International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 7 (2017) pp. 1334-1339

Predictive Data Mining of Chronic Diseases Using Decision Tree: A Case Study of Health Insurance Company in Indonesia

Dini Hidayatul Qudsi
Department of Information System
Politeknik Caltex Riau, Pekanbaru, Indonesia.

Mira Kartiwi
Department of Information System, KICT
International Islamic University Malaysia, Kuala Lumpur, Malaysia.

Nurliyana Binte Saleh
Department of Information System, KICT
International Islamic University Malaysia, Kuala Lumpur, Malaysia.

Abstract:

This study aims to identify the potential benefits that data mining can bring to the health sector, using Indonesian Health Insurance company data as case study. The most commonly data mining technique, decision tree, was used to generate the prediction model by visualizing the tree to perform predictive analysis of chronic diseases. All the steps in data mining process have been performed by a data mining tool, named WEKA. Additionally, WEKA also was utilized to evaluate the prediction performance by measuring the accuracy, the specificity and the sensitivity. Among the result found in this study shows some factors that the health insurance can take into account when predicting the treatment cost of a patient.

Keywords: data mining, prediction model, chronic disease, decision tree, sensitivity analysis, WEKA

INTRODUCTION

Badan Penyelenggara Jaminan Sosial (BPJS), Healthcare and Social Security Agency, is one of the health insurance companies in Indonesia. This company handles health insurance for all kinds of diseases. An example of the latter is their handling of chronic diseases which are the most deadly diseases all over the world [1]. BPJS Health insurance company has been implementing IT systems for many years, to perform several business tasks. In Indonesia, Information System has assisted health practitioners in performing their duties related to decision making (Decision Support System) [2]. Not only in decision making, but the benefits of IT can also be seen in the implementation of e-health, tele nursing, etc. that can improve Indonesian public health services [3].Thus, it is necessary for BPJS Health insurance company to have a good data management.

Still the organization meets some challenges during data management process, especially in the health sector where huge amounts of data needs to be organized and stored. Price et al. (2013) stated that various efforts such as case management implementation, utilization review, and disease management, have been made by the health care data management practitioners to control the cost of the healthcare and handle the utilization of services [4]. However, all of these programs do not appear to work in controlling the cost. They suggested different methods to identify patients with chronic disease (since it has higher risk for readmission) to control cost in medical professional manner and predictive models built by data mining become one of its solutions [4].

Thus, data mining can be one of the solutions by offering its technique to extract information from the huge amount of data that may improve the quality of data decision making management [5]. The crucial objective of data mining is prediction. Predictive data mining is the most common type of data mining and one that has the most straight business applications [6]. Decision tree, as one of the data mining techniques has proven to become the most accurate predictor among other techniques, namely artificial neural network and regression model [7].

As can be seen from the benefits of data mining above, it is the aim of this study to identify the potential benefits that data mining can bring to the health sector in Indonesia. The study will utilize the health insurance data owned by BPJS Health insurance company to predict factors that influence chronic diseases by using decision tree as the data mining technique.

RESEARCH METHODOLOGY

This research adopted Data Mining as the methodology. It is said that data mining itself is a methodology and technology that has been developed which has become famous since 1994 [8]. The methodology and technology of data mining transform a large amount of data into meaningful information to support decision making [9]. In this research, a data mining classification algorithm, C4.5 (J4.8) algorithm, was used to predict factors that influences chronic disease based on several factors such as age, gender, length of stays and disease.

WEKA would process the mining by using the decision tree which is used to help the organization making decisions based on the factors classified. The data gathered was obtained through some sessions of interview with the IT Analyst and the IT manager of BPJS Health insurance company via face to face interactions, phone or email. The main objective of these interview sessions was to get an insight about the kind of data they have in order to have some overviews of the probability how data mining could be beneficial for BPJS Health insurance company. Figure 1 shows the Research Design.

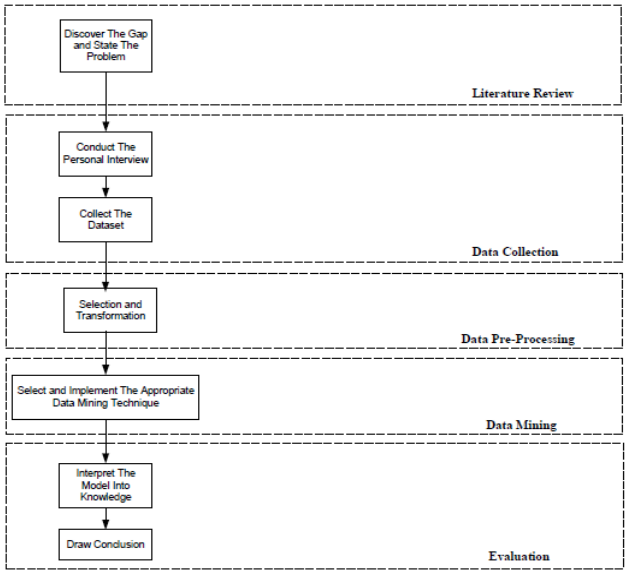

Figure 1. The Research Design

MATERIALS AND METHODS

A. Data Collection and Data Understanding

Patients billing data of the Bengkalis Hospital collected from BPJS Health insurance company in Indonesia from the period of February 1, 2014 to March 15, 2014 was used as the raw data for this research. The total amount of data is about 2352 data patients, consisting of 605 inpatients and 1747 outpatients' data eligible to be analyzed. Out of the 2352 data, 976 are chronic diseases and 1376 are non-chronic diseases patient's data. These data would be used as the input features to process the implementation of C4.5 decision tree algorithm in classification and prediction of the model. The output of the predictive model is "Chronic Disease" or "non-Chronic".

B. Data Integration

In this research, raw data is stored in Comma Separated Value (CSV) format. Three files of the inpatients' database and one file of outpatients' database were integrated into one table which has been designed in Microsoft Excel and has been converted to CSV format in order to be compatible and processed easily by WEKA data mining tool.

C. Data Cleaning

After the process of data integration was completed, the data cleaning was performed, the irrelevant data, the missing value (the data that does not have contain value), and the redundant data (the data that contains more than one records with the same values) have been removed because its existence may reduce the quality or accuracy of the data mining result.

One table has been produced after performing data integration and data cleaning which consists of four columns (Age, Gender, LOS (Length of Stay), and Disease). The description of the attributes can be seen in Table I.

TABLE I. THE DESCRIPTION OF THE ATTRIBUTES IN THE DATASET

Attributes	Description	Type	Possible Value
Age	Patient's Age	Numeric	(1,2,3,…..) years
Gender	Patient's Gender	Nominal	Male, Female
Length of Stay	Length of Stay of the patient in the hospital	Numeric	0 day, 1 day, 2 days,….
Disease	Whether the patients got chronic disease or not	Nominal	Chronic, Non-Chronic

EXPERIMENTAL ANALYSIS

A. Data Preprocessing

Experimental analysis was performed on the patients' data by using WEKA software to get the visualization tree and to measure the classification accuracy, the confusion matrix, and the TP-rate and TN rate. After having loaded the file into the Explorer Window of WEKA, the histograms were produced. It shows that the number of female patients is more than male with the nominal number of 1348 patients. Male patients suffer from chronic diseases nearly as much as those who do not suffer from chronic diseases, whereas the number of female patients suffering from chronic diseases is less than those not suffering from them. Also, it shows that the minimum of LOS of the patient is 1 day and the maximum is about 37 days and the number of patients who suffer from a chronic disease is less than those with a non-chronic disease, with the nominal number of 976 patients compared to 1376.

B. Data Mining Process

After the process of preparing the data is completed, the classification models were built. Decision tree technique provided by WEKA which is C4.5 (J4.8) algorithm classifier has been implemented on the dataset to generate the classification tree. The classifier will be evaluated by training data set.

C. Decision Tree

The generated classification tree is shown in the Figure 2. A number of nodes were shown in Figure 2 with the length of stay (LOS) is identified to be the most critical factor to predict a patient has chronic diseases. Length of stay (LOS) becomes the root node of the tree since it got the highest gain information among the other attributes. Hence, the classification rules generated from the decision tree will be explained in two parts, namely inpatient (for patients whose stay is contained within one day) and outpatient (for patients who stay for more than one day).

D. The Summary Evaluation

The performance of the algorithm is examined by utilizing the decision tree visualization, the outpatient and inpatient analysis, the classification accuracy, the confusion matrix, and the TP-rate and TN rate.

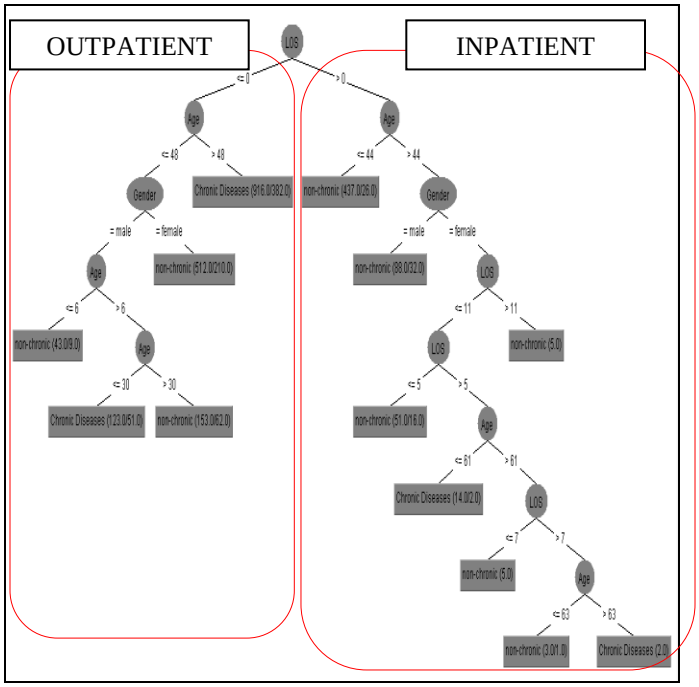

Figure 2. The Decision Tree The Decision Tree where LOS as The Most Critical Factor

1) The Decision Tree Summary Analysis

From the experiment above, it can be summarized that the Length of Stay (LOS) is the most crucial factor that has an impact on the status of the diseases which enable the company to identify the duration of stays of patient depending on several factors, such as age, gender, and kind of diseases. Table II shows the factors that influence chronic disease, which can be concluded that most patients with chronic diseases are outpatient with age more than 48.

TABLE II. THE FACTORS THAT INFLUENCE CHRONIC DISEASES

Type	Age Group (years old)	Gender	LOS (days)	Number
Outpatients	7-30	Male	0	72
Outpatients	>48	Both	0	534
Inpatients	45-61	Female	6-11	12
Inpatients	>63	Female	8-11	2

2) The Accuracy of Prediction Performance Analysis

The value of classification accuracy for predicting performance relatively above the average is about 66.37%. 1561 instances out of 2352 were correctly classified where the category percentage of non-chronic (68.4%) is higher than chronic disease (63.5%). The result were not too much different from the previous research whereas their precision's result is about 67%; he has predicted the chronic disease through the anomaly happened with the patients based on vital signs like ECG result [10].

The Kappa statistic value is 0.3154 which assess the agreement among values predicted by the model. The following error values (mean absolute error, root mean squared error, relative absolute error and root relative squared error) estimate the error of the prediction. In addition, ROC area for both classes of the decision tree is 0.713, which indicates that the validity of the classifier is high. Since, the closer the ROC to 1, the higher the discriminating power of the classifier [11].

WEKA allows visualizing the classification errors. Classifier error indicated that a number of factors needed to be analyzed that should be taken into consideration extensively. For example, there were younger patients with non-chronic diseases to be more likely incorrectly classified as chronic disease. This is perhaps due to the lack of data patients with chronic diseases especially for the younger patients considering the ratio between young patients (below 25 years) with chronic and non-chronic diseases is 140: 387. Consequently, the algorithm does not have enough data to predict chronic diseases for the younger patients.

Data mining is a naturally iterative process, where several steps need to be repeated many times [12]. In this case, considering the accuracy which is not high, data preprocessing were performed again in order to increase the accuracy, such as removing the error one by one (based on the classification error), removing the noisy data, detecting the outlier and minimizing the numeric data by classifying the age into toddlers, kiddy, early teens, late teens, early adulthood, late adult, initial elderly, late elderly, and elderly (based on the age grouping according to Health Department Republic of Indonesia). But, even after performing data processing for the second time, still it could not raise the accuracy.

E. The Confusion Matrix

Table III shows the confusion matrix for every classifier to interpret and understand the results.

TABLE III. THE CONFUSION MATRIX

		Actual Classes	
		A	***B***
Predicted Classess	***a = Non-Chronic***	TP **941**	FP 435
	b = Chronic Diseases	FN 356	TN **620**
J4.8	Total	1297	1055
	Accuracy	72.55 %	58.76 %

A confusion matrix provides the information that is used to describe in numbers how well the performance of the classification model is, which makes it easier to compare the performance of different models [13]. In addition, the confusion matrix does not only show how well the model

predicts, but also the details that might go wrong during the data mining process. The confusion matrix is shown in Table III where the columns show the actual classes and the rows show the predicted classes. From Table III, the number of 941 and 620 indicates the number of cases where the actual and predicted values are similar. 941 represents the patients who are predicted with non-chronic disease are truly non-chronic patients while 620 represents the patients who are predicted with chronic diseases are truly chronic patients. In other words, the diagonal shows all the right predictions. While the number of 435 represents the number of cases where the actual outcome was a non-chronic disease but was predicted as being a chronic disease and the number of 356 represents the number of cases where the outcome was a chronic disease but it was predicted as non- chronic. Also, it can be analyzed that the percentage of the accuracy for non-chronic category is higher (72.55 %) than chronic disease (58.76%).

1) Sensitivity and Specificity

The reliability of the prediction test can be measured by calculating the sensitivity and specificity [14]. In Table III, the value of TP (True Positive) was initiated to be 941, and the value of FP (False Positive) was initiated to be 435. Moreover, 356 instances of FN (False Negative) were found and 620 instances of TN (True Negative) were initiated. The cells labeled with TP (True Positive) are the number of actual deception cases which are exactly predicted by the classifier, while the other cells, FP, FN, and TN are interpreted in similar ways. The performance of the algorithm not only can be evaluated by calculating the accuracy, but also sensitivity and specificity.

Sensitivity refers to the ability of the prediction test to be correctly identifying those patients with the non-chronic disease. The sensitivity is identical to the true positive rate:

Sensitivity/Recall = TP/ (TP + FN)

$$941/ (941+356) = 0.73$$

Specificity refers to the ability of the prediction test to be correctly identifying those patients with the chronic disease. The specificity is identical to the true negative rate:

Specificity/Precision = TN / (TN+FP)

$$620 / (620+435) = 0.59$$

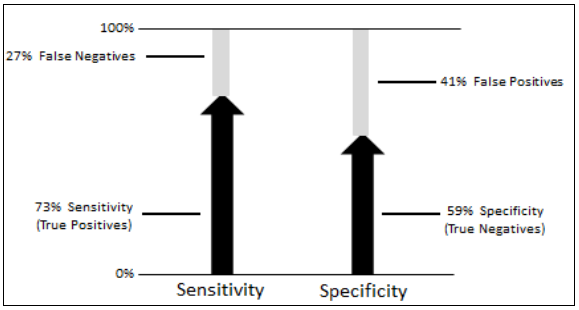

Figure 3. Sensitivity and Specificity Proportion

Figure 3 shows a high sensitivity but low specificity (higher False Positives) explaining the number of cases where the actual outcome was a non-chronic but it was predicted as chronic disease. It might be caused by the imbalanced data where a number of non-chronic data is much higher than chronic. The issue of imbalance data has spread and occurs universally which affect the result of the data mining process [15]. For example, in this research, a number of outpatients are much larger than inpatients data with the ratio of 1747: 605 and a number of patients with chronic diseases are much fewer than non-chronic disease with the ratio of 976: 1376. Since the TP-rate/sensitivity up to 0.73, the prediction of all the instances by the model is true, considering the result of the higher TP-rate/sensitivity.

CONCLUSION AND DISCUSSION

A. Data Mining Performance

The initial purpose of this research is to predict the demographic factors affecting chronic disease which will be classified based on chronic or non-chronic. But, after performing data mining process, the length of stay (LOS), as the knowledge obtained from the prediction model, emerges as the most important factor which strongly predicts a person has a chronic disease. In other word, the Length of Stay (LOS) is a strongly predictor of chronic diseases, which enables the company to predict the duration of stays (LOS) for the patients depending on disease status, gender, and age. Therefore, data mining has been proven to be more than a matter of statistics, but also a predictive analysis tool.

Moreover, C4.5 algorithm successfully generated a classification model in the form of visualization (decision tree) that can be read and interpreted easily to predict the chronic disease. The prediction accuracy was 66.37%. However, the sensitivity measurement up to 0.73 was achieved, which means the reliability of the prediction test is good and the data has been proven 100% valid. Thus, despite the low accuracy of the model derived, the result shows some factors that the company can take into account when predicting the treatment cost of a patient. For example, the model shows that gender or age, strongly affect the duration of hospitalization for a patient. Older age patient with chronic diseases have to be admitted in hospital longer than the younger ones. In conclusion, decision tree is a useful and an informative technique to perform predictive data mining.

B. Data Mining Benefit for Health Insurance Company

The knowledge that could be identified from the patterns generated after performing the data mining can probably be taken into consideration as a new policy for BPJS Health insurance company. The visualization of the decision tree that is easy to read and interpret also supports the people from the company to simply draw the information. Below are the suggestions that can be adopted in BPJS Health insurance company.

1) BPJS Health insurance company might be able to identify the duration stays of the patient with chronic disease, to allocate more funds for health services and medicines against outpatients of chronic diseases, to estimate how much premium package and coverage that they should decide for one inpatient and to estimate how much premium package the participants should pay every month

2) Conduct health counseling, especially for women who are prone to chronic diseases compared to the male counterparts.
3) Increase the knowledge, ability, awareness of chronic diseases and understanding of the participants
4) It is necessary to perform medical record screening of participants above 40 years, at least once every year. The purpose is to detect risk factors for chronic diseases.

C. The Importance of Data Quality in Health Sector

1) The Proportion of the Sample

Before performing data mining, it is important to make sure that the proportion of the sample is balanced. It does not have to be an equal number but adequate for the machine, to learn how to produce the classification for each class. The data used in this study is largely dominated by one class of the samples. As a result, the imbalanced data caused low accuracy and incorrectly classified, due to the number of non-chronic diseases data that are more than the chronic disease

2) The Utilization of Different Techniques for Different Analysis

The organization can consider the different methods usage to enhance the accuracy. For example, one of the techniques is random forest or Support Vector Machine (SVM) which is one of the options available to enhance the accuracy [16]. However, other techniques might have better accuracy but lack of visualization like what decision tree offers. Especially, for the company where the people are not familiar with data mining, it will be easier for them to draw the information directly from the visualization of the tree.

3) The Quality of Data

Based on this study, it is found that the quality of the data from the company is not up to the expected quality. As a result, a series of data cleaning techniques were adopted. It indicated that data mining techniques performance are depending on the quality of the data where the company must take it into consideration. The data used in this research contains many of missing value, inconsistent data or incomplete information which will affect the quality of the data that have to be cleaned (data cleaning) before performing data mining process. Therefore, some mechanisms are needed to ensure that the data being kept are clean. So, the data miner can easily process the data without having to be intervened by certain kind of personal justification or personal intuition which can be prevented if the organization has anticipated having clean data from the beginning.

4) Data Management Maturity

Having a good data management is one of the approaches to maintain the quality of the data. The quality of the data can be maintained when the organization is willing to invest funds to conduct gap analysis to understand the level of maturity and make improvement so that the procedures and methods on how the data is managed can be improved. The data management maturity level can be used to assist to assess the current state of the data management of the organization. Based on the findings of this research, the organization has no process to know whether the data is clean or not. Thus, based on the five data quality maturity levels stated by Larry English's approach [17], the quality of the data does not reflect the maturity at the second level where the organization is not only require to perform data profiling and data cleansing, but also an enterprise-wide support to improve data quality.

REFERENCES

[1] Cmcd.sph.umich.edu,. (2014). *What is Chronic Disease? | Center for Managing Chronic Disease - Putting People at the Center of Solutions*. Retrieved 7 November 2014, from http://cmcd.sph.umich.edu/what-is-chronic-disease.html

[2] Oberty, E. (2012). Efektifitas dalam penerapan teknologi PDA (Personal Digital Assistant) di Pelanyanan Keperawatan. *Faculty of Nursing, University of Indonesia*.

[3] Murdiyanti, D. (2012). *Penerapan Telenursing Sebagai Salah Satu Cara Menyediakan Pelayanan Keperawatan Dalam Era Teknologi Informasi*. Indonesia.

[4] Price, S., Dobbs, D., Oliveira, J., Beidas, S., Burkhart, T., & Sharp, J. (2013). Clinical & Business Intelligence : Data Management – A Foundation for Analytics. Retrieved from http://www.himss.org/files/himssorg/content/files/201304_data_enrichment_enhancement_final.pdf.

[5] Milovic, B., & Milovic, M. (2012). Prediction and Decision Making in Health Care using Data Mining Corresponding Author :, *1*(2), 69–76.

[6] Statsoft.com,. (2014). *What is Data Mining, Predictive Analytics, Big Data*. Retrieved 15 December 2014, from http://www.statsoft.com/textbook/data-mining-techniques

[7] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*, 113–127. doi:10.1016/j.artmed.2004.07.002

[8] Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science and Technology (ARIST)*, *32*, 197–229.

[9] Koh, H., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management—Vol*, *19*, 64–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22585098\nhttps://www.himss.org/ASP/ContentRedirector.asp?ContentID=76467

[10] Tseng, V. S., Chen, L. C., Lee, C. H., Wu, J. S., & Hsu, Y. C. (2008). Development of a vital sign data mining system for chronic patient monitoring. In *Proceedings - CISIS 2008: 2nd International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 649–654). doi:10.1109/CISIS.2008.140

[11] Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health Care Management Science*, *11*(3), 275–287. doi:10.1007/s10729-007-9045-4

[12] Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management* (2nd ed., p. 7). Indianapolis, Indiana: Wiley.

[13] Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Classification : Basic Concepts , Decision Trees , and. In *Introduction to Data Mining* (Vol. 67, pp. 145–205).

Boston: MA: Addison Wesley. doi:10.1016/0022-4405(81)90007-8

[14] Patil, B. M., Toshniwal, D., & Joshi, R. C. (2009). Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment. *IEEE International Advance Computing Conference*, (March), 6–7.

[15] Chawla, N. V. N. (2005). Data Mining for Imbalanced Datasets- An Overview. *Data Mining and Knowledge Discovery Handbook*, 853–867. doi:10.1007/0-387-25465-X_40

[16] Edeki, C. A., & Shardul., P. (2012). A Comparative Study Of Data Mining And Statistical Learning Techniques For Prediction Of Cancer Survivability. *International Journal of Computer Science and Information Security.*, *10*(June).

[17] Adelman, S., Moss, L., & Abai, M. (2005). Data Quality. In Data Strategy (1st ed., p. 384). Addison-Wesley Professional. Retrieved from http://ptgmedia.pearsoncmg.com/images/0321240995/samplechapter/Adelman_ch03.pdf