, MSE1, MSE2, …, MSEk. The average of these values represents the estimation of k-fold cross validation value as shown in Fig. 2.

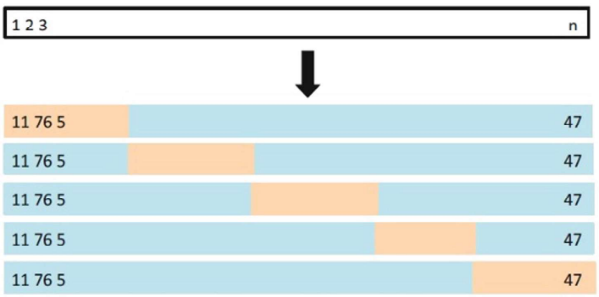

Fig. 2. The k-fold cross validation approach [11]

2.6 Evaluation Model

Confusion matrix [2] is used as an indicator of the classification rule. It contains the number of elements that is correct or incorrect classification for each class. The model performance evaluation is shown in Table 1.

Table 1. Confusion matrix [2]

	Predicted negative	Predicted positive
Negative examples	A	B
Positive examples	C	D

The model evaluation performance is evaluated as follows: Accuracy, Precision, Recall, and F-Measure. They are calculated according to Eqs. 1–4. It is based on the Confusion matrix as shown in Table 1 to evaluate the forecast results compared with the actual results, as the following equations:

- Accuracy

$$Accuracy = ((A + D)/(A + B + C + D)) \times 100 \quad (1)$$

- Precision

$$Precision = (\ D/(B + D)) \times 100 \quad (2)$$

- Recall

$$Recall = (D/(C + D)) \times 100 \quad (3)$$

- F-Measure

$$F - Measure = 2 \times ((Precision \times Recall)/(Precision + Recall)) \quad (4)$$

Where A is the number that the model correctly predicts the negative class. B is the number that the model incorrectly predicts the positive class. C is the number that the model incorrectly predicts the negative class. D is the number that the model correctly predicts the positive class.

2.7 Related Works

In 2011, a decision tree technique was used for medical diagnosis based on information of diabetes, hepatitis and heart disease to develop the decision support system as medical expert. It is based on the C4.5 algorithm, the ID3 algorithm, and the CART algorithm to classify diseases and compare their performances [12]. In 2016, a decision support system was developed to analyze the risk of chronic disease in diabetes and high blood pressure using the decision tree technique from Rapid miner and K-nearest neighbor programs. The model was tested using 10 fold cross validation technique. It was found that the prediction accuracy was 96.45% [13]. In 2017, the medical specialist system was proposed to diagnose common diseases using decision tree technique via mobile application. The J48 algorithm and the Weka program were used to test the model that gives a processing accuracy of 100% [14]. In 2018, a decision tree technique was applied for the diagnosis of disease in cow on mobile phone. The diagnostic models were created using three decision tree algorithms. Cross validation was used to test each model performance to get the best diagnostic model. It was found that the Random Tree algorithm had the most efficient with 99.47% accuracy [15]. In 2019, the research was to examine the relationship of factors affecting pathogenesis and disease risk model of elderly with data mining technique, the result found that the Decision Tree: J48 technique provided the most accurate value with Correctly at 99.796% and generating 10 rules from this model which is an effective and acceptable value [16]. and in 2020, the research aims to analyze coding errors in 43 files database systems for medical records department using data mining techniques, compare performance of decision tree (C4.5) and Naïve Bayes algorithms in analyzing and classifying the errors. The results, when evaluating the classification efficiency with 10-folds cross validation method, showed that the accuracy of the decision tree was 90.16% and the Naïve Bayes algorithm was 89.87% [17]. The decision tree technique was applied in medicine to help support decision making in medicine and other areas such as decision making of loan application by debtor [3], expert system for rice disease diagnosis [4], and decision support systems for longan leaf disease diagnosis with decision tree technical [5].

3 Research Method

The three steps of the research process include the Data Collection and Preparation, Model Construction, and Evaluation Decision Tree Model, as shown in Fig. 3. In the Fig. 3 shows the three steps in the research process as the following:

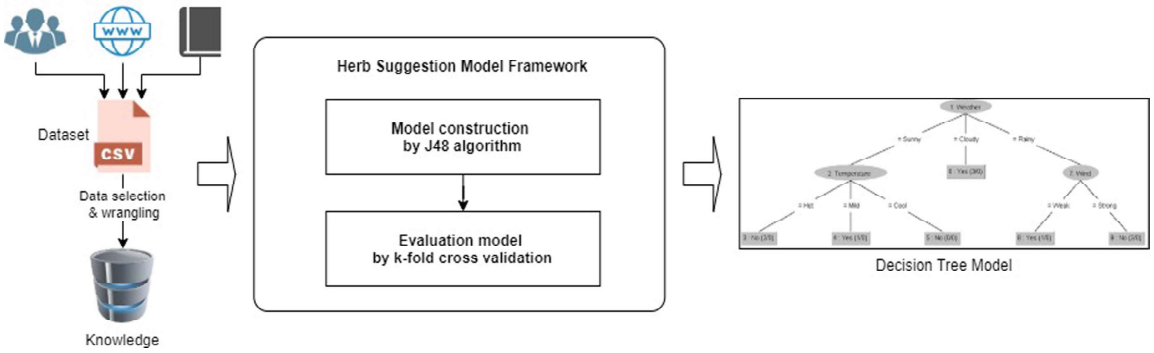

Fig. 3. Three steps in research process: Data Collection and Preparation, Model Construction, and Evaluation Decision Tree Model

3.1 Data Collection and Preparation

In this paper, we collected data set from the book "Herbs for Primary Health" [7] and "Herbs used in Primary Health" website [18]. This will be used as the initial dataset for specifying the data attributes to create a model for the introduction of herbs for the treatment of diseases according to disease and symptoms. It consists of 38 data attributes (including label class). The details of data attributes are described as shown in Table 2.

Table 2. The data attributes used in modeling

No	Code	Attribute	Value	No	Code	Attribute	Value
1	A_0	Colic	yes, no	20	F_0	Dysuria	yes, no
2	A_1	Abdominal wind and other symptoms	yes, no	21	F_1	Mixed blood	yes, no
3	A_2	Stomachache after eating fatty foods	yes, no	22	F_2	Pain when urinating	yes, no
4	A_3	Stomachache before or after meals, cardialgia	yes, no	23	G_0	Toothache	yes, no
5	B_0	Constipation	yes, no	24	H_0	Helminthiasis	yes, no
6	B_Age	Age	<= 50, >50	25	H_1	Malnutrition	yes, no
7	B_1	Symptoms of heart disease	yes, no	26	H_2	Stomachache or vomiting one more	yes, no
8	B_2	How often constipation?	normal, sometime, none	27	H_3	Ascariasis, mouth or anus	yes, no
9	C_0	Diarrhea	yes, no	28	H_4	Stools segmental flat or a long flat line	yes, no

(*continued*)

Table 2. (*continued*)

No	Code	Attribute	Value	No	Code	Attribute	Value
10	C_1	Liquid transfusions more than 3 times	yes, no	29	H_5	Blister hard under the skin	yes, no
11	C_2	Foamy stool with a strong smell	yes, no	30	H_6	Epilepsy	yes, no
12	D_0	Cough	yes, no	31	H_7	Buttocks or vagina itchy at night	yes, no
13	D_1	Cough and sore throat	yes, no	32	H_8	Bloody diarrhea	yes, no
14	D_2	Dry cough	yes, no	33	H_9	Pale disc	yes, no
15	D_3	Cough and white sputum	yes, no	34	H_10	Fever, diarrhea, and muscle pain	yes, no
16	D_4	Cough and green or yellow sputum	yes, no	35	H_11	Jaundice and enlarged liver	yes, no
17	E_0	Vomit	yes, no	36	H_12	Swollen, itchy on the skin	yes, no
18	E_1	Vomiting once more than and mixed blood	yes, no	37	H_13	Eating a lot but not gaining weight or weight loss (children)	yes, no
19	E_2	Dizziness	yes, no	38	Herb	Herb	class

In the Table 2, there are the following data attributes: no, attribute code (code), attribute, and attribute value (value) respectively. For learning the data attributes, we provide data set for next phase that includes the diseases groups and symptoms, along with the results of herbs for treatments by symptoms of diseases.

3.2 Model Construction

We use the data set that is collected and stored as shown in Table 2. After that we input these data set into model construction process for herb suggestion model using decision tree techniques with J48 algorithm [2] in Weka implementation, we want to predict the data attribute "Herbs to treat disease" based on its measured attributes like abdominal wind and other symptoms or stomachache after eating fatty foods and among others (shown in Table 2). The model result shows in Fig. 4.

Figure 4, shows decision tree model which is constructed from the Weka software. As a result, there are 14 rules form for classifying an herbs to treat diseases by disease groups and symptoms, as shown in Fig. 4.

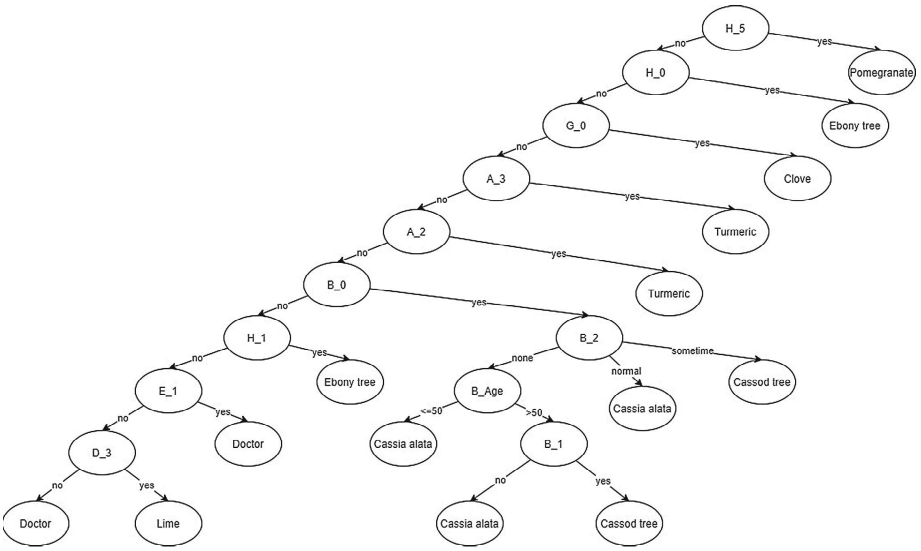

Fig. 4. The herb suggestion decision tree model for treat disease by symptoms using J48 algorithm

3.3 Evaluation Decision Tree Model

This is a step after the model construction step. Evaluation decision tree model step evaluate model performance and find the best performing model. We assess the effectiveness with k-fold cross validation method, splitting the data into 2 to 10 folds. In addition to determining the method of measuring efficiency with accuracy, precision, recall and F-measure values are also taken into consideration of the model's performance. Table 3 shows a comparison of the J48, RandomTree, and REPTree algorithms with fold cross validation, accuracy, precision, recall and F-measure.

Table 3. The performance evaluation results of decision tree model

Algorithm	Fold cross validation	Accuracy	Precision	Recall	F-Measure
J48	9	94.40	0.950	0.944	0.938
RandomTree	10	93.54	0.939	0.935	0.936
REPTree	10	84.05	0.840	0.841	0.835

From Table 3, it is found that J48 algorithm is the best solution with 9 folds (9-Fold Cross Validation). J48 algorithm has accuracy 94.40%, precision 0.950, recall 0.944, and f-measure 0.938.

4 Research Results

The result in Fig. 4 is used to develop as a rule to support the application development for disease analysis by symptoms. The rule condition refers to records in terms of a selected data attributes (disease groups and symptoms). It is presented in the form of conditions

```
IF B_0 = yes AND B_2 = none And B_Age > 50 AND B_1 = yes Then
   Herb = Cassod Tree
End IF
IF B_0 = yes AND B_2 = none And B_Age > 50 AND B_1 = no Then
   Herb = Cassia alata
End IF
```

Fig. 5. The rules for herb suggestion to treat symptoms of "constipation"

IF… THEN…. After that we apply these rules into application. Figure 5 shows rules for the diagnosis of "constipation" symptom.

From Fig. 5, the attribute code B_0, B_2, B_Age, and B_1 represent attributes in table 2. The rules condition for the diagnosis of "constipation" symptom will predicted it based on its measured attributes like having constipation symptom (B_0 = yes), the symptoms are: not frequently constipated (B_2 = none), age over 50 years (B_Age > 50), and heart disease symptoms (B_1 = yes) or no heart disease symptoms (B_1 = no) is the one that contains much more information and for this reason it has been selected as the "Cassod Tree" and "Cassia alata" herbs by "heart disease symptoms" to treat "constipation" symptoms are respectively.

5 Conclusions

When considering the decision tree model and herbal suggestion result by symptoms, it can summarize that the rules from the decision tree classification is 14 rules. Data attribute is related to the disease and the symptoms of the 8 disease groups and symptoms as follows: colic, diarrhea, constipation, cough, vomit, dysuria, toothache, and helminths. Each disease and symptoms group can be identified disease features and symptoms and can be linked to medical herbs for the correct treatment of disease and symptoms. The model can be used to support usage on the application for users to choose to analyze by disease and symptoms. Eventually, this model provides an alternative tool for accessing knowledge in the field of disease treatment by herbal.

6 Suggestion and Future Works

In this research, our main findings are as follows: Firstly, this model should be customized the dataset (as well as attributes) to be used in the design and development of models that are more accurate and applicable in a wider range of applications. Secondly, an application should demonstrate to interact with users to expand other information.

References

1. Food and Drug Administration: The Pathway to Sustainable Thai Herb Development. In: Varavoot, S. (ed.) Med.&Herb Journal 2015, vol. 2, no. 5, pp. 1–16. ACFT, Nontaburi (2015)
2. Rokach, L., Maimon, O.: Data Mining with Decision Trees Theory and Applications, 2nd edn. World Scientific Publishing Co. Pte. Ltd., Singapore (2015)

3. Amin, R.K., Indwiarti, Sibaroni, Y.: Implementation of decision tree using c4.5 algorithm in decision making of loan application by debtor (case study: Bank Pasar of Yogyakarta Special Region). In: 3rd International Conference on Information and Communication Technology (IcoICT), Indonesia, pp. 75–80. IEEE (2015)
4. Patcharaporn, R., Chantana, C.: Expert system for rice disease diagnosis. Veridian E-J. Sci. Technol. Silpakorn Univ. **6**(1), 904–913 (2013)
5. Chidchanok, S., Phaisarn, T., Suradech, B.: Decision support systems for longan leaf disease diagnosis with decision tree technical. Veridian E-J. Sci. Technol. Silpakorn Univ. **1**(6), 1–14 (2014)
6. Weka Software Homepage. https://www.cs.waikato.ac.nz/ml/weka. Accessed 20 Nov 2020
7. Primary Health Care Administrator: Herbal medicine for primary health care. WVO Officer of Printing Mill, Thailand (1997)
8. RapidMiner Homepage. https://rapidminer.com. Accessed 29 Nov 2020
9. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
10. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, United States of America (1993)
11. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: With Application in R. Springer, New York (2013)
12. Kumar, D.S., Sathyadevi, G., Sivanesh, S.: Decision support system for medical di-agnosis using data mining. Int. J. Comput. Sci. Issue (IJCSI) **8**(3), 147–153 (2011)
13. Nicha, N.J.: A decision support system for analyzing the risk of two chronic diseases: diabetes mellitus and hypertension. Ind. Technol. Lampang Rajabhat Univ. J. **9**(2), 11–19 (2016)
14. Thanida, S., Pheerasut, B., Sukanchalika, B., Weerachai, K.: The general medical diagnosis expert system by decision tree technique on the smartphone application. J. Innov. Technol. Manag. **4**(2), 84–93 (2017)
15. Nattavadee, H., Phongnarin, S.: Applying decision tree classification techniques for diagnose the disease in cow on mobile phone. J. Sci. Technol. Ubon Ratchathani Univ. **20**(1), 44–58 (2018)
16. Aporn, M., Nithit, S., Bussayamat, H.: A study of disease risk of elderly with data mining technique. Narkbhutparitat J. Nakhon Si Thammarat Rajabhat Univ. **11**(3), 29–34 (2019)
17. Sakarin, H., Nadh, D.: Analysis of ICD-10 coding errors in 43 files database systems for medical record department using data mining techniques. KKU Sci. J. **48**(1), 142–155 (2020)
18. The Herbs in Primary Healthcare Homepage. http://medplant.mahidol.ac.th/pubhealth. Accessed 08 Dec 2020