Comparison of C4.5 and Naïve Bayes Algorithm for Mustahik Classification

1st Hendra Bayu Suseno[1], 2nd Awliya Wanhari[1], 3rd Siti Ummi Masruroh[1]
{hendra.bayu@uinjkt.ac.id[1], awliya.wanhari14@mhs.uinjkt.ac.id[1], ummi.masruroh@uinjkt.ac.id[1]}

Universitas Islam Negeri Syarif Hidayatullah, Jakarta Indonesia[1]

Abstract. Zakat is one of the pillars of Islam that must be done for all Muslims. People who fulfill zakat are called muzakki and people who receive zakat are called mustahik. However there's main problems of the Amil Zakat Agency related to distribution of zakat funds which are sometimes not fulfill on the target needed. Based on these problems, this study aims to analyze the comparison of the C4.5 and Naive Bayes algorithms for the classification of mustahik determination. The results of the study concluded that it was found that the use of the C4.5 algorithm is better than the Naive Bayes algorithm, proved by the level of accuracy starting from 75% - 100%, while using the Naive Bayes algorithm the level of accuracy starts from 50% - 100% and the execution time of each algorithm was similar which is 0 s. The results of the accuracy in this research are obtained through RapidMinerStudio tools using split validation.

Keywords: Zakat, Mustahik, Muzakki, C4.5 Algorithm, Naive Bayes Algorithm, Accuracy, Execution Time, Split Validation

1 Introduction

Zakat according to language is growing, developing, fertile or increasing. People who pay zakat are called *muzakki,* and people who receive zakat are called *mustahik* . The order obliged to fulfill the zakat is regulated in the Koran surat at-Taubah ayat 60 which means, " *Indeed the zakat is only for the needy, the poor, the administrators of zakat, the converts, to (liberate) slaves, those who are in debt, for the way of Allah, and for those who are on their way as a decree that is required by Allah, and Allah is Knower, Wise.*

According to the Indonesian Zakat Outlook published by the National Amil Zakat Agency Study Center (PUSKABAZNAS) the problems and challenges to fixing national zakat are weak quality and quantity of human resource (HR) zakat, uneven performance of Zakat Management Organizations (OPZ) in all regions in Indonesia, lack of structuring of zakat systems and institutions, limited synergy, integration, and cooperation in managing zakat nationally and the lack of studies, research, and integration of national zakat data [1].

The research conducted by Ai Nur Bayinah entitled *Role of Zakat as Social Finance Catalyst to Islamic Banking and Economic Growth* (2017) shows that zakat has a significant impact on Islamic banking, so this institution will contribute to economic growth both in the short and long term , besides that zakat also has a positive impact on the economy through increasing Islamic bank financing [2].

Subsequent research conducted by Salman Ahmed Shaikh and Abdul Ghafar Ismail entitled *Role of Zakat in Sustainable Development Goals* shows that zakat can play an

ICIIS 2019, November 07-08, Jakarta, Indonesia

DOI 10.4108/eai.7-11-2019.2294560

important role in meeting the goals of sustainable development related to poverty, hunger, health and global welfare, quality education, decent work, economic growth, and income inequality [3].

Along with the advances in technology, many studies have done this in order to make it easier for these institutions to determine the appropriate classification of *mustahik*.

In the classification there are several algorithms used such as *C4.5, Naive Bayes, K-Nearest Neighbor (KNN), Artificial Neural Network (ANN*), and *Support Vector Machine (SVM)* . The most widely used algorithm is the *C4.5* algorithm and the *Naive Bayes* algorithm . The results of the classification will produce the *accuracy* value and *execution time* . However, most research only reaches the value of *accuracy* and does not reach the *execution time* value and does not use data validation. Several studies from Laurensia Maria Nindia Bernita (2017) entitled *Classification of Normal Labor or Caesar Using C4.5 Algorithm* [4] , Rizky Haqmanullah Pambudi (2018) entitled *Application of C4.5 Algorithms in Programs to Predict Middle School Student Performance* [5], Larissa Navia Rani (2016) whose *Customer Classification Uses C4.5 Algorithm As the Basics of Crediting* [6] *,* they focus on finding the *accuracy* value of the algorithms they use and not using data validation. Even if you use validation data, the test will be more accurate because the data is generated by the system automatically. Another study from Rohmanul Galby Isyroqi (2018) entitled *Decision Support System for Mustahik Distribution of Alms Funds Using the VIKOR and Entropy Methods* [7] and Dennis Oktavianto (2016), entitled *Implementation Method of Weighted Product (WP) In Determining The amount of Zakat Funds Distribution Against Mustahik* [8], they are focusing on the object under study is the amount of zakat which Y by any *mustahik mustahik* and in what priority. So with the *accuracy* value *, the execution time* and data validation will be more accurate to test the classification algorithm.

2 Similar Research

There are five studies that the author makes as a similar study in this study , namely , the *Role of Zakat as a Social Finance Catalyst to Islamic Banking and Economic Growth* . Furthermore, the second research is , *Implementation of Weighted Product Method (WP) in Determining the Amount of Distribution of Zakat Funds to Mustahik* . The third study, namely *The Comparison between Consumption and Production-based Zakat Distribution Programs for Povety Alleviation and Income Inequality Reduction* . The fourth study is *the Decision Support System for Mustahik Distribution of Alms Funds Using the VIKOR and Entropy Methods* . The fifth research is *Classification of Normal Labor or Caesar Using C4.5 Algorithm*.

3 Research Method

In this study, using a simulation method consisting of problem formulations *(Problem Formulation)*, the conceptual model *(Conceptual Model)*, input / output data *(Input / Output Data)*, modeling *(Modeling)*, simulation *(Simulation)*, verification and validation *(Verification and Validation)*, experimentation *(Experimentation)*, and output analysis *(Output Analysis)*.[9] The following is the research flow used by the author:

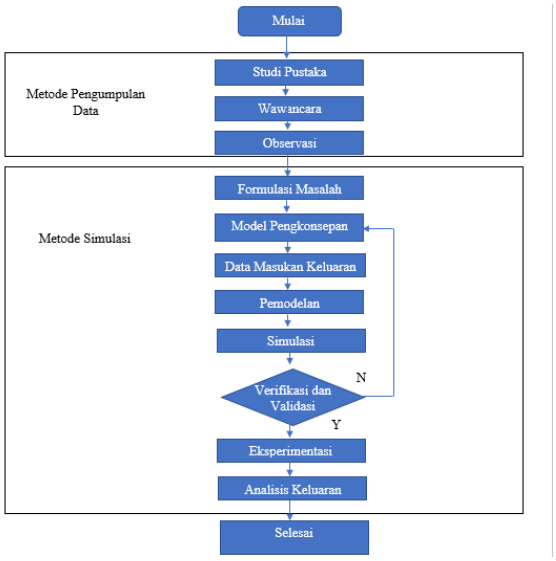

Fig. 1. Research Flow

1. Problem Formulation

After collecting data through literature studies, the author can formulate a problem that is a comparison of the *C4.5* and *Naive Bayes* algorithms for *mustahik* classification . This aims to determine the level of *accuracy* which is the highest of the two algorithms in the *Mustahik* classification.

2. Conceptual Model

At this stage, the authors make a conceptual model of *mustahik* classification using the *C4.5* and *Naive Bayes* algorithms. This conceptual model uses simulation methods with the help of *RapidMinerStudio tools* by using *split validation* , the value of the *split ratio* is 0.7 and the *tools* will produce the *accuracy* value and *execution time* . These two values will be a comparison.

3. Data Input/ Output

Input is needed on this simulation is the data *mustahiq* as *mustahik* either Y or N. *The output* obtained in this simulation is *accuracy* and *execution time*

4. Modelling

At this stage, the writer determines the scenario model that will be used in this stage of the simulation. The model is 18 scenarios. The author uses 18 different scenarios, starting with 15 to 100 data (increasing 5 data per scenario).

Table 1 Modelling

Scenario	Amount of Data	Split Ratio	Output
1	15	*0,7*	*Accuracy, Execution Time*
2	20	*0,7*	*Accuracy, Execution Time*
3	25	*0,7*	*Accuracy, Execution Time*
4	30	*0,7*	*Accuracy, Execution Time*
5	35	*0,7*	*Accuracy, Execution Time*
6	40	*0,7*	*Accuracy, Execution Time*
7	45	*0,7*	*Accuracy, Execution Time*
8	50	*0,7*	*Accuracy, Execution Time*
9	55	*0,7*	*Accuracy, Execution Time*
10	60	*0,7*	*Accuracy, Execution Time*
11	65	*0,7*	*Accuracy, Execution Time*
12	70	*0,7*	*Accuracy, Execution Time*
13	75	*0,7*	*Accuracy, Execution Time*
14	80	*0,7*	*Accuracy, Execution Time*
15	85	*0,7*	*Accuracy, Execution Time*
16	90	*0,7*	*Accuracy, Execution Time*
17	95	*0,7*	*Accuracy, Execution Time*
18	100	*0,7*	*Accuracy, Execution Time*

5. Simulation

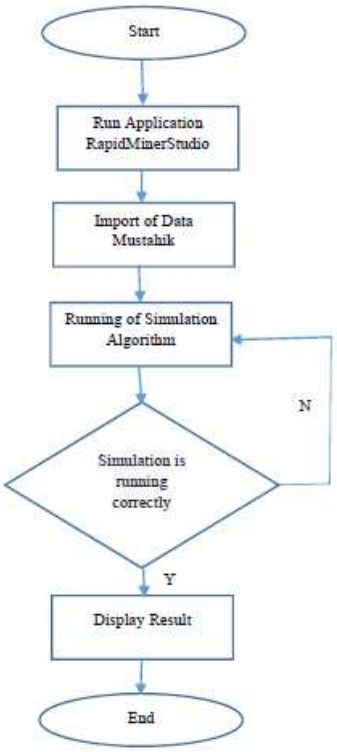

Fig. 2. Flowchart Simulation

6. Verification and Validation

In the sixth stage, the author conducted a verification and validation process of the simulations that had been carried out previously. This verification is done to ensure that there are no *bugs* or errors that occur when calculating the *C4.5,* and *Naive Bayes* Algorithms , while this validation is done to ensure the suitability of the simulation made based on the conceptual model that has been made. If it has not been fulfilled, it will return to the phasing out of the conceptual model to make a new conceptual model, and if it is fulfilled, it can proceed to the next stage.

7. Experiment

At this stage, the author conducts experiments according to the scenario model created at the modeling stage. 18 scenarios have been determined in the previous stage. Each scenario was conducted three times.

8. Ouput Analysis

In this last stage, the author analyzes the results of the *output* obtained from this simulation method. The *output* results are implemented in the form of tables and graphs. The *output* results are in the form of *accuracy* and *execution time*. *Accuracy* is the level of closeness between predictive value and actual value, while *execution time* is the amount of time needed to process data.

4 Result

The following results from the simulation that the author has done:

Table 2 Simulation Results

	Amount of Data	***C4.5* Algorithm**	***Naive Bayes* Algorithm**
Accuracy	15	75%	50%
	20	83.33%	66.67%
	25	85.71%	100%
	30	77.78%	55.56%
	35	80%	70%
	40	100%	83.33%
	45	100%	100%
	50	100%	100%
	55	100%	100%
	60	100%	100%
	65	100%	84.21%
	70	100%	100%
	75	100%	100%
	80	100%	100%
	85	100%	100%
	90	100%	100%

	95	100%	100%
	100	100%	100%
Execution Time	15	0 s	0 s
	20	0 s	0 s
	25	0 s	0 s
	30	0 s	0 s
	35	0 s	0 s
	40	0 s	0 s
	45	0 s	0 s
	50	0 s	0 s
	55	0 s	0 s
	60	0 s	0 s
	65	0 s	0 s
	70	0 s	0 s
	75	0 s	0 s
	80	0 s	0 s
	85	0 s	0 s
	90	0 s	0 s
	95	0 s	0 s
	100	0 s	0 s

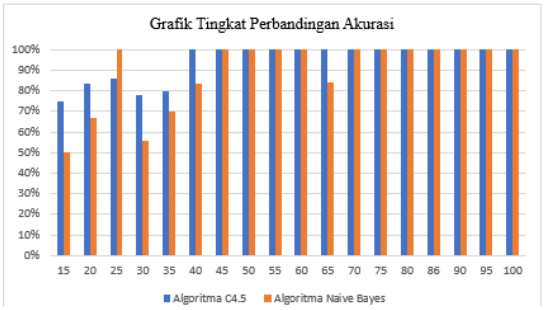

Fig.3 Comparison of *Accuracy*

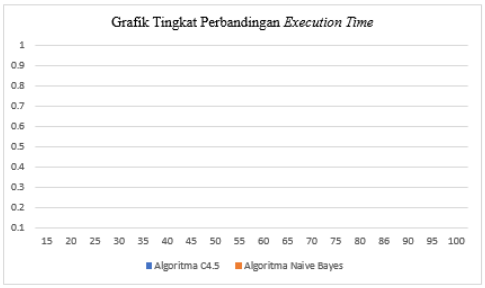

Fig 4. Comparison of *Execution Time*

The calculation steps for the two algorithms are as follows:

Table 3 Data Mustahik

No	Residence	Income	Dependent	Ameneties	Status
1	STL	A little	Ordinary	STM	Y
2	TL	Enough	Ordinary	TM	Y
3	Ordinary	A little	SB	TM	Y
4	Ordinary	Many	Ordinary	STM	N
5	Ordinary	SB	SB	Luxury	N
6	Worthy	Enough	Ordinary	TM	Y
7	TL	SB	SB	Ordinary	Y
8	Ordinary	Many	Ordinary	Ordinary	N
9	Worthy	Very A little	Ordinary	Ordinary	N
10	Ordinary	Enough	Ordinary	Ordinary	N
11	Very Worthy	A Little	Ordinary	TM	N
12	Worthy	Many	Many	SM	N
13	Worthy	Enough	Ordinary	TM	Y
14	Worthy	Enough	Many	Ordinary	N
15	TL	Many	Many	TM	Y

STL = Very Unworthy
TL = Not Eligible
SB = Very Many
STM = Very Luxury
TM = Unusual
SM = Very Luxury
Y = Accepted
N = Not

1. *C4.5* Algorithm

a. Determine the *Entropy* value :

$$Entropy\ (S) = \sum_{i=1}^{n} - p_i\ x\ log_2 p_i$$

$$Entropy\ (S) = \left(-\frac{7}{15} x\ log_2\left(\frac{7}{15}\right)\right) + \left(-\frac{8}{15}\ x\ log_2\left(\frac{8}{15}\right)\right) = 0{,}997$$

b. Determines the *Entropy* value of each attribute

It is known from table 4.2 that there are 4 attributes that make accepted or rejected as *mustahik,* among them are residence, income, dependents, and facilities. Each of these

attributes has 5 criteria each. So, we calculate *Entropy* one by one with the same formula, namely:

$$Entropy\ (S) = \sum_{i=1}^{n} - p_i\ x\ log_2 p_i$$

This is an example of calculating *Entropy* from Income attributes with a Little criterion.

$$Entropy\ (Income,\ \ Little)$$
$$= \left(-\frac{2}{3} x\ log_2\left(\frac{2}{3}\right)\right) + \left(-\frac{1}{3}\ x\ log_2\left(\frac{1}{3}\right)\right)$$
$$= 0,918$$

Next we calculate all the criteria on all the attributes in table 4.2.

c. Calculate the *gain* value for each attribut

To calculate the *Gain* value using the formula:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}\ Entropy\ (S_1)$$

With S: Set of cases, A: Attribute, n: Number of partitions attribute A, | S $_i$ | : Number of cases on partition i, and | S | : Number of cases in S.

Example of calculating the *Gain* value on the income attribute:

$$Gain\ (Income) =$$
$$0{,}997 - \left(\left(\frac{1}{15}\right) * 0\right) - \left(\left(\frac{3}{15}\right) * 0{,}918\right) -$$
$$\left(\left(\frac{5}{15}\right) * 0{,}971\right) - \left(\left(\frac{4}{15}\right) * 0{,}811\right) - \left(\left(\frac{2}{15}\right) * 1{,}000\right) = 0{,}140$$

After getting the *Gain* value , look for the highest *Gain* value . The highest *Gain* value will be *root / node 1* . Below is the final result implemented in the form of a decision tree.

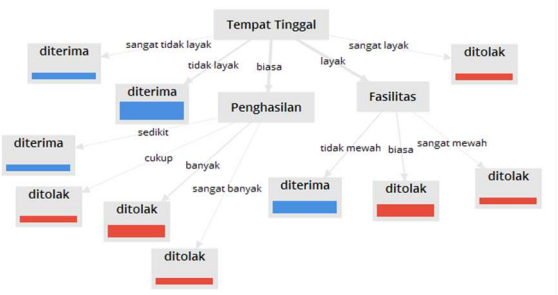

Fig. 5. Final Results of the Decision Tree

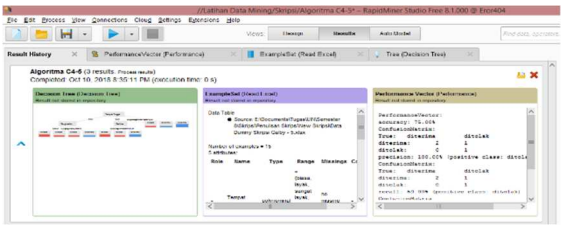

Fig. 6 *Accuracy* Value of *C4.5* Algorithm In Data 15

2. *Naive Bayes* Algorithm

Naive Bayes algorithm is one classification method that uses probability and statistical calculations. The advantage of the *Naive Bayes* algorithm is that it only requires a small amount of data to determine the estimated parameters needed in the classification process. The stages for calculating the *Naive Bayes* algorithm are as follows:

1. Calculate the number of classes / labels.
2. Count the number of cases per class.

The formulas used are:

$$P(H \mid X) = \frac{P(X \mid H)\, P(H)}{P(X)}$$

a. Calculating the Number of Classes / Labels

Known :

Amount of data (S) = 15

Amount of data received (S1) = 7

Number of data rejected (S2) = 8

P (Status = “Accepted”) = $\frac{7}{15}$ = 0,467

P (Status = “Not”) = $\frac{8}{15}$ = 0,533

b. Calculating the Number of Cases per Class

P (Income = "Very A Little" | Status = "Accepted") = $\frac{0}{7}$ = 0

P (Income = "Very A Little" | Status = "Not")= $\frac{1}{8}$ = 0,125

P (Income = "Little" | Status = "Accepted")= $\frac{2}{7}$ = 0,286

P (Income = "Little" | Status = "Not")= $\frac{1}{8}$ = 0,125

P (Income = “Enough” | Status = “Accepted”) = $\frac{3}{7}$ = 0,429

P (Income = “Enough” | Status = “Not”) = $\frac{2}{8}$ = 0,250

P (Income = “Many” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Income = “Many” | Status = “Not”) = $\frac{3}{8}$ = 0,375

P (Income = “SB” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Income = “SB” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Residence = “STM” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Residence = “STM” | Status = “Not”) = $\frac{0}{8}$ = 0

P (Residence = “TL” | Status = “Accepted”) = $\frac{3}{7}$ = 0,429

P (Residence = “TL” | Status = “Not”) = $\frac{0}{8}$ = 0

P (Residence = “Ordinary” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Residence = “Ordinary” | Status = “Not”) = $\frac{4}{8}$ = 0,500

P (Residence = “Worthy” | Status = “Accepted”) = $\frac{2}{7}$ = 0,286

P (Residence = “Worthy” | Status = “Not”) = $\frac{3}{8}$ = 0,375

P (Residence = “Very Worthy” | Status = “Accepted”) = $\frac{0}{7}$ = 0

P (Residence = “Very Worthy” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Amenities = “STM” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Amenities = “STM” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Amenities = “TM” | Status = “Accepted”) = $\frac{5}{7}$ = 0,714

P (Amenities = “TM” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Amenities = “Ordinary” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Amenities = “Ordinary” | Status = “Not”) = $\frac{4}{8}$ = 0,500

P (Amenities = “Luxury” | Status = “Accepted”) = $\frac{0}{7}$ = 0

P (Amenities = “Luxury” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Amenities = “SM” | Status = “Accepted”) = $\frac{0}{7}$ = 0

P (Amenities = “SM” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Dependent = “SB” | Status = “Accepted”) = $\frac{2}{7}$ = 0,286

P (Dependent = “SB” | Status = “Not”) = $\frac{1}{8}$ = 0,125

P (Dependent = “Many” | Status = “Accepted”) = $\frac{1}{7}$ = 0,143

P (Dependent = “Many” | Status = “Not”) = $\frac{2}{8}$ = 0,250

P (Dependent = “Ordinary” | Status = “Accepted”) = $\frac{4}{7}$ = 0,571

P (Dependent = “Ordinary” | Status = “Not”) = $\frac{5}{8}$ = 0,625

P (Dependent = “Little” | Status = “Accepted”) = $\frac{0}{7}$ = 0

P (Dependent = “Little” | Status = “Not”) = $\frac{0}{8}$ = 0

P (Dependent = “Very Little” | Status = “Accepted”)
= $\frac{0}{7}$ = 0

P (Dependent = “Very Little” | Status = “Not”) = $\frac{0}{8} = 0$

Below is a picture of the *accuracy* value of the *Naive Bayes* algorithm in 15 data.

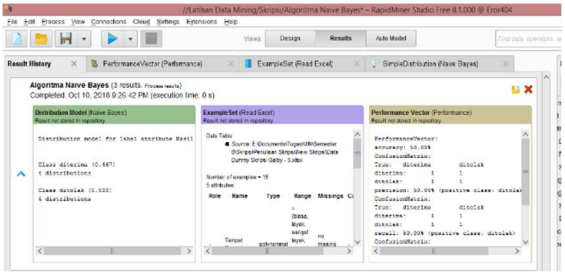

Fig. 7. *Accuracy* Value of *Naive Bayes* Algorithm At 15 data

5 Conclusion

The comparison algorithm *C4.5* and *Naive Bayes* on *mustahik* classification using simulation method which consists of eight stages, namely *the problem formulation, conceptual models, input and output of data, modeling, simulation, verification and validation, experimentation and output analysis* shows that the algorithm *C4 .5* has a high *accuracy* value compared to the *Naive Bayes* Algorithm . This is evidenced by the first scenario with the value of *accuracy* of 75% while the *Naive Bayes* algorithm has a value of *accuracy of* 50%.

Acknowledgements. This research supported by UIN Syarif Hidayatullah Conference Grant for ICONQUHAS 2019

Reference

[1] Publication and Network Division of BAZNAS Strategic Study Center (PUSKAS). *Outloook Zakat Indonesia 2017* . Central Jakarta: BAZNAS. (2017)

[2] Nur Bayinah Ai. *Role of Zakat as Social Finance Catalyst to Islamic Banking and Economic Growth* . Depok.STEI SEBI. (2017)

[3] A hmed Shaikh Salam, Ghafar Ismail Abdul. *Role of Zakat for Sustainable Development Goals* . Malaysia: Universiti Kebangsaan Malaysia. (2017)

[4] Bernita Laurensia Maria Nindia. *Normal Classification of Labor or Caesar Using C4.5 Algorithm.* Yogyakarta: Sanata Dharma University. (2017)

[5] Hakmanullah Pambudi Rizky. : *Application of the C4.5 Algorithm in the Program Predicts the Performance of Middle School Students* . Poor: Universitas Brawijaya. (2018)

[6] Navia Rani Larissa. *Customer Classification Using C4.5 Algorithm as the basis for giving credit* . Padang: Universitas Putra Indonesia Padang Computer College Foundation. (2016)

[7] Rohmanul Galby Isyroqi. *Decision Support System for Distribution of Zakat Funds to Mustahik Using the Vikor and Entropy Methods* . Jakarta: Jakarta State Islamic University. (2018)

[8] Oktavianto , Dennis. *implemented M etode Weighted Product (WP) in M enentukan B esaran P enyaluran D ana Z AKAT T erhadap M ustahik.* Jakarta: Jakarta State Islamic University. (2016)

[9] Madani, S., Kazmi, J., & Mahlknecht, S. (2011).Wireless sensor *network*s: modeling and simulation. *Intech*, (2004)