JOIN (Jurnal Online Informatika)
Volume 3 No. 1 | Juni 2018 : 22-28
DOI: 10.15575/join.v3i1.171

C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education

Erlan Darmawan
Asia E University
c701011700002@aeu.edu.my

Abstract-Data mining has background with the condition of an abundance of data (the overload data) and the explosion information faced by companies, institutions or organizations that are stored for many years. This situation is also faced in several universities that stores various kinds of data, especially new admissions database. But the abundant data has not been widely used in digging the information or knowledge that can help university management in making strategic plans. Every year there are new students who retire that do not register,therefore, it takes an application that can process a lot of data to find out the possible retirement for new students. To find out the prediction retirement prospective students, this paper uses C.45 algorithm. The method can change the a very large fact into a decision tree that represents the rule. The result of this research is tthe application can classify the new students in tree structure in order that it can produce a rule. This application is able to predict the possibility of the retirement of new student. With this application, it is expected that the possibility of a prospective student will retire from college can be known at an early stage, so the management can make a decision easily. Development of this application built uses PHP as the interface application system and MySql in database processing. System development methodology is used the waterfall model.

Keywords : C4.5 Algorithm, Retirement New Student Candidate, prediction Waterfall

I. INTRODUCTION

A. Background of Study

Often used interchangeably to explain the process of extracting hidden information in a large database[1] .Data mining is based on the abundance of data (overload of data) and explosion of information experienced by companies, institutions or organizations stored for years. The situation is also experienced by several universities that store various data on the database. However, this abundance of data has not been widely used in exploring information or knowledge that can help university leaders in making strategic plans. One of the data stored is the data of new student candidates are always increasing every year so that data accumulation occurs.

As an example of the new admissions data of a college 2010/2011 academic year the number of new students who graduated is 1548, but prospective new students who register is 1312. So there are 236 new students who resigned by not registering. There are 15.2% of potential new students who may be potentially untenable by the college. If the new student's resignation can be known earlier, then the leadership can anticipate by creating a strategic plan to retain prospective new students, given the increasing competition in the world of education. However, the problem until now, some universities do not have a standard that can be used as a tool to analyze the possibility of the resignation of new student candidates so that required a supporting system according to the research[2].

To overcome these problems, then one of the efforts that can be done, namely to make an application to perform analysis of possible resignation of prospective students by implementing data mining classification techniques in the form of decision trees. The algorithm used in making decision trees was C4.5 made by J. Ross Quinlan in 1992. This application can be used to analyze the possibility of resigning new prospective students based on pre-existing data.

Based on the above, then conducted a study entitled " C4.5 ALGORITHM APPLICATION FOR PREDICTING THE RESIGNATION OF NEW

STUDENTS CANDIDATE IN HIGHER EDUCATION".

B. The Research Questions

From the background above problems can be formulated into several problems as follows:

1. How to classify new prospective students' data to produce a decision whether the prospective student to register or not?
2. How to make an application to predict the resignation of a new students candidate?
3. How to implement C4.5 algorithm on application to predict the resignation of new students candidate?

C. The Scope of The Research

In the research undertaken determined some limitations as follows:

1. Algorithm of decision tree formation using C4.5 algorithm.
2. The application is built with structured systems development approach technique, using PHP programming language and MySQL database.
3. The output generated by this application is the pattern used to predict the resignation of new student candidates in the form of decision tree formed using C4.5 algorithm.
4. The variables used as the determinant variable in decision tree formation are gender, moving status, wave, graduate, ladder and class.
5. The data used in this study comes from the data admissions new college students who have passed the year 2010 to 2015.

D. The Objective Of Research

The purposes of this research is as follows:

1. Process the data stack to generate useful information.
2. Creating an application program predicates the resignation of new student candidates by implementing the C4.5 algorithm.
3. With this application is expected the possibility of a prospective student will resign from a college can be known early, so it can help leaders in making decisions.

II. RESEARCH METHOD

A. Implementation

According to this research in his book titled Tesaurus Bahasa Indonesia that implementation is the application, implementation, implementation, experience, manifestation, practice, engineering[3].

B. Prediction

According to Endarmoko in his book entitled The Indonesian Tesaurus that predictions are anticipated, shadows, guesses, estimates, approximate, forecasts, forecasts, projections, forecasts[3].

B. Resignation of Prospective Students New Student

According to Endarmoko in his book titled Tesaurus Bahasa Indonesia that "the withdrawal is a withdrawal. While the self is the person, the crew, the body. Candidates are aspirants, will, cadres, cadets, candidates, interns, champions[3]. New is actual, brand new (Jw), brand, warm, fresh. "Students are people who study in college [4].

So it can be concluded that the resignation of the new student candidate is a personal-withdrawal of student candidates studying at a college.

D. Data Mining

Turban in [5] suggests that:

"Data mining is a term used to describe the discovery of knowledge in a database. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases. "Garther Group in [5]suggests that:

"Data mining is a process of finding meaningful relationships, patterns, and trends by checking in a large set of data stored in storage using pattern recognition techniques such as statistical and mathematical techniques."

As stated by Larose in that: "The continuing progress in the field of data mining is driven by several factors, among others[6]:

1. Rapid growth in the data set.
2. Data storage in the data warehouse, so that all companies have access to a reliable database.
3. Increased data access via web and intranet navigation.
4. The pressure of business competition to increase market share in economic globalization.
5. The development of software technology for data mining (availability of technology).
6. Great development in computing capability and capacity building of storage media

Pramudiono in suggests that[7]:

"Data mining is not an entirely new field. One difficulty to define data mining is the fact that data mining inherits many aspects and techniques from established fields of science first. Figure 1 shows that data mining has long roots from fields of science such as artificial intellegent, mechine learning, statistics, databases, and also information retrieval. "

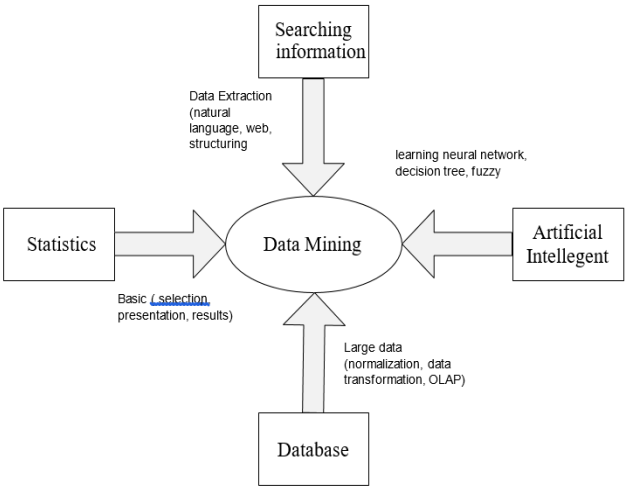

Figure 1 Field of Data Mining Science
Data Mining Algorithm [5]

E. C4.5 Algorithm

1. Decision Tree

As Kusrini states: "The decision tree is a very powerful and well known method of classification and prediction[8]. The decision method turns a very large fact into a decision tree that represents the rule. Rules can be easily understood with natural language. And they can also be expressed in database form such as Structure Query Language to search for records in certain categories.

Decision trees are also useful for exploring data, finding the hidden relationship between a number of potential input variables with target variables.

Because decision trees combine data exploration and modeling, it is very good as a first step in the modeling process even when used as a final model of some other technique. "

Berry and Linoff in [5]suggests that: "A decision tree is a structure that can be used to divide large datasets into smaller record sets by applying a set of decision rules. With each set of divisions, the members of the result set become similar to each other.

A decision tree model consists of a set of rules to divide a heterogeneous set of populations into smaller, more homogeneous ones by taking into account its destination variables.

A decision tree may be built carefully manually or it can grow automatically by applying one or more decision tree algorithms to model unclassified data sets.

Objective variables are usually grouped with definite and decision tree models more lead to the probability calculation of each record against these categories or to classify records by grouping them in one class.

Decision trees can also be used to estimate the values of the continuing variables although there are some techniques that are more appropriate for this case.

According to Larose in suggests that: "Many algorithms can be used in the formation of decision trees, such as ID3, CART, and C4.5"[6].

The C4.5 algorithm is development of the ID3 algorithm. According to Basuki & Sharif in [5] suggests that:

"Data in a decision tree is usually expressed in tabular form with attributes and records. The attribute states a parameter created as a criterion in the formation of a tree. Suppose to determine the main tennis, the criteria to note are the weather, wind, and temperature. One of the attributes is an attribute that states the data solution per data item called the target attribute. The attribute has values named with the instance. Let's say the weather attribute has instance in the form of sunny, cloudy, and rainy.The process on the decision tree is to transform the data form (table) into a tree model, change the tree model into a rule, and simplify the rule. "

2. Algorithm

As stated by Kusrini that: "In general the algorithm C4.5 to build a decision tree is as follows[8]:

a. Select attribute as root.
b. Create a branch for each value.
c. divide case in branch.
d. Repeat the process for each branch until all the cases on the branch have the same class.

To select an attribute as a root, based on the highest gain value of the attributes. To calculate the gain used the formula is given in the following equation.

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

Noted:
S : set of case A: attribute n: number of attribute partition A | Si | : the number of cases on the i-th partition
|S| : number of cases in S

Meanwhile, the calculation of entropy values can be seen in the following equation 2:

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

Noted :
S : set case A: attribute n:
number of partitions S
pi: the proportion of Si to S"

III. RESULT AND DISCUSSION

A. Flowchart C 4.5 Algoritma

The following is a flowchart of C4.5 algorithm shown in Figure 2.

ISSN 2527-1682 (Print)
ISSN 2527-9165 (Online)

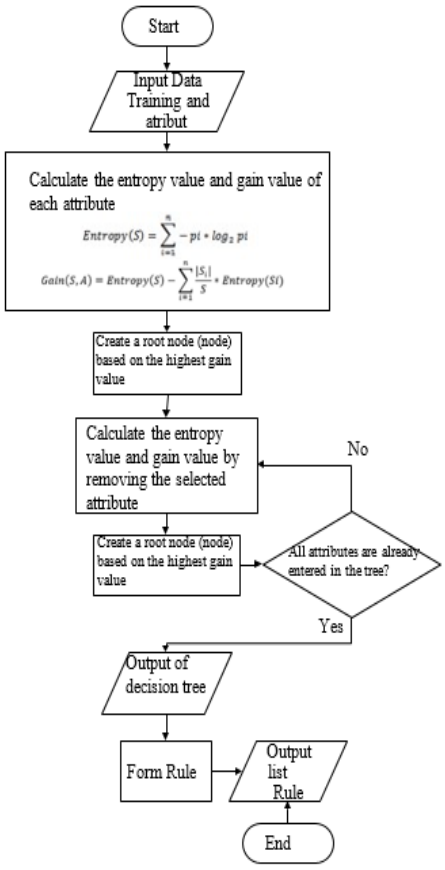

Figure 2. Flowchart Algorithm C4.5

B. Process Design

1. Context Diagram

The context diagram of the Prediction Application of New Student Withdrawal in one of the universities can be seen in Figure 3.

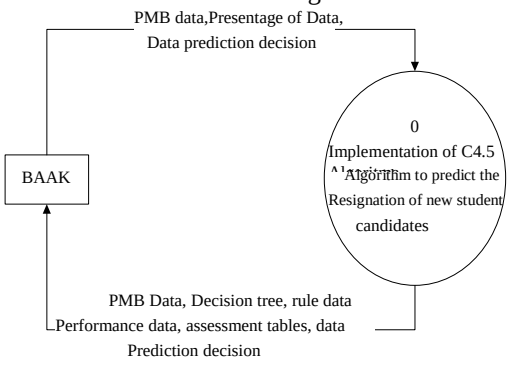

Figure 3. Context Diagram

2. DFD Level 0

Here is a Level 0 DFD from the New Students Withdrawal Prediction . Prediction Application in one of the universities can be seen in Figure 4.

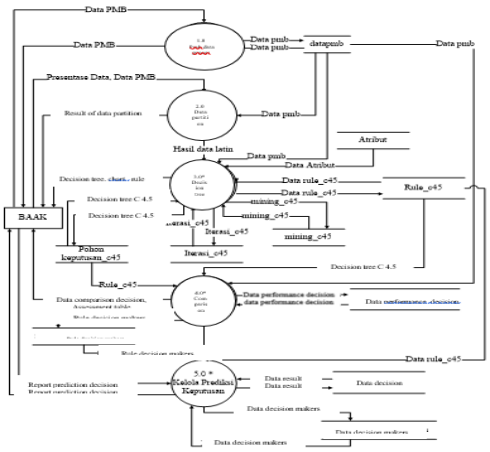

Figure 4. *DFD Level* 0

noted:

The main process that occurs in Level 0 DFD is processing the pmb data, performing data partitioning, process mining, process performance and decision making process. In addition there is one external entity, namely BAAK. Process process involves some datastore, datastore datapmb, datastore attribute, datastore iteration_c45,
datastore mining_c45, datastore
tree_decision_c45, datastore
rule_c45, datastore
data_performance_data, datastore
rule_determines, datastore
data_decision, data_preferent_decision. datastore

3. DFD Level 1

a. DFD Level 1 Process 3 Mining process is developed into several processes that can be done, namely to form a decision tree, display the rule. External entity involved is BAAK. The datastore involved are attributes, datapmb, pohon_keputusan_c45, rule_c45. Here is a DFD Level 1 Process 3 can be seen in Figure 5.

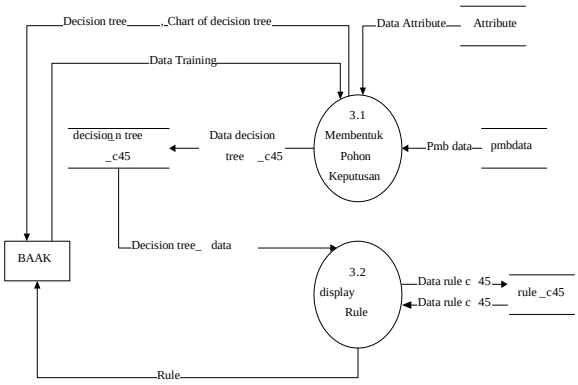

Figure 5. *DFD Level* 1 process 3

b. DFD *Level* 1 Process 4

Performance process is developed into several processes that can be done, which displays performance results and display assessment table.

External entity involved is BAAK. Attached datastore are attributes, tree_decision_c45, data_decision_performance, rule_preferent_decision, datapmb. Here is a DFD Level 1 Process 4 can be seen in Figure 6.

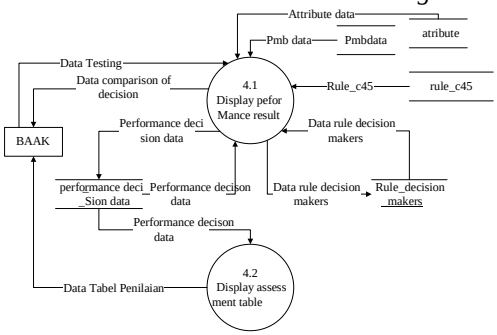

Figure 6. *DFD Level* 1 Process 4

c. *DFD Level* 1 Process 5

Decision prediction process is developed into several processes that can be done, namely to make the decision prediction process and display prediction results. External entity involved is BAAK. The involved datastore is tree_decision_c45, data_decision, data_preferent_decision, rule_preference_decision. Here is a DFD Level 1 Process 5 can be seen in Figure 7.

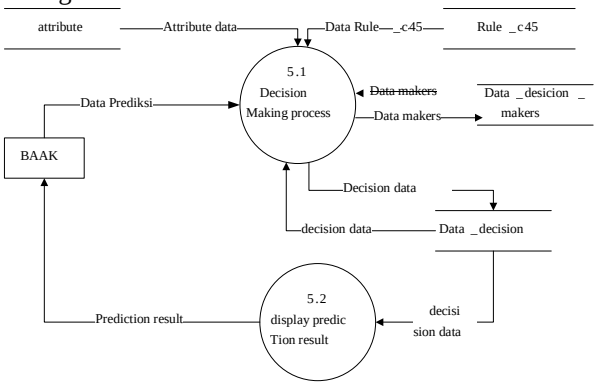

Figure 7. *DFD Level* 1 Process 5

4. DFD Level 2

Performance process is developed into several processes that can be done, which displays performance results and display assessment table. External entity involved is BAAK. The datastore involved are attributes, tree_decision_c45, data_performance_decision, rule_preference_decision, data PMB. Here is a DFD Level 1 Process 4 can be seen in Figure 8.

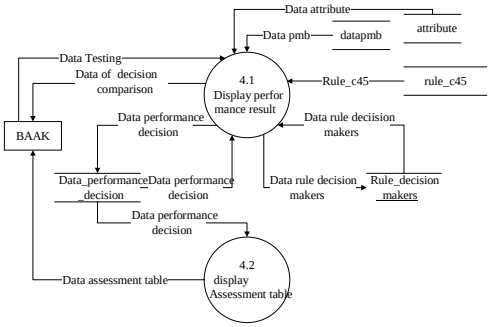

Figure 8. *DFD Level* 1 Process 4

C. Database Design

1. Entity Relationship Diagaram (ERD)

Here is an ERD from the Prediction Application of New Student Withdrawal At one of the universities can be seen in Figure 9.

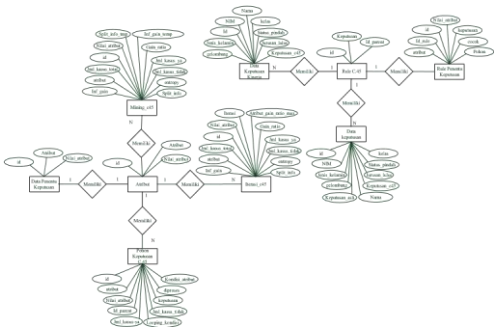

Figure 9. *Entity Relationship Diagram* (ERD)

D. Program Manual

The following is a manual of the use of C4.5 Algorithm Implementation Program For New Student Withdrawal Prediction:

1. Login page Login page is used to input user's username and password in order to access the main page of C4.5 application. Enter the username and password, to run the C4.5 program. The Login page can be seen in Figure 10.

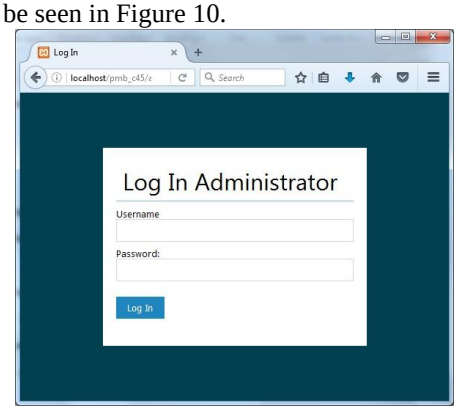

Figure 10. Log In Page

2. Home Page The second view of this C4.5 application program is the Home Page. This Home page is used as the parent of all menus present in

ISSN 2527-1682 (Print)
ISSN 2527-9165 (Online)

this app. Where when successfully log in then the page to be accessed is the home page. So when you want to access other menu can be selected from this home page. The Home Page can be seen in Figure 11.

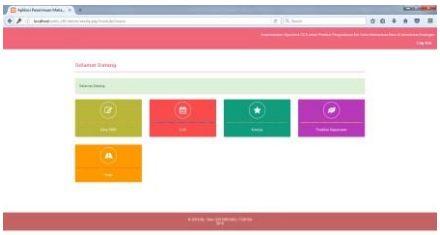

Figure 11. Home Page

PMB Data Input Page
Page Input *PMB* Data is used to input *PMB* Data for training data or data testing. Enter the pmb data you want to use for training or testing. The *PMB* Data Input Page can be seen in Figure 12.

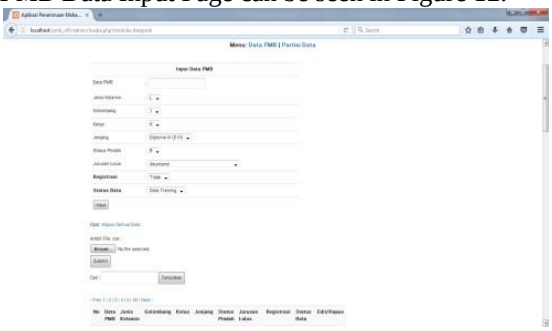

Figure 12. Input Data PMB Page

3. Page Partition Data Page Partition Data is used to divide PMB data into data for training and data for testing. To partition the data we can input into the Data Set Data Set (All Data). The view of Partition Data Page can be seen in Figure 13.

Figure 13. Page Partition Data

4. Decision Tree Page Decision Tree page is a page that is used to display decision tree, decision tree chart and rule formed from mining process. To access this page can choose menu C4.5 The Decision Tree page can be seen in Figure 14.

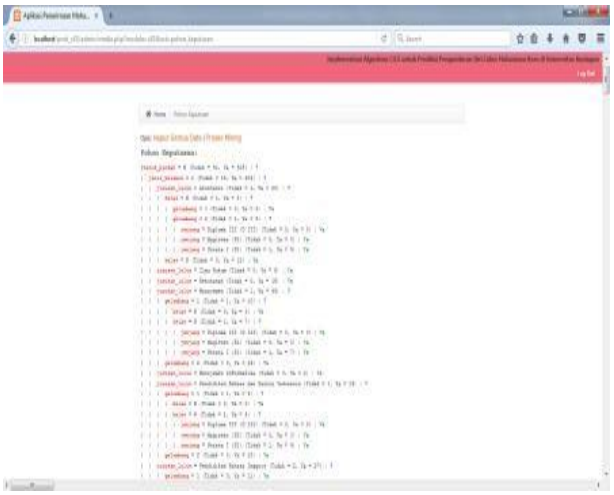

Figure 14. Decision Tree Page

5. Performance Page Performance page is used to display the results of the comparison of the original decision and decision based on the decision tree, to be able to see the performance results then have to perform the performance process first. As for Page Performance can be seen in Figure 15.

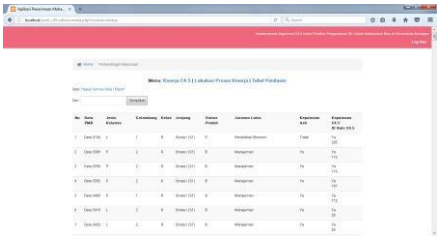

Figure 15. Performance Page

6. Figure 4.6 Performance Page Page Table Assessment Page The Rating Table is used to display the assessment results from the original decision comparison and decision based on the decision tree, on this page showing the value of precision, recall, accuracy. The Page Table of Assessment can be seen in Figure 16.

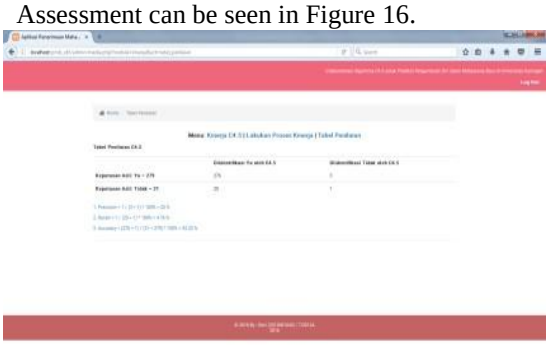

Figure 16. Page Assessment Table

7. Page Prediction Decision Page Prediction Decision is the interface (interface) used to input

the PMB data to be tested based on the decision tree formed. To perform PMB data input can be done by entering PMB data one by one or by uploading PMB data with csv file type. The Decision Prediction Page can be seen in Figure 17.

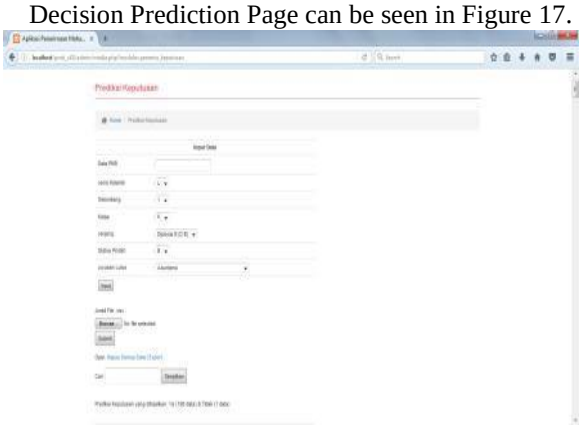

Figure 17. Decision Prediction Page

IV. CONCLUSIONS

A. Conclusion

The conclusion of this research is to achieve all research objectives as follows:

1. Can process the stack of new students admissions data into useful information.
2. Applications created can classify data of new student candidates that can generate predictions for the resignation of new student candidates.
3. Algorithm C4.5 can be implemented for the case of new student data to know the number of new student candidates who will do the registration and do not do the registration.

B. Suggestion

In the research C4.5 algorithm implementation to predict the resignation of new student candidates are still many shortcomings, so require further refinement. As for some suggestions among others:

1. It is expected that further development can improve the processing time of mining, performance and decision prediction.
2. The interface of this application is simple, so it is expected that further development can be made even more interesting.

V. REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, vol. 3918. 1996.

[2] E. Darmawan, *Sistem Penunjang Keputusan Penerimaan Beasiswa Menggunakan Fuzzy Multiple Attribute Decision Making(FMADM)*. Nuansa Informatika, 2013.

[3] E. Endarmoko, *Tesaurus Bahasa Indonesia*. Gramedia, 2006.

[4] Pusat Bahasa Departemen Pendidikan Nasional, *Kamus Besar Bahasa Indonesia (Fourth Edition)*. 2008.

[5] Kusrini and E. T. Luthfi, *Algoritma Data Mining*. Yogyakarta: Andi, 2009.

[6] D. T. Larose, *Discovering knowledge in data: an introduction to data mining*, vol. 28, no. 1. 2005.

[7] I. Pramudiono, "Apa Itu Data Mining?," 2006. [Online]. Available: https://datamining.japati.net/cgibin/indodm.cgi?bacaarsip&1155 52767&artikel.

[8] H. . Kusrini, Hartati.S, Wardoyo.R, "Perbandingan metode nearest neighbor dan algoritma c4.5 untuk menganalisis kemungkinan pengunduran diri calon mahasiswa di stmik amikom yogyakarta," *J. Dasi*, vol. 10(1), no. 1, pp. 1–132, 2009.