Comparative performance between C4.5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment

Azizah, Erwina Nurul; Pujianto, Utomo; Nugraha, Eki; Darusalam

DOI
10.1109/ICEAT.2018.8693928

Publication date
2019

Document Version
Accepted author manuscript

Published in
2018 4th International Conference on Education and Technology, ICET 2018

Accepted Author Manuscript. Link to published article (IEEE) https://doi.org/10.1109/ICEAT.2018.8693928

Comparative performance between C4.5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment

1st Erwina Nurul Azizah
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
erawina25@gmail.com

2nd Utomo Pujianto
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
utomo.pujianto.ft@um.ac.id

3rd Eki Nugraha
Department of Computer Science
Universitas Pendidikan Indonesia
Bandung, Indonesia
ekinugraha@upi.edu

4th Darusalam
Faculty of Technology, Policy and Management
Delft University of Technology
Delft, Netherlands
d.darusalam@tudelft.nl

***Abstract*— People now trying to maximizing function of Virtual Learning Environment. Virtual Learning Environment, not only as a place to help learning system but now has become a place of learning itself. But with the change of the learning system, teacher now have difficulty to monitor the activity of the student and the learning material. Although there is data that is considered capable become a benchmark for students and the interaction with Virtual Learning Activity. This paper will make a data prediction using Naïve bayes and C4.5 Algorithm using the Web History data and the sum of webpage interaction of the students in Virtual Learning Environment.**

***Keywords*— Data Mining, Prediction, Naive Bayes, C4.5**

I. INTRODUCTION

Virtual learning environment is a dimension created to get information, and this information is information related to education. In the Virtual learning environment , students not only actively use, but also as actors who build the learning space [1]. Virtual learning environment are mostly used only as supporting learning systems, such as academic information system for example. In fact, in other countries the Virtual learning environment has been used as an alternative lecture for several universities [2]. But from the Virtual learning environment , we are not able to monitor student behavior during the learning period. We know in education not only the value of the learning subjects is important, but also the enthusiasm and attitude of students in the learning period. To be able to measure the activity of a student in a virtual learning environment, we can see a list of the virtual learning environment web pages that students have accessed and their interactions. In this study, the prediction method will be used to see whether activities on the Virtual learning Activity web page can be used as a reference in graduating students.

Virtual Learning Environment doesn't refer to website related to the theme of education. According to [3] Virtual Learning Environment is a social environment that is centered around education-themed interaction, where students are not only active, but also as actors, they help build the scope. Virtual Learning Environment certainly require computers and the internet as a major component in the learning system. The purpose of developing the Virtual Learning Environment is to maximize learning using the internet. The internet really provides a profitable potential to make flexible access to connecting communication between students and teachers. The Virtual Learning Environment provides the opportunity for teachers to make learning materials quickly without requiring additional work, because the internet function can be used in the Virtual Learning Environment.

Many classification algorithms are used to predict student graduation. One of them is the C4.5 algorithm which produces an accuracy of 87.5% [4]. There are also predictions with the Neural Network Algorithm and optimized with Particle Swarm, this method produces accuracy of 78.26% [5]. Evaluation of academic performance predicted by the Naive Bayes algorithm has a high accuracy rate of 82.8% [6]. Using two algorithms, C4.5 and CART. The C4.5 algorithm looks better with 85.61% accuracy [7]. From the algorithm described, it looks C4.5 and also Naive Bayes has a high level of accuracy. Therefore this research will use these two algorithms.

II. METHODS

A. Research Design

The design of the research will be carried out as in Fig. 1.

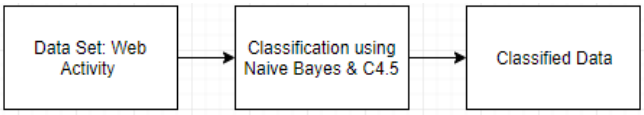

Fig. 1. Research design

Naive Bayes Classifiation is an simple classification algorithm which is able to calculate the probability, By calculating the frequency and combination of entries in the dataset. This algorithm uses Bayes theorem, assuming all attributes are independent variable values [9].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

The Formula Above make H is a Class and X is the Attribute of data. P(H) is a prior probability of H, and P(X) is the prior probability of X Attribute.

Naive Bayes is the most efficient Algorithm, which the complexity of O(*tn*) for classification induction of dataset with *t* and *n* variabel. As well as complexity of O(*cn*) to classify new data , *c* is the number of classes.

The main advantage of Naive Bayes is its ease of construction, without the need to make difficult iterative parameter schemes. In addition, Naive Bayes is also strong in issues that are irrelevant and attributes [10].

The C4.5 algorithm was presented in 1993 by Ross Quinlan, the goal is as a better method than ID3. To solve problems regarding ID3 weaknesses that are difficult to use for large data, C4.5 have "*Information Gain*" function. Where not only makes a new ratio, but also measures it [11].

The Gain Ratio C4.5 formula is as follows:

$$GainRatio(p,T) = \frac{Gain(p,T)}{SplitInfo(p,T)} \tag{2}$$

$$Splitinfo(p,test) = -\sum_{j=1}^{n} \left(\frac{j}{p}\right) X\ log\left(p'\left(\frac{j}{p}\right)\right) \tag{3}$$

P'(j/p) is the proportion of elements in p position. Take values from-j. Unlike entropy, the definition above does not depend on the distribution in different classes.

B. Research Data

The data is from The Open University England, in the form of an anonymus record of student learning activities in interacting with virtual learning environment [12]. The dataset used is grouped into 7 attributes, where one attribute is the class to be predicted. Explanation of the attributes to be used is as follows.

TABLE I. VARIABLE DESCRIPTION

Name	*Description*	*Attribute Type*
kategorial_region	Province origin of students	Varchar
highest_education	Student's last education	Varchar
kategorial_credit	Credit taken by students	Range numeric
disability	Student disability	Boolean
kategorial_web	Number of web pages accessed	Numeric
kategori_klik	Number of clicks on the web page	Numeric
Final_result	The end result whether students graduate or not	Boolean

To facilitate the mining process, the data that has already been converted becomes categorical that can be processed by the instrument. Data that needs to be categorized is "kategorial_region", "kategorial_credit". Here is a series of data that has been processed.

TABLE II. KATEGORIAL_REGION DATA

No	*Label*	*Count*
1	Another Country	4288
2	Middle Region	4746
3	North Region	3794
4	South Region	4172

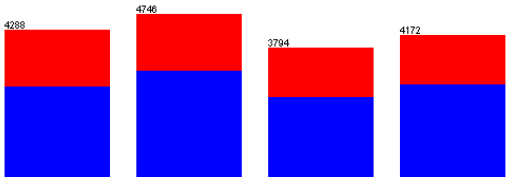

Fig. 2. Kategorial_Region

Table II and fig.2 shows the categorical data student region at UK. Another country label refer to nearest country like Scotland and Belgium.

TABLE III. HIGHEST_EDUCATION DATA

No	*Label*	*Count*
1	A Level or Equivalent	7558
2	HE Qualification	2470
3	Lower Than A Level	6675
4	No Formal quals	166
5	Post Graduate Qualification	131

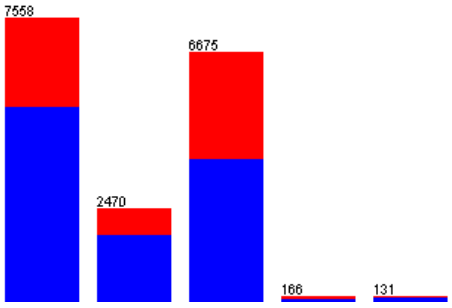

Fig. 3. Kategorial_Region

Table III and fig. 3 shows the categorial data of student degree. The UK has five educational qualifications.

TABLE IV. KATEGORIAL_CREDIT DATA

No	*Label*	*Count*
1	100 to 150	3319
2	50 to 100	10908
3	Above 150	496
4	Below 50	2277

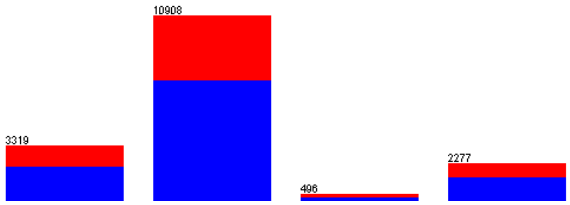

Fig. 4. Kategorial_Credit

Table IV and fig. 4 shows the categorical data of student credit. Categories divided into 4 with a maximum value of 150.

TABLE V. DISABILITY DATA

No	*Label*	*Count*
1	N	15561
2	Y	1439

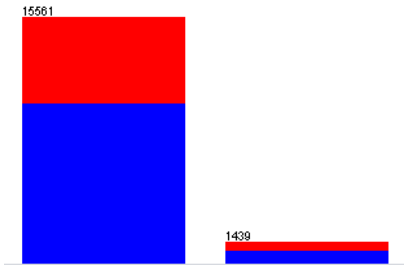

Fig. 5. Disability

Table V and fig. 5 shows the shows that students have physical disabilities or not.

TABLE VI. KATEGORIAL_WEB DATA

No	*Label*	*Count*
1	1000 to 2000	1215
2	Under 1000	15731
3	Above 2000	54

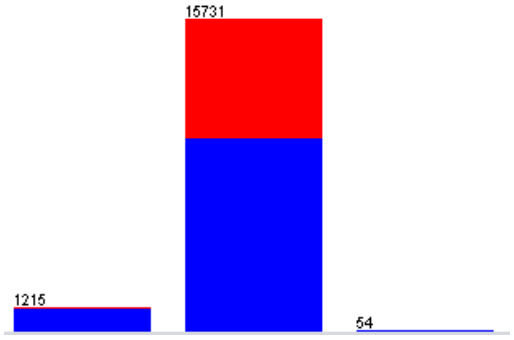

Fig. 6. Kategorial_web

Table VI and fig. 6 shows the categorical data of accumulated web pages accessed by students for each subject.

TABLE VII. KATEGORIAL_KLIKDATA

No	*Label*	*Count*
1	8000 to 16000	143
2	under 8000	16847
3	Above 16000	10

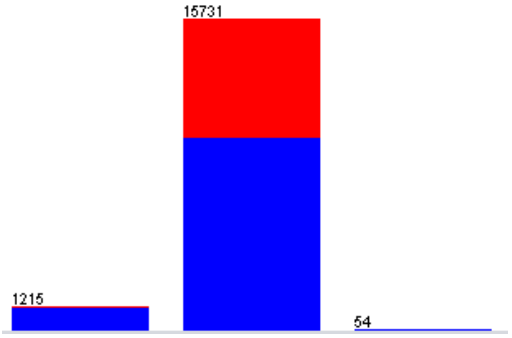

Fig. 7. Kategorial_klik

Table VII and fig. 7 shows the categorical data of accumulated click from students for each subject.

TABLE VIII. FINAL_RESULT DATA

No	*Label*	*Count*
1	Pass	10926
2	Fail	6074

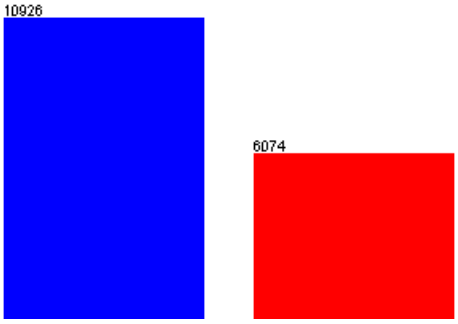

Fig. 8. Final_result

Table VIII and fig. 8 shows information whether graduating or not form their respective subjects.

C. *Evaluation*

From the data collected, there are a total of 17,000 data that are ready to be processed. We uses 10 Fold-Cross Validation for performance evaluation. The 10 Fold-Cross Validation is one of the K fold that is recommended for selecting the best model because it tends to provide estimates of accuracy that are less biased than the usual Cross Validation. The evaluation that will be carried out to test accuracy is *confusion matrix*. To find accuracy, precision, and recall using confusion matrix, the following formula is used.

$$Accurasy = \frac{TP+TN}{TP+TN+FN+FP} \quad (4)$$

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

Where the description of the formula is as follows.

- TP : *True Positive*, the correct amount of positive predictive data.
- FP : *False Positive*, the wrong number of positive predictive data.
- FN : *False Negative*, Incorrect negative prediction data.
- TN : *True Negative*, correct negative prediction data.

III. Results and Discussion

A. Naïve Baves classifier performance

Training data was calculated and taken 1700 data sets to see the performance of the prediction of the algorithm. The results of the algorithm performance are as follows.

TABLE IX. Classification Result

Correctly Classified Instances	1084	63.7647 %
Incorrectly Classified Instances	616	36.2353 %
Kappa statistic	0.0468	
Mean absolute error	0.4344	
Root mean squared error	0.4679	
Relative absolute error	94.0971 %	
Root relative squared error	97.0829 %	

TABLE X. Confusion Matrix Naïve Bayes

a	b	
1037	40	**a = Pass**
576	47	**b = Fail**

```
=== Predictions on test split ===

 inst#     actual  predicted error prediction
    1     2:Fail     2:Fail       0.517
    2     1:Pass     1:Pass       0.953
    3     1:Pass     1:Pass       0.686
    4     2:Fail     1:Pass   +   0.548
    5     1:Pass     1:Pass       0.686
    6     2:Fail     1:Pass   +   0.635
    7     1:Pass     1:Pass       0.663
    8     1:Pass     1:Pass       0.668
    9     1:Pass     1:Pass       0.643
   10     2:Fail     1:Pass   +   0.517
   11     1:Pass     1:Pass       0.69
   12     1:Pass     1:Pass       0.686
   13     1:Pass     1:Pass       0.548
   14     1:Pass     1:Pass       0.579
   15     1:Pass     1:Pass       0.64
```

Fig. 9. Comparison of Actual Data with Naïve bayes predictions

B. C4.5 classifier performance

TABLE XI. Confusion Matrix Naïve Bayes

Correctly Classified Instances	1081	63.5882 %
Incorrectly Classified Instances	619	36.4118 %
Kappa statistic	0.0232	
Mean absolute error	0.4353	
Root mean squared error	0.4689	
Relative absolute error	94.2885 %	
Root relative squared error	97.2894 %	

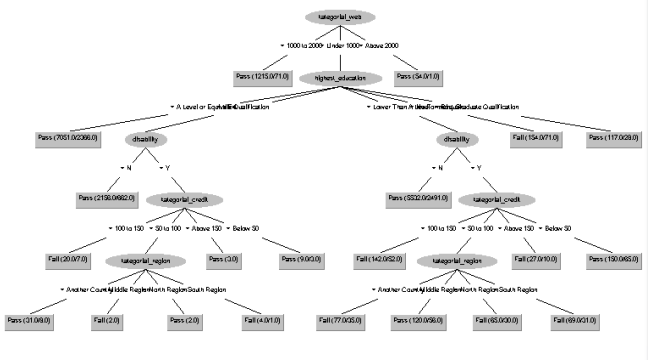

Fig. 10. C4.5 Tree generated from the training data

TABLE XII. Confusion Matrix C4.5

a	b	
1059	18	**a = Pass**
601	22	**b = Fail**

```
=== Predictions on test split ===

    inst#     actual  predicted error prediction
        1     2:Fail     2:Fail       0.536
        2     1:Pass     1:Pass       0.939
        3     1:Pass     1:Pass       0.691
        4     2:Fail     1:Pass   +   0.549
        5     1:Pass     1:Pass       0.691
        6     2:Fail     1:Pass   +   0.669
        7     1:Pass     1:Pass       0.669
        8     1:Pass     1:Pass       0.669
        9     1:Pass     1:Pass       0.669
       10     2:Fail     2:Fail       0.554
       11     1:Pass     1:Pass       0.669
       12     1:Pass     1:Pass       0.691
       13     1:Pass     1:Pass       0.549
       14     1:Pass     1:Pass       0.669
       15     1:Pass     1:Pass       0.669
```

Fig. 11. Comparison of actual and predictive data C4.5

C. Equations

The results of the two algorithm's confusion matrix are calculated Accuracy, Precision and Recall.

TABLE XIII. COMPARISON OF ACCURATION, PRECISION AND RECALL

Comparison	**Naive Bayes**	**C4.5**
Accurasy	63,8 %	63,6 %
Presisi	64,3 %	63,8 %
Recall	96,3 %	98,3 %

IV. CONCLUSION

From the data that has been processed using two algorithms. The results obtained, both algorithms have almost the same level of accuracy. The accuracy of Naive Bayes is superior to 63.8% even though the results from C4.5 only differ 0.2% from the accuracy of Naive Bayes. Even so, Recall from C4.5 Algorithm is better with a value of 98.3%. From the results of the C4.5 decision tree we know that web pages are an important factor in student graduation, therefore these variables can be used as supporting variables for the value of students in the Virtual Learning Environment. Concerns about the trees formed, it was seen that students who go to the website in 1000-2000 times were categorized as students who passed the lecture. This data can be used as proof that students' VLE activities are able to be used as a reference whether students can graduate or not.

REFERENCES

[1] Dillenbourg, Pierre, Daniel Schneider, and Paraskevi Synteta. "Virtual learning environments." *3rd Hellenic Conference" Information & Communication Technologies in Education"*. Kastaniotis Editions, Greece, 2002.

[2] Distance Education and Virtual Learning Environments." *The University of Edinburgh*, The University of Edinburgh, 8 Sept. 2015, www.ed.ac.uk/records-management/data-protection/guidance-policies/distance-education.

[3] Kamagi, David Hartanto, and Seng Hansun. "Implementasi Data Mining dengan Algoritma C4. 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa." *vol. VI* 1 (2014): 15-20.

[4] Kusumawati, Dewi, Wing Wahyu Winarno, and M. Rudyanto Arief. "Prediksi Kelulusan Mahasiswa Menggunakan Metode Neural Network dan Particle Swarm Optimization." *SEMNASTEKNOMEDIA ONLINE* 3.1 (2015): 3-8.

[5] Nugroho, Yuda Septian. "Data Mining Menggunakan Algoritma Naive Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro." *Dian Nuswantoro Fakultas Ilmu Komputer Skripsi* (2014).

[6] Rahmayuni, Indri. "Perbandingan Performansi Algoritma C4. 5 dan Cart Dalam Klasifiksi Data Nilai Mahasiswa Prodi Teknik komputer POLITEKNIK NEGERI PADANG." *Jurnal TeknoIf*2.1 (2014).

[7] Dillenbourg, Pierre, Daniel Schneider, and Paraskevi Synteta. "Virtual learning environments." 3rd Hellenic Conference" Information & Communication Technologies in Education". Kastaniotis Editions, Greece, 2002.

[8] D. T. Larose, "Data Mining Methods and Models," p. 385, 2007.

[9] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *International journal of computer science and applications* 6.2 (2013): 256-261.

[10] Soria, Daniele, et al. "A 'non-parametric'version of the naive Bayes classifier." *Knowledge-Based Systems* 24.6 (2011): 775-784.

[11] Hssina, Badr, et al. "A comparative study of decision tree ID3 and C4. 5." *International Journal of Advanced Computer Science and Applications* 4.2 (2014).

[12] Kuzilek, Jakub, Martin Hlosta, and Zdenek Zdrahal. "Open university learning analytics dataset." *Scientific data* 4 (2017): 170171.