# Comparison Between WEKA and Salford Systemin Data Mining Software

Wirda Fitriani[1],Andysah Putera Utama Siahaan[2]
[1,2]*Faculty of Computer Science*
[1,2]*Universitas Pembangunan Panca Budi*
[1,2]*Jl. Jend. Gatot Subroto Km. 4,5 Sei Sikambing, 20122, Medan, Sumatera Utara, Indonesia*

***Abstract***

*Processing of the data generated from transactions that occur every day which resulted in nearly thousands of data per day requires software capable of enabling users to conduct a search of the necessary data. Data mining becomes a solution for the problem. To that end, many large industries began creating software that can perform data processing. Due to the high cost to obtain data mining software that comes from the big industry, then eventually some communities such as universities eventually provide convenience for users who want just to learn or to deepen the data mining to create software based on open source. Meanwhile, many commercial vendors market their products respectively. WEKA and Salford System are both of data mining software. They have the advantages and the disadvantages. This study is to compare them by using several attributes. The users can select which software is more suitable for their daily activities.*

**Keywords –***Data Mining, Decision Tree, Software*

## I. INTRODUCTION

Data is one of the valuable assets. The input data is a set of sequences, called datasequences [5]. To be able to get useful information, then the required data to produce good quality reliable information and real time. The excellent information will provide useful results for the user information. It is undeniable that in the current era of globalization, the transactions that occur each day will produce data whose numbers are also very much [1, 8]. Moreover, such data should be processed for any interests. Imagine if the data is processed manually, it takes a long time to finish. The system must handle thousands of data every day. It also requires a proper machine to process thousands of data stored in the database so that the information received fast, accurate, reliable and of course available when needed [8]. Because every day, the collected information in a huge number, are the result of some transactions that occur in next time. So it takes data mining to analyze large data sets. Moreover, of course, tools can analyze and process thousands of data per day. With the tools available for data mining, the data processing jobs will be easier to explain.

Currently, there is various software available for data mining. The open source and the commercial applications  Each of the software certainly has advantages and disadvantages, besides indeed paid apps will get better support from the provider. Even so, it does not mean open source software quality is not better than the commercial one. Moreover, the open source software will support the learning process with no charge at all. Only, when it was discovered bug during use of the software, users can not directly get improvements. The users have to wait for the open source community fix it then.In this paper, this research tries to test the contribution of each software. In some case, people selects the specific software based on some criteria. It involves some tests of the decision tree to compare which one is better for the overall process.

## II. THEORIES

### A. *Data Mining.*

Data mining is a method that uses statistical techniques, mathematic, artificial intelligence and machine learning to extract and identify useful information and knowledge. It is assembled from a variety of large databases [2]. Also, data mining is also often referred to as Knowledge Discovery from Data or KDD [1]. Relationships are sought in data mining. It could be a connection between two or more in one dimension. For example in the dimensions of the products, the linkages purchase of a product from other products. Relationships arise between two or more attributes and two or more objects [9].

### B. *C.45 Algorithm*

The decision tree is a classification and prediction method that is powerful and famous. Decision tree method changes a huge fact into a decision tree that represents the rule. Moreover, also can be explained in the form of database language such as Structured Query Language to find records in a particular category. The decision tree is also useful to explore the data, find hidden relationships between some potential input variables with a target variable.A decision tree is a structure that divides large datasets into the sets of records that are smaller by applying a set of decision rules. With each of the division series, members of the result set to be similar to each other [1].Several algorithms can be used in the formation of the decision tree, such as ID3, CART, and C4.5. The C4.5 algorithm is a development of the algorithm ID3 [4][7].

### C. *CART (Classification and Regresion Trees)*

CART is a method or algorithm of decision tree technique. CART is a nonparametric statistical method to describe the relationship between the response variable (dependent variable) with one or more predictor variables [4]. CART method was first proposed by Leo Breiman in 1984. The resulting CART decision tree is a binary tree, where each node required to have two branches. CART is recursively dividing the records into the training data into subsets that have the value of the target attribute (class) of the same. CART algorithm build a decision tree to select the most optimal branch for each node. Selection works by counting all possibilities at each variable [5].

### D. *Salford System*

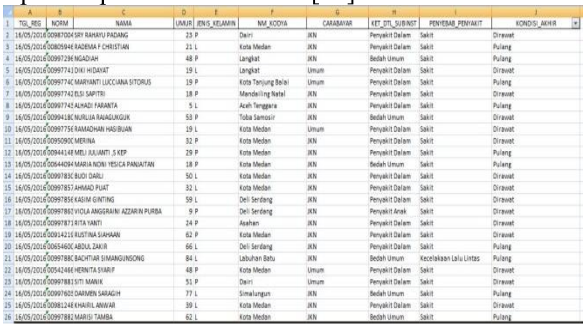
Salford Systems was founded in 1983; Salford Systems specializes in providing new-generation data mining and modeling software and consulting services selection. Applications are provided, both software and consulting range of market segmentation research, direct sales, fraud detection, credit scoring, risk management, biomedical research and the quality control of the manufacturer. Industrial use of products and consulting services from Salford Systems, including telecommunications, transportation, banking, financial services, insurance, healthcare, manufacturing, retail and catalog sales, and education. Salford Software Systems have been installed in more than 3,500 locations worldwide, including 300 major universities. The main customers include AT & T Universal Card Services, Pfizer Pharmaceuticals, General Motors, and Sears, Roebuck and Co.

### E. *WEKA*

Waikato Environment for Knowledge Analysis (WEKA) is a machine learning software written in Java are popular, developed at the University of Waikato in New Zealand [3]. WEKA is free software available under the GNU General Public License. Weka provides the use classification technique using J48 decision tree algorithm [6]. Classification techniques and algorithms used in the WEKA called classifier.

## III. PROPOSED WORK

This paper uses the patient registration data in the Emergency Room (IRD) on Pirngadi Hospital Medan, where the data is taken from patient registration on May 16, 2016. From these data, it is known that the patients admitted and registered at the IRD are not only due to illness, but there are also recorded as a traffic accident patients. Data mining softwares are suitable for this situation. It produces complete reports and decisions [10].



|  | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TGL_REG | NORM | NAMA | UMUR | JENIS_KELAMIN | NM_KODYA | CARABAYAR | KET_DTL_SUBINST | PENYEBAB_PENYAKIT | KONDISI_AKHIR |
| 2 | 16/05/2016 | 00987004 | SRY RAHAYU PADANG | 23 | P | Dairi | JKN | Penyakit Dalam | Sakit | Dirawat |
| 3 | 16/05/2016 | 0080594E | RADEMA F CHRISTIAN | 21 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Pulang |
| 4 | 16/05/2016 | 0099729E | NGADIAH | 48 | P | Langkat | JKN | Bedah Umum | Sakit | Pulang |
| 5 | 16/05/2016 | 00997741 | DIKI HIDAYAT | 19 | L | Langkat | Umum | Penyakit Dalam | Sakit | Dirawat |
| 6 | 16/05/2016 | 0099774C | MARYANTI LUCCIANA SITORUS | 19 | P | Kota Tanjung Balai | Umum | Penyakit Dalam | Sakit | Pulang |
| 7 | 16/05/2016 | 00997742 | ELSI SAPITRI | 18 | P | Mandailing Natal | JKN | Penyakit Dalam | Sakit | Dirawat |
| 8 | 16/05/2016 | 00997745 | ALHADI FARANTA | 5 | L | Aceh Tenggara | JKN | Penyakit Dalam | Sakit | Pulang |
| 9 | 16/05/2016 | 0099418C | NURLIJA RAJAGUKGUK | 53 | P | Toba Samosir | JKN | Bedah Umum | Sakit | Dirawat |
| 10 | 16/05/2016 | 0099775E | RAMADHAN HASIBUAN | 19 | L | Kota Medan | Umum | Penyakit Dalam | Sakit | Dirawat |
| 11 | 16/05/2016 | 0095090C | MERINA | 32 | P | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 12 | 16/05/2016 | 0094414E | MELI JULIANTI ,S KEP | 29 | P | Kota Medan | JKN | Penyakit Dalam | Sakit | Pulang |
| 13 | 16/05/2016 | 00644094 | MARIA NONI YESICA PANJAITAN | 18 | P | Kota Medan | JKN | Bedah Umum | Sakit | Pulang |
| 14 | 16/05/2016 | 0099783C | BUDI DARLI | 50 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 15 | 16/05/2016 | 00997857 | AHMAD PUAT | 32 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 16 | 16/05/2016 | 0099785E | KASIM GINTING | 59 | L | Deli Serdang | JKN | Penyakit Dalam | Sakit | Dirawat |
| 17 | 16/05/2016 | 00997862 | VIOLA ANGGRAINI AZZARIN PURBA | 9 | P | Deli Serdang | JKN | Penyakit Anak | Sakit | Dirawat |
| 18 | 16/05/2016 | 00997871 | RITA YANTI | 24 | P | Asahan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 19 | 16/05/2016 | 00914216 | RUSTINA SIAHAAN | 62 | P | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 20 | 16/05/2016 | 0065460C | ABDUL ZAKIR | 66 | L | Deli Serdang | JKN | Penyakit Dalam | Sakit | Pulang |
| 21 | 16/05/2016 | 0099788C | BACHTIAR SIMANGUNSONG | 84 | L | Labuhan Batu | JKN | Bedah Umum | Kecelakaan Lalu Lintas | Pulang |
| 22 | 16/05/2016 | 0054246E | HERNITA SYARIF | 48 | P | Kota Medan | Umum | Penyakit Dalam | Sakit | Dirawat |
| 23 | 16/05/2016 | 00997881 | SITI MANIK | 51 | P | Dairi | Umum | Penyakit Dalam | Sakit | Dirawat |
| 24 | 16/05/2016 | 00997605 | DARMEN SARAGIH | 77 | L | Simalungun | JKN | Bedah Umum | Sakit | Pulang |
| 25 | 16/05/2016 | 0098124E | KHAIRIL ANWAR | 39 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 26 | 16/05/2016 | 00997882 | MARISI TAMBA | 62 | L | Kota Medan | JKN | Bedah Umum | Sakit | Pulang |

**Fig. 1 The Screenshot of the Patients List**

Figure 1 shows the patients list captured by the computer. The data is a raw data before making the decision. It works on both WEKA and Salford System.

## IV. TESTING AND IMPLEMENTATION

### A. *Several Attributes*

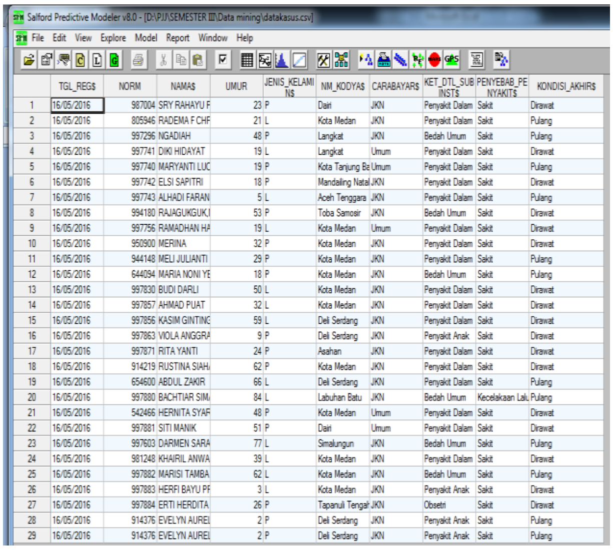
This research proposes the comparison between WEKA and Salford System. The test covers several attributes such as installation, price, configuration, interface and operation system. Table 1 illustrates several attributes that compare both software.

TABLE 1 : THE SEVERAL COMPARISONS

| Attribute | WEKA | Salford |
|---|---|---|
| Installation | Easy | Easy |
| Price | Open Source | Commercial |
| Configuration | Easy | Easy |
| Interface | Complicated | Simple |
| Operating System | Windows | Windows |
| Operation | Hard | Easy |

### B. *Evaluation of Salford System*

The test on data in Figure 1, using the Salford Predictive Modeler version 8. Data use cases will compare the attributes of the last condition, where these attributes will determine whether the patients had received treatment at IRD allowed to go straight home, or hospitalized. In Figure 2, there are ten attributes which will be compared then. Each attribute is connected to the other. The algorithm provided by the software calculates the attributes and results the decision tree.

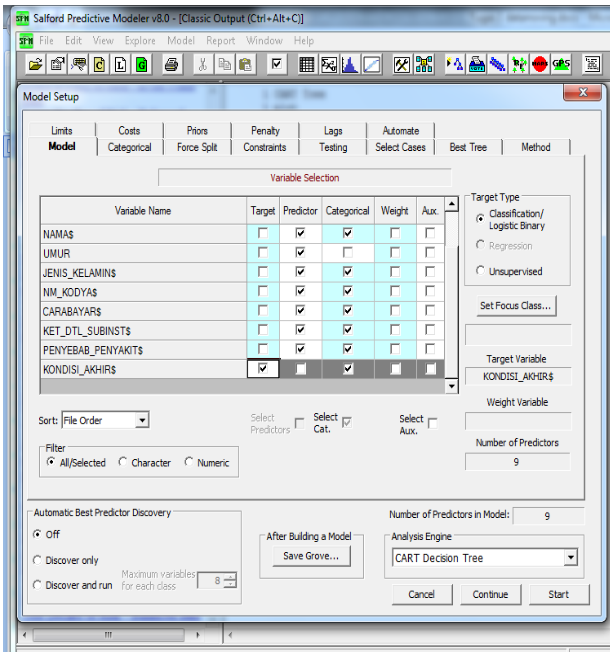


Salford Predictive Modeler v8.0 - [D:\PJJ\SEMESTER III\Data mining\datakasus.csv]
File Edit View Explore Model Report Window Help

| | TGL_REG$ | NORM | NAMA$ | UMUR | JENIS_KELAMIN$ | NM_KODYA$ | CARABAYAR$ | KET_DTL_SUBINST$ | PENYEBAB_PENYAKIT$ | KONDISI_AKHIR$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 16/05/2016 | 987004 | SRY RAHAYU F | 23 | P | Dairi | JKN | Penyakit Dalam | Sakit | Dirawat |
| 2 | 16/05/2016 | 805946 | RADEMA F CHF | 21 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Pulang |
| 3 | 16/05/2016 | 997296 | NGADIAH | 48 | P | Langkat | JKN | Bedah Umum | Sakit | Pulang |
| 4 | 16/05/2016 | 997741 | DIKI HIDAYAT | 19 | L | Langkat | Umum | Penyakit Dalam | Sakit | Dirawat |
| 5 | 16/05/2016 | 997740 | MARYANTI LUC | 19 | P | Kota Tanjung Ba | Umum | Penyakit Dalam | Sakit | Pulang |
| 6 | 16/05/2016 | 997742 | ELSI SAPITRI | 18 | P | Mandailing Natal | JKN | Penyakit Dalam | Sakit | Dirawat |
| 7 | 16/05/2016 | 997743 | ALHADI FARAN | 5 | L | Aceh Tenggara | JKN | Penyakit Dalam | Sakit | Pulang |
| 8 | 16/05/2016 | 994180 | RAJAGUKGUK,I | 53 | P | Toba Samosir | JKN | Bedah Umum | Sakit | Dirawat |
| 9 | 16/05/2016 | 997756 | RAMADHAN HA | 19 | L | Kota Medan | Umum | Penyakit Dalam | Sakit | Dirawat |
| 10 | 16/05/2016 | 950900 | MERINA | 32 | P | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 11 | 16/05/2016 | 944148 | MELI JULIANTI | 29 | P | Kota Medan | JKN | Penyakit Dalam | Sakit | Pulang |
| 12 | 16/05/2016 | 644094 | MARIA NONI YE | 18 | P | Kota Medan | JKN | Bedah Umum | Sakit | Pulang |
| 13 | 16/05/2016 | 997830 | BUDI DARLI | 50 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 14 | 16/05/2016 | 997857 | AHMAD PUAT | 32 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 15 | 16/05/2016 | 997856 | KASIM GINTING | 59 | L | Deli Serdang | JKN | Penyakit Dalam | Sakit | Dirawat |
| 16 | 16/05/2016 | 997863 | VIOLA ANGGRA | 9 | P | Deli Serdang | JKN | Penyakit Anak | Sakit | Dirawat |
| 17 | 16/05/2016 | 997871 | RITA YANTI | 24 | P | Asahan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 18 | 16/05/2016 | 914219 | RUSTINA SIAH | 62 | P | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 19 | 16/05/2016 | 654600 | ABDUL ZAKIR | 66 | L | Deli Serdang | JKN | Penyakit Dalam | Sakit | Pulang |
| 20 | 16/05/2016 | 997880 | BACHTIAR SIM | 84 | L | Labuhan Batu | JKN | Bedah Umum | Kecelakaan Lalu | Pulang |
| 21 | 16/05/2016 | 542466 | HERNITA SYAF | 48 | P | Kota Medan | Umum | Penyakit Dalam | Sakit | Dirawat |
| 22 | 16/05/2016 | 997881 | SITI MANIK | 51 | P | Dairi | Umum | Penyakit Dalam | Sakit | Dirawat |
| 23 | 16/05/2016 | 997603 | DARMEN SARA | 77 | L | Simalungun | JKN | Bedah Umum | Sakit | Pulang |
| 24 | 16/05/2016 | 981248 | KHAIRIL ANWA | 39 | L | Kota Medan | JKN | Penyakit Dalam | Sakit | Dirawat |
| 25 | 16/05/2016 | 997882 | MARISI TAMBA | 62 | L | Kota Medan | JKN | Bedah Umum | Sakit | Pulang |
| 26 | 16/05/2016 | 997883 | HERFI BAYU PF | 3 | L | Kota Medan | JKN | Penyakit Anak | Sakit | Dirawat |
| 27 | 16/05/2016 | 997884 | ERTI HERDITA | 26 | P | Tapanuli Tengah | JKN | Obsetri | Sakit | Dirawat |
| 28 | 16/05/2016 | 914376 | EVELYN AUREL | 2 | P | Deli Serdang | JKN | Penyakit Anak | Sakit | Pulang |
| 29 | 16/05/2016 | 914376 | EVELYN AUREL | 2 | P | Deli Serdang | JKN | Penyakit Anak | Sakit | Pulang |

**Fig. 2 The Attributes**

After configuring the data on the model by specifying the attributes to be compared to the target, and then determine the Class Value on Categorical and finally press the Start button to determine the results and see the tree configuration arising from the configuration. Figure 3 shows the configuration window and the attributes. These checkboxes must be selected to determine what attributes are used.



| Variable Name | Target | Predictor | Categorical | Weight | Aux. |
|---|---|---|---|---|---|
| NAMA$ | ☐ | ☑ | ☑ | ☐ | ☐ |
| UMUR | ☐ | ☑ | ☐ | ☐ | ☐ |
| JENIS_KELAMIN$ | ☐ | ☑ | ☑ | ☐ | ☐ |
| NM_KODYA$ | ☐ | ☑ | ☑ | ☐ | ☐ |
| CARABAYAR$ | ☐ | ☑ | ☑ | ☐ | ☐ |
| KET_DTL_SUBINST$ | ☐ | ☑ | ☑ | ☐ | ☐ |
| PENYEBAB_PENYAKIT$ | ☐ | ☑ | ☑ | ☐ | ☐ |
| KONDISI_AKHIR$ | ☑ | ☐ | ☑ | ☐ | ☐ |

Target Type: ● Classification/ Logistic Binary ○ Regression ○ Unsupervised

Target Variable: KONDISI_AKHIR$

Weight Variable:

Number of Predictors: 9

Sort: File Order

Select Predictors ☐ Select Cat. ☑ Select Aux. ☐

Filter: ● All/Selected ○ Character ○ Numeric

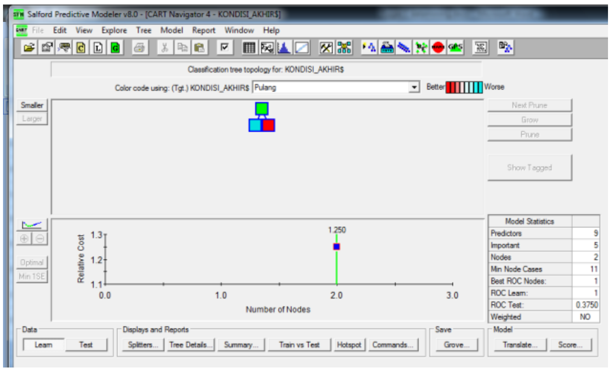
Automatic Best Predictor Discovery: ● Off ○ Discover only ○ Discover and run Maximum variables for each class 8

Number of Predictors in Model: 9

Analysis Engine: CART Decision Tree

**Fig. 3 The Attributes Configuration**

The continue buttonwill process the calculation. It produces two pieces of nodes on the tree using decision tree CART. The decision tree comes with several colored-nodes. There is also a chart to represent the node value and the legend of each node.

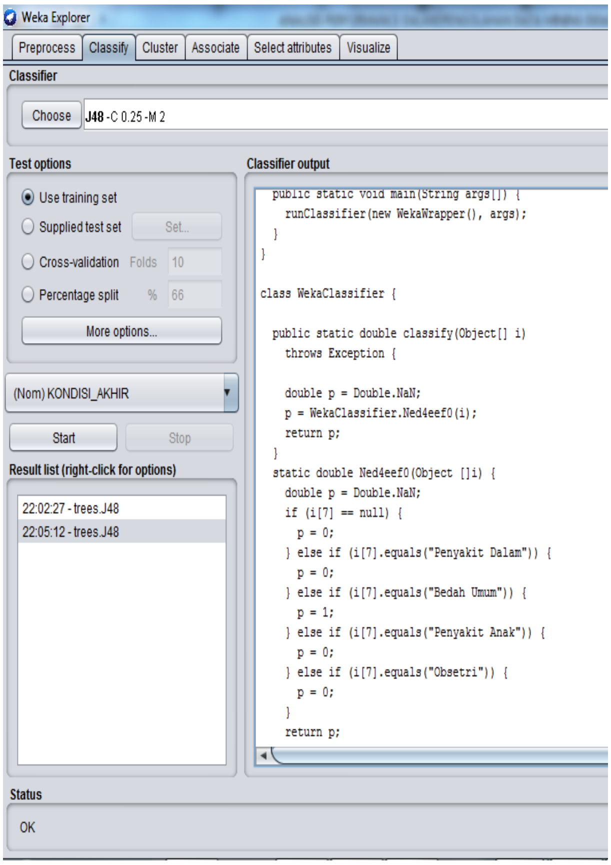


| Model Statistics | |
|---|---|
| Predictors | 9 |
| Important | 5 |
| Nodes | 2 |
| Min Node Cases | 11 |
| Best ROC Nodes: | 1 |
| ROC Learn: | 1 |
| ROC Test: | 0.3750 |
| Weighted | NO |

**Fig. 4 The Decision Tree**

### C. *Evaluation of WEKA*

With the same attributes and the same test sample data in Figure 1, the data will be tested by using Weka version 3.8.0. Data with 29 cases that used to be comparing the attributes of the last condition as well as the comparison attribute. Figure 5 shows the configuration attributes using the J48 algorithm. In open source apps, the script can be modified to improve the result better.



J48 -C 0.25 -M 2

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) KONDISI_AKHIR

22:02:27 - trees.J48
22:05:12 - trees.J48

```
  public static void main(String args[]) {
    runClassifier(new WekaWrapper(), args);
  }
}

class WekaClassifier {

  public static double classify(Object[] i)
    throws Exception {

    double p = Double.NaN;
    p = WekaClassifier.Ned4eef0(i);
    return p;
  }
  static double Ned4eef0(Object []i) {
    double p = Double.NaN;
    if (i[7] == null) {
      p = 0;
    } else if (i[7].equals("Penyakit Dalam")) {
      p = 0;
    } else if (i[7].equals("Bedah Umum")) {
      p = 1;
    } else if (i[7].equals("Penyakit Anak")) {
      p = 0;
    } else if (i[7].equals("Obsetri")) {
      p = 0;
    }
    return p;
```

Status: OK

**Fig. 5 The WEKA Configuration Attributes**

After configuring the data attribute case with the last condition for comparison by using the J48 decision tree, then the result output that appears with the source code. After the configuration is done, it shows the visualization tree consisting of four pieces of leaves as showed in Figure 6.

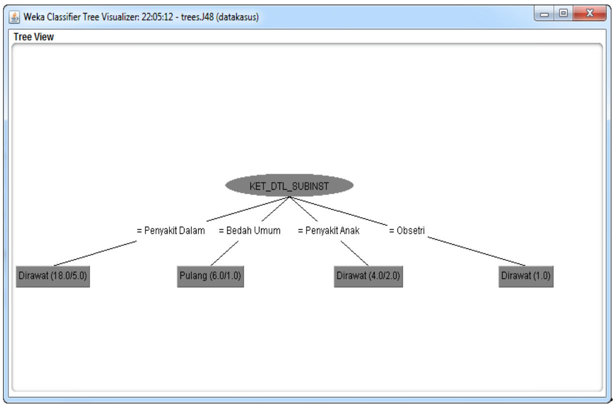


**Fig. 6 The Tree Visualizer**

## V. CONCLUSION

In the global network, data is valuable assets to maintain. Many software offers the facilities to keep data structured. From the comparison of both the software, it is concluded that the configuration data on the use of Weka software easier than on the configuration of the Salford Predictive Modeler. In Salford System Predictive Modeler, the name attribute that consists of 29 cases become the root. Meanwhile, in WEKA, the KET_DTL_SUBINST attribute becomes the root. The tree structure looks clearer than in Salford System. Use of Salford Predictive Modeler data processing cases more detail above. From this comparison, the solution of the problem more complex the better use of Salford Systems. For the case of small and simple, the use of the software is more focused on Weka.

## REFERENCES

[1] M. J. Berry, G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, New York: John Wiley & Sons, Inc, 1997.

[2] A. Kumar, O. Singh, V. Rishiwal, R. K. Dwivedi, R. Kumar, “Association Rule Mining On Web Logs For Extracting Interesting Patterns Through Weka Tool,” International Journal of Advanced Technology In Engineering And Science, vol. 3, no. 1, pp. 134-140, 2015.

[3] D. T. Larose, Data Mining Methods and Models, Canada: A John Wiley & Sons, Inc, 2006.

[4] C. D., Discovering Knowledge in Data: An Introduction to Data Mining, Canada: John Wiley & Sons, 2014.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, vol. 11, no. 1, pp. 10-18, 2015.

[6] T. Krishna, D. Vasumathi, “A Study of Mining Software Engineering Data and Software Testing,” Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 11, 2011.

[7] T. Silwattananusarn, A. D. KulthidaTuamsuk, “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012,” International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 5, 2012.

[8] S. Rajagopal, “Customer Data Clustering Using Data Mining Technique,” International Journal of Database Management Systems, vol. 3, no. 4, pp. 1-11, 2011.

[9] D. Tomar, S. Agarwal, “A survey on Data Mining approaches for Healthcare,” International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241-266, 2013.

[10] A. P. U. Siahaan, “Various Patterns of Data Mining Techniques,” 2011. [Online]. Available: http://www.academia.edu/download/46339017/Various_Patterns_of_Data_Mining_Techniques.doc. [Access: 10 7 2016].