Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 3 No. 1 (2023) 40 - 44 E-ISSN: 2809-8064

Comparison of C4.5 and Naive Bayes Algorithm Methods in Prediction of Student Graduation on Time (Case Study: Information Systems Study Program)

Disty Dikriani[1*], Alvina Tahta Indal Karim[2]

[1,2*]Information Systems, Faculty of Informatics, Telkom Purwokerto Institute of Technology

[1*]19103047@ittelkom-pwt.ac.id, [2]19103075@ittelkom-pwt.ac.id

Abstract

In tertiary institutions, students become one of the important parameters in the evaluation of study program organizers. Prediction of student graduation is a special concern to know, early identification for students is needed as an important action. Information processing to predict student graduation is by implementing data mining. The implementation of data mining can be applied if a university, especially a study program, does not yet have an early classification in achieving student graduation on time. The ITTP Information System study program is one of the study programs that does not have an early identification of student graduation on time. Determination of graduation for SI ITTP Study Program students includes GPA, TOEFL scores, and total credits. The purpose of this research is to find out which attributes have the most influence in predicting graduation of ITTP IS Study Program students. The method used in this prediction is by using the classification of the C4.5 Algorithm and Naïve Bayes. The classification is used to determine which attributes influence predicting student graduation on time and to compare the two classification methods. The results obtained are the training set size 70% which has the best accuracy when compared to other training set sizes. Comparing the accuracy between the two methods, it is known that the C4.5 algorithm has good accuracy when training set size is 70% and Naïve Bayes has higher accuracy when training set size is 75%. Decision tree C4.5 interprets that the most influential attribute is the GPA as the root of the decision tree to predict student graduation on time. The research is expected to be used as a reference for the ITTP IS Study Program in formulating student graduation policies on time and as a reference for further researchers in predicting in the same field.

Keywords: Data Mining, Classification, Graduation, C4.5 Algorithm, Naïve Bayes

1. Introduction

Education is the learning of knowledge, skills and habits of a group of people that are passed down from one generation to the next through teaching, training, or research. Education is education in English which is absorbed in Indonesian into Education [1]. Education is important for all human beings from children to adults. The stages of education level start from pre-school such as Early Childhood Education (PAUD) to the level of lectures or colleges.

In tertiary institutions, students become one of the important parameters in the evaluation of the organizers of the study program. Monitoring of student attendance, student achievements, increasing student competence, graduation ratio to the total number of students and achievement of graduate profiles, should receive serious attention. The quota of students accepted every year is increasing, but not all students can graduate on time according to the specified study period, resulting in an accumulation of the number of students who do not pass according to their study period [2]. These problems make information about student graduation predictions a special concern to know, early identification for students is needed as an action in overcoming these problems.

Information processing to predict student graduation can be done by implementing data mining. Data Mining is an analysis of reviewing data sets to find unexpected relationships and can summarize data in a way that is different from the previous one that is understandable and useful for data owners. Another definition of data mining is a term used to describe the discovery of knowledge in data mining. Grouping in data mining, namely, Description, Estimation, Classification, Prediction, Association and Clustering [3],[4]. Many studies use data mining techniques. One of the

Received: 18-08-2022 | Revised: 01-02-2023 | Published: 04-02-2023

universities that use this technique is IT Telkom Purwokerto. IT Telkom Purwokerto has 3 faculties, one of which is the Faculty of Informatics. The Faculty of Informatics has several study programs, namely Information Systems, Informatics Engineering, Software Engineering and Data Science. This research focuses on Information Systems study program.

The problems that exist in the Information Systems study program are that it is known that currently there are still batches of 2017 and 2018 who have not graduated, this happens due to the lack of carefulness of students who feel confused about the Final Project, lack of consultation with supervisors, students who are tired of many revisions., as well as students who are trapped in a comfort zone. The Information Systems Study Program has been identified as having never applied data mining classification techniques in determining student graduation factors on time, so this research and prediction is expected to be able to find factors that dominate or affect Information System graduation on time.

Many studies have been carried out on student graduation, one of which is "Application of Application of the C45 Algorithm to Predict Web-Based Student Graduation" which concludes that Data mining using the C4.5 Algorithm can be used in predicting early graduation on time or not for students and can be used as program evaluation material. studies in improving student learning systems. This study uses a comparison of 2 methods between the C4.5 Algorithm and Naïve Bayes methods with the aim of knowing the level of accuracy of the two methods.

2. Research Methods

The flow chart of this research can be seen in Figure 1 as follows.

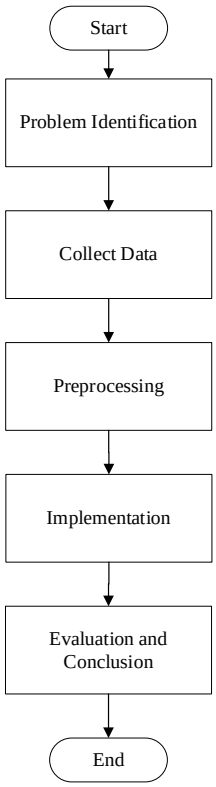

Figure 1. Research Flow

2.1. Identify the Scope of the Problem

The problems experienced by the Information Systems Study Program (IS Study Program) of the Telkom Institute of Technology Purwokerto (ITTP) were identified by conducting interviews with the Chair of the ITTP SI Study Program, namely Mrs. Dwi Mustika Kusumawardhani, S.Kom., M.Kom. The results of interviews conducted with the head of the ITTP SI Study Program, it is known that the reason that greatly affects students not graduating on time is the students themselves. Among them are students who disappeared after receiving many revisions, students who were confused but did not ask their supervisors, and students who were trapped in their comfort zones. In addition to these reasons, it is known that from a number of students who are currently carrying out their final assignments, only a few students have completed the obligations of Final Projects 1 and 2.

2.2. Data Collection

Data for ITTP IS Study Program students was obtained from the ITTP Faculty of Informatics Academic Information System (SIA). The data will be processed using two classification methods, namely student data from the 2017 and 2018 batches who have completed

the Final Project as training data, as well as student data from the 2019 batch who are carrying out the Final Project to be used as testing data. The data collected is about 64 student data who have carried out graduation, some of the data can be seen in Table 1.

Table 1. Judicium Student Data

Count SKS	IPK	TOEFL	Length of Study
144	3,74	464	yes
144	3,3	557	no
144	3,33	627	no
144	3,44	454	no
144	3,17	547	no
144	3,24	464	no
144	3,12	534	no
144	3,27	507	no
144	3,6	534	yes
144	3,58	467	yes
144	3,77	524	yes
144	3,51	494	yes
144	3,5	530	yes
144	3,76	474	yes
144	3,86	520	yes
144	3,88	484	yes
144	3,48	457	yes
144	3,47	624	no
144	3,57	560	no
144	3,61	457	yes

2.3. Analysis Method

The data that has been collected from the SIA FIF ITTP will be analyzed using two classification methods, namely the C4.5 Algorithm and Naïve Bayes.

The process of the C4.5 Algorithm is as follows:

a. Select root attribute.
   Determination of the attribute as the root by calculating the value of entropy and gain. The Entropy formula for calculating each attribute can be seen in Formula 1 [7].

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

   Dimana:
   Si = number of attributes
   S = total number
   n = number of partitions
   Pi = proportion of Si to S

   The entropy calculation in each attribute determines the highest gain value so which attribute has the highest gain value will be the root of the C4.5 decision tree [5],[6]. The Gain formula for the relationship between 2 variables can be seen in Formula 2.

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

   Dimana:
   S = Case Set
   A = Attribute
   n = Number of attributes a
   |Si| = number of cases on partition i
   |S| = number of cases in S

b. Create a branch for each value.
c. Divide cases in branches.
d. Repeat the process for each branch until all cases on the branch have the same class.

The Bayesian theoretical process that is calculated is P(H|X), which is the probability of the hypothesis H based on condition X in Formula 3 [8],[9].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (3)$$

Dimana:
X = Sample data with unknown class (label)
H = Hypothesis that X is data with class (label) C.
P(H)= Probability of the hypothesis H
P(X)= Probability of X observed
P(X|H)= Probability X, based on the conditions in hypothesis H.

2.4. Results and Discussion

The results obtained from the comparison of the two methods are the results of accuracy and the results of which attributes are the most influential in predicting student graduation on time from the two methods.

2.5. Testing with Software

Testing is done with one of the Data Mining software, namely Orange. In the Orange software, the data that has been analyzed is then tested to get the accuracy value of each method to find out what the formation of the C4.5 decision tree looks like, and to find out which method is the most appropriate for predicting the graduation of ITTP IS Study Program students.

3. Results and Discussion

Data that has passed the C4.5 and Nave Bayes algorithm processes, then tested using the orange software.

3.1. C4.5 Algorithm Decision Tree

The data collected and after being processed using the Data Mining process, then tested with the orange software, following the interpretation of the C4.5 decision tree in Figure 2.

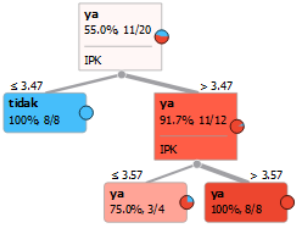

Figure 2. Decision Tree C4.5

The attribute that is the root of the decision tree is the GPA and shows that the data that has been collected can predict that SI Study Program students can graduate on time using training data, namely student data who have graduated.

3.2. Results of Accuracy, Precision, Naïve Bayes Recall and C4.5 Algorithm

Accuracy is how accurate the model or method is in classifying correctly. Precision shows the accuracy between the requested data and the prediction results provided by the model or method used. Recall describes the success of the model in retrieving information. The display of accuracy, precision, and recall results using two classification methods, namely the C4.5 Algorithm and Naïve Bayes can be seen in Figures 3, 4, and 5. Training set size 66% shows that the accuracy of Naïve Bayes is higher than Tree, but when Training set size 70%, Tree accuracy value is higher than Naïve Bayes.

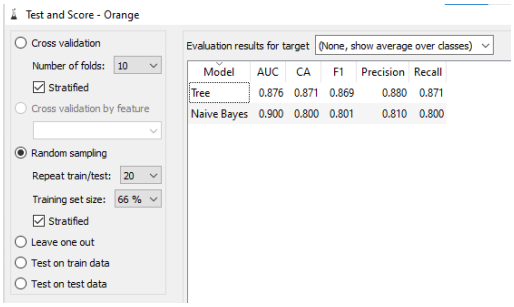

Figure 3. Training Set Size 66%

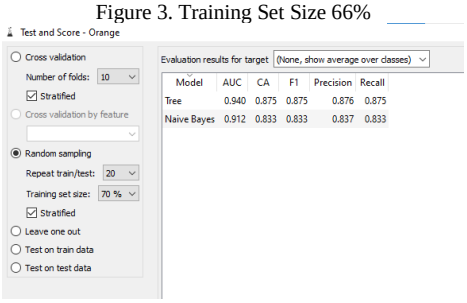

Figure 4. Training Set Size 70%

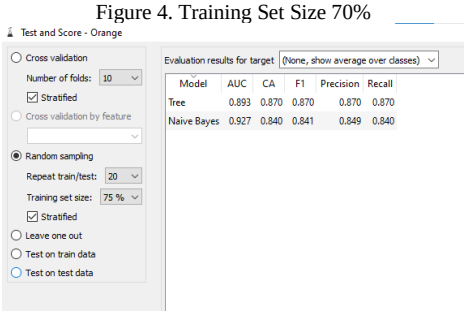

Figure 5. Training Set Size 75%

3.3. Confusion Matrix

Confusion Matrix is a measurement to produce two or more classes with a table of 4 different combinations of predicted values and actual values. There are 4 terms that represent the results of the classification process in the confusion matrix, namely True Positive, True Negative, False Positive, and False Negative.

Calculation of accuracy, precision, and recall values based on confusion matrix is by using Formula 3 as follows [10].

Accuracy = (TP + TN) / (TP+FP+FN+TN)

= (33 + 54) / (33+7+4+54)

= 0,887 * 100% = 88,7%

Precision= (TP) / (TP + FP)

= 33 / (33 + 7)

= 0,825 * 100% = 82,5%

Recall = TP / (TP + FN)

= 33 / (33+ 6)

= 0,846 * 100% = 84,6%

F-1 = (2 * Recall * Precision) /(Recall + Precision)

= (2 * 0,84 * 0,82) / (0,84 + 0,82)

= 1,3776 / 1,66

= 0,829 * 100% = 82,9%.

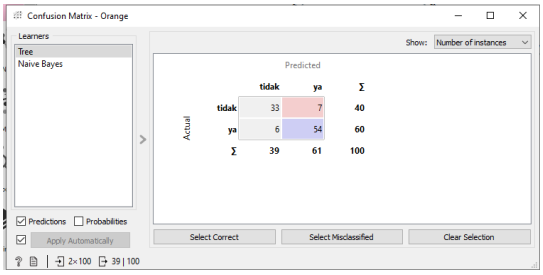

Figure 6. Confussion Matrix Decision Tree C4.5

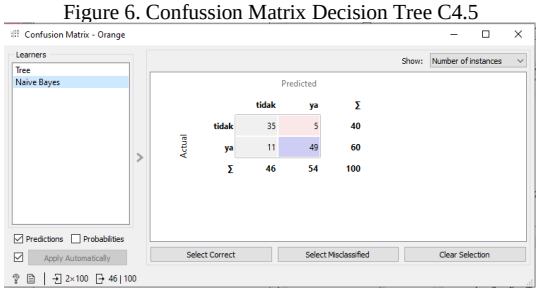

Figure 7. Confussion Matrix Naive Bayes

4. Conclusion

From the results of research and implementation that has been carried out, it can be concluded that the application of the c4.5 and Naïve Bayes algorithms in predicting student graduation on time has a fairly good level of accuracy and is said to be suitable by knowing the attributes that affect student graduation such as GPA, number of credits and TOEFL. It is known that the Naïve Bayes method has a higher accuracy value than the C4.5 algorithm method with a training set size of 66% and 75%. The most influential attribute in predicting student graduation is the GPA attribute seen from the C4.5 decision tree on orange software. This research is expected to be a reference for further research and can be developed again in a comparison of two or more classification methods in prediction in the same field as this research.

References

[1] Zakky, "ZonaReferensi.com," Zona Referensi Ilmu Pengetahuan Umum, 26 Maret 2020. [Online]. Available: https://www.zonareferensi.com/pengertian-pendidikan/#:~:text=Pengertian%20pendidikan%20%E2%80%93%20Pendidikan%20adalah%20pembelajaran%20pengetahuan%2C%20keterampilan%2C,yang%20juga%20diserap%20dalam%20bahasa%20Indonesia%20menjadi%20edukasi[Acesso em 2022 Juli 31].

[2] M. Ridwan, "Sistem Rekomendasi Proses Kelulusan Mahasiswa Berbasis Algoritma Klasifikasi C4.5," Jurnal Informatika, vol. 2, nº 1, pp. 105-111, 2017.

[3] F. S. d. R. M. A. Sulaeman, "Aplikasi Penerapan Algoritma C45 untuk Meprediksi Kelulusan Mahasiswa Berbasis Web," Jurnal Media Teknik & Sistem Industri, vol. V, nº 1, pp. 41-54, 2021.

[4] R. T. Vulandari, Data mining : teori dan aplikasi rapidminer, Yogyakarta: Gava Media, 2017.

[5] M. M. A. e. a, Data Mining Algoritma C4.5, Semarang: Universitas Negri Semarang, 2019.

[6] Y. P. S. Hutagaol, "Penerapan Metode Algoritma C4.5 untuk Menentukan Kualitas Telur Ayam Australia Terbaik," J-Com (Journal of Computer), vol. I, no 3, pp. 159-166, 2021.

[7] Charisma, Rifqi Alfinnur, et al. "Analisis Penerapan Metode Ensembled Learning Decision Tree Pada Klasifikasi Virus Hepatitis C." Journal of Computer System and Informatics (JoSYC), vol. 3, no. 4, pp. 405-409, 2022.

[8] Nugraheni, Annisa, et al. "Perbandingan Performa Antara Algoritma Naive Bayes Dan K-Nearest Neighbour Pada Klasifikasi Kanker Payudara." *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 2, no. 1, pp. 11-20, 2022.

[9] Alamsyah, Rizki, et al. "Sentiment Analysis Destinasi Wisata Berdasarkan Opini Masyarakat Menggunakan Naive Bayes." *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 2, no. 2 pp. 64-74, 2022.

[10] Ramadhan, Nur Ghaniaviyanto, Aji Gautama Putrada, and Maman Abdurohman. "Improving smart lighting with activity recognition using hierarchical hidden markov model." Indonesia Journal on Computing (Indo-JC), vol. 4, no. 2, pp. 43-54, 2019.