SEAN INSTITUTE
Sharing Knowledge

http://infor.seaninstitute.org/index.php/infokum/index

JURNAL INFOKUM, Volume 7, No.1, Desember 2018 ISSN 2302-9706

TESTING DECISION TREE ALGORITHM USING TANAGRA APPLICATIONS

Soni Bahagia Sinaga
Pogram studi Teknik Informatika
AMIK Stiekom Sumatera Utara Indonesia, 20154
bahagiainaga86@gmail.com

Abstract

Decision tree is one algorithm that is used to classify segmentation or grouping which is predictive, Decision Tree Algorithm has the Advantages in processing numerical (continuous) and discrete data, can handle missing attribute values, produces rules that are easily interpreted and the fastest among other algorithms. Prediction accuracy is the ability of the model to be able to predict class labels against new or previously unknown data well. In terms of speed or computational time efficiency needed to create and use a model. The application used is Tanagra because the application is available for Decision tree architecture.

Keywords: Decision Tree, Data Mining, tanagra

1. Introduction

With the development of information technology today, the need for accurate information is needed in daily life, so that information will become an important element in the development of society today and the future. However, high information needs are sometimes not balanced by the presentation of adequate information, often the information must still be reviewed from very large amounts of data. The ability of information technology to collect and store various types of data far leaves the ability to analyze, summarize and extract knowledge from data. The traditional method for analyzing existing data, cannot handle large amounts of data.

Utilizing existing data in information systems to support activities taking activities, it is not enough to rely solely on operational data, we need a data analysis to explore the potential of existing information. Decision-makers try to utilize the data warehouse that is already owned to dig up useful information to help retrieve the data needed, this encourages the emergence of new branches of science to overcome the problem of extracting information or patterns that are important or interesting from large amounts of data, called data mining. The use of data mining techniques is expected to provide knowledge previously hidden in the data warehouse so that it becomes valuable information.Decision Tree Method is a method for determining the main factors based on a comparison between one factor with another. In determining a graduation level, of course various considerations need attention.

2. Literature Review

2.1 Data Mining

Data Mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically or a series of processes to explore the added value of a data set in the form of knowledge that has not been known manually. It is worth remembering that the word mining itself means an attempt to get a few valuable items from a large amount of basic material. Therefore Data Mining has long roots in the fields of science such as artificial intelligence, machine learning, statistics, and databases. Data mining is the process of applying this method to data to uncover hidden patterns. In other words Data mining is a process for extracting patterns from data. Data mining is becoming an increasingly important tool for turning that data into information. This is often used in various profile practices, such as marketing, surveillance, fraud detection, and scientific discovery. It has been used for years by businesses, scientists, and governments to filter out volumes of data such as flight passenger travel records, census data, and supermarket scanner data to produce a market research report. [1]–[3]

The main reason for using data mining is to assist in the analysis of a collection of observational behaviors. The data is susceptible to collinearity because a link is known. The inevitable fact of data mining is that the subsets of data sets analyzed may not represent the entire domain, and therefore may not contain examples of certain critical relationships and behaviors that exist in other parts of the domain. To overcome this kind of problem, analysis can be augmented using trial-based and other approaches, such as Choice Modeling for human-generated data. In this situation, inherent can be controlled correlation to, or removed altogether, during design construction. [4], [5]

2.2 Decision Tree

Decision tree is a classification method that uses tree structure representations where each node represents an attribute, its branches represent the values of the attributes, and leaves represent the class. The top node of the decision tree is called root. A decision tree is the most popular classification method used. Besides being relatively fast construction, the results of the model built are easy to understand

In the decision tree there are 3 types of nodes, namely: [6]–[8]

1. Root Node, is the top node, at this node there is no input and can have no output or have more than one output.
2. Internal Node, is a branching node, at this node there is only one input and has a minimum output of two.
3. Leaf node or terminal node is the final node, at this node, there is only one input and has no output

2.3 Algorithm C 4.5

C4.5 algorithm is one method for making decision trees based on the training data provided. Some of the developments that have been carried out at C45 include being able to overcome missing values, being able to overcome continuous data, and pruning. Here is the basic algorithm from C4.5: [9]–[12]

1. Build a decision tree from the training set
2. Pruning to simplify the tree.
3. Change the tree generated in several rules. The number of rules is equal to the number of paths that might be built from the root to the leaf node.

To select the attribute as the root, based on the highest gain value of the existing attributes. To calculate the gain the formula is used as stated in formula 1:

Gain(S,A)=Entropy(S)-$\sum_{i-}^{n} \frac{\mathrm{S_i}}{\mathrm{S}}$xEntropy($S_i$)..(1)

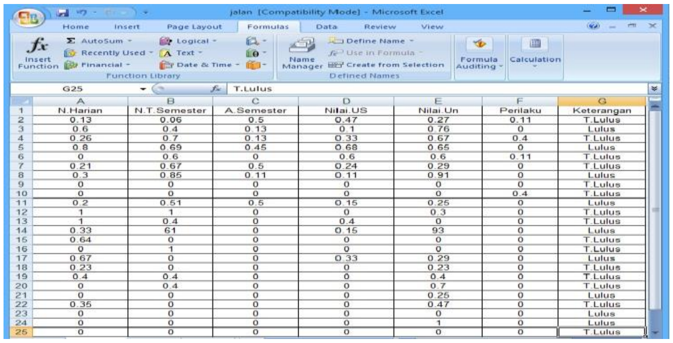
With:
S: The set of cases
A: Attributes
n: Number of attribute attributes A
[Si]: number of cases on partition i
[S]: number of S cases
While the entropy value calculation can be seen in the following formula 2:
Entropy(S) =$\sum_{i-1}^{a} -pi * log_2 pi$..(2)
With:
S: Case Set
a: Feature
I: Number of partitions S
Pi: Proportion from Si to S

3. Results and Discussion

Software testing is an investigation conducted to obtain information about the quality of the product or service being tested. Software testing also provides an objective and independent perspective on software, which is useful in business operations to understand the level of risk in its implementation. Testing techniques include, but are not limited to the process of executing a part of the program or the entire application to find errors or other defects. The testing in Tanagra 1.4 software can be done in a computer device.

1. *Infut File*

Setelah semua kebutuhan *hardware* dan so*ftwa*re dilakukan, selanjutnya melakukan penginputan nilai data siswa kedalam Microsoft Excel.



G25 T.Lulus

	A	B	C	D	E	F	G
1	N.Harian	N.T.Semester	A.Semester	Nilai.US	Nilai.Un	Perilaku	Keterangan
2	0.13	0.06	0.5	0.47	0.27	0.11	T.Lulus
3	0.6	0.4	0.13	0.1	0.76	0	Lulus
4	0.26	0.7	0.13	0.33	0.67	0.4	T.Lulus
5	0.8	0.69	0.45	0.68	0.65	0	Lulus
6	0	0.6	0	0.6	0.6	0.11	T.Lulus
7	0.21	0.67	0.5	0.24	0.29	0	T.Lulus
8	0.3	0.85	0.11	0.11	0.91	0	Lulus
9	0	0	0	0	0	0	T.Lulus
10	0	0	0	0	0	0.4	T.Lulus
11	0.2	0.51	0.5	0.15	0.25	0	Lulus
12	1	1	0	0	0.3	0	T.Lulus
13	1	0.4	0	0.4	0	0	T.Lulus
14	0.33	61	0	0.15	93	0	Lulus
15	0.64	0	0	0	0	0	T.Lulus
16	0	1	0	0	0	0	T.Lulus
17	0.67	0	0	0.33	0.29	0	Lulus
18	0.23	0	0	0	0.23	0	T.Lulus
19	0.4	0.4	0	0	0.4	0	T.Lulus
20	0	0.4	0	0	0.7	0	T.Lulus
21	0	0	0	0	0.25	0	Lulus
22	0.35	0	0	0	0.47	0	T.Lulus
23	0	0	0	0	0	0	Lulus
24	0	0	0	0	1	0	Lulus
25	0	0	0	0	0	0	T.Lulus

Figure 1 Student data in Entropy calculations

After the student data is distorted in Microsoft Excel, then saved in Text (Tab Delimited) format.

2. Tanagra Initial Display 1.4

The display of the Tanagra software that has not been inputted with data or is empty can be seen in the following image display:

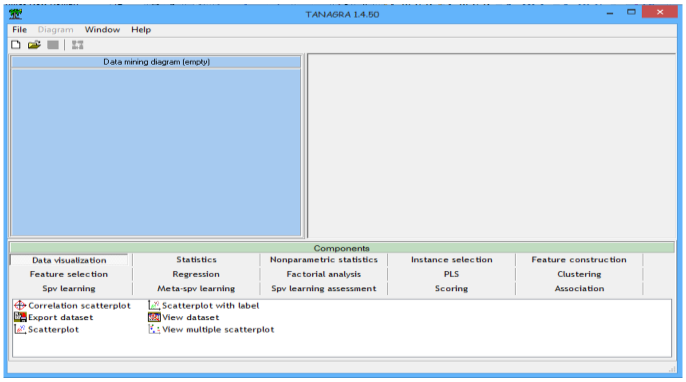
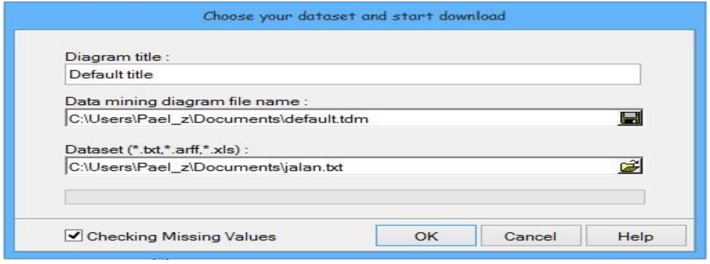


Figure 2 Tanagra Display 1.4

3. Display Download

Before the data is processed, the data must first be selected or downloaded as shown in the following image:



Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Users\Pael_z\Documents\default.tdm

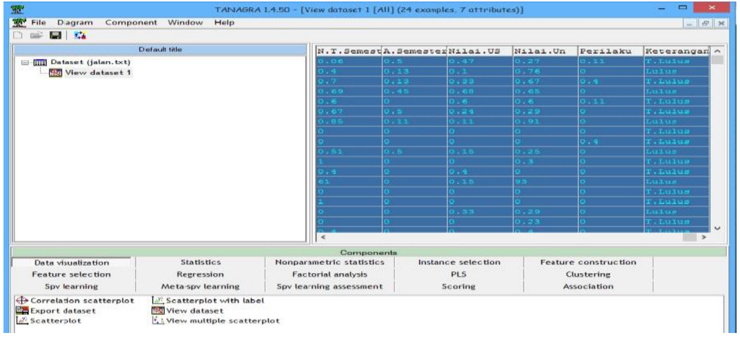
Dataset (*.txt,*.arff,*.xls) :
C:\Users\Pael_z\Documents\jalan.txt

☑ Checking Missing Values

Figure 3 Display Download

4. Display View Database

To display the data that has been selected, the Database view tool is used, while the display is as follows:

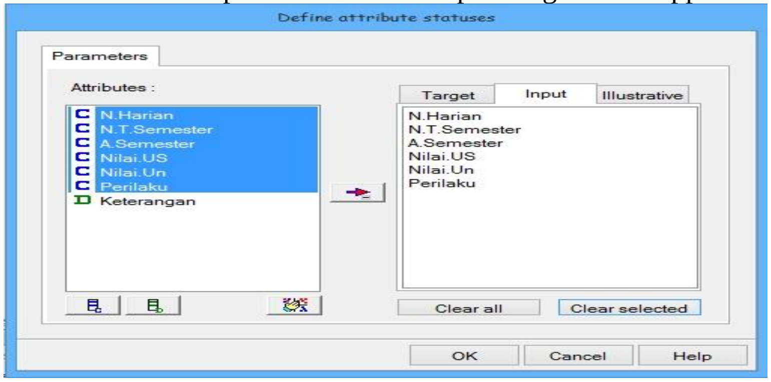


N.T.Semest	A.Semester	Nilai.US	Nilai.Un	Perilaku	Keterangan
0.06	0.5	0.47	0.27	0.11	T.Lulus
0.4	0.13	0.1	0.76	0	Lulus
0.7	0.13	0.33	0.67	0.4	T.Lulus
0.69	0.45	0.68	0.65	0	Lulus
0.6	0	0.6	0.6	0.11	T.Lulus
0.67	0.5	0.24	0.29	0	T.Lulus
0.85	0.11	0.11	0.91	0	Lulus
0	0	0	0	0	T.Lulus
0	0	0	0	0.4	T.Lulus
0.51	0.5	0.15	0.25	0	Lulus
1	0	0	0.3	0	T.Lulus
0.4	0	0.4	0	0	T.Lulus
61	0	0.15	93	0	Lulus
0	0	0	0	0	T.Lulus
1	0	0	0	0	T.Lulus
0	0	0.33	0.29	0	Lulus
0	0	0	0.23	0	T.Lulus

Figure 3 Display data

5. Display Define Attribute Statuses

Define Attribute Statuses are used to input attributes and input Targets. The appearance is as follows:



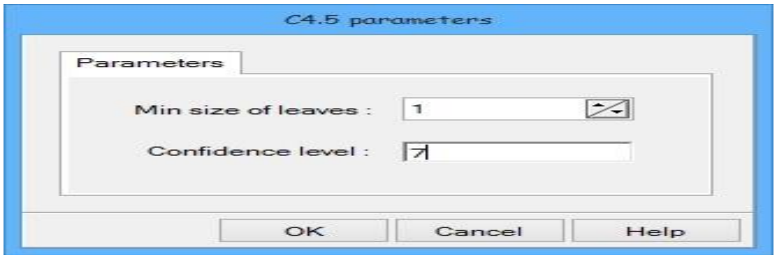
Attributes:
- C N.Harian
- C N.T.Semester
- C A.Semester
- C Nilai.US
- C Nilai.Un
- C Perilaku
- D Keterangan

Input:
- N.Harian
- N.T.Semester
- A.Semester
- Nilai.US
- Nilai.Un
- Perilaku

Figure 5 Display Define Attribute Statuses

6. Display C 4.5 Parameters

C 4.5 Parameters are used to input a minimum size of leaves and a Confidancel level, while the appearance is as follows:



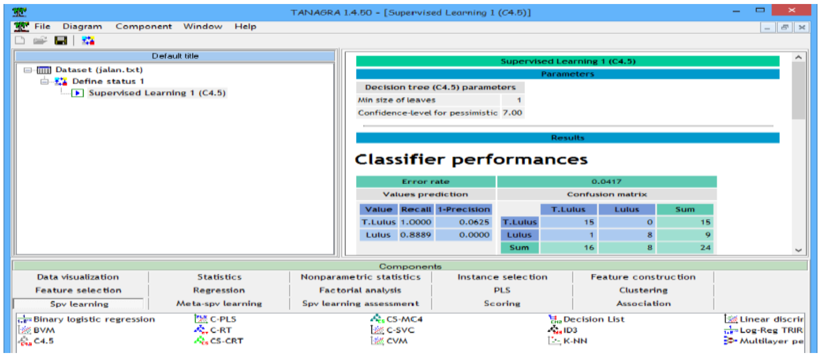
Min size of leaves : 1

Confidence level : 7

Figure 6 Display C 4.5 Parameters

7. Display Supervised Learning

Supervised Learning is the final stage to find out the Decision Tree. The appearance is as follows:



Supervised Learning 1 (C4.5)

Parameters

Decision tree (C4.5) parameters	
Min size of leaves	1
Confidence-level for pessimistic	7.00

Results

Classifier performances

Error rate: 0.0417

Values prediction

Value	Recall	1-Precision
T.Lulus	1.0000	0.0625
Lulus	0.8889	0.0000

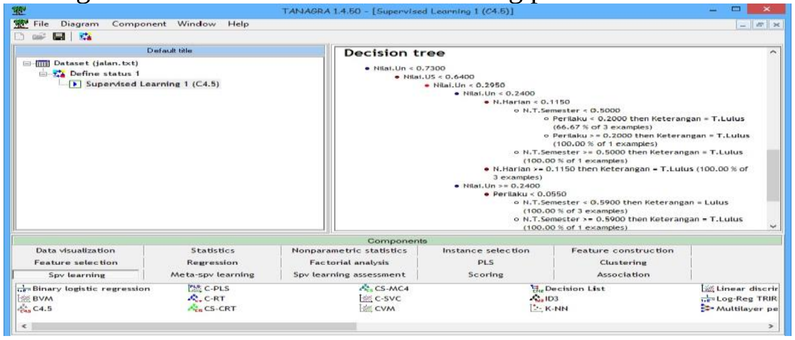
Confusion matrix

	T.Lulus	Lulus	Sum
T.Lulus	15	0	15
Lulus	1	8	9
Sum	16	8	24

Figure 7 Display of Supervised Learning

8. Display the results of testing C 4.5 algorithm

The results of testing the C 4.5 algorithm can be seen in the following figure, which is also the final result of the process in the Tanagra software. As seen in the following picture:



Decision tree

```
• Nilai.Un < 0.7300
    • Nilai.US < 0.6400
        • Nilai.Un < 0.2950
            • Nilai.Un < 0.2400
                • N.Harian < 0.1150
                    o N.T.Semester < 0.5000
                        o Perilaku < 0.2000 then Keterangan = T.Lulus
                          (66.67 % of 3 examples)
                        o Perilaku >= 0.2000 then Keterangan = T.Lulus
                          (100.00 % of 1 examples)
                    o N.T.Semester >= 0.5000 then Keterangan = T.Lulus
                      (100.00 % of 1 examples)
                • N.Harian >= 0.1150 then Keterangan = T.Lulus (100.00 % of
                  3 examples)
            • Nilai.Un >= 0.2400
                • Perilaku < 0.0550
                    o N.T.Semester < 0.5900 then Keterangan = Lulus
                      (100.00 % of 3 examples)
                    o N.T.Semester >= 0.5900 then Keterangan = T.Lulus
                      (100.00 % of 1 examples)
```

Figure 8 Display of C 4.5 Algorithm with Decision Tree

5. Conclusions

The results of data mining using the Decision Tree method is a sequence of activities that support each other in the student assessment process so that it is easier to understand by looking at the stages of the decision tree image.

Reference

[1] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Trans. Knowl. Data Eng.*, 2014.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[3] S. Agarwal, “Data mining: Data mining concepts and techniques,” in *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*, 2014.

[4] D. Mining and T. Mining, “A Programmers Guide to Data Mining,” *Text*, 2017.

[5] J. Apostolakis, “An introduction to data mining,” *Struct. Bond.*, 2010.

[6] Y. Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” *Shanghai Arch. Psychiatry*, 2015.

[7] Y. Ben-Haim and E. Tom-Tov, “A streaming parallel decision tree algorithm,” *J. Mach. Learn. Res.*, 2010.

[8] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artificial Intelligence Review*. 2013.

[9] A. Mukminin and D. Riana, “Komparasi Algoritma C4 . 5 , Naïve Bayes Dan Neural Network Untuk Klasifikasi Tanah,” *J. Inform.*, 2017.

[10] S. Hawani, A. P. Windarto, S. Solikhun, and D. Hartama, “PENERAPAN C 4.5 UNTUK

MENENTUKAN CALON SUAMI TERBAIK DALAM PERNIKAHAN PADA KANTOR KUA SIANTAR MARTOBA PEMATANGSIANTAR,” *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, 2017.

[11] E. R. Ariyanto, Wijanarto, and Sudaryanto, “Klasifikasi Citra Porno dengan Algoritma C 4.5 Berbasis Model Warna YCbCr dan Shape Detector,” *Techno.COM*, 2016.

[12] B. Santoso, “Perancangan Aplikasi Data Mining Penjualan Laptop Pada Sinergi Komputer Lubuklinggau Menggunakan Algoritma C 4.5,” *J. Ilm. Betrik*, 2017.