Versi Online tersedia di : http://jurnal.buddhidharma.ac.id/index.php/te

JURNAL TECH-E

| 2581-1916 (Online) |

Artical

Prototype of Timely Graduated Prediction for Buddhi Dharma University Using Comparison of C4.5 and Naïve Bayes Algorithms

Suwitno[1]

[1]*Universitas Buddhi Dharma, Manajemen Infomatika, Banten, Indonesia*

JEJAK PENGIRIMAN

Diterima: 15 Agustus 2017
Revisi Akhir: 20 Agustus 2017
Tersedia Online: 15 September 2017

KATA KUNCI

C4.5, Naïve Bayes, Timely graduated, Prediction

KORESPONDENSI

Telepon: 081311190089

E-mail: suwitno@ubd.ac.id

A B S T R A K

Set of data on database in academic information system of Higher Education is often not used effectively, whereas from set of data with technique of data mining can give knowledge that not yet known before. The purpose in this research is to know how to form a prediction model of students timely graduated rate at Buddhi Dharma University, Tangerang through students' timely graduated data. Prediction of student timely graduated using comparison of algorithm C4.5 and Naïve Bayes done with stage of data selection stage, data transformation, data mining and interpretation. This study retrieve of 300 training data and 90 data testing. Then the process of classification technique using decision tree method using C4.5 algorithm and posterior calculation using Naïve Bayes algorithm. Evaluation of classification performance is done to know how well the accuracy of a model is formed. Based on the research that has been done, the model is formed with the help of Rapidminer software, and calculated the average value of k-fold cross validation on testing up to k = 9 for algorithm C4.5 and Naïve Bayes. Testing is done with Confusion Matrix and ROC curves. Accuracy results obtained prove that Algorithm C4.5 yields 88.33% accuracy percentage and Naïve Bayes yield 86% accuracy percentage. Thus the C4.5 algorithm has a higher accuracy value than Naïve Bayes. This C4.5 algorithm can be used to prototype predictions of students timely graduated at Buddhi Dharma University Tangerang.

Introduction

The development of information technology is so advanced nowdays, cause the level of accuracy of data was needed in everyday life. Every information that exists becomes an important thing to determine every decision in certain situations. This causes the provision of information to be a means to be analyzed and summarized into knowledge of useful data when making a decision. In education system, student are important assets for an

educational institution and for this reason its important to pay attention to student's timely graduated. The percentage rise and fall of the students' ability to complete a timely study is one of the elements of university accreditation assessment. Therefore, it is necessary to monitor and evaluate students' tendency to pass on time or not.

I. METHOD

C4.5 Algorithm

The C4.5 algorithm was designed by J. Ross Quinlan, be named C4.5 because was descended from the ID3 approach to build decision trees. C4.5 was an algorithm that was suitable for classification problems in data mining and machine learning. C4.5 maps the attributes of the classes so that it can be used to find predictions for data that have not yet appeared. In the decision tree, the central node is an attribute of the tested data (tuple), the branch is the result of the attribute test, and the leaf will be a class that is formed.

Stages in the C4.5 algorithm were:

a. Note the label on the data, if it is all the same, then the leaves will be formed with the value of the entire data label.

b. Calculating the total value of the information (Entropy)

$$\text{Entropy} = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

c. Calculates the info value of each attribute (Info)

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j)$$

d. Calculates the gain value of each attribute (Gain)

$$\text{Gain}(A) = \text{Entropy} - \text{Info}_A(D)$$

e. After the decision of branch treee is formed, the calculation was performed again as in steps a through d. However if the branch has reached the maximum allowable branches, the leaf would be formed with the majority value of the data value.

Naïve Bayes

Bayesian approach is used to determine the likelihood of assumptions around it. In Bayesian statistics, parameters considered against random variables and data to be weighed against the likely outcome. Bayesian approach was first performed by the Reverend Thomas Bayes (1702-1761) on "Essay Towards Solving a Problem in the Doctrine of Chances" published in 1763 [2].

Phases in Naïve Bayes methods, ie:

a. Calculating the amount of data

b. Searching probability value (P)

$$P(x) = \frac{E}{n}$$

c. Searching a mean value (μ)

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

d. Finding the value of the standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

e. Classifying continuous data by the formula Density Gauss

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

f. Finding the value of posterior

$$\text{Posterior}(X) = \frac{P(X)\ P(\text{Atribut1}|X)\ P(\text{Atribut2}|X)\ P(\text{atribut..n}|X)}{p(X)\ P(\text{Atribut1}|X)\ P(\text{Atribut2}|X)\ P(\text{atribut..n}|X) + p(Y)\ P(\text{Atribut1}|Y)\ P(\text{Atribut2}|Y)\ P(\text{atribut..n}|Y)}$$

$$\text{Posterior}(Y) = \frac{P(Y)\ P(\text{Atribut1}|Y)\ P(\text{Atribut2}|Y)\ P(\text{atribut..n}|Y)}{p(X)\ P(\text{Atribut1}|X)\ P(\text{Atribut2}|X)\ P(\text{atribut..n}|X) + p(Y)\ P(\text{Atribut1}|Y)\ P(\text{Atribut2}|Y)\ P(\text{atribut..n}|Y)}$$

Accurate research will be obtained if a study has a large number of samples in a population. In this study, application testing for predictions of the timeliness of passing students and sampling methods using systematic sampling method. The graduation data used as the sample in this study was obtained from the database of Academic Information System (IAS) of Higher Education. The data used are 390 student data that have passed with 300 data as training data and 90 data as data testing. The purpose of the

classification algorithm is to find the relation between several variables belonging to the same class. The relation will be illustrated by rules in order to predict the class of data whose attributes are known. Classification C4.5 and Naïve Bayes are selected because this method has a high degree of accuracy and speed when applied to large amounts of data and can be used to predict the probability of membership of a class.

II. RESULT

An important step in this research is the use of C4.5 and Naïve Bayes algorithms to form a model. The result modeling would be comparative to find the best level of accuracy that would be used to determine the pattern of the ability of students who have the ability to pass timely or not. In this research, validation process was done to find and convert data to be used in algorithm of data mining method and got better accuracy and performance. In the dataset to be used this, the validation of data used was deleted incomplete or empty data that had no value (null). After that, attribute selection was done to select which attributes were needed from the dataset used in the process of analyzing the student's timely graduated at Buddhi Dharma University.

Table 1. Attribute List and Description

No	Attribute	Description
1.	*Waktu_Kuliah*	Time Session
2.	*Jenis_Kelamin*	Gender
3.	*Prodi*	Study Program
4.	*IPS1*	Grade Point (GP 1)
5.	*IPS2*	Grade Point (GP 2)
6.	*IPS3*	Grade Point (GP 3)
7.	*IPS4*	Grade Point (GP 4)
8.	*IPK_4*	Grade Point Average (GPA 4)
9.	*Total_ SKS_ Lulus4*	Amount of SKS that has passed until the 4th semester
10.	*Jur_Asl_Sekolah*	Major of school
11.	*Status_Asal_Sklh*	Graduated school status
12.	*Status_Pek_Ortu*	Parent's job status
13.	*Cuti*	Leave of absence amount

Graduation data of training and testing data collected has 390 records and 13 attributes. All of these attributes was collected and analyzed to view dominant data patterns and data types to assist in the process of selecting appropriate data mining methods and algorithms.

Table 2. Comparison of Accuracy and AUC

Prediction	C4.5 Algorithm	Naïve Bayes Algorithm
Success prediction on-time	170	167
Success prediction not on-time	95	91
Level of Accuracy	88,33%	86%
AUC	0,874	0,500

By looking at the comparison of accuracy and AUC, it could be seen that the C4.5 algorithm had the best accuracy and performance, so the rule generated by C4.5 algorithm serve as the rule for prototype making which can facilitate the prediction of the student's graduated timely.

Table View Plot View

accuracy: 88.33% +/- 2.33% (mikro: 88.33%)

	true NO	true YES	class precision
pred. NO	95	16	85.59%
pred. YES	19	170	89.95%
class recall	81.33%	91.40%	

Figure 1: Accuracy of C4.5 Algorithm

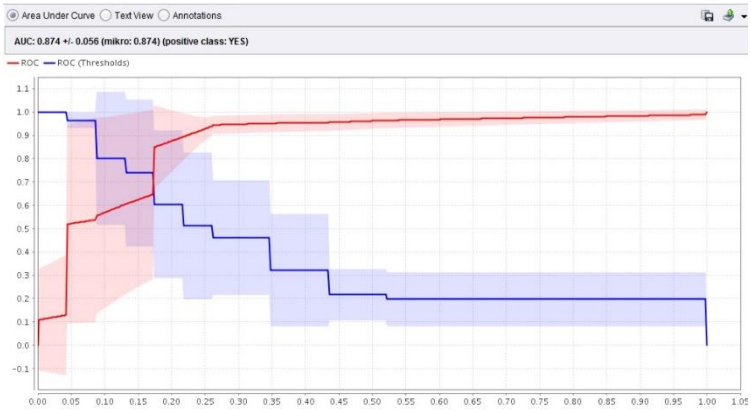

Figure 2: AUC Level of C4.5 Algorithm

Multiclass Classification Performance Annotations

Table View Plot View

accuracy: 86.00% +/- 1.63% (mikro: 86.00%)

	true NO	true YES	class precision
pred. NO	91	17	84.26%
pred. YES	25	167	86.98%
class recall	78.49%	83.08%	

Figure 3: Accuracy of Naïve Bayes Algorithm

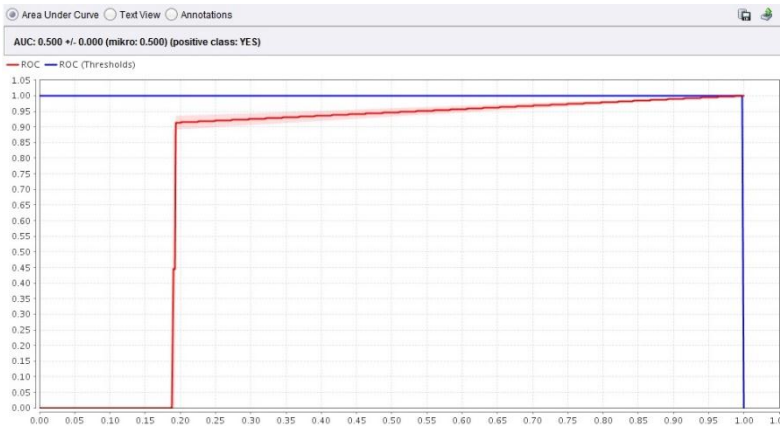

Figure 4: AUC Level of Naïve Bayes Algorithm

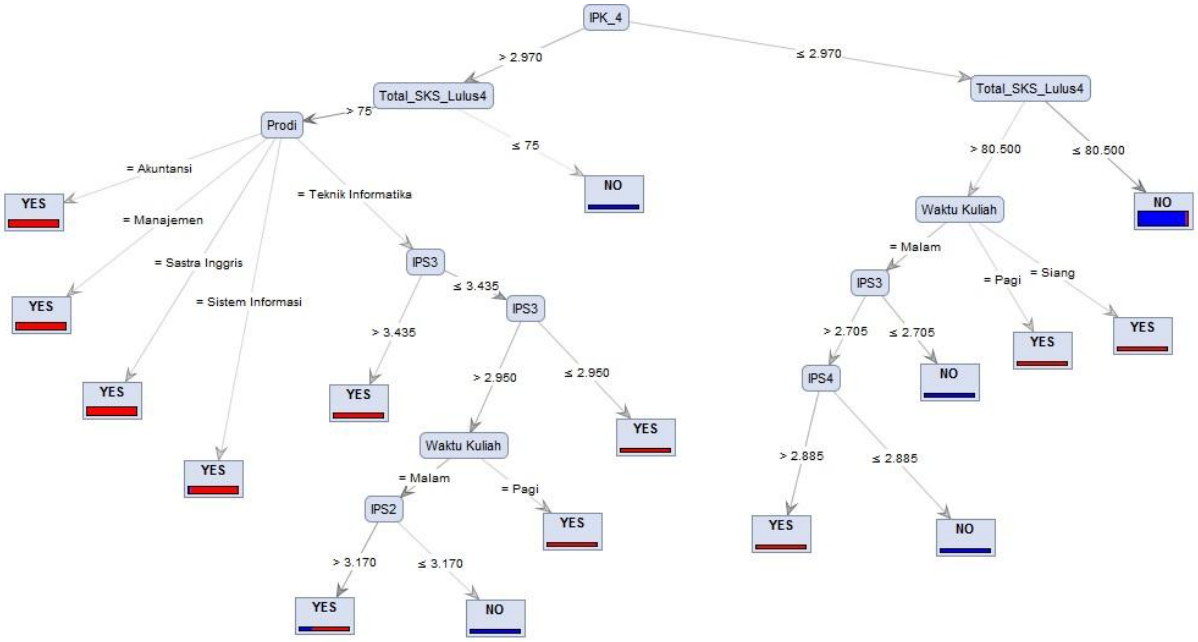

Figure 5: Decision Tree of C4.5 Algorithm

Rules generated from the decision tree based on training data are as follows:

```
TREE

IPK_4 > 2.970
|   Total_SKS_Lulus4 > 75
|   |   Prodi = Akuntansi: YES {NO=0, YES=32}
|   |   Prodi = Manajemen: YES {NO=1, YES=29}
|   |   Prodi = Sastra Inggris: YES {NO=0, YES=47}
|   |   Prodi = Sistem Informasi: YES {NO=2, YES=30}
|   |   Prodi = Teknik Informatika
|   |   |   IPS3 > 3.435: YES {NO=0, YES=13}
|   |   |   IPS3 ≤ 3.435
|   |   |   |   IPS3 > 2.950
|   |   |   |   |   Waktu Kuliah = Malam
|   |   |   |   |   |   IPS2 > 3.170: YES {NO=2, YES=5}
|   |   |   |   |   |   IPS2 ≤ 3.170: NO {NO=2, YES=0}
|   |   |   |   |   Waktu Kuliah = Pagi: YES {NO=0, YES=4}
|   |   |   |   IPS3 ≤ 2.950: YES {NO=0, YES=9}
|   Total_SKS_Lulus4 ≤ 75: NO {NO=3, YES=0}
IPK_4 ≤ 2.970
|   Total_SKS_Lulus4 > 80.500
|   |   Waktu Kuliah = Malam
|   |   |   IPS3 > 2.705
|   |   |   |   IPS4 > 2.885: YES {NO=0, YES=3}
|   |   |   |   IPS4 ≤ 2.885: NO {NO=2, YES=0}
|   |   |   IPS3 ≤ 2.705: NO {NO=7, YES=0}
|   |   Waktu Kuliah = Pagi: YES {NO=0, YES=5}
|   |   Waktu Kuliah = Siang: YES {NO=0, YES=5}
|   Total_SKS_Lulus4 ≤ 80.500: NO {NO=96, YES=3}
```

III. DISCUSSION

One of the most important things to determine the errors or deficiencies in the developed prediction application is to test. The test was conducted nine times which involved 90 data in addition to training data and using the confusion matrix method, ie the table used as a useful measuring tool to analyze how well the classification is right and wrong from the predictions made. Accuracy obtained can be calculated by the formula:

$$Accuracy = \frac{\sum correct\ predictions}{\sum correct\ and\ incorect\ predictions}$$

Whereas to calculate the Error rate can be calculated by the formula:

$$Error\ rate = \frac{\sum incorrect\ predictions}{\sum correct\ and\ incorect\ predictions}$$

Table 3. Testing Result

Testing	accuracy	error rate
K-1	90%	10%
K-2	100%	0%
K-3	80%	20%
K-4	80%	20%
K-5	90%	10%
K-6	80%	20%
K-7	100%	0%
K-8	80%	20%
K-9	90%	10%

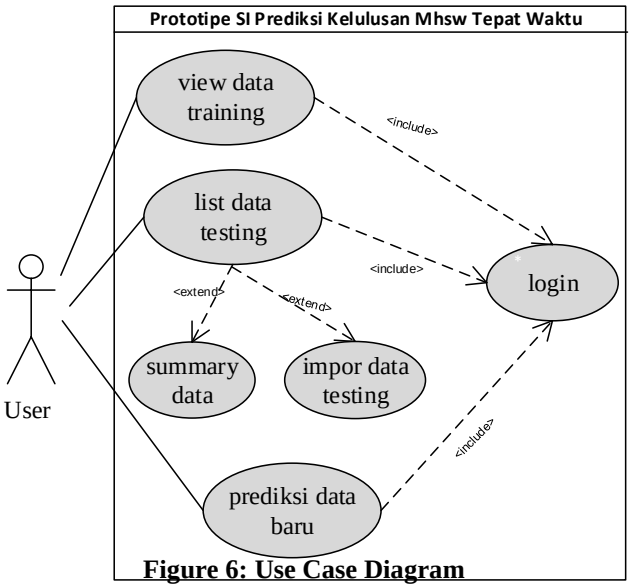

Figure 6: Use Case Diagram

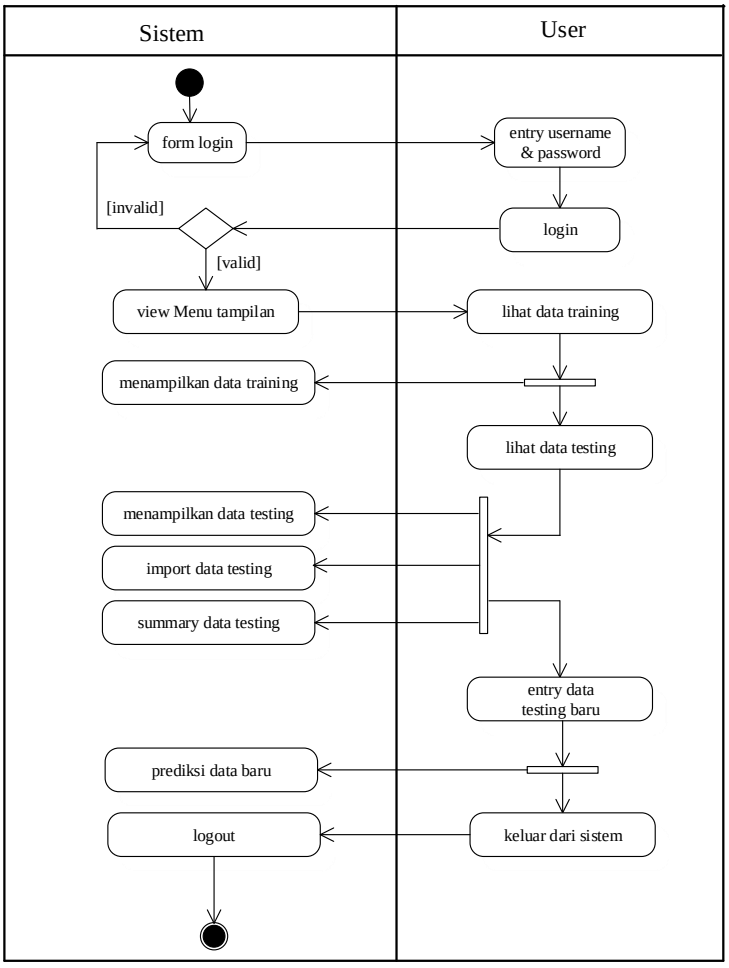

Figure 7: Activity Diagram

Of the nine experiments that have been done, then got the summary that is:
for accuracy level of:

$$\frac{90 + 100 + 80 + 80 + 90 + 80 + 100 + 80 + 90}{9} x\ 100\%$$
$$= \frac{790}{9} x\ 100\%$$
$$= \mathbf{87,77}\%$$

and for the rate of error rate of:

$$\frac{10 + 0 + 20 + 20 + 10 + 20 + 0 + 20 + 10}{9} x\ 100\%$$
$$= \frac{110}{9} x\ 100\%$$
$$= \mathbf{12,23}\%$$

IV. CONCLUSION

From the performance of measurements that had been done on two classification of method algorithms, the result of this research could be concluded that:

1. Classification of data mining method was appropriate to be implemented into the prototype of student predictions information system on time.
2. The C4.5 algorithm has the best accuracy between the two classification algorithms. So this algorithm will be implemented into the prototype timely graduated students prediction. It could be seen that the C4.5 algorithm had an accuracy value of 88.33% and AUC value of 0.874 which belongs to the category of good classification.
3. With this research helps the management of universities in conducting evaluation and monitoring of students who graduated timely or not.

REFERENCES

Alpaydın, E., *Introduction to Machine Learning. Second*., London: The MIT Press, 2012**.**

Chapman, P., *CRISP-DM 1.0, Step-by-step data mining guide*, 2000.

Gorunescu, F., *Data Mining Concepts, Models and Techniques*, Springer, 2011.

Hall, T., *A Systematic Literature Review on Fault Prediction Performance in Software Engineering*, 2011.

Vercellis, C., *Business Intelligence : Data Mining and Optimization for Decision Making*, John Wiley & Sons, Inc, 2009.

Witten et al., *Data Mining Practical Machine Learning Tools and Techniques 3rd*, Burlington: Elsevier Inc, 2011.

Wu, X. et al., *Top 10 algorithms in data mining*, A Chapman & Hall Book, 2008.

BIOGRAPHY

Suwitno received his Bachelor degree in Information System (S.Kom) from Buddhi Dharma University, Indonesia and Master Degree in Computer Sciences (M.Kom) concentration in Applied Computing Engineering from Budi Luhur University, Indonesia. He is a lecturer at the Department of Information Management, Faculty of Information Technology, Buddhi Dharma University.