Implementation of C4.5 and PAPI Kostick to Predict Students Potential as Organization Caretaker

Ricky Jiandy, Yustinus Widya Wiratama, Seng Hansun
Computer Science Department
Universitas Multimedia Nusantara
Tangerang, Indonesia
jiandy.ricky@gmail.com, yustinus.widya@umn.ac.id, hansun@umn.ac.id

***Abstract*—Student organization usually requires qualified students in leadership as its caretakers. There are some difficulties in identifying and classifying potential students as organization caretakers. Therefore it is important to have an approach for identification and classification of students as a potential organization caretaker. Then to identify and classify the student potential, C4.5 Algorithm can be used. C4.5 Algorithm can be implemented with PAPI Kostick testing tool to predict the students with the aim to identify potential students as organization caretaker. The processing of the data obtained from PAPI Kostick can be done with C4.5 Algorithm. Based on the implementation of C4.5 Algorithm to predict potential of students as organization caretaker with PAPI Kostick test tool, it can be concluded that the PAPI Kostick test tool doesn't match well with C4.5 Algorithm because the results from predicting student's potential as organization caretaker don't have a significant accuracy.**

Keywords—C4.5 algorithm, data mining, organization caretaker, PAPI Kostick, predictive analysis

I. INTRODUCTION

Organizations in college have an important role to support the ability of students, where the main function of student organizations in college is to develop the ability of students, especially in developing the leadership skills of students so that the required figure of appropriate student as a member of the organization is achieved [1]. However, in college life, there are many types of students who have different abilities and different daily activities. The differences in these characteristics proof to be a challenge to find students who could potentially become the organization's caretakers that meets the requirements of leadership [2]. One of leadership requirements is capabilities to interact with members or colleagues and others [3]. Then, there are various problems in finding people qualified in leadership [4]. The results of few studies concluded that many companies experiencing a shortage of human resources who have the talent as a leader. In addition, the capacity of the leadership at that time was far from required, and every few years, the requirements of leadership keep growing [4].

	Now	Future
1	73% - Leading people	89% - Leading people
2	64% - Strategic planning	86% - Strategic planning
3	63% - Managing change	86% - Inspiring commitment
4	64% - Resourcefulness	82% - Managing change
5	64% - Doing whatever it takes	82% - Resourcefulness
6	62% - Inspiring commitment	81% - Participative management
7	60% - Being a quick learner	79% - Being a quick learner
8	60% - Decisiveness	79% - Employee development
9	57% - Building and mending relationships	77% - Doing whatever it takes
10	57% - Composure	76% - Balancing personal life and work

Fig. 1. Leadership's Requirement Ratio [4]

Causes of talents deficiency in leadership occurred because the characteristics of human resources are unknown so that the necessary approach to classify the human resources into multiple levels from which have the ability to leads up to that which does not have the ability to lead, and through the results of the classification, the data processing can be done in certain populations to find resources with potential or in accordance with the terms of the classification that has been built [4]. In addition, identification of the potential as a leader in an organization becomes an important consideration [2]. This of course will be a problem if the identification of potential leaders is experiencing various difficulties which required the classification of leadership in an organization to facilitate potential identification. The leadership crisis was influenced by two factors, namely first the loss of the main characters in the leadership structure caused by the retirement of baby boomers and the second factor which is the lack of capacity and expertise of the next generation [5]. Then from the various problems that exist, it's important to have an approach in predicting or identifying potential leaders or executives of the organization based on their characteristics. One of the approaches is using C4.5 Algorithm for classification and prediction of a certain students who has potential as organization caretaker. In determining the characteristics and personality of the students, a personality test can be used. One of the personality test is Personality and Preferences Inventory (PAPI) Kostick test. This personality test in form of a questionnaire was discovered by Dr. Max Martin Kostick in 1960 could be used to determine a person's personality from

various aspects such as leadership and work style. Through PAPI Kostick tests, one's personality elaboration of measurement needs and perceptions of individuals in the work environment can be described [6].

C4.5 algorithm itself uses a decision tree classification techniques, where the decision tree facilitate decision-making becomes more clear and understandable [7]. C4.5 algorithm has good accuracy in the classification and has a faster speed in computation [8]. Therefore C4.5 algorithms can be combined with PAPI Kostick to predict. Then there are other studies that use this algorithm, namely "Analisa dan Penerapan Metode C4.5 untuk Prediksi Loyalitas Pelanggan" by Santoso [9] and "Implementasi Algoritma C4.5 untuk Menentukan Tingkat Bahaya Tsunami" by Abidin [10]. Various forms of research using the algorithm C4.5 became an inspiration in doing this research by using this algorithm solving a different sphere, which is social sphere, especially in the classification of an individual nature and opportunities by using psychological tests PAPI Kostick. In addition, the implementation of C4.5 algorithm was developed using the Java programming language. The advantage of the Java programming language is it's an object-oriented programming and supports applications built to run on a variety operating systems.

II. PAPI KOSTICK

Personality and Preference Inventory (PAPI) Kostick is a test tool in the field of psychology is used to determine a person's personality. This test tool invented by Dr. Max Kostick originated in America in 1960 [6]. PAPI Kostick describes one's personality in 20 aspects with each representing a need or another role in a comprehensive manner personality system based on perception of individuals who conduct tests on himself [11]. Based on research's paper by Cemani, Soebroto, and Wicaksono in 2013 entitled "Sistem Pakar Tes Kepribadian PAPI Kostick untuk Seleksi dan Penempatan Kerja" [12], PAPI Kostick aspect of the test is as follows.

1. Work direction with sub aspects:
   a. Need to finish task (N) which is the need to finish a task independently,
   b. Hard intense worked (G) which is a role as a hard worker, and
   c. Need to achieve (A) which is the needs in achieving prestige.
2. Leadership with sub aspects:
   a. Leadership role (L) which is the role in leading,
   b. Need to control others (P) which is the needs to manage or control others, and
   c. Ease in decision making (I) which is the role in making a decision.
3. Activity with sub aspects:
   a. Pace (T) which is having a role with high pace, and
   b. Vigorous type (V) which is having a role with high vigor.
4. Social nature with sub aspects:
   a. Need for closeness and affection (O) which is the needs to feel closeness and affection,
   b. Need to belong to groups (B) which is the needs to feel accepted in a group,
   c. Social extension (S) which is the needs to have a social life, and
   d. Need to be noticed (X) which is the needs to be noticed.
5. Work style with sub aspects:
   a. Organized type (C) which is the capabilities in organizing things,
   b. Interest in working with details (D) which is the skills in doing details, and
   c. Theoretical type (R) which is the skills in working with theory.
6. Temperament with sub aspects:
   a. Need for change (Z) which is the needs to make a change,
   b. Emotional resistant (E) which is the role in controlling emotions, and
   c. Need to be forceful (K) which is the needs to act aggressively.
7. Followership with sub aspects:
   a. Need to support authority (F) which is the needs to help authority, and
   b. Need for rules and supervision (W) which is the needs for rules and supervision.

The participants of the PAPI Kostick test will be given a sheet with 90 pieces statements to be answered. In general PAPI Kostick can be done by filling in a sheet of paper, but with the development of Internet technology, PAPI Kostick can be done online as well as on the website psikologi.umn.ac.id. Conclusion of PAPI Kostick can be formed as a circular diagram [6]. The hypothesis of PAPI Kostick test results can be used to determine the main ability of an individual and can be used by certain parties to develop their personality [6].

III. PREDICTIVE ANALYTICS

In information technology, there are statistical models that can be used to conduct an analysis of existing data. Statistical models can be combined with a model derived from observation or also called empirical models with the aim of producing a prediction models empirically [13]. Extraction of information from a set of data can be done using predictive analytics [13]. This form of representation predictive analytics can be used to predict existing potential [14]. Then predictive analytics has two components, namely empirical predictive models which designed to perform data mining, and the second one to make predictions of an analysis as well as a method to evaluate the predictive ability of the model [13].

Some steps are needed to form a predictive models, which is defining objectives, data collection and instructional design, data preparation, perform analysis in a comprehensive manner, choosing allowed variables, choose a potentially method, conduct evaluation, validation and model selection, and using the model. In conclusion it can be seen that predictive analytics is an approach that can be used for the implementation of a prediction algorithm or classification algorithm.

IV. C4.5 ALGORITHM

Data mining has two basic types of learning methods, first is clustering and the second is classification. According to Rokach and Maimon in their research in 2005, clustering is a method of learning that is unsupervised where the processed data with unknown type or category, while classification takes supervised learning method and work by processing the data that already have a class or category [15].

C4.5 algorithm is an extended algorithm of ID3 algorithm that used to perform classification or prediction. This algorithm applies the concept of classification. Additionally, the C4.5 algorithm is also used to construct a decision tree. C4.5 algorithm has a basic algorithm devised by Quinlan namely Decision Tree Induced 3, or often called ID3 [16]. The basic idea of ID3 is selecting attributes with the highest information gain based entropy as the axis attribute classification and then recursively expand the branches of the decision tree until the entire tree is formed [17]. Entropy according to the dictionary IGI (International Publisher of Progressive Academic Research) Global is the amount of data that isn't relevant to the information from a data set or can be explained that entropy is 1 - information while the information gain is the information obtained from the change in entropy on a set of data through observation or it can also concluded with a partition to a set of data T into T1 and T0 based on certain characteristics then this can be regarded as information gain [18].

In this algorithm, there is a process similar to ID3, which is a decision tree building process. This process can be done by dividing the data recursively until each part consists of data derived from the same class. As described by the Crows in Abidin et al. research's paper [19], the development process decision tree consist of four steps, namely

1. select the attributes as root,
2. create a branch for each value,
3. divide each case in the branch, and
4. repeat the process in each branch to all the cases in the branch have the same class

In the use of C4.5 algorithm, calculating information gain is necessary and can be obtained by first calculating entropy values [10]. Entropy is a measure of the probability of the uncertainty of the data [20]. Then the information gain for calculation of the C4.5 algorithm serves as a predictive calculation based on the data and the reduction of entropy that can produce a simple conclusion [21].

$$Entropy\ (S)\ =\ \sum_{i=1}^{n} -Pi\ *\ log2\ (Pi) \qquad (1)$$

where

S : The set of cases

n : The number of S partitions

Pi : The number of cases in partition i

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

where

S : The set of cases

A : Attribute

n : The number of attribute A Partitions

$|Si|$: The number of cases in the partition i

$|S|$: The set of cases in S

After calculating the entropy and information gain of all the attributes from the data obtained, the decision tree can be form by following C4.5 algorithm steps. The decision tree will be used as a predictive rule to predict a data.

V. CROSS VALIDATION

Cross validation is a statistical method used to conduct the evaluation and comparison to a learning algorithm. The evaluation method is done by dividing the data into two segments, where one segment is used for testing and other segments used for training [22]. One type of cross validation base is k-fold validation. This validation is done by dividing the entire data sample into as many segments k evenly. Total segment adjusted by the number of tests performed, as many as k [22]. In each validation, the segment k-th is selected to be a testing data and other segments of data are used as training data. The trial will be done for each segment until the last segment of testing data is done. For an illustration of the trial k-fold cross validation can be seen in Fig. 2.

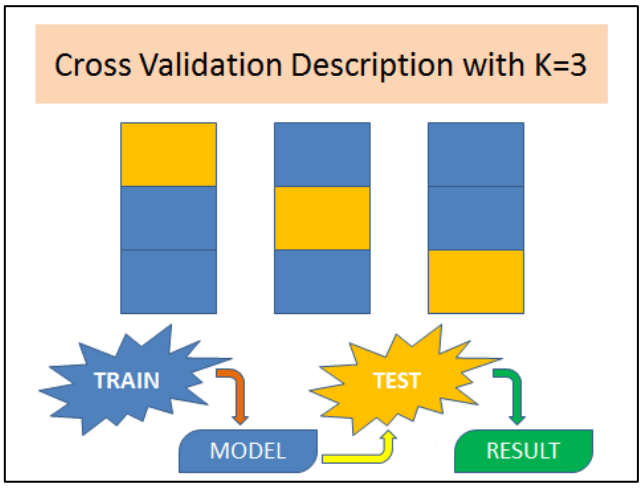

Fig. 2. Cross Validation Example with k = 3

In addition to k-fold cross validation, there are also another type method of cross validation according to Refaielzadeh et al. [22], such as re-substitution validation which is done by testing the entire training data as testing, hold-out validation which is done by dividing the data into two distinct parts or not overlapping so that it can produce estimates with good

accuracy, and leave- one- out cross validation performed like k-fold validation with some changes where k corresponding to the entire amount of data and iterated on all the data to be observed.

VI. APPLICATION DESIGN

Application starts by selecting a prediction method using a set of data or prediction of a data entered by the user.

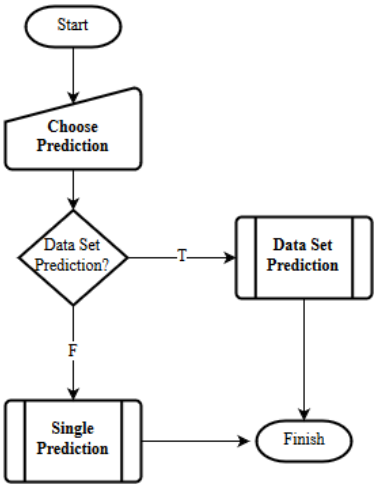

Fig. 3. Flowchart of Application Flow

Fig. 3 descried the order of the application process in general. After the prediction method is done, the application executes a process based on the selected prediction method. When the user selects a prediction using the data set, then the application will perform the procedures prediction data sets. However, if the user selects a prediction method to a single piece of data, then the application will run a single prediction procedure data.

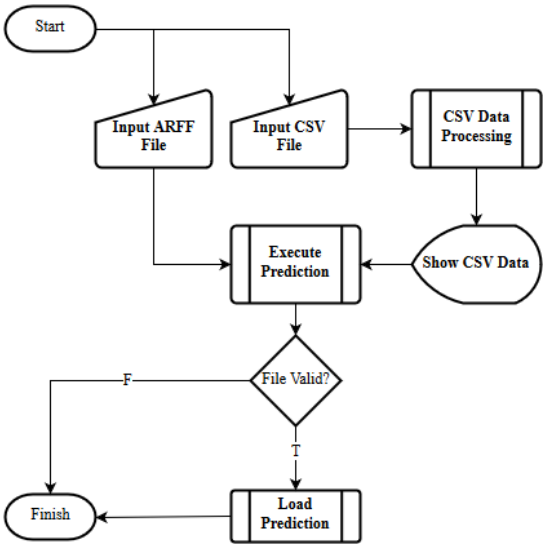

Fig. 4. Flowchart of Data Set Prediction

Fig. 4 described the sequence of the application process to make predictions using data sets. For data testing with csv format, it required a simple processing so that when users enter the csv data as data testing, the can be displayed directly. Once the data from csv displayed, predictions can be executed. In the process of execution of this prediction, arff format data and csv format will be used in a new page which is a result page. Data with arff format will be processed using the C4.5 Algorithm. If data from csv and arff is invalid in terms of both file attributes is mismatch or differences that can't be read, the application will bring up a notification stating that the file is invalid and the application will exit or complete. However, when the two files turn right, then the results page will be displayed along with the results of the predictions that have been made.

In the procedure of single data prediction, data is first entered by selecting each attribute provided by the application. This data is based on test results from doing PAPI Kostick test. Then, after the data is entered, all the attributes will be processed and will be used in the prediction process data.

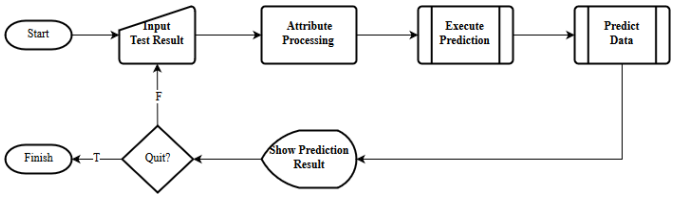

Fig. 5. Flowchart of Single Prediction

Based on Fig. 5 after processing the entered attribute, the next process is to execute a prediction where in this process a prediction rule is formed and based on the existing data sets. After the prediction rule was formed, the prediction based on rule and users input is displayed. Prediction result data is displayed in the form of a simple notification and predictions can be repeated until the user exits the application.

In the general procedure of applications, data entered in csv format processed into a table. The treatment process is to take each data in csv files put it into a table model.

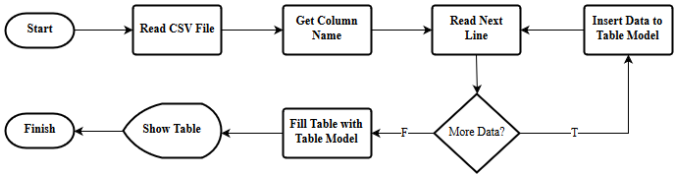

Fig. 6. Flowchart of .csv Processing

Fig. 6 outlines that the first process is done after reading the file and specifies the name of the column is inserted into the table. Then each row of data in the csv file is read and processed into a model table until all the data runs out or no data available to read. Then after all the data in csv file is read and entered into the table model, applications will include a model table into a table to be displayed on the home page of the application. The procedure of this application occurred on the main page right after selecting the file to csv format.

As described in the general procedure execution of the application when the button is pressed, the application will run prediction execution procedure. In this procedure the reading of arff file that has been included with an aim to produce a prediction rule.

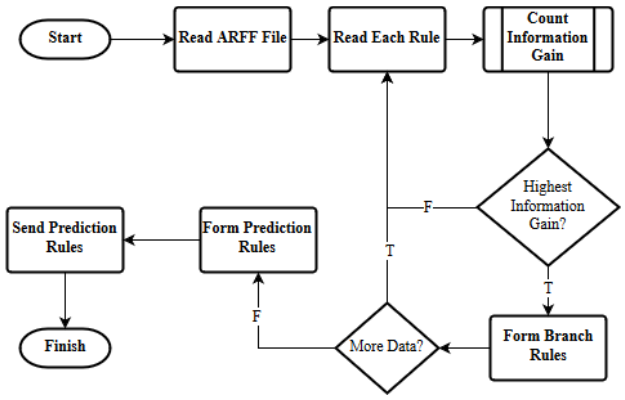

Fig. 7. Flowchart of Execute Prediction

Fig. 7 shows that each of the data read from the file arff will enter a calculation to produce information gain and entropy. Then each attribute is read and counted will be checked whether it has the highest information gain, if it turns out that attribute has the highest information gain, then created a new branch node or rule. This procedure will continue to run so that each attribute has been checked. When forming the rules of the branch, if it turns out attribute can't be concluded, then the attributes will not be listed as a branch prediction rule so that the rule does not become complex. Once all the data is complete, the application will send the rules to be processed on the results page.

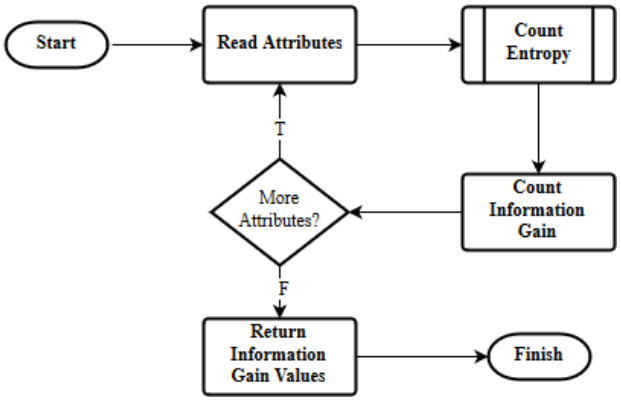

Fig. 8. Flowchart of Information Gain Calculation

Calculations information gain done by calculating the entropy of each attribute. This process is called when any data is read on a prediction execution procedure. However, the calculation process is only performed for each of the attributes of the data read. On the Read Attributes, calculation of the number of cases associated with each attribute based on data from the arff file contained in the execution process prediction.

Then in the process of calculation of information gain, there are sub entropy calculation process that has been described previously. Entropy calculation process is done by reading each value contained in the attribute and the category of that value. The selected category is the category of potential and non-potential.

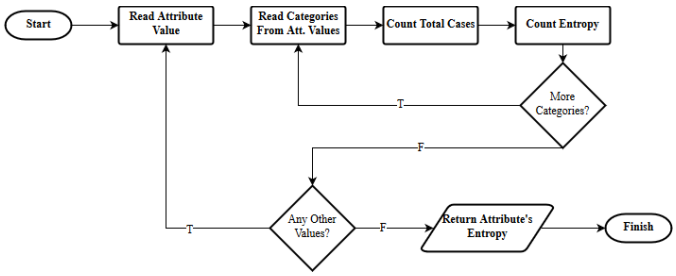

Fig. 9. Flowchart of Entropy Calculation

Fig. 9 describes the process of calculating the entropy. In this process, each value in the attribute is read and followed by reading category of the attribute value. As an example, attribute A with "baik" (eng: good) value and seen its category along with the number of cases are will be calculated to get its entropy. This process is repeated until the entire value of entropy is generated and summed.

Predictions loading done by reading all rules resulting from the execution of predictions procedure. Before displaying the prediction, the included testing data will be sorted first so the predictions can begin. Each of the data contained in the data testing file will processed based on rules drawn from the prediction execution procedure.

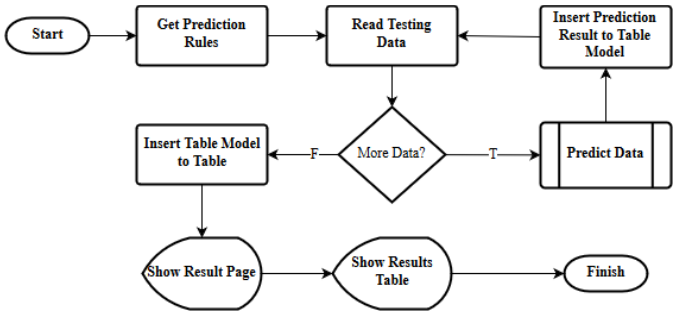

Fig. 10. Flowchart of Loading Prediction

The process of reading data from the file data is repeated until there are no more left. Then after all the data and the predicted results incorporated into a model table is used as a main table, the results page is displayed along with the predicted results. For the prediction data alone, any rules must be traced so that all the attributes and data to be read can be predicted.

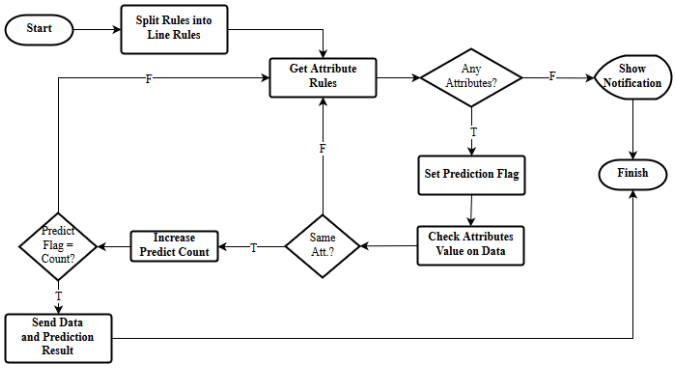

Fig. 11. Flowchart of Data Prediction

Fig. 11 described the process of solving of prediction rules obtained. In this process, every rule is broken down into each row and then processed so that it can process the data entered. When a testing file is read, the prediction process is carried out based on the rules that have been processed. Then each row of data will be modified by adding the prediction results to the data and put in a table model that will be displayed.

Each line rule when obtained, a processing will be performed where attributes and attribute values of the attributes and values are matched with the data. In each line of the rules, there are some attributes so that it's required to checks repeatedly by noting how many attributes that must be matched on a single line. If one of the rules is met with the terms of each attribute on the line fit, then returns a new data along with the predicted results.

VII. Application Testing

Once the application is successfully implemented in accordance with the design that has been conceived, a test is carried out to find out the functions of the application. The first process of testing used the predictive data sets method involving two sets of data. The first set of data is the result of respondents of all samples and will be used as training data to generate a prediction rules. Then for the second data set, dummy data is compiled and used to test the prediction rules that have been formed by the application.

The attributes used for the training and testing are in Bahasa Indonesia since it was tested and used by the real Student Organization Caretaker under the Guidance of Vice Rector of Student Affairs in Universitas Multimedia Nusantara. Here, "Kurang" equals to "Poor", "Cukup" equals to "Average", and "Baik" equals to "Good". Therefore, the attributes shown on Fig. 12 up to Fig. 15 will refer to the above meanings.

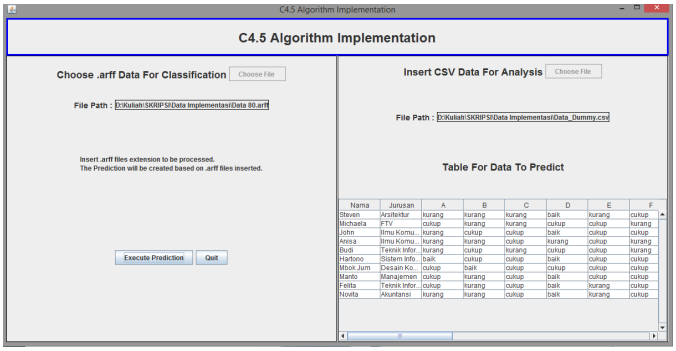

Fig. 12. Dummy Data Testing

In Fig. 12, trials conducted by entering data from a sample of 80 pieces of data as the data to be processed and a dummy data to test the performance of applications. When the Execute Prediction button is executed, the application process arff file format into a stipulation. The provision will be used to classify and predict data based on entered csv file.

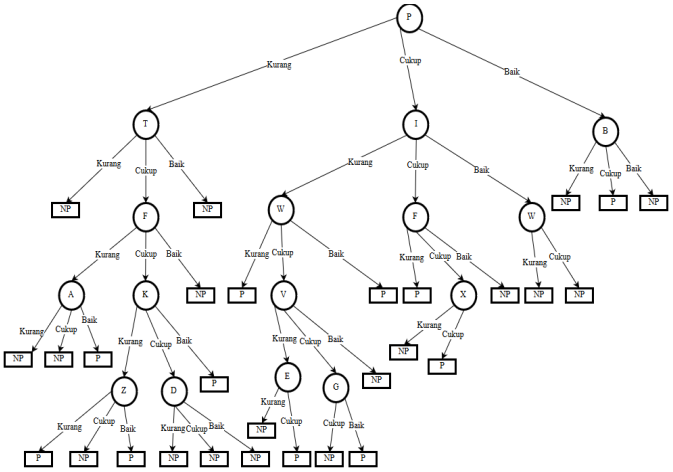

Fig. 13. Decision Tree Based on 80 Sample Data

Fig. 13 lays out the rules obtained from a sample of 80 pieces of data as training data in the form of a decision tree. The decision tree is used to determine the classification of the data entered for testing. Then the results of 80 samples of data is used as a model to predict data sets and also to perform a single data prediction.

Next we choose ten arbitrary data for the testing phase. Based on ten pieces of sample data, the prediction results are in accordance with the provisions which have been discussed previously. Every prediction result is adapted with the prediction rules from the decision tree in Fig. 13. Then the next trial is to do a single data prediction which is entered manually. Trials using the single data is performed twice, namely to generate predictions with non-potential and potential results based on a decision tree that has been described.

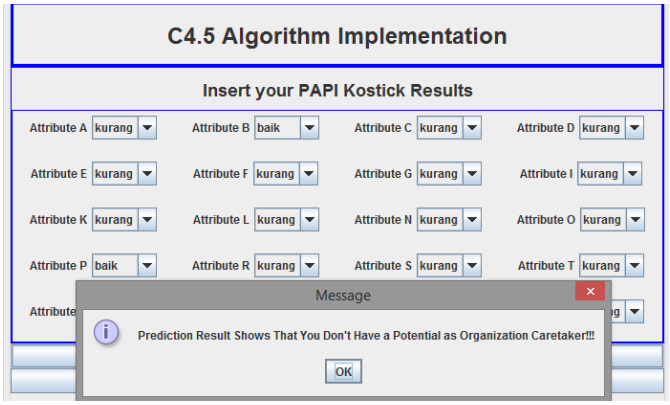

Fig. 14. Single Data Prediction Result

In Fig. 14, the results of trials conducted in compliance with the rules of good predictions P attributes and attribute B is sufficient to produce a prediction no potential as an organization caretaker. Results from the trial are in accordance with the one shown in the decision tree in Figure 13, where the data used included the category of non-potential.

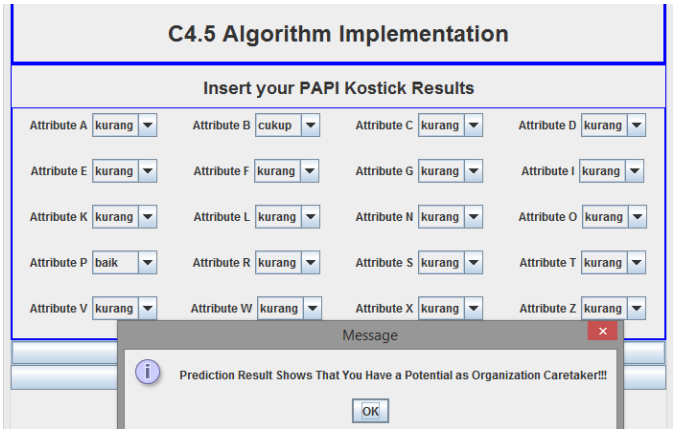

Fig. 15. Single Data Prediction Results 2

Then in Fig. 15 conducted similar trials to generate predictive potential as an organization caretaker. Single data is entered for the test adjusted to a decision tree where P is the good attributes and attribute B is enough so that potentially categorized as an organization caretaker. In conclusion it can be seen that the implementation of the application is running properly and in accordance with a predetermined function.

VIII. Results Analysis

From the application trial, a calculation is held to get the prediction accuracy based on the results. The calculations were performed by analyzing the results of the application using the k-fold cross validation as described in the literature review. The test conducted a 10-fold cross validation with sample of 80 pieces data that have been collected.

TABLE I. 10-Fold Cross Validation Result

Trial	Numbers of Data Testing		Correct Prediction	Faulty Prediction	Accuracy
	P	NP			
1	4	4	5	3	63%
2	4	4	3	5	38%
3	4	4	4	4	50%
4	4	4	5	3	63%
5	4	4	2	6	25%
6	4	4	5	3	63%
7	4	4	4	4	50%
8	4	4	4	4	50%
9	4	4	4	4	50%
10	2	6	6	2	75%
Total Accuracy					**53%**

Table 1 outlines the results of the calculation of the cross validation is performed. For the third column and the fourth column, are the results of a prediction of the prediction results by organization caretaker students and non-organization caretaker students and its prediction results. The entirety of accuracy to predict accurately generated by cross validation trials is 53 %.

Accuracy of cross validation test results turned out to be insignificant. This happened because the data collected based on the results of respondents using PAPI Kostick does not have a structured pattern. Based on data from research conducted by Santoso [9], these data have a clear pattern where the price will determine the loyalty of customers and several other attributes. The exact pattern can be seen from some of the attributes that have a value of zero entropy. Santoso's research [9] could produces significant accuracy up to 97%. Unlike the sample data obtained using PAPI Kostick. The sample data of PAPI Kostick has no clear pattern for a different answer from each sample so that the data becomes scattered and does not have a clear pattern. In addition, the combination of a possible answer using PAPI Kostick is 320 which made the data sample has a scattered pattern and causes the accuracy of the C4.5 algorithm to predict declines.

IX. Conclusion

Implementation of C4.5 algorithm to predict the potential of students as organization caretaker using PAPI Kostick results is a success. However, the result of the predictions is not significant with 53% accuracy. This happens because the data sample results obtained using psychological tests PAPI Kostick do not have a structured pattern. Therefore, it can be concluded that the implementation of C4.5 algorithm to predict the potential of students as the organization's caretaker does not match well with the psychological tests PAPI Kostick tool.

References

[1] Fitriyatun, Makalah Pengembangan Organisasi Kemahasiswaan dan Peran & Tanggung Jawab Mahasiswa. 2014.

[2] Corbett, J., A Study of Community College Leadership Practices in Response to the Impending Leadership Crisis (Doctoral Dissertation, CAPELLA UNIVERSITY), 2012.

[3] Winston, E.B. and Patterson, An Integrative Definition of Leadership, *International Journal of Leadership Studies, 1(2)*, pp. 6-66, 2006.

[4] Leslie, B.J., The Leadership Gap, *Center for Creative Leadership*, 2009.

[5] Holinsworth, S.R., Case study: Henrico County, Virginia: Succession Management: A Developmental Approach. *Public Personnel Management, 33(4)*, p. 475-486, 2004.

[6] Cubiks, PAPI The Complete Personality Assessment. December, 2015.

[7] Wang, J. and Wang, D., Decision Trees, *Encyclopedia of Information Technology Curriculum Integration*. IGI Global, pp. 203-208, 2008.

[8] Ruggieri, S., Efficient C4.5, *IEEE Transaction on Knowledge and Data Engineering,* 14(2), pp. 438-444, 2002.

[9] Santoso, B.T., Analisa dan Penerapan Metode C4.5 untuk Prediksi Loyalitas Pelanggan, *Jurnal Ilmiah Fakultas Teknik LIMIT'S, 10(1)*, 2012.

[10] Abidin, A.Z.Z., Implementasi Algoritma C4.5 untuk Menentukan Tingkat Bahaya Tsunami. *In Proceedings of Seminar Nasional Informatika 2011 UPN 'Veteran' Yogyakarta*, 2011.

[11] Tirtawinata, M.C., Mengenal dan Menemukan Diri Melalui Kebersamaan dengan Orang Lain. Character Building Development Center Binus University, 2013.

[12] Cemani, P.D., Soebroto, A.A., and Wicaksono, A.S., Sistem Pakar Tes Kepribadian PAPI Kostick untuk Seleksi dan Penempatan Tenaga Kerja, 2013.

[13] Shmueli, G. and Koppius, R.O., Predictive Analytics in Information Systems Research, *MIS Quarterly,* 35(3), pp. 553-572, 2011.

[14] Predictive Analytics, Extending the Value of Your Data Warehousing Investment, 2007.

[15] Rokach, L. and Maimon, O., Clustering Methods, Data Mining and Knowledge Discovery Handbook, pp. 321-352, 2005.

[16] Quinlan, J.R., Induction of Decision Trees. *Machine Learning*, *1(1)*, pp.81-106, 1986.

[17] Chourasia, S., Survey Paper on Improved Methods of ID3 Decision Tree Classification. *International Journal of Scientific and Research Publications*, pp. 88, 2013.

[18] IGI Global Dictionary, What is Information Gain, December, 2015.

[19] Abidin, A.Z.Z., Implementasi Algoritma C4.5 untuk Menentukan Rekomendasi Model Pengajaran pada Sistem Pemberi Rekomendasi Ajar. *In Proceedings of SNTE PNJ 2010*, pp. 124-132, 2010.

[20] Principia Cybernetica Web, Entropy and Information, 2001.

[21] School of Computer Science Carnegie Mellon University, Information Gain, 2001.

[22] Refaeilzadeh, P., Tang, L. and Liu, H., Cross-validation, *Encyclopedia of Database Systems*, pp. 532-538, 2009.