CLASSIFICATION NEEDS TEACHERS USING ALGORITHM C4.5

Siti Suhada[1)] Edi Setiawan[2)]
[1) 2)]Lecturer of Computer Science, Engineering Faculty.
Gorontalo State University
Gorontalo, Indonesia
Email: sitisuhada@gmail.com[1)] elioedi@gmail.com[2)]

Abstract - The problem in education is the lack of teachers , the teachers are not in accordance with the educational background (mismatch) , low qualifications , competence disparities , and the uneven distribution of teachers .
This study aims to (1) To design a classification algorithm model distribution needs of teachers by using C4.5 algorithm (2) the accuracy of the model for equalization algorithm C4.5 to teachers' needs .
This study is a historical research using experimental methods is to perform design and modeling systems . Data was collected through library research methods (library research) and methods of data collection (field research) and application development based on analysis of the results of data mining methods the algorithm C4.5 .
The results of this study is the classification of state information needs of teachers of subjects is more , or less enough at each school . C4.5 algorithm accuracy rate reaches 83 % .

Keywords - component ; data mining , algorithm C4.5 algorithm Introduction

I. INTRODUCTION

Development of education today has shown significant results for national development . Education is seen as one of a variety of investments that are considered crucial in improving the quality of human resources .

In the world of education , the role and function of the teacher is one very significant factor . Teachers are the most important part in the learning process , according to the Law of the Republic of Indonesia Number 14 Year 2005 on Teachers and Lecturers . Therefore , in any attempt to improve the quality of education in this country can not be separated from a variety of matters relating to the existence of teachers themselves [1] .

Subject teachers is one of the important factors in the implementation of the curriculum . However ideally supported by a curriculum without the teacher's ability to implement it , then it would not be meaningful curriculum as an educational tool , and instead of learning without curriculum as a guideline would be ineffective [2] .

On the other hand , the condition of the education world today is faced with complex problems such as the classical problem , namely the lack of teachers , the teachers are not in accordance with the educational background (mismatch) , low qualifications , competence disparities , and the distribution of teachers who are not effective . This can be evidenced by the current situation in Indonesia is still shortage of 200,000 teachers (DG PMPTK , 2010) .

Equitable distribution of teachers needs to be proven uneven at SMAN 1 Makassar for math teachers there are 8 (eight) number of teachers , while teachers are required for mathematics courses only seven (7) teachers , so the number of teachers over the needs of teachers . On the other side of the SMA Negeri 3 Lau Maros there are 5 (five) number of teachers while teachers are required for mathematics courses should be 6 (six) teachers , so the number of teachers is less than the needs of teachers .

In connection with the above conditions, the necessary scientific data mining to classify teachers' needs. Classification model that is used is to use the method of C4.5 algorithm to obtain equity classification accuracy and the model needs the right teacher.

It is hoped that the results of this study will be able to become one of the reference materials in the local government to provide policy decision-making in the distribution of teachers according to the needs of teachers in each school in order to improve the quality of education.

II. BASIS THEORY AND CONCEPTS

1 . *Basic Concept of Data Mining*

1.1 *Data Mining*

Data mining is a term used to describe the knowledge discovery in databases . Data mining is a process that uses statistical techniques , mathematics , artificial intelligence and machine learning to extract and identify useful information and relevant knowledge from a variety of large databases [3] .

Data mining is the analysis of the survey data set to discover unexpected relationships and summarize the data in a way that is different from before, which is understandable and useful to the data owner [3] .

In general, the measurement model of data mining refers to three criteria : accuracy (Accuracy) , reliability (Reliability) and usefulness (Usefulness) . Balance among

the three is necessary because not necessarily a reliable model that is accurate , reliable or accurate and is not necessarily useful [3] .

The method used in data mining is a method of learning (supervised learning) , and no learning method (unsupervised learning) . Learning methods include the role of estimation , prediction , classification and association while without learning methods include clustering [4] .

1.3 *Algorithm C4.5*

C4.5 algorithm is an algorithm that is used to form a decision tree . This algorithm is a classification and prediction methods are very powerful and famous . Decision tree is useful to explore the data , find the hidden relationship between the number of candidate input variables to the target variables . There are two variables used in determining the root of the decision tree and the entropy value gain value [3] .

In general the C4.5 algorithm to build decision tree is as follows [3] :

a. Select attributes as the root
b . Create a branch for each value
c . For cases in the branch
d . Repeat the process for each branch until all cases the branches have the same class

To select attributes as the root , based on the value of the highest gain of existing attributes .

Entropy values obtained from the formula

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

…(2.1)

Description:

S = Set Case
n = number of partitions S
pi = proportion of Si to S

Gain value is obtained from the formula:

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

.(2.2)

Description:

S = Set Case
A = Attributes
n = number of partitions attribute A
| Si | = number of cases in the i-th partition
| S | = number of cases in S

2 . *Toncept Teachers*

Teachers are professional educators with the primary task of educating , teaching , guiding , directing , training , assessing , and evaluating students in early childhood education , formal education , elementary education , and secondary education (Government Regulation No. 74 , 2008 , about the Guru , Article 1 , paragraph 1) [1] .

2.1 *Calculation Formulas High School Teacher Needs*

The calculation formula is based on the needs of high school teachers of technical instructions decree with five ministers , namely :

$$KG = \frac{JTM}{24} = \frac{(MP1 \times \sum K1 + (MP2 \times \sum K2) + (MP3 \times \sum K3)}{24} \ldots$$

…(2.3)

Description :

KG = Needs Teacher
JTM = amount per type of face-to -face teacher per week
MP = Allocation hours per week on the subject 's eyes particular subject at one level
ΣK = number of classes at a level that follows the eye certain subjects
24 = Mandatory teaching per week , use the number 24
1,2,3 = level / grade 1 , 2 and 3

2 . *Unified Modeling Language (UML)*

Modeling (modeling) is a stage in the process of designing a software before doing the encoding process (coding) . Models could be analogous to making blueprints on construction of a building . Using the model , the software development can be expected to meet all the needs of users with a complete and precise .

Unified Modeling Language (UML) is a " language " that has become the standard for visualizing, designing and documenting software systems . With UML models can be created for all types of software applications that will be able to run the application on the hardware , operating system and any network (multi- platform) and can be written in any programming language .

B. *Review of Related Research*

There is some previous research that examines teachers' use of data mining , among others :

1. Clustering teacher competence using a portfolio assessment of teacher certification with data mining methods . Ari Kurniawan , Mochamad Hariadi [5] , S2 Electrical Engineering (Telematics) , Surabaya Institute of Technology . This research applies the data mining process for the processing of teacher portfolios with K - mean clustering method to classify a relatively homogeneous competence of teachers .
2 . Mapping of Education (Education Mapping) For Improving Basic Education Services . By Surya Priadi [

6] , Yogyakarta State University published in the paper ICEMAL (International Conference Educational Management , Administration and Leadership) , 4-5 July 2012 in Malang . East Java . Indonesia . This study focused on the author develops the concept mapping study of the concept of school mapping . Various aspects of education within the scope of the school was brought to a wider sphere . Both in terms of the scope and extent of the study area are discussed . Mapping education to adopt and adapt the concept mapping of geography .

3 . Mapping the quality of teachers and education personnel based spatial . By Arif Hukmiah [7] , S2 Electrical Engineering (Computer Science) This study aimed to map the condition of teachers and education staff based on the quality and quantity spatial . The research was conducted in one district Rappocini Makassar , by taking schools kindergarten, elementary , middle and high school . The results showed that the geographic database system can load the location of the school on the map , the identity of the school, the teachers and identity in detail .

4 . Sms -based decision support system to determine the nutritional status of the k - nearest neighbor method by Ninki Hermaduanti , Sri Kusumadewi [8] . Department of Informatics , Faculty of Industrial Technology Islamic University of Indonesia . K - nearest neighbor method is a classification method that can be used to determine a person's nutritional status based on data that have been obtained previously . Sms technology can be utilized to develop the system as needed with regard to the aspects of speed and cost performance of the system reached 90.41 % .

5 . Comparison of nearest neighbor methods and algorithms to analyze the possibility of C45 resignation STMIK AMIKOM prospective students in Yogyakarta , by Kusrini , Sri Hartati , Retantyo Ward , Agus harjoko [9] of STMIK AMIKOM Yogyakarta , published in the Journal of TIE Vol . No. 10 . March 1, 2009 , ISSN : 1411-3201 . This study emphasizes the classification and comparison of the accuracy of Nearest Neighbor algorithm and C4.5 algorithm for decision making in the process likely candidate STMIK AMIKOM resigned in Yogyakarta .

6 . Insurance Customer Data classification using an algorithm C4.5 , by Sunjana [10] , Widyatama University , published in the National Seminar on Information Technology 2010 (SNATI 2010) , Yogyakarta , June 19, 2010 . This study focused on the customer grouping and grade to grade lancer lancer not use C4.5 algorithm . The results are used by insurers to predict new customers who want to join .

7 . Data Mining for customer classification with Ant Colony Optmization by Maulani Kapiudin [11] Department of Informatics, Faculty of Industrial Technology , Petra Christian University , who published the paper informatics Petra . The study focused on the potential customer classification system is designed to perform rule based on classification of extract raw data with specific criteria . Search process using the customer database of a bank with data mining techniques with ant colony optimization . Conducted an experiment with the variety and phenomone min_case_per_rule updating on a certain time period . The result is a group of customers that are based on the class rules are built with ant and modified with phenomone updating , problem areas to be widened . Provide information on the potential customers of the Bank so that it can be classified by protoptype of software .

Above studies did not cover the entire study may have been done before because of the limitations of the writer in search of the literature .

C. Conceptual Framework

To further clarify the conceptual framework will be presented, it is depicted in the following diagram:

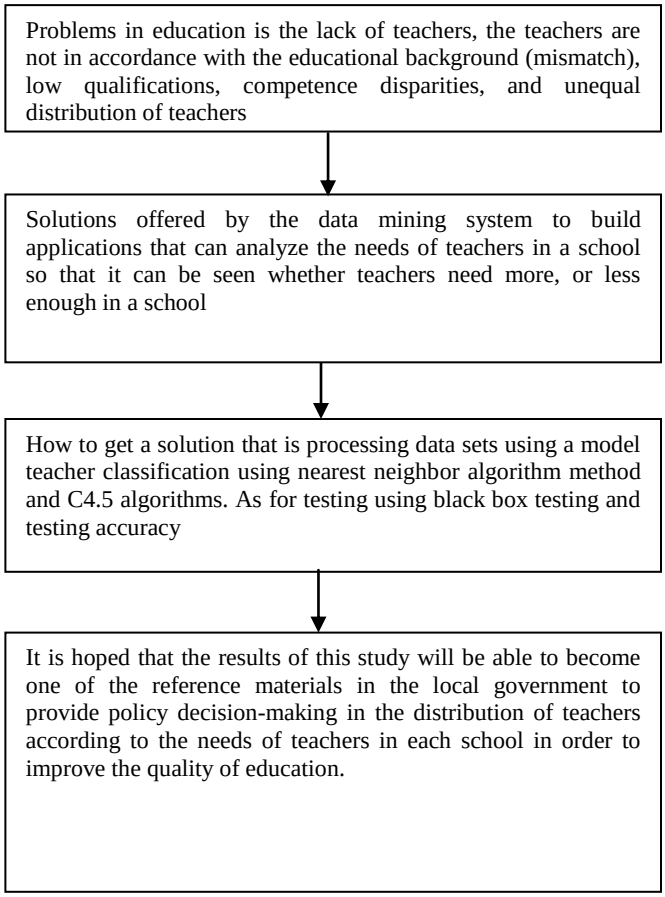

Figure 2.1 Conceptual Framework Chart

III. METHODS

A. Stages of research design

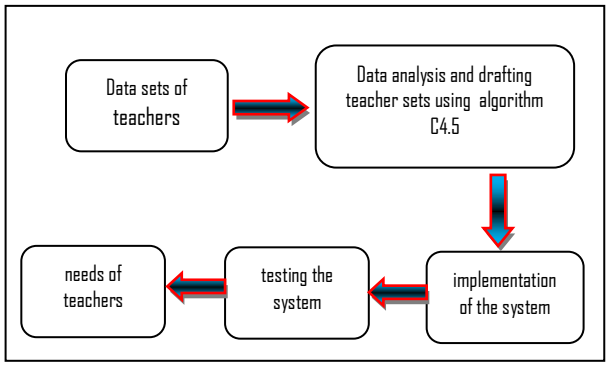

Figure 3.1 Stages of Research Design

The figure 3.1 is the design stage of a study conducted beginning stages of requirements analysis, design , system implementation and system testing .

1 . *Needs Analysis Phase*

At this stage will be the analysis and specification requirements (Requirement Analysis and Specification) for the problem to be solved . Starts by identifying the data set that includes data school teachers , teacher data , the data subject , the data in the study groups in each school .

2 . *Stage of system design*

2.1 At this stage, the design of software generally used in making system design diagram of Unified Modeling Language (UML) to describe the design of the system to be designed . Here are some diagrams that will explain the system to be designed . In this study , the Unified Modeling Language (UML) consists of a use case diagram , activity diagram and class diagram .

A. *Use case diagrams*

Use case diagrams describe the expected functionality of a system . A use case represents an interaction between the actors in the system . Use case diagram of the system is contained in Figure 3.2 below :

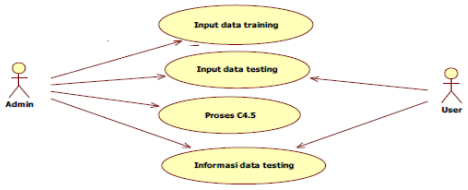

Figure 3.2 Use case diagrams

Figure 3.2 a use case diagram there are two main actors are admin and user. Both actors can perform input training data, namely input, add, change and delete data that further training will be in training. Input data testing, the testing process input data that will be processed by the algorithm C4.5.

A. *activity diagram*

Activity diagrams describe the flow of activities in a variety of systems that are being designed, how each groove begins, a decision that may occur and how they ended. Activity diagram of the system is contained in Figure 3.3 below:Gambar 3.3 *Activity* diagram

A. Class diagram

Class diagrams describe the state (attribute / property) of a system as well as offering services to manipulate the methods / functions. Class diagrams describe the structure and description of the class, along with the package and object relationships to one another. Activity diagram of the system is contained in Figure 3.4 below:

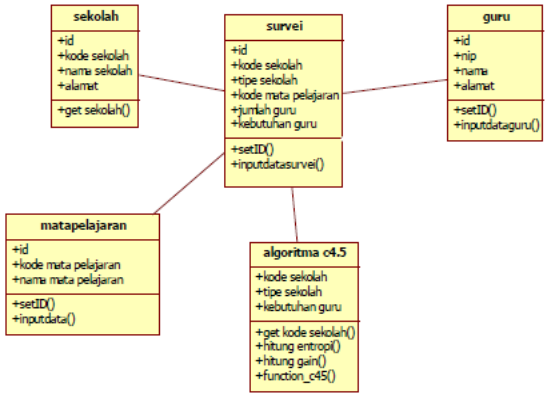

Figure 3.4 Class Diagram

2.2 *Algorithm flowchart design*

a classification algorithm C4.5processes

The steps of the process of formation of the Algoritmma C4.5 decision tree , among others , namely :

1 . Started training input data
2 . Determining Entropy corresponded to the equation 2.2 , of all cases
3 . After that , do the calculation according to the equation 2.3 Gain for each attribute .
4 . From the results of the calculation can be seen that the attribute with the highest gain which would later become the root node will form a tree .
5 . Attributes that already classify cases into one class that performed further calculations , but for the value of the class attribute is classified 2 then still need to be calculated again .
6 . From these results we can digambarkann interim decision tree .

7 . Then do the calculations again as in steps 1 through 5 , to note that all cases are entered in one class and one will form the final decision tree .. More detail can be seen in Figure 3.5 Flowchart C4.5 classification algorithm

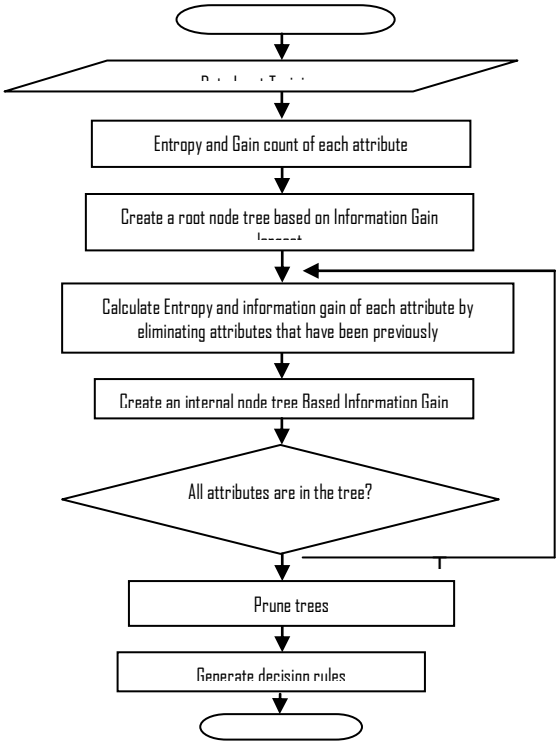

Figure 3.5 Flowchart classification algorithm. C4.5

Implementation Phase System

Implementation phase of the system or the process of coding is the process of translating from the system design .

2 . System Testing Phase

Stages of testing or evaluation system using the test methods used in this study consisted of a black box testing and testing accuracy . On functional testing will use a black box testing method , black box testing methods focus on the functional needs of the software . Therefore , black box testing method makes it possible to create a set of input conditions that will train all the functional requirements of a program .

A. *types of Research*

This study is a historical research using experimental methods is to perform design and modeling systems . Data was collected through library research methods (library research) and methods of data collection (field research) and application development based on analysis of the results of data mining methods the algorithm C4.5 .

IV. RESULTS AND DISCUSSION

A. *System Overview*

General description of the system to be developed in this research can be seen in figure 4.1 below:

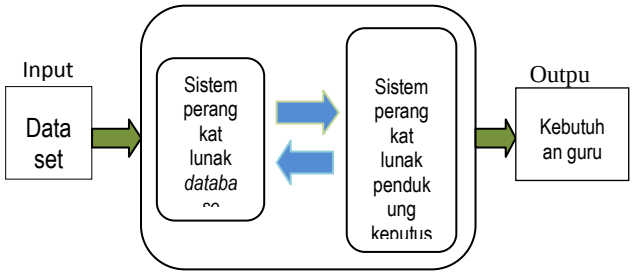

Figure 4.1 Overview of the system

The teacher sets the data will be inputted into a database software system. Teacher data that has been input and the training data will be stored in the database. Furthermore, the data is processed using a decision support system software with data mining technology using the algorithm C4.5 decision tree algorithm.

The processing of the results obtained information needs of teachers in each school as shown in figure 4.2 below:

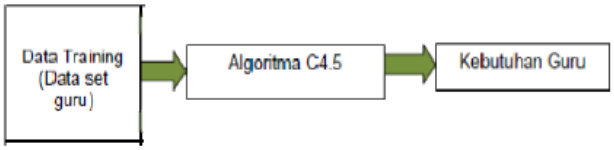

Figure 4.2 Process Data Classification

In Figure 4.2 shows that the process of data classification using the training data are processed using the C4.5 decision tree algorithm. Of these methods will result in rules needs of teachers at each school.

B. *Results Analysis and Discussion*

2. *Application of Algorithm C4.5*

In general the C4.5 algorithm to build decision trees select attributes as the root and create branches for each value. To select attributes as the root, based on the value of the highest gain of existing attributes. Formula used to calculate the gain as given in equation 1 and the calculation of entropy. The following case example in Table 4.3. output to determine the needs of teachers.

Table. 4.3 Case Needs Teacher (source: Office of National Education Makassar)

Sekolah	No	Mata Pelajaran	Kebutuhan Guru	Jumlah Guru	Output
SMU Negeri 2	1	Agama Islam	Lebih	lebih	kurang

	2	Bahasa Indonesia	Cukup	Kurang	Kurang
	3	PPKN	Cukup	kurang	cukup
	4	Bahasa Inggris	Lebih	kurang	kurang
	5	Matematika	cukup	lebih	lebih
	6	Biologi	cukup	lebih	lebih

In the case shown in Table 4.3. decision tree will be made to determine whether the needs of teachers are worth more, or less enough with the steps are as follows:

Step 1. *Select attributes as the root*

To select attributes as the root, based on the highest of the highest gain of existing attributes. Formula used to calculate the gain as shown in Equation 2.3 and the previous calculation of entropy values as given in equation 2.2.

Step 2. *Calculation of level 0*

Calculating the number of cases to need less teachers, enough and more and entropy for all cases and the cases are divided based on the attributes and needs of teachers of teachers. Gain After that do the calculations for each attribute. Calculation results are shown in table 4.4 calculation of level 0.

Table 4.4. Calculation of level 0

Atribut	Nilai atribut	Jumlah kasus	Jumlah lebih	Jumlah cukup	Jumlah kurang	Entropy	Gain
		6	2	1	3	0.187635764	
Kebutuhan Guru							0.520969097
	lebih	2	0	0	2	0	
	cukup	4	2	1	1	0.5	
	kurang	0	0	0	0	0	
Jumlah Guru							0.187635764
	kurang	3	0	1	2	0	
	lebih	3	2	0	1	0	

Table 4.4. it can be seen that the attribute with the highest gain is thus the Teacher Needs teacher's needs menjad roots level. There are three value attributes of the teachers need more, less and reasonably. Less the value of the attribute is absent in the case and the value 0. So just do further calculations to value more and enough atrbut, attribute values more is classified into one class so that no further peritungan done enough and attributes will be calculated. From these results, while portrayed as a decision tree in Figure 4.3.

Figure 4.3. Level 0 decision tree

Counting the number of cases to determine the needs of teachers less, and pretty and entropy of all cases and cases divided by the number of teachers that could attribute to the value of the root attribute value enough. Calculation results are in Table 4.5

Table 4.5. Calculation Level 1.1

Atribut	Nilai atribut	Jumlah kasus	Jumlah lebih	Jumlah cukup	Jumlah kurang	Entropy	Gain
Kebutuhan Guru – cukup		4	2	1	1	0.125	
Jumlah Guru							
	kurang	2	0	1	1	0	
	lebih	2	2	0	0	0	

From the results of table 4.5 calculation level 1.1 decision tree can be created images. The decision tree image can be seen in Figure 4.4 1 level decision tree.

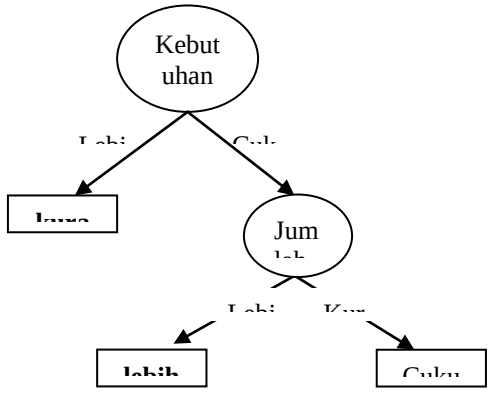

Figure 4.4 Decision tree level 1

Figure 4.4 is a final decision tree for each attribute value are classified into one class and case. Established rules of algorithm C4.5 is:

1. If the teacher needs ore then the output is less.
2. If the teacher needs sufficient and the number of teachers over the output is more.
3. If the needs of teachers and the number of teachers lacking enough then the output is sufficient.

C. *user Interface*

V.CONCLUSIONS AND RECOMMENDATIONS

By applying C4.5 algorithm on the data subject teachers can produce a pattern classification that produces uneven distribution of teachers in each subject at each school with the state allocation per subject teachers more, or less enough. It is hoped that the results of this study will be able to become one of the reference materials in the local government to provide policy decision-making in the distribution of teachers according to the needs of teachers in each school in order to improve the quality of education.

In order to get better results, suggestions can be given related to this research to further the development of neighbors is determined weight value to be calculated in order to produce a more efficient pattern classification. Using other classification algorithms such as decision tree algorithm C4.5.

REFERENCES

[1] Technical Guidance (technical guidelines) Five Joint Implementation Regulation of the Minister of Restructuring and Equity Master PNS , November 2011 .

[2] Sanjaya , Vienna . DR.M.Pd.2006 . In the Learning Competency-Based Curriculum Implementation . Jakarta : Kencana .

[3] Kusrini , Emha T. Lutfi . " Data Mining Algorithms " . Publisher ANDI , 2009

[4] Budi Santosa , " Data Mining Techniques Data Utilization for Business Purposes " . Publisher 2007.

[5] Ari Kurniawan , Mochamad Hariadi . " Clustering competence of teachers use assessment results to the teacher certification portfolio data mining methods " . Journal of Electrical Engineering (Telematics) , Surabaya Institute of Technology .

[6] Priadi Surya . " Mapping of Education (Education Mapping) For Improving Basic Education Service " . Yogyakarta State University in a paper published ICEMAL (International Conference Educational Management , Administration and Leadership) , 4-5 July 2012 .

[7] Hukmiah Arif . " Mapping the Quality of Teachers and Education Personnel -Based Spatial " . Thesis Graduate Program Hasanuddin University, Makassar . , 2011 .

[8] Kusrini , Sri Hartati , Retantyo Ward , Agus harjoko " Nearest Neighbor Comparison of methods and algorithms to analyze the possibility of C45 resignation STMIK AMIKOM prospective students in Yogyakarta " , Yogyakarta , TIE Journal Vol . No. 10 . March 1, 2009 , ISSN : 1411-3201 .

[9] Sunjana . " Customer Data Classification using an algorithm C4.5 Insurance " . SNATI 2010 Yogyakarta , June 19, 2010 .

[10] Maulani Kapiudin . " Data mining for customer classification with Ant Colony Optimization " . Department of Informatics, Faculty of Industrial Technology , Petra Christian University , who published the paper informatics Petra .

[11] Roger S. Pressman . " Software Engineering Practitioner Approach (book one) " . Publisher Andi Yogyakarta . , 2002.

[12] Yuni Sugiarti . " Analysis and Design UML (Unified Modeling Language) Generated VB.6 " . Graha Science Publishers Yogyakarta 2013.

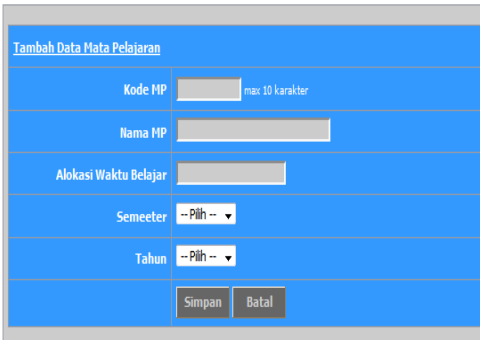

Figure 4.5 Form input on schools

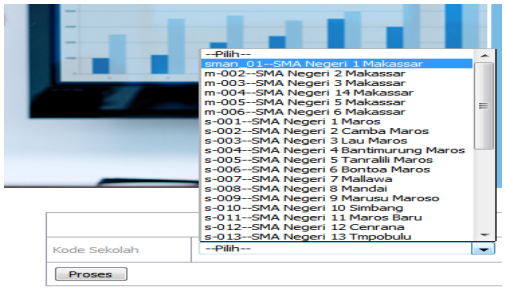

Figure 4.6 Form Input Data Subject

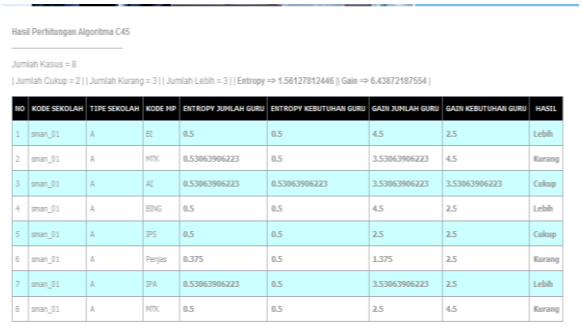

Figure 4.9 Output System algorithm C4.5