Implementation of C4.5 Algorithm and Forward Chaining Method for Higher Education Performance Analysis

Mochamad Idris[1,*], *Mustafid* [2], and *Jatmiko Endro* Suseno[3]

[1] Magister Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang - Indonesia
[2] Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang – Indonesia
[3] Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Semarang - Indonesia

Abstract. Higher education has an important role to develop human resources in the economic growth and development of the country. One of specific way of evaluating and analyze data in education is to use data mining techniques. C4.5 algorithm as one of the data mining techniques that have good performance is very relevant used for data analysis tools. In this research using data on the performance of lecturers in college, there are 100 records with a 6 variable that affects individual factors in the productivity of lecturers including age, employment, attendance, certification, position, Education, and additional duties. In the end of the mining result, the forward chaining method is used to extract the rules that are generated by C4.5 algorithm. The input premises are examined by forwarding chaining to generate the prediction result.

Keywords: **Data Mining; Decision Tree; C4.5 Algorithm; Forward Chaining Method; Higher Education Performace.**

1 Introduction

Several performances measuring frameworks derived from the public sector have been developed for performance measurements in colleges. One method of improving the performance of a college organization that is aligned with the measurement of individual performance in colleges, this measurement has the risk that the public organization is worried may not be fully Take the indicator of the performance key of the college (key performance indicator) [1].

Higher education has an important role to develop human resources in the economic growth and development of the country. Over the last few decades, the number of colleges has increased, therefore college – colleges in international competitions have highlighted the importance of improving human resource performance in this regard, performance assessments that can help organizations to plan future strategies and set employee performance targets to reach the final target of the entire organization [2].

Staff at colleges who are often evaluated their performance of faculty or lecturers, wherever colleges have done a lot of performance evaluation of their faculty staff based on key performance indicators (KPI) such as teaching, research, and publications [3]. The first step of this research is identifying KPI that colleges use to evaluate and measure the performance of their organizations. From many KPI, this study focuses more on the performance of lecturers in college.

2 Literature Review

2.1 Higher education performance

Before learning how to analyze the performance of the following colleges there are several definitions of performance. performance is defined as multidimensional construction and common factors often associated with performance organizations such as efficiency, quality, responsiveness, cost and overall effectiveness [4].

The indicator is a tribute to the performance contributions of academic staff such as benefits, performance money, etc. The population of the case study was determined by purposive sampling and consisted of staff, professors, senior lecturers and public university lecturers in all categories of Malaysia. The results showed that the performance indicators had positive and significant implications for academic staff [3].

2.2 Data Mining and Educational Data Mining

Data Mining is the analysis step to discover the knowledge in database process. With data mining process analysis of data from various perspectives and to summarize it into useful information. Then in the process of data mining required software as one of number of analytical tools to analyze the data. With the help of this software, it is possible that users can analyze data from different dimensions or angles, categorize it, and

* Corresponding author: idrez.mochamad@gmail.com

summarize the identified relationships. In recent years, there has been an increased interest in the use of data mining to investigate some of the problems – scientific problems in the educational world. [5]. In the data mining research, it implements several techniques such as K-nearest neighbor, decision Tree, Naïve Bayes, Neural Network, Fuzzy and others [6].

While educational data Mining (EDM) can be defined as the application of traditional data mining techniques for the analysis of educational data aimed at solving problems in the context of education. Some EDM applications consist of developing e-learning systems, pedagogic support, Clustering educational data, and student achievement predictions [7]. Because of previous explanation, this research focuses are on the performance analysis of lecturers in higher education because writers are interested in understand the effect of individual performance factors (social, personal and academic) tp their performance in higher education.

2.3 Decision Tree

As widely used for research in classification problem, decision tree is such as the famous algorithm in data mining methodology. Decision tree rules itself is represented by decision tree method, big data applied decision tree can be transformed into smaller records by decision tree structure along with its rules. The heterogeneous data can be transformed to homogeneous data with unique target variable by applying decision tree models [8].

As supervised learning classification method, decision tree has class labels or categories label, they are set in the beginning and in middle of the model making process to be used for the training data to classify the new data. Decision tree model contains subpart or node [9] :

- Root node is a top subpart or top node of decision tree, There are no incoming branches in this subpart or node and it has branches (one or more than one branch). This node represents the greatest influence on the decision tree model.
- Internal node, this node is set after the root node, this node has only one incoming branch (root node), and has one or more than one branch.
- Leaf node, this is the end node that has no branch. The decision model stops in this node. This node also represents the class label in decision tree model.

2.4 C4.5 Algorithm

As the decision tree algorithm, J. Ross Quinlan develops C4.5 algorithm to continue ID3 algorithm as the previous development algorithm [8]. Basically the ID3 and C4.5 algorithms are no different principles. According to Han [10], there are three steps in the C4.5 working algorithm:

- The first step to be done is construct the decision tree, this step is to create a model of a set of training data that will be used to predict the class of new data.
- The second step is the prunning process. Prunning is used to simplify the decision tree result, so the decision tree construction can be easier to read, the pruning is based on the value level of confidence.
- The last step is the rules making of the constructed decision tree. Basically, the rule forms are extracted from the decision tree are if-then forms, they are instance of the decision tree by exploring from the root node to the leaf node.

Basically, the decision tree used by C4.5 algorithm is built by greedy algorithm, it is processed recursively from top to bottom by divide and conquer technique [10].

In general, the algorithm C 4.5 used to build the decision tree is the following [11]:

1. Counts the total cases, the total of cases each class, then the entropy of all cases and cases divided by attribute values. Entropy is used to determine how informative an input attribute to generate the output attribute. The basic formula of the entropy is as follows:

$$Entropy(S) = \sum_{i=1}^{k} - p_i \, log_2 \, p_i \qquad (1)$$

$Entropy(S)$ describes the probability of the cases, and p_i is a proportion of the cases. They are used to suspect the candidate split, which partitioned the training dataset into numerous parts.

2. Calculates the information gain for each attribute. The root attribute selection is using the highest Gain value of the existing attributes. The gain calculation is performed by formulas as listed as follows:

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

$Gain\ (S, A)$ inform how much influence of attribute A to total cases S, and (S_i) is the number of probabilities of S_i against S

3. Calculating the SplitInfo for each attribute.

$$SplitInfo\ (S, A) = - \sum_{i=1}^{n} \frac{S_i}{S} \, log_2 \, \frac{S_i}{S} \qquad (3)$$

The partitioning training data S generate SplitInfo to describe entropy or potential information, it expresses into different variables which belong to attribute A.

4. Calculating the GainRatio for each attribute.

$$GainRatio\ (A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \qquad (4)$$

The differences between ID3 and C4.5 algorithm is gain ratio, it improves the disadvantages of information gain use only.

5. Select the attributes with the largest gain ratio as a node.
6. Divide the data by attribute value of the selected attribute. Then use it to take the next step
7. Repeat steps 1 through 6 until all the attributes are used or meet a stopped condition.

2.5 Forward Chaining

Forward chaining is one of the methods of an expert system that seeks or searches for solutions through problems [12]. In other words this method does the consideration of the facts that then culminate in a

conclusion based on the facts. This method is the inverse of a backward chaining method that performs a search that originates from the hypothesis to the facts to support the hypothesis.

Meanwhile, according to Gaag and Lida [13], Forward chaining is a computing model from the bottom up. The reasoning Model begins with a series of known facts and is applied to the rules for generating new facts in accordance with known facts, and continuing this reasoning process until it reaches a goal that has Set, or until there are further facts that can be obtained that correspond to known facts. This reasoning method examines the facts against a predetermined request or goal and shows that the conclusion moves forward towards the intended facts. Forward chaining is also used to improve and develop modeling of Expert Systems (ES) and modeling of human brains in the Artificial Intelligence (AI) realm.

2.6 Confusion Matrix

The performance measurement of a classification algorithm is very important because it illustrates how well the system is running data classification. The performance measurement in this study uses confusion matrix to measure the performance of the C 4.5 algorithm. Confusion matrix works by comparing the classification results that be done by the system with the results it should be. The process in the classification of confusion matrix has four forms to represent the result, the four forms describe the results as True Positive (TP) is the total of positive data that is correctly classified by the system, the True Negative (TN) is the total of negative data and detected correctly by the system, False Positive (FP) is a positive data but is detected as the negative, it is detected incorrectly by the system and False Negative (FN)) is the total of the negative data but detected as the positive, it is detected incorrectly by the system. Based on True Negative (TN), False Positive (FP), False Negative (FN) and True Positive values can be obtained accuracy, precision and recall values. The four forms above describe the accurate classification of the system, the division of total of positive data and the total of classified positive data describes the precision value of the classification of system. The percentage of the positive classified data to real positive data is shown by the recalled formula. Confusion matrix is indicated with Table 1. While the equation for accuracy, precision, and recall values are as follows:

Table 1. Confusion matrix

		Real Data	
		Positive (1)	**Negative (0)**
Classified Data	**Positive (1)**	TP	FP
	Negative (0)	FN	TN

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\% \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} * 100\% \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} * 100\% \quad (7)$$

2.7 10 Fold Cross-Validation

Performance evaluation is needed for the classification algorithm working process. The performance evaluation in this study is using the 10-fold cross-validation method to evaluate a model C 4.5 algorithm. The 10 fold cross-validation split into two datasets the first one is part of the test/evaluation dataset and the other one is the exercise dataset. 10 fold cross-validation is one of the best model selection because it tends to provide better and measurable accuracy estimation. Because in this research use 10 fold Cross-validation, the dataset that has been divided into two parts then carried out training 10 times and the result of the training is calculated the average value of accuracy of each Iteration to get the accuracy value. Figure 1. is the validation illustration.

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Training Data   Testing Data

Fig. 1. 10 Fold Cross-Validation.

The 10 folds cross-validation works as follows:

1. The Total instance is divided into 10 parts or folds.
2. The 1st fold starts by the 1st part becomes data test and the rest become training data. Next, calculate the accuracy based on the portion of the data. Calculation of such accuracy by using confusion matrix accuracy equation.
3. Go on the 2nd fold, the 2nd fold start by the 2nd part becomes the data testing and the rest to the training data. Next, calculate the accuracy based on the portion of the data.
4. Continue the process until it reaches the 10th fold. After the process gets all the accuracy results, then perform calculation to count the average accuracy of all the 10 pieces accuracy above.

3 Implementation

Input of the system to be built is a lecturer data in the form of a table that contains several attributes include: Age, employment, attendance, certification, position, education, additional duties, and the faculty's origin. The lecturer Data will be classified based on the target specified and calculated using the Decision Tree method that is C 4.5 algorithm to find the value of Entropy and Gain information. After the calculation process is

completed it will generate a rule or condition used in determining the decision in the prediction process.

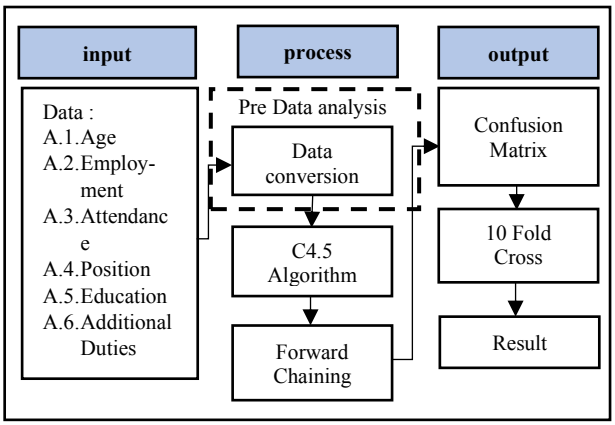

Fig. 2. Analysis process flow.

This research aims to conduct analysis of factors affecting the level of performance of lecturers in this study using data on the employee's performance system. The achievement of this performance is a target to achieve lecturers to be a guide to payment of incentives received by lecturers. In these performance achievements, there are two conditions that are fulfilling and not fulfilling.

The input variable that will be used in the process of data mining using the decision tree with algorithm C 4.5 is a variable that affects individual factors in the productivity of lecturers including age, employment, attendance, certification, position, education, and additional duties. In this research, there are 100 lecturers with 75 lecturers got fulfilling status and 25 are not fulfilling. Below are the data used for this research. Each of these variables is presented in the Table 2.

Table 2. Event number recapitulation data

Var	Desc	Category	Case	Fulfilling	Not Fulfilling
Total			100	75	25
Age	Age during the assignment period	U1 : < 31 y	3	2	1
		U2 : 31 – 40 y	12	3	9
		U3 : 41 – 50 y	40	37	3
		U4 : > 50 y	45	33	12
Employment	The teaching period began to be appointed	M1 : < 11 y	9	2	7
		M2 : 11 – 20 y	7	2	5
		M3 : 21 – 30 y	24	21	3
		M4 : > 30 y	60	50	10
Functional position	Functional position of the lecturer	Pengajar	5	0	5
		Asisten Ahli	9	4	5
		Lektor	35	30	5
		Lektor Kepala	40	32	8
		Guru Besar	11	9	2
Education	Education level that has been pursued by lecturers	S2	49	30	19
		S3	51	45	6
Additional duties	Whether the lecturer has an additional duties or is not	Yes	46	37	9
		No	54	38	16
Attendance	Percentage of attendance	K1 : 91% - 100%	24	24	0
		K2 : 51% - 90%	67	47	20
		K3 : < 51%	9	4	5

4 Result and Discussion

Based on the provided data, the next step is the mining process. The mining process that we will perform is using one of the popular data mining tools called rapid miner. In the figure below we can see that the generated decision tree is quite compact because the created nodes are not many. So the rules that can be extracted from the decision tree is quite simple.

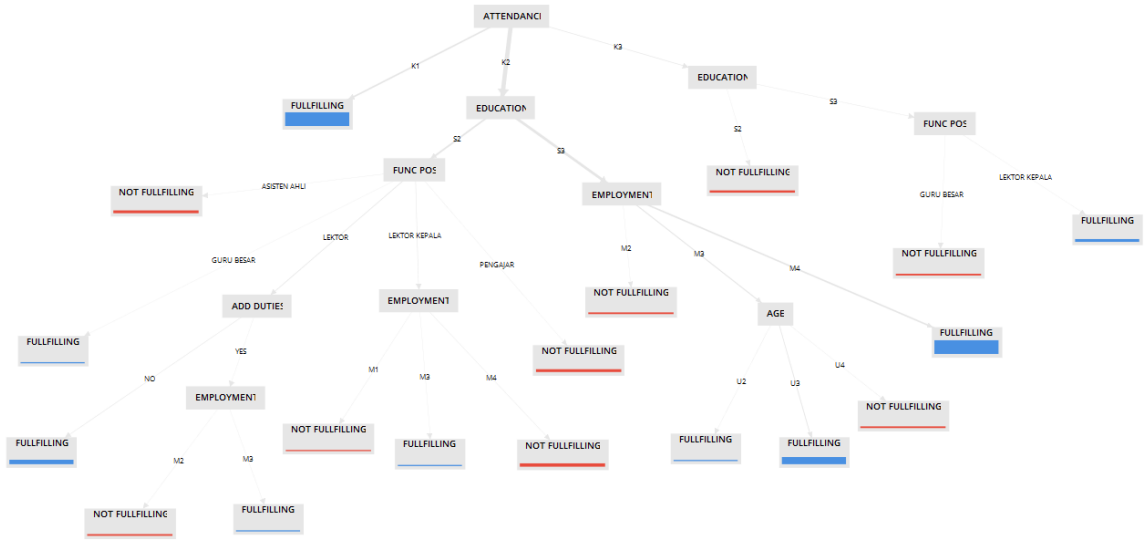

Fig. 3. Generated decision tree model

In forward-chaining the algorithm compares the data with the conditions of each rule in the rule base, using the ordering of that rule base. If the information provides a rule is firing, the outcome is placed in memory and the procedure continues to the next rule. Assume there is one condition with the following details applying for predict, and the details are:

1. The percentage of attendance is 80%
2. The lecturer education is S3
3. The employment period is 25 years
4. The age of the person is 45 years old.
5. The functional position of the lecturer is LEKTOR
6. The lecturer has Additional duties

In forward-chaining reasoning, we will start by examining the premises of all rules in order to see what information is. the first premise is The percentage of attendance is 80% so we can assume the category of data is K2 the next premise is the lecturer education is S3, based on rule that made by C4.5 algorithm with these premises we have 3 conditions of employment but the next premise is the employment period is 25 years so the appropriate condition is employment M3.

The next premise is about the age, the input premise is the age of the lecturer is 45 years old so we can say the person belongs to category U3. According to the rules are made by C4.5 algorithm the conditions that provided match with following rule.

RULE :
IF ATTENDANCE = K2 **AND** EDUCATION = S3 **AND** EMPLOYMENT = M3 **AND** AGE = U3
THEN FULLFILLING

According to rule above the forward-chaining process is stopped and the best prediction from the conditions are provided is FULFILLING.

After constructing the decision tree model, we enter testing data set into the model. Table below shows the accuracy of model based on C4.5 decision tree. From the table, can be seen that decision tree model is quite accurate to do the evaluation based on lecturer's performance.

Table 3. Confusion matrix result

		Actual		**Class Precision**
		Positive (1)	**Negative (0)**	
Predicted	**Positive (1)**	71	6	92.21%
	Negative (0)	4	19	82.61%
Class Recall		94.67%	76%	

$$\text{Accuracy} = \frac{71 + 19}{71 + 19 + 4 + 6} * 100\% = 90\%$$

5 Conclusion

The result is showed in the above table proves that the accuracy of the prediction is generated from C4.5 algorithm is 90%. The accuracy is quite good enough to predict the result from the premises that input to system. The rules from decision tree collaborate with forward-chaining method is used to examine the input premises. The application of decision tree is inseparable from forward-chaining method, which is the forward-chaining is one of the ways to extract the inference from decision tree result.

After the result of decision tree is generated, one of the ways to validate and measure the decision tree performance by using 10 folds cross-validation, in the validation process 10 fold validation using confusion matrix to calculate the accuracy. The bigger value of accuracy we get, more accurate system can predict the output result.

References

1. Xiaocheng, Wang. Report Research, University of Twente (2010).
2. Neda Jalaliyoon, Hamed Taherdoost. Procedia - Social and Behavioral Sciences **46**, pp 5682 – 5686 (2012).
3. T. A. Masron, Zamri Ahmad, N. Baba Rahim. Procedia - Social and Behavioral Sciences, **56**, pp 494 – 503 (2012).
4. Lockett, J. Effective Performance Management, Kogan Page, London (1992).
5. Crist´obal Romero. IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications and Reviews, **40**, 6 (2010).
6. Carlos Márquez Vera, Cristóbal Romero Morales and Sebastián Ventura Soto. The IEEE Journal of Latin of American Learning Technologies (IEEE-RITA) , **8**, 1: 7-14 (2013).
7. F. Eduardo, H. Maristela, V. Marcio, B. Vinicius, C. Rommel, V.E, Gustavo. Journal of Business Research, **94**, 335-343 (2019).
8. M. J. Berry and G. S. Linoff, Data Mining Techniques For Marketing, Sales, Customer Relationship Management, Second ed., Indianapolis, Indiana: Wiley Publishing, Inc. (2004).
9. P.N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, **1**, Boston: Pearson Addison-Wesley (2006).
10. J. Han, M. Kamber and J. Pei. Waltham: Elsevier Inc (2014).
11. J. R. Quinlan. USA: Morgan Kaufmann (1993).
12. A. Al-Ajlan. International Journal of Machine Learning and Computing, **5**, No. 2.4 (2015).
13. I. Gaag and V. Lida. Principles of Expert Systems, Addison-Wesley. ch. **3**, pp 131-139 (1991).