Jurnal Politeknik Caltex Riau
http://jurnal.pcr.ac.id

# Predictive Analytics Data Mining in Imbalanced Medical Dataset

**Dini Hidayatul Qudsi**[1]

[1]Politeknik Caltex Riau, email: dinihq@pcr.ac.id

**Abstract**

*The development of information and communication technology has rapidly penetrated to several sectors including health sector. A good data management has become necessity for a healthcare company since it will provide better control of the costs and mitigate risks. However, to develop a good quality data management is complex. Therefore, data mining as one of the advancements of science and technology development offers its technique (such as decision tree) to mine the hidden information from the large amounts of medical data that may improve the decision making. But still, the previous studies founded that the issue of imbalanced datasets has spread and occurs universally which affect the result of the data mining process. It is the aim of this study to identify the potential impact of the imbalanced medical data mining. 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 such as data collection, data preprocessing and data mining have been performed by a data mining tool, named WEKA. Additionally, WEKA also was utilized to evaluate the prediction performance by measuring the accuracy. Among the result found in this study shows that imbalanced datasets caused low accuracy and incorrectly class classified.*

***Kata kunci:*** *predictive data mining, imbalanced dataset, accuracy*

## 1. Introduction

Money and people have long been considered to be assets, but nowadays, many organizations rely on their data to make more informed and effective decisions which help the organizations to achieve their goals. Hence, data needs to be managed seriously [1]. But developing a good quality data management is not easy. Still sometimes, the organization meets some challenges during data management process, especially in the health sector, huge amounts of data need to be organized and stored.

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. 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) or to predict disease progression and health status have been considered by the health insurers and health systems to control cost in medical professional. Therefore, predictive models built by data mining could be one of the solutions.

Dokumen diterima pada 23 Oktober 2016
Dipublikasikan pada 25 November, 2016

Data mining is becoming more well-known day by day, since it strengthens the companies to discover profitable patterns and trends from their existing databases [2]. Still, [3] found that the issue of imbalanced data has spread and occurs universally which affect the result of the data mining process. The data is called imbalanced if a number of instances of one class is not approximately equal to another class [4].

As can be seen from the statements above, it is the aim of this study to identify the potential impact of imbalanced medical data mining. The study will utilize the health insurance data owned by BPJS Health insurance company and decision tree as the data mining technique.

## 2. Research Methodology

The framework of this research which is based on data mining tasks starts from stating the problem statement and formulating the hypothesis, collecting the raw data from several sources for data mining process, selecting and transforming the data, selecting and implementing the proper data mining technique, then finally, interpreting and drawing the conclusion from extracting knowledge after the data mining process.

### 2.1 Problem Statement and Hypothesis Formulation

The purpose of this phase is to collect some previous information related to this study obtained from the Internet, several books, articles and journals. Those collected materials are investigated and reviewed to identify the specific knowledge and experience or to discover a gap from previous works as a meaningful problem statement. Specifically, reviewing the previous studies about the challenge of data management in the health sector.

### 2.2 Data Collection

The purpose of this phase is to collect the dataset in order to fulfill the requirement. A personal interview session was conducted with the IT analyst in order to obtain the dataset and to gather the information about the kind of data they have and the possibility to get the data, especially, to get the data of the patient diagnosis which will be processed using the data mining technique. Additionally, personal interviews were also conducted in order to get the relevant data for the data mining process and to get the overview of the possible way to deal with the issues discovered in the healthcare data.

### 2.3 Data Pre-Processing / Data Preparation

Typically, the necessary data for data mining is not only from one database but from multiple databases or text files. Similar to this study, data integration was conducted since data needed for data mining process for this research is not only derived from one source but it, too, also is derived from different sources / text files. Moreover, the data obtained, either from the database of a company or from the experimental results, usually generate missing or invalid data. It is better to eliminate the irrelevant data because it can reduce the quality or accuracy of data mining results. Therefore, transformation and data cleaning have been carried out in the data integration, such as the elimination of unnecessary columns of the table or the replacement of the missing values in the data. During the transformation stage all collected data is transformed to the formats that are used for process visualization and analysis so that it will be easy to understand for an analysis.

### 2.4 Data Mining

The activities conducted in this phase include:

1. Implementing data mining technique to the dataset that is classification rules, particularly the decision tree. This model was selected in this study considering people can interpret and understand easily by visualizing the tree, in order to extract the meaningful information and also as mentioned in the previous study, it is the most

accurate prediction technique among other techniques [5]. C4.5, known as J4.8 in WEKA, was used as the algorithm to generate the decision tree.

2. Using Microsoft Excel to store the dataset and WEKA 3.6.9 (data mining software) to process the data mining.

### 2.5 Evaluation

In this phase the results of data mining techniques in the form of distinctive patterns and prediction models were evaluated. After that, the extracted knowledge from data mining process was formulated into a decision or action in order to generate some useful information**.**

## 3. Results and Findings

### 3.1 Decision Tree

This study carried out the data mining process of 2352 patients' data from the period of February, 1 2014 to March, 15 2014. A group of attributes has been selected to be processed in WEKA to predict the factors that influence chronic disease based on 4 criteria which are age, gender, Length of Stay (LOS) and disease. After the the data processed, the classification models are built (See Figure 4.5).

A number of nodes were shown in Figure 1 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.

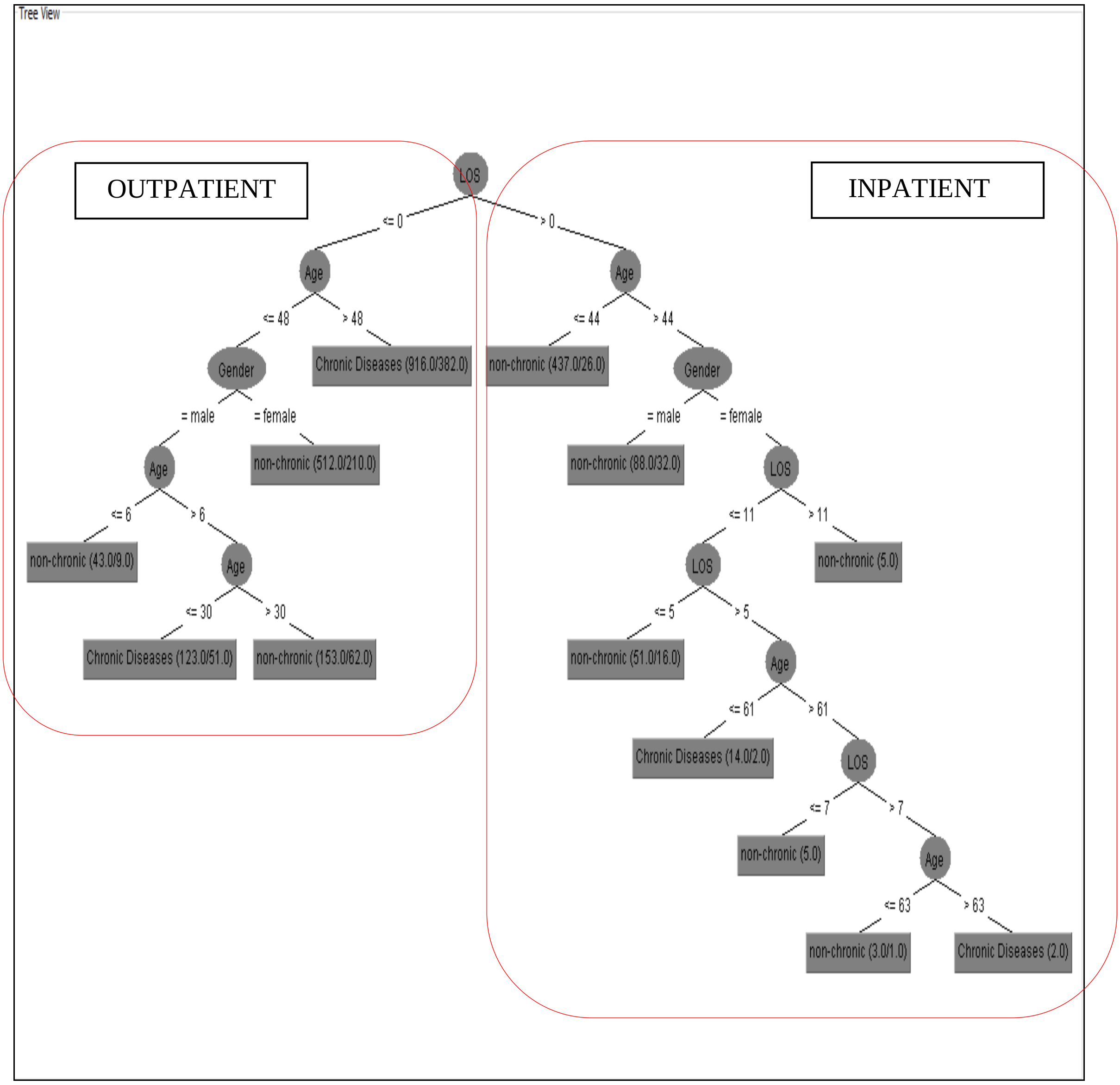
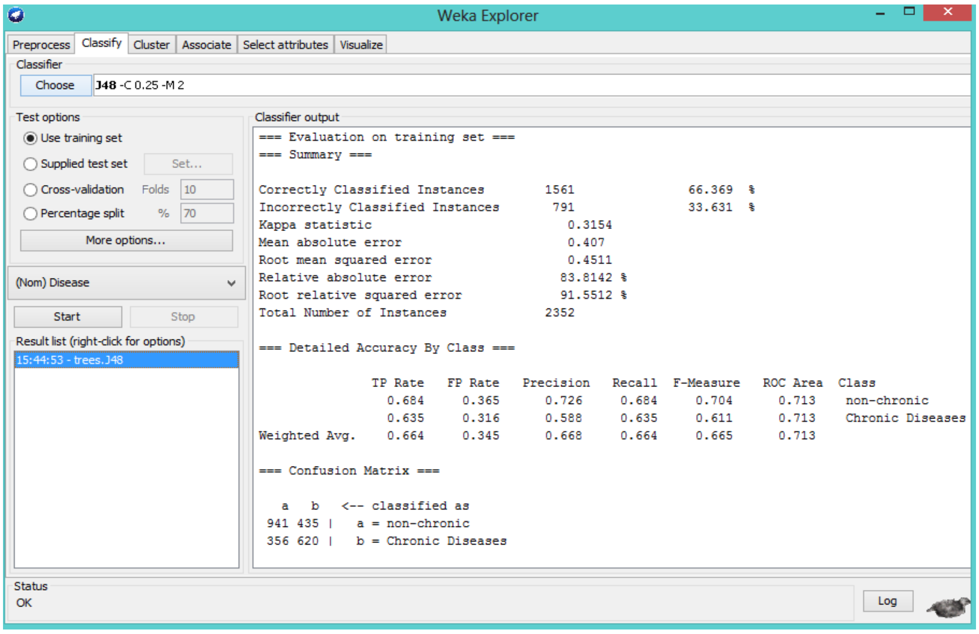


**Figure 1. The Decision Tree where LOS as The Most Critical Factor**

### 3.2 The Accuracy of Prediction Performance Analysis

The classifier output of Decision Tree (Figure 2) generated from WEKA showed that the value of classification accuracy for predicting performance relatively above the average which 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%).



Classifier: J48 -C 0.25 -M 2

Folds: 10

%: 70

(Nom) Disease

15:44:53 - trees.J48

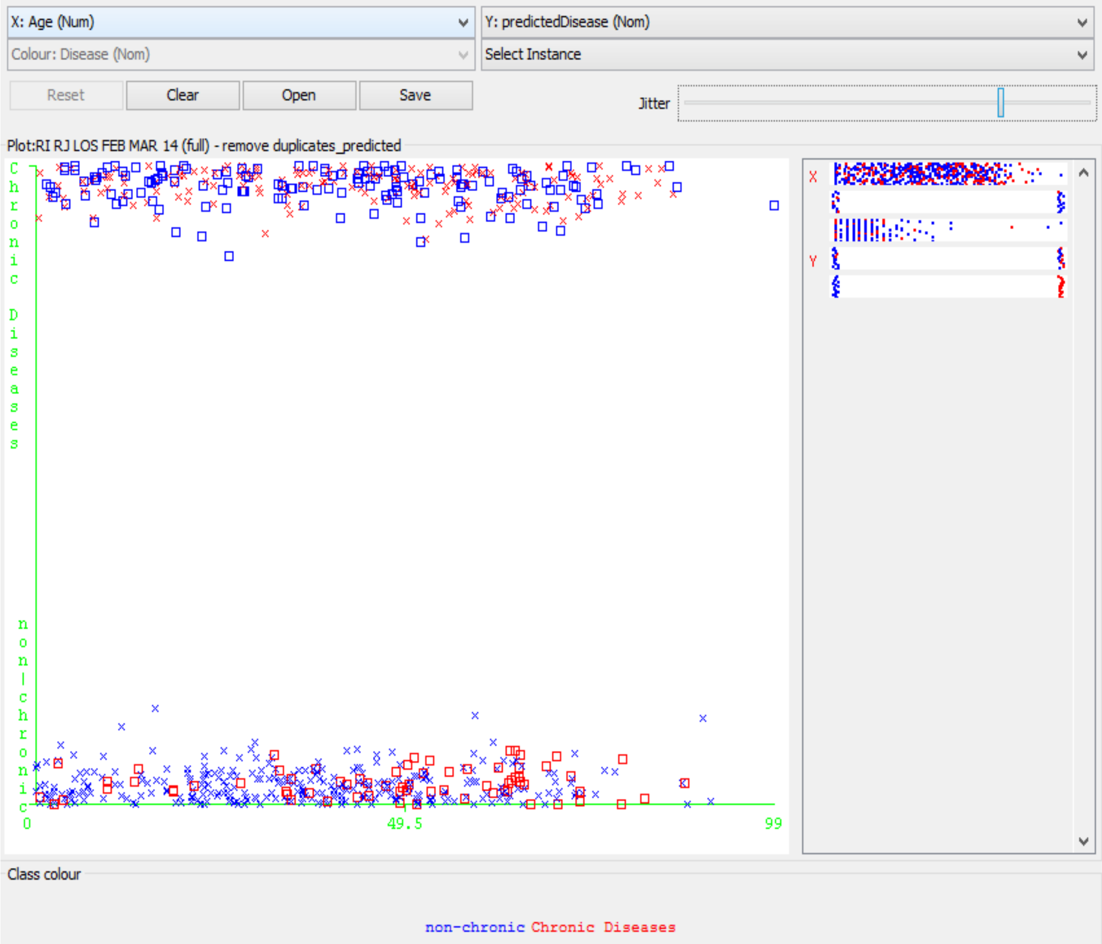
```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1561               66.369  %
Incorrectly Classified Instances       791               33.631  %
Kappa statistic                          0.3154
Mean absolute error                      0.407
Root mean squared error                  0.4511
Relative absolute error                 83.8142 %
Root relative squared error             91.5512 %
Total Number of Instances             2352

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.684     0.365      0.726     0.684     0.704      0.713    non-chronic
                 0.635     0.316      0.588     0.635     0.611      0.713    Chronic Diseases
Weighted Avg.    0.664     0.345      0.668     0.664     0.665      0.713

=== Confusion Matrix ===

   a   b   <-- classified as
 941 435 |   a = non-chronic
 356 620 |   b = Chronic Diseases
```

Status: OK

**Figure 2. The Classifier Output of Decision Tree**

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 [6].

WEKA allows visualizing the classification errors, as shown in Figure 3. Correctly classification instances are represented as crosses, while the incorrectly classification instances are represented as squares. The blue cross in the left lower corner indicates correctly classified instances. To identify the errors, click on the blue squares in the upper left corner or in the red squares in the lower left corner. The details of the errors would appear, as can be seen in Figure 4.

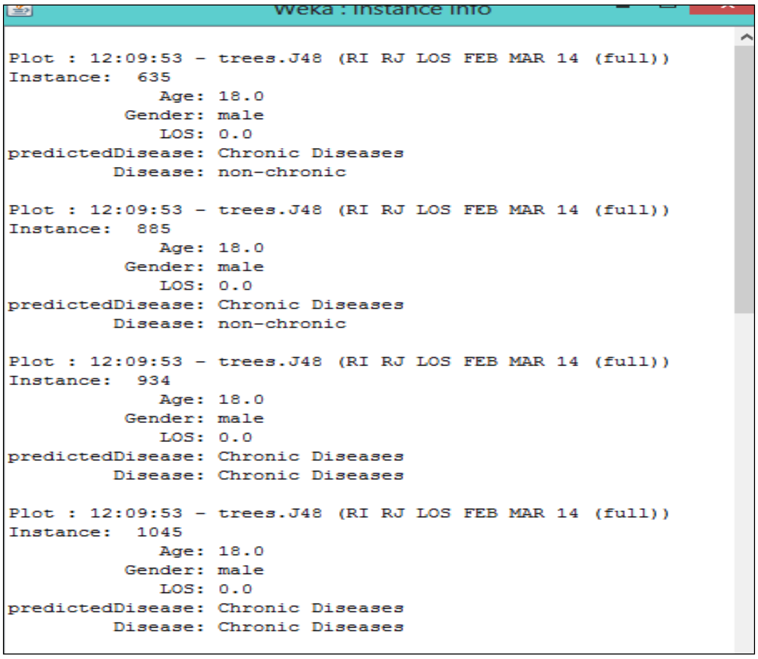


X: Age (Num)

Y: predictedDisease (Nom)

Colour: Disease (Nom)

**Figure 3. The Classifier Error Graph of Prediction Model**



```
Plot : 12:09:53 - trees.J48 (RI RJ LOS FEB MAR 14 (full))
Instance:   635
             Age: 18.0
          Gender: male
             LOS: 0.0
predictedDisease: Chronic Diseases
         Disease: non-chronic

Plot : 12:09:53 - trees.J48 (RI RJ LOS FEB MAR 14 (full))
Instance:   885
             Age: 18.0
          Gender: male
             LOS: 0.0
predictedDisease: Chronic Diseases
         Disease: non-chronic

Plot : 12:09:53 - trees.J48 (RI RJ LOS FEB MAR 14 (full))
Instance:   934
             Age: 18.0
          Gender: male
             LOS: 0.0
predictedDisease: Chronic Diseases
         Disease: Chronic Diseases

Plot : 12:09:53 - trees.J48 (RI RJ LOS FEB MAR 14 (full))
Instance:  1045
             Age: 18.0
          Gender: male
             LOS: 0.0
predictedDisease: Chronic Diseases
         Disease: Chronic Diseases
```

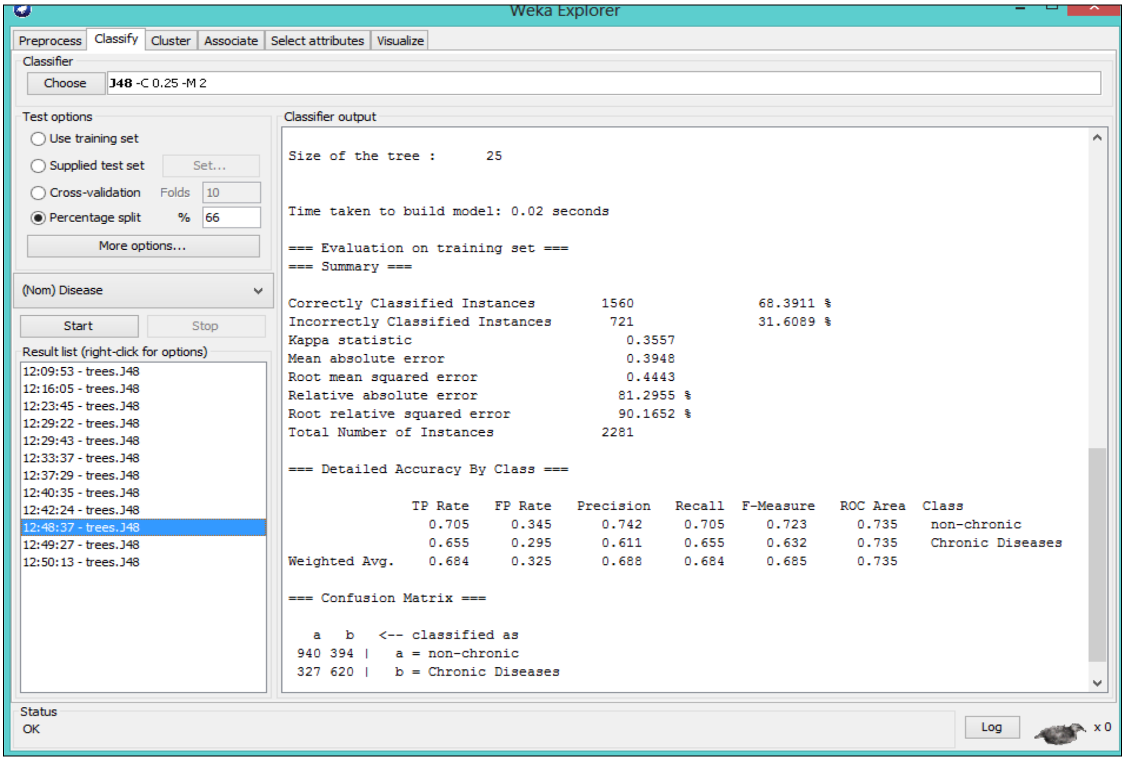
**Figure 4. The Classifier Error Information**

Classifier error indicated that a number of factors needed to be analyzed that should be taken into consideration extensively. For example, Figure 4 shows that, it was 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. In this case, considering the accuracy which is not that high, data preprocessing would be performed again in order to increase the accuracy. Below are some efforts, as follow:

1. Remove the error one by one.

By removing some of the noisy data based on as seen in the Figure 4. As a result, the accuracy of prediction performance increases and the prediction model is correctly classified 68.39 % which will continue to increase if all of the error data is removed (See Figure 5). But, it is not an effective way to enhance the accuracy, because, by removing the data, it will reduce the contained information from the data itself.



J48 -C 0.25 -M 2

Folds 10

% 66

(Nom) Disease

12:09:53 - trees.J48
12:16:05 - trees.J48
12:23:45 - trees.J48
12:29:22 - trees.J48
12:29:43 - trees.J48
12:33:37 - trees.J48
12:37:29 - trees.J48
12:40:35 - trees.J48
12:42:24 - trees.J48
12:48:37 - trees.J48
12:49:27 - trees.J48
12:50:13 - trees.J48

```
Size of the tree :      25


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1560               68.3911 %
Incorrectly Classified Instances       721               31.6089 %
Kappa statistic                          0.3557
Mean absolute error                      0.3948
Root mean squared error                  0.4443
Relative absolute error                 81.2955 %
Root relative squared error             90.1652 %
Total Number of Instances             2281

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.705     0.345      0.742     0.705     0.723      0.735    non-chronic
                 0.655     0.295      0.611     0.655     0.632      0.735    Chronic Diseases
Weighted Avg.    0.684     0.325      0.688     0.684     0.685      0.735

=== Confusion Matrix ===

   a   b   <-- classified as
 940 394 |   a = non-chronic
 327 620 |   b = Chronic Diseases
```

OK

**Figure 5. The Accuracy of Prediction Performance After Removing the Errors**

2. Remove the noisy data.

WEKA, a data mining tool, provides a feature to remove instances which are incorrectly classified by using the following method:

*After load the file > Choose Button > WEKA > Filters > unsupervised > instance > removeMisclassField > ok > apply button > save.*

Yet, instead of removing the incorrectly classified data (the noisy data), it removed all the chronic disease instances.

3. Detect the outlier

The following is the method to find the outlier, as follow:

*After load the file > Choose Button > WEKA > Filters > unsupervised > attribute > inter quartile range > ok > apply button > save.*

The result shows that the accuracy is not much different, even lower, about 65.7%.

4. Minimize the numeric data by classifying the age

The age has been classified 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. The result shows the accuracy which is not much different, with the percentage of 64.92%. Moreover, the decision tree generated (See Figure 6) did not provide the specific information, for example, about how long the patients with chronic disease stay or how old the patient with chronic disease who stays for 6 days.

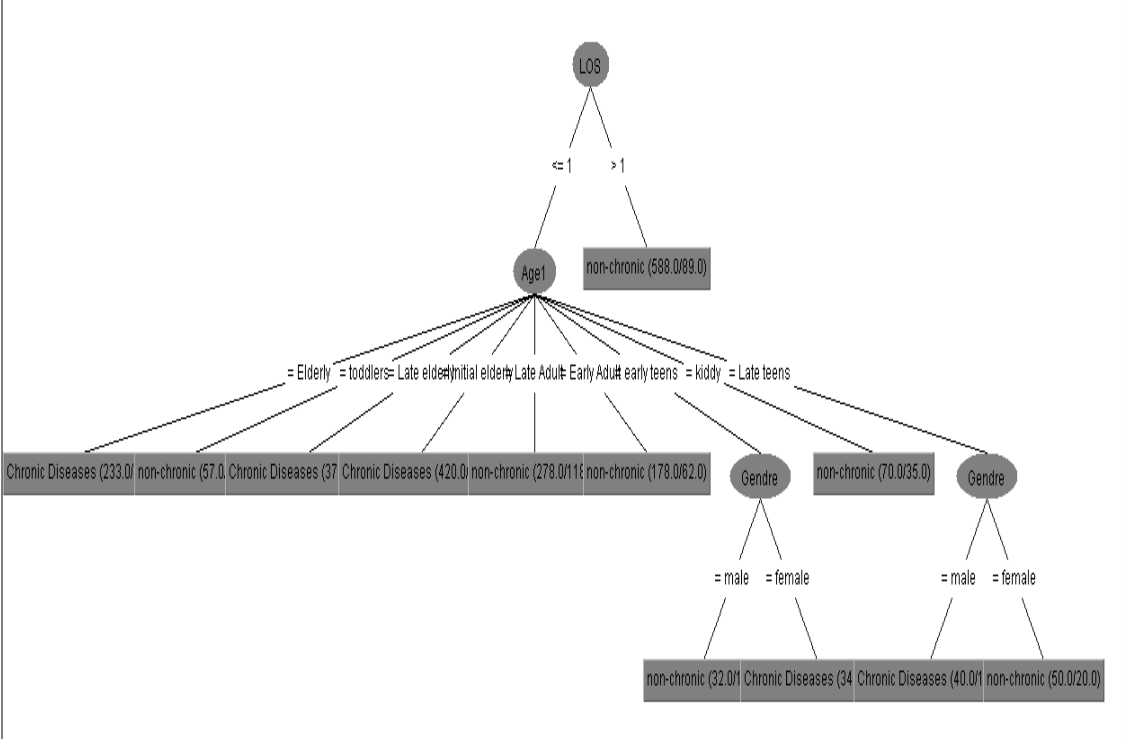


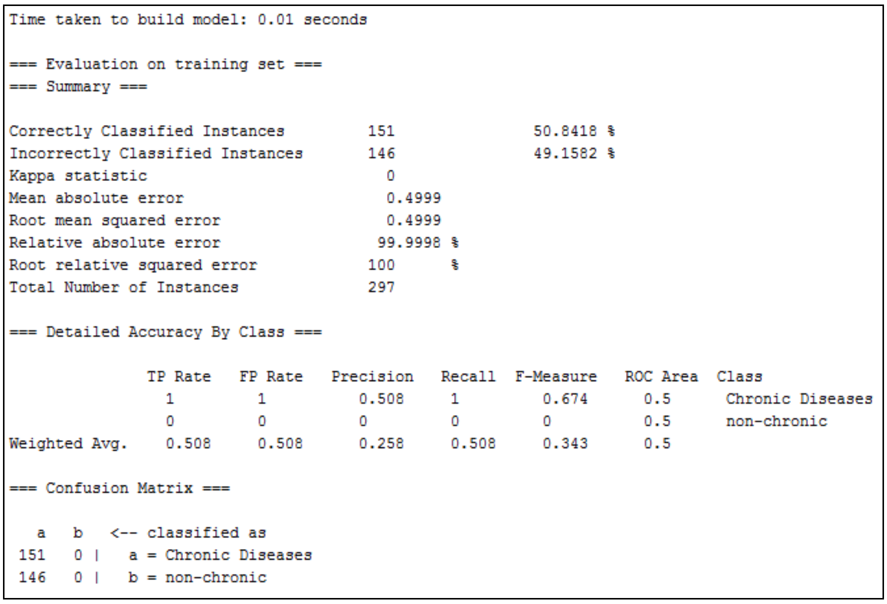
**Figure 6. The Decision Tree Based on Age Grouping**

Therefore, based on this research, it has justified that several causes why the accuracy result is not too high, as followed:

1. Data quality.

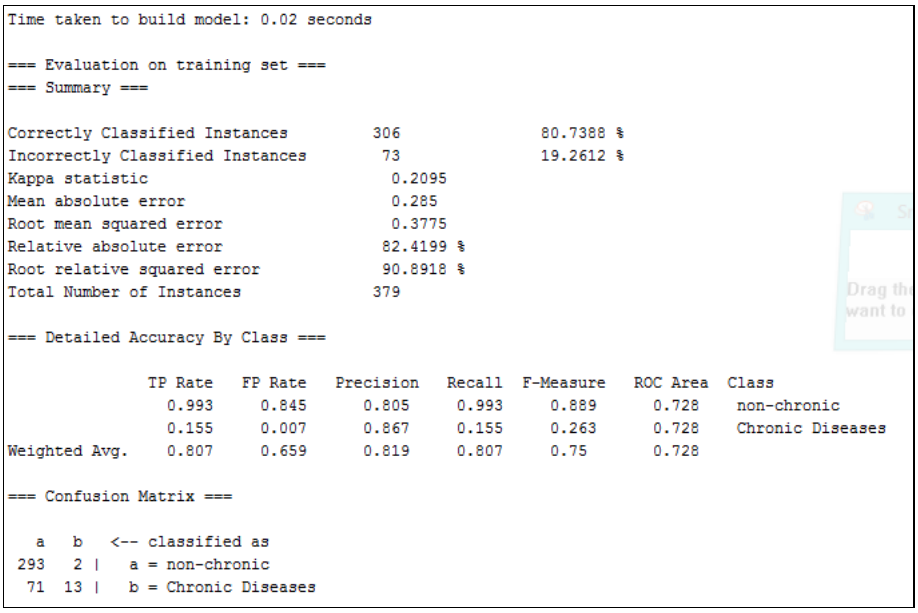
Most of the incorrectly classified data comes from outpatient. This probably happened because of the incorrectly data entry by the operators. For example, if a patient stayed for 3days, from Monday to Wednesday. She is going home on Thursday, but the disease relapses on Friday, thus, she has to visit again on Friday to be treated by the doctor. Probably, the staff will record it as outpatient, but actually it supposed to be inpatient since she has stayed at hospital for 3 days before.

Moreover, it is probably occurred because some classes are identified incorrectly. Therefore, an experiment has been conducted to check the accuracy of prediction model for inpatient and outpatient, respectively. And it can be seen from the result that the accuracy for outpatient is under average which is about 50.481 % (See Figure 7), while the accuracy for the inpatient only, relatively high, and is about 80.738% (See Figure 8).



```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         151               50.8418 %
Incorrectly Classified Instances       146               49.1582 %
Kappa statistic                          0
Mean absolute error                      0.4999
Root mean squared error                  0.4999
Relative absolute error                 99.9998 %
Root relative squared error            100      %
Total Number of Instances              297

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1         0.508     1         0.674      0.5       Chronic Diseases
                 0         0         0         0         0          0.5       non-chronic
Weighted Avg.    0.508     0.508     0.258     0.508     0.343      0.5

=== Confusion Matrix ===

   a   b   <-- classified as
 151   0 |   a = Chronic Diseases
 146   0 |   b = non-chronic
```

**Figure 7. The Accuracy of Prediction Performance for Outpatient Only**



```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         306               80.7388 %
Incorrectly Classified Instances        73               19.2612 %
Kappa statistic                          0.2095
Mean absolute error                      0.285
Root mean squared error                  0.3775
Relative absolute error                 82.4199 %
Root relative squared error             90.8918 %
Total Number of Instances              379

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.993     0.845     0.805     0.993     0.889      0.728     non-chronic
                 0.155     0.007     0.867     0.155     0.263      0.728     Chronic Diseases
Weighted Avg.    0.807     0.659     0.819     0.807     0.75       0.728

=== Confusion Matrix ===

   a   b   <-- classified as
 293   2 |   a = non-chronic
  71  13 |   b = Chronic Diseases
```

**Figure 8. The Accuracy of Prediction Performance for Inpatient Only**

2. Imbalanced data.

The issue of imbalance data has spread and occurs universally which affect the result of the data mining process [2]. The data is called imbalanced if a number of instances of one class is not approximately equal to another class [3]. 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. Such imbalanced data may results in low accuracy, which in this case is 68.4 % of the model derived. [3] stated that this could occur because of the standard algorithms used to execute mainly on balanced class distribution.

However, the data used in this prediction model has gone through a conscientious validation process. The data has been examined using SPSS software. The result shows that the reliability of data is 100% valid.

## 4. Conclusion

Based on the findings, it can be concluded that the imbalanced data caused low accuracy and incorrectly class classified. The data used in this study is largely dominated by one class of the samples. As a result, due to the number of non-chronic diseases data that are largely dominated than the chronic diseases, there are younger patients with chronic diseases who were incorrectly classified as non-chronic. Thus, it is important to make sure that the proportion of the sample is balanced before performing data mining. It does not have to be an equal number but adequate enough for the machine, to learn how to produce the classification for each class.

## 5. Bibliography

[1] Searchdatamanagement.techtarget.com,. (2013). *The importance of managing data assets*. Retrieved 15 December 2014, from http://searchdatamanagement.techtarget.com/feature/The-importance-of-managing-data-assets

[2] Larose, D. T. (2005). *Discovering Knowledge in Data, An Introduction to Data Mining* (1st ed., pp. 1–26). Hoboken, N.J.: Wiley-Interscience. doi:10.1002/0471687545.ch1

[3] 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

[4] Haibo He, & Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions On Knowledge And Data Engineering*, *21*(9), 1263-1284. http://dx.doi.org/10.1109/tkde.2008.239

[5] 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

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