IT FOR SOCIETY, Vol. 08, No. 02
ISSN 2503-2224

J48 Tree Classification and Simple K-Means Clusterization on the Behavior of Academic Community in Accessing Internet for Bandwidth Management Plan

Imron Iswanto
Faculty of Computer Science
President University
Bekasi, West Java
imron@president.ac.id

Abdul Ghofir
Faculty of Computer Science
President University
Bekasi, West Java
geoff@president.ac.id

***Abstract* – The purpose of the research was to review the improvements in internet usage behavior by the academic community at a university which consisted of favorite website variables, the amount of internet bandwidth capacity used and the time spent accessing the website. The research was conducted in a population of 200 lecturers and 5267 students. Internet bandwidth mapping is carried out using two methods of the J48 tree classification algorithm and simple K-Means clustering contained in the WEKA software. The results of this study are for the purpose of mapping internet bandwidth, determining internet service provider load balancing, classification and clustering of favorite websites based on time duration which can affect the use of existing internet bandwidth capacity systems.**

***Keywords* – bandwidth, internet behavior, WEKA, J48 Tree, K-Means.**

I. INTRODUCTION

University is a large entity in which there are various units. Each unit has different business and activities. In carrying out their activities, they use both intranet and internet networks. With a very large usage capacity, they need a way that regulates how the network runs properly and smoothly so that work can be done quickly.

In general, networks at universities are used as the main media to support the operational work of lecturers, students, computer laboratories, servers, employee work operations, student dormitories and other business units. Internet is the main requirement to improve university data processing that is integrated with various existing systems.

In the research conducted, there were 200 lecturers, 5267 students, and 300 employees (educational staff). Each type of user has a different bandwidth allocation. 10 Mbps for lecturers, 15 Mbps for students and 5 Mbps for employees. The average number of internet users is 1183 users, while daily bandwidth usage is 329.3 Mbps for inbound and 355.4 Mbps for outbound.

The determination of bandwidth allocation that has been carried out still does not meet the needs, because there are still several complaints such as speed, stability, reliability when using the internet as a work support and learning resource for all academics.

II. RELATED RESEARCH

Research entitled *"Analisis Web Usage Mining Menggunakan Teknik K-Means Clustering Dan Association Rule"* authorized by Fauzanu, Darwiyanto, Wisudiawan on 2017 has discussed about how to explore and understand the behavior of visitors of website which can then be used as a reference for optimization. In this study, it was explained that one of the things that is part of the behavior of visitors to a website can be seen from the access patterns used by each individual or visitor. The log data will then go through the pre-processing stage to be carried out by clustering techniques using the K-Means algorithm and association rules. This process will produce an analysis of individual behavioral tendencies when accessing the website.

By applying the concept of web usage mining, the research will be conducted in this study presents the results of research on the similarities in the activities carried out by users in accessing e-learning websites at a university by considering the distance between existing users or Euclidean Distance. The user activity is then stored in a log file which will be processed using a clustering technique using the Hierarchical Agglomerative Clustering algorithm as a web usage mining technique. This log data will be processed by taking several pieces of data needed such as IP address, user Id, page Id and user access time.

III. RESEARCH METHODS

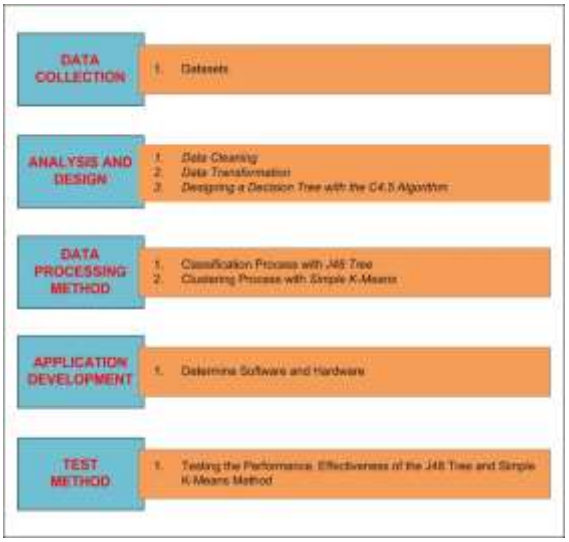

Fig 1. Research method stages

In this study, the method will be described in a research stage scheme to present a more detailed, organized and systematic guide or description. The scheme of these stages can be seen in the figure 1. It will be divided into five stages, namely data collection, analysis and design, data processing method, application development, method testing.

A. Data Collection

The data set was obtained from internet traffic data usage at the University. The data resulted can be seen in Table I.

B. Analysis and Design

Analysis and design will divided by three steps, namely data cleaning, data transformation, and designing a decision tree with the C4.5 algorithm.

1. Data Cleaning
   It is the process of detecting and repairing data that will be used as data training. The data used as elements of the decision tree model that will be designed using the C4.5 algorithm is Src Ip Address, Dst Ip Address, Website, Server Hosting, Time, Bandwidth, Website Classification, and Academic Community.
2. Data Transformation
   Based on the data training, the data will be transformed is destination IP address, website, server hosting, bandwidth, and academic community. These data are data that will be used as input attributes for designing the decision tree model and the simple k-means model. Transformation is done by making a classification on each input attribute, as displayed in Table II. It has been determined that five attributes are used as input and each attribute has a classification so that it can help the mining process become easier to determine internet bandwidth usage according to these five attributes

TABLE II
CLASSIFICATION OF INPUT ATTRIBUTE

No	Attribute	Class
1	Bandwidth	< 5120 (low)
		5120 <= Bandwidth < 10240 (Middle)
		> = 10240 (High)
2	Destination IP Address	IX (International Exchange)
		IIX (Indonesia Internet Exchange)
3	Website	Government websites
		Education websites
		Seo websites
		Streaming
		Social media
		Email
		Ecommerce
		News websites
		Another websites
4	Server Hosting	US
		INA
		SG
		CN
5	Academic community	Lecturer
		Staff
		Student

Furthermore, the five input attributes will be analyzed using a decision tree and simple k-means to determine the classification and clustering of internet bandwidth usage based on low, medium and high, according to the target attributes of the data as shown in Table III.

TABLE III
TARGET ATTRIBURE

Target Attribute	Description
Upgrade	Bandwidth usage >= 10240 Kbps
No upgrade	Bandwidth usage < 10240 Kbps

Table III shows that there are two classifications of desired target attributes, namely Upgrade and No Upgrade for campus internet bandwidth usage. The data used to design a decision tree and simple k-means are 10 input attributes and 1 target attribute. For data analysis using the C4.5 tree and simple k-means algorithm, 20 data samples will be used which will be taken randomly from the campus internet network traffic data, then the 20 data samples are transformed to produce low, medium, and high internet bandwidth usage decisions that are used by academic community for mapping plans or allocating internet bandwidth at the university as shown in Table IV.

3. Designing a Decision Tree using C4.5 Algorithm
   Before extracting data into a tree, there are several steps that need to be considered in the formation of a tree structure, which are as follows:
   - Choose the root based on the largest gain ratio.
   - Choose the internal root / root branch based on the largest gain ratio after deleting the attribute that has been selected as the root.
   - Then repeat the steps until all the attributes are calculated by the value of the gain ratio.

Based on Table IV, it is necessary to calculate entropy and gain to determine the roots of the decision tree which can help determine the bandwidth plan.

$$Entropy(S) = \sum_{i}^{c} -pi \log_2 pi \qquad (1)$$

C = Number of values on the target attribute
Pi = Number of class i samples

$$Gain(S, A) = Entropy - \sum_{i=1}^{k} \frac{|Si|}{|S|} * Entropy\ (Si) \qquad (2)$$

S = the sample data space used for training
A = Attribute
|Si| = the number of samples for the value of V
|S| = the total number of data samples

$$SplitInfo\,(S,A) = \sum_{i=1}^{k} \frac{Si}{S} \log_2 \frac{Si}{S} \quad (3)$$

$$Gain\ Ratio\ (S,A) = \frac{Gain\,(S,A)}{SplitInfo\,(S,A} \quad (4)$$

$$ACC = \frac{((TP + TN) * N}{DS} \quad (5)$$

ACC = Accuracy
TP = True Positive
TN = True Negative
N = 100 %
DS = Data Set

$$P = \frac{((TN)) * N}{FP + TN} \quad (6)$$

$$R = \frac{((TP)) * N}{FN + TN} \quad (7)$$

FP = False Positive
FN = False Negative

Furthermore, the bandwidth can be used as the root node. Based on the results of calculating the entropy and gain values, it can be illustrated that the temporary decision tree as shown in figure 2.

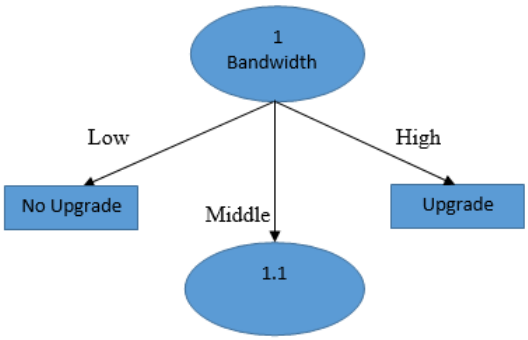

Fig 2. Decision tree node 1

Based on figure 2, there are three bandwidth attribute values, namely low, medium, and high. The low attribute on collectability classifies bandwidth as not upgrading, and the high attribute on collectability classifies bandwidth upgrade. For bandwidth that has a high value, it has classified the value into 1 decision, namely "upgrade", while for bandwidth that has a medium value, it has not classified the value into one decision, so it is necessary to carry out further entropy and gain calculation processes, as explained in Table V.

From Table V is known that the attribute with the highest gain value is the website, which is 0.200861153. Then the website becomes the next node as shown in figure 3.

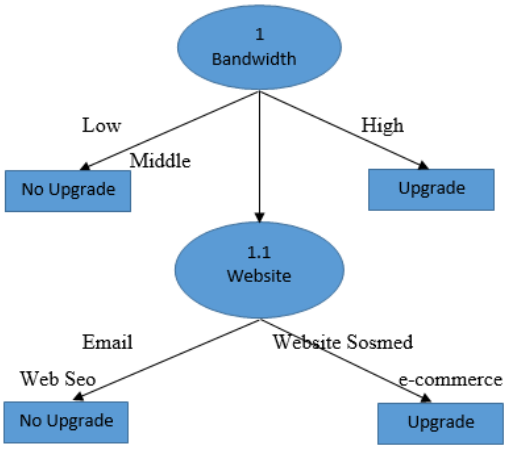

Fig 3. Decision tree as the result of node 1.1

The website attribute has 9 attribute values, namely government websites, education websites, Seo websites, streaming websites, email, e-commerce, news websites, and other websites. The attribute values of social media websites, streaming websites, and other websites have classified cases into 1 decision, namely "Upgrade". Attributes of government websites, education websites, email, e-commerce, and news websites have also classified cases into 1 decision, namely "No Upgrades". Then to determine the attributes of node 1.2, a recalculation process is carried out to find the highest gain and entropy values. The result of the calculation can be seen in Table VI.

From the calculation results in the table VI, it can be seen that the attribute that has the highest gain value is the "Academic community" attribute, with a gain value of 0.016455402. Thus the "Academic community" attribute can be used as a branch node of the "No Upgrade" attribute website. The academic community attribute has 3 attribute values, namely "Lecturer", "Student", and "Staff". Then the "Staff" attribute value has classified the case into 1 decision, namely "No Upgrade".

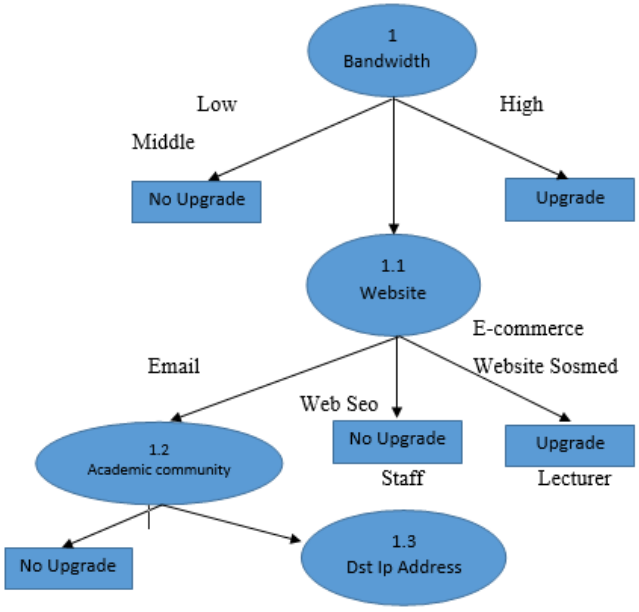

Fig 4. Decision tree as the resul of node 1.2

Furthermore, the value of the attribute "lecturer" and "student" needs to be recalculated by trying further training data. Because the value of one of the attributes "Academic

community" classifies the case as 1 and the value of the attribute "Destination IP Address" with the value of the attribute "Server Hosting" has the same gain value, resulting in entropy and gain values at node 1.2 which can be described in the decision tree as figure 4.

The confusion matrix to calculate the value of accuracy, precision, and recall for training data can be seen table VII.

TABLE VII
CONFUSION MATRIX CLASSIFICATION

Classification	**Class Prediction**	
	Class = Upgrade	**Class = No Upgrade**
Class = Upgrade	8 (True Positive – TP)	12 (False Negative – FN)
Class = No Upgrade	0 (False Positive – FP)	10 (True Negative – TN)

By using the formula for accuracy, precision, and recall, it can be calculated as follow:

$$\text{Accuracy} = \frac{(8+10)}{20} * 100\% = 90\%$$

$$\text{Precision} = \frac{(10)}{0+10} * 100\% = 100\%$$

$$\text{Recall} = \frac{(10)}{12+10} * 100\% = 45.45\%$$

Based on the above calculations, it can be concluded that the accuracy is 90%, precision is 100%, and the recall is 45.45%. From these values, it shows that the accuracy value is close to accurate, the precision value is very accurate and the recall value is included in the good category for bandwidth management plans.

C. Data Processing Method

The stages of data processing to be used are divided into two process stages:

1. Classification Process Using Tree J48
   Training data that will be used as a classification process by Weka J48 Tree is some tests such as Use Training Set, Supplied Test Set or Test Sets Provided, Cross Validation or Cross Validation, and Percentage Split in order to find out the amount of data that has been classified correctly. Classified training data includes attributes such as: Destination Ip Address classification, website classification, server hosting classification, academic community classification, Bandwidth classification, and bandwidth management classification.
2. Clustering Process Using Simple K-Means
   The data training will be clustered in this process is destination IP address, website, server hosting, academic community, bandwidth class, and bandwidth management. The data used for the processing stage includes training data and testing data

D. Application Development

At this stage, the determination of requirements such as software and hardware will be carried out to implement the system. In this discussion, the software used by the author is the Weka 3.8.2 application, and the hardware that manages internet network traffic is Cloud Core and radius manager server

E. Method Testing

The next stage is testing the method which aims to measure the accuracy of the results of each classification and clustering model that has been proposed in the Weka 3.8.2 application program

IV. RESULT AND DISCUSSION

The object of this study is to test the two methods of the J48 Tree and Simple K-Means algorithms classify and clustering the data which aims to find out the appropriate algorithm in determining the value of accuracy and prediction. Furthermore it can be used as a decision making in planning bandwidth management improvements. The training data used in this study is data from internet usage on campus in the last 1 year. The data used in this study include several variables such as: source ip address, destination ip address, website, hosting, time, bandwidth, website classification, destination ip address class, academic community, bandwidth class, and bandwidth management.

A. Research Result

Classification method is a method used to process large amounts of data which can then be grouped into smaller data. Each of these data will be processed using Weka and displayed in a description of several tables. Table VIII is the tests that have been carried out for data classification using J48 Tree Algorithm. *Supplied Test Set* has the highest accuracy value (100%), *Error Mean* value is 0.0022 and the *time* used in the test is 0.02 seconds.

TABLE VIII
THE RESULT OF THE CLASSIFICATION OF THE J48 TREE ALGORITHM

Test Option	**Correctly Classified**	**Incorrectly Classified**	**Accuracy**	**Time**	**Error Mean**
Use Training Set	1204	2	99.83%	0	0.1658
Supplied Test Set	372	0	100%	0.02	0.0022
Cross Validation	1200	6	99.50%	0	0.0704
Percentage Split	408	2	99.51%	0	0.0697

In Table IX shows a comparison of accuracy, precision and recall data values. Where testing uses the supplied test set has 100% accuracy, 1% precision and 1% recall. The supplied

test set is better than the test such as use training set, cross validation and percentage split.

TABLE IX

COMPARATIVE VALUE OF ACCURACY, PRECISION AND RECALL

	Use Training Set	**Supplied Test Set**	**Cross Validation**	**Percentage Split**
Accuracy	99.83%	100%	99.50%	99.51%
Precision	99.8%	1%	99.5%	99.5%
Recall	99.8%	1%	99.5%	99.5%

Table X is the results of clustering test using the simple k-means algorithm. It is known that supplied test set have lowest value for iteration and SSE.

TABLE X
RESULT OF SIMPLE K-MEANS ALGORITHM

Test Option	**Iteration (Time)**	**SSE (Sum of Squared Error)**	**Time**	**Clusters**
User Training Set	6	5602	0	Not Accordance
Supplied Test Set	4	1668	0	Accordance
Percentage Split	4	4413	0	Accordance
Classes to Clusters Evaluation	7	5522	0.03	Not Accordance

B. Research Analysis

Based on the result in the tables, J48 Tree Algorithm uses the *entropy*, *information gain* and *gain ratio* values where the highest value can be used as the attribute of choice which is the most appropriate or best part of the class and label determined to increase the value of *accuracy*, *precision* and *recall.* Whereas in the Simple K-Means algorithm, according to the results of the tests that have been carried out, there are cluster results that change so that the cluster results will be less than optimal, then the level of sensitivity to the data being tested can also result in an error in determining a cluster.

V. CONCLUSION

Based on the results of the research and discussion previously described, it can be concluded that behavioral data on academic community in accessing the campus internet uses J48 Tree and Simple K-Means algorithms by using several test options such as use training set, supplied test set, cross validation, percentage split and classes to clusters evaluation, resulted in the following conclusions:

1. J48 Tree Classification
   Using the *Supplied Test Set* produces data by the highest accuracy value is 100%, while the *Mean Error* value is at least 0.0022, and the time used in processing classification data is 0.02 seconds.
2. Simple K-Means Clustering
   Using the *Supplied Test Set* by the number of iterations carried out is 4 times, resulting number of SSE (Sum of Squared Error) is 1668 and the time used to create the cluster model is 0 seconds.

So by using J48 Tree Classification algorithm method is more appropriate rather than the Simple K-Means Clustering algorithm method when processing data in this study.

REFERENCES

[1] Adam Prayogo Kuncoro, Dwi Presetyo Hutomo, Muhammad Zulfadhilah "Grouping internet users using the K-Means method on server access log data" ,CITISEE 2017

[2] Adi Suryaputra, Febrillian Samopa, Bekti Cahyo Hindayanto, "Internet Traffic Clustering and Analysis Using Fuzzy C Mean With Data Feature Extraction", Informatics Journal Vol. 12, No. 1, pp 33-99, Mei 2014

[3] Fadillah Sarah, "Implementation of Data Mining to Recognice Customer Transaction Characteristic using The C4.5 Algorithm", Pelita Informatics Budi Darma Vol. 5, No. 3, 2301-9425,2013

[4] I Komang Kompyang Agus Subrata, I Made Oka Widyantara, Linawati, "Classification of the use of Communication Protocols in Network Traffic using the K-Nearest Neighbor Algorithm", Electricity Technology Vol. 16, No.1, January-April 2017

[5] Jantan, Hamidah, Razak Hamdan Abdul, Ali Othman, Zulaiha, "Human Talent Prediction in HRM using C4.5 Classification Algorithm", International Journal on Computer Science and Engineering, Vol.02, No.08, 2526-2534,2010

[6] J. Zhang, C. Cheng, Y. Xiang, W. Zhou and Y. Xiang, "Internet Traffic Classification by Agregating Correlated Naïve Bayes Predictions", IEEI Trans, Inf. Forensics Secure., Vol.8 No.1, pp 5-15, January 2013.

[7] Lerrisa Navia Rani, "Classification of Customers using the C4.5 Algorithm as the Basis for Granting Credit", Journal of INOVTEK POLBENG – INFORMATICS SERIES

[8] Rahmani, Md. Khalid ImamRahmani, Naina Pal, Kamiya Arora, "Clustering of Image Data Using K-Means and Fuzzy", Int J Adv. Comput, Sci. Appl., Vol.5,No.7, pp 160-163, 2015

[9] Rani Wahyu Ningrat, Budi Santosa, "Selection of a Nutritional Diet for Hypertension Sufferers using the Decision Tree Classification Method (Case Study of RSUD Syarifah Ambami Rato Ebu Bangkalan)", ITS Technic Journal, Vol.1,No.1, A536-A539, 2012

[10] Rastri Prathivi, "Classification of Internet Traffic Data Using the Bayes Network Method (Case Study of the University of Semarang Internet Network)", Informatics Journal, Vol.12,No.2, January 2015

[11] R.S.Anu Gowsalva, S. Miruna Joe Amali, "Naïve Bayes Based Network Traffic Classification using Corellation Information", IJARCSSE Vol.4, Issue ty, pp 8-14, March 2014

[12] Riadi, 2016, A Iswardani and I . Riadi, "Denial of Service Log Analysis Using Density K-Means Method", Vol.83, No.2, pp. 299-302, 2016

[13] Rismayanti, "Decision Tree Determining the Study Periode of Informatics Engineering Study Program Students (Case Study of the Faculty of Engineering and Computers, Harapan University, Medan)", Information System Journal, Vol.2, No.1, April 2018

[14] Swastina, Liliana, "Application of the C4.5 Algorithm for Determining Student study programs", GEMA AKTUALITA Journal, No.1, Vol.2, 93-98,201

TABLE I
DATASET

No	Source Ip Address	Destination Ip Address	Download	Dimensions	Upload	Dimension	Times
1	10.10.0.20	74.125.24.154	2	Kbps	2,1	Kbps	30
2	10.10.0.20	216.58.205.238	10	Mbps	6,9	Mbps	90
3	10.10.0.20	157.240.7.20	8	Mbps	779	Kbps	120
4	10.10.0.20	172.217.23.169	221	Kbps	2,9	Mbps	60
5	10.10.0.20	103.6.117.3	161	Kbps	1,8	Mbps	30
6	10.10.0.91	115.124.93.136	15,5	Mbps	441,1	Kbps	360
7	10.10.0.91	74.125.24.154	216	Kbps	6,9	Kbps	30
8	10.10.0.125	157.240.7.20	4,1	Mbps	9,4	Kbps	36
9	10.10.0.125	54.89.31.191	5	Mbps	823	Kbps	90
10	10.10.0.125	47.89.94.5	2	Mbps	620	Kbps	60
11	10.10.0.125	104.24.97.3	4	Mbps	309	kbps	60
12	10.10.0.125	104.20.19.254	47,7	kbps	36,4	kbps	30
13	10.10.0.125	169.55.60.170	1641	kbps	49	kbps	18
14	10.10.0.125	151.101.194.152	2,5	Mbps	96	kbps	25
15	10.10.0.125	13.107.21.200	6,3	kbps	3,1	kbps	11
16	10.10.0.125	140.82.118.3	1108	kbps	20	kbps	35
17	10.10.0.128	172.217.194.119	343,9	kbps	15,9	kbps	22
18	10.10.0.128	43.240.231.13	634,7	kbps	41	kbps	22
19	10.10.0.189	117.18.232.102	645,8	kbps	33	kbps	29
20	10.10.0.189	104.244.42.2	158,4	kbps	26,9	kbps	31
21	10.10.0.237	157.240.7.20	2,7	kbps	5,3	kbps	33
22	10.10.0.237	74.125.24.154	635	kbps	19	kbps	15
23	10.10.0.237	65.55.253.93	484	kbps	10	kbps	21
24	10.10.0.237	151.101.192.84	685	kbps	12	kbps	42
25	10.10.0.237	108.174.10.19	2,3	Mbps	69	kbps	60
26	10.10.0.252	43.240.231.13	5,7	Mbps	278,8	kbps	180
27	10.10.0.252	157.240.7.20	42,2	kbps	18,6	kbps	11
28	10.10.0.252	13.107.246.10	1287	kbps	30	kbps	22
29	10.10.0.252	104.27.155.237	479	kbps	11	kbps	37
30	10.10.3.44	172.217.194.119	25,2	kbps	20,1	kbps	26

TABEL IV
DATA TRANSFORMATION OF INTERNET BANDWIDTH USAGE IN UNIVERSITY

Src Address	Dst Address	Website	Server Hosting	Time	Bandwidth	Website Classification	Dst Address Class	Academic Community	Bandwidth Class	Bandwidth Management
10.10.0.20	74.125.24.154	www.google.com	US	30	61,5	Web SEO	IX	Student	Low	No Upgrade
10.10.0.20	216.58.205.238	www.youtube.com	US	90	3555	Web Steaming	IX	Student	Low	No Upgrade
10.10.0.20	157.240.7.20	www.facebook.com	US	120	47220	Web SosMed	IX	Student	High	Upgrade
10.10.0.20	172.217.23.169	www.Blogspot.com	US	60	6717	Another web	IX	Student	Sedang	No Upgrade
10.10.0.20	103.6.117.3	www.Kaskus.co.id	INA	30	2442	Another web	IIX	Student	Low	No Upgrade
10.10.0.91	115.124.93.136	www.tachyon.net.id	INA	360	82188	Another web	IIX	Lecturer	High	Upgrade
10.10.0.91	74.125.24.154	www.google.com	US	30	3343,5	Web SEO	IX	Lecturer	Low	No Upgrade
10.10.0.125	157.240.7.20	www.facebook.com	US	36	243	Web SosMed	IX	Student	Low	No Upgrade
10.10.0.125	54.89.31.191	www.Instagram.com	US	90	37260	Web SosMed	IX	Student	Tinggi	Upgrade
10.10.0.125	47.89.94.5	www.Lazada.co.id	US	60	18660	Ecommerce	IX	Student	Tinggi	Upgrade
10.10.0.125	104.24.97.3	www.Layarkaca21.tv	US	60	9390	Web Streaming	IX	Student	Sedang	Upgrade
10.10.0.125	104.20.19.254	www.Adf.ly	US	30	1261,5	Another web	IX	Student	Low	No Upgrade
10.10.0.125	169.55.60.170	www.Whatsapp.com	US	18	15210	Web SosMed	IX	Student	Tinggi	Upgrade
10.10.0.125	151.101.194.152	www.Scribd.com	US	25	1231,25	Another web	IX	Student	Low	No Upgrade
10.10.0.125	13.107.21.200	www.Bing.com	US	11	51,7	Web Seo	IX	Student	Low	No Upgrade
10.10.0.125	140.82.118.3	www.Github.com	US	35	19740	Another web	IX	Student	Tinggi	Upgrade
10.10.0.128	172.217.194.119	www.gmail.com	US	22	3957,8	Web Email	IX	Lecturer	Low	No Upgrade
10.10.0.128	43.240.231.13	www.sumberdata.co.id	INA	22	7432,7	Another web	IIX	Lecturer	Sedang	No Upgrade
10.10.0.189	103.117.82.24	www.bukalapak.com	INA	29	9842,6	Another web	IIX	Staff	Sedang	Upgrade
10.10.0.189	104.244.42.2	www.twitter.com	US	31	2872,15	Web SosMed	IX	Staff	Low	No Upgrade

TABLE V

RESULT OF GAIN AND ENTROPY CALCULATION

	Class	Data Training	Upgrade	No Upgrade	Entropy	Information Gain	Split Info	Gain Ratio
TOTAL		20	8	12	0,970950594			
Bandwidth						0,770950594	1,485475297	**0,518992538**
	Low	10		10	0			
	Middle	4	2	2	1			
	High	6	6		0			
Dst Ip Address						0,007403392	0,721928095	0,010255027
	IX	16	6	10	0,954434003			
	IIX	4	2	2	1			
Website						0,465957321	2,319798101	0,200861153
	Website Goverment							
	Website Education							
	Website SEO	3		3	0			
	Website Streaming	2	1	1	0,5			
	Website Sosmed	5	3	2	0,970950594			
	Website Email	1		1	0			
	Ecommerce	2	2		0			
	Website News							
	Website others	8	3	5	0,530639062			
Server Hosting						0,007403392	0,721928095	0,010255027
	US	16	6	10	0,954434003			
	INA	4	2	2	1			
Academic Community						0,019035274	1,156779649	0,016455402
	Lecturer	4	1	3	0,811278124			
	Staff	2	1	1	1			

	Student	14	6	8	0,985228136			

TABLE VI
RESULT OF GAIN AND ENTROPY CALCULATION FOR NODE 1.1

	Class	Data Training	Upgrade	No Upgrade	Entropy	Information Gain	Split Info	Gain Ratio
TOTAL		20	8	12	0,970950594			
BANDWIDTH								
	Low	4	2	2	1			
DST IP ADDRESS						0,007403392	0,721928095	0,010255027
	IX	16	6	10	0,954434003			
	IIX	4	2	2	1			
WEBSITE						0,465957321	2,319798101	**0,200861153**
	Website Government							
	Website Education							
	Website SEO	3		3	0			
	Website Streaming	2	1	1	1			
	Website Sosmed	5	3	2	0,970950594			
	Email	1		1	0			
	Ecommerce	2	2		0			
	News							
	Website Others	8	3	5	0,530639062			
SERVER HOSTING						0,007403392	0,721928095	0,010255027
	US	16	6	10	0,954434003			
	INA	4	2	2	1			
Academic Community						0,019035274	1,156779649	0,016455402
	Lecturer	4	1	3	0,811278124			

	Staff	2	1	1	1			
	Student	14	6	8	0,985228136			