Journal of Soft Computing Exploration

Homepage: htttps://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Prediction of PTIK students' study success in the first year using the c4.5 algorithm

Asri Astuti[1*], Dwi Maryono[2], Febri Liantoni[3]
[1, 2, 3]Department of Informatics and Computer Engineering Education, Universitas Sebelas Maret, Indonesia

Article Info

Article history:

Received November 1, 2023
Revised December 28, 2023
Accepted March 4, 2024

Keywords:

Algorithm C4.5
Data mining
Study success

ABSTRACT

The purpose of this study is to determine the factors that influence the success of student studies in the first year through data mining research using the C4.5 algorithm. This research is a type of quantitative research. This research uses student data of a study program as much as 85 data which will be processed using the Weka application. The data obtained will then be processed using the C4.5 data mining method to produce a decision tree containing rules to predict the success of student studies in the first year. The best result using percentage-split 80% obtained an accuracy of 82.35% as well as the rules contained in the decision tree. The most important factor in determining the success of studies in first-year students is the selection of college entrance pathways. Other factors that become other determinants are education before college, intensity of communication with friends, class year, intensity of off-campus organizations, and plans to change study programs.

Corresponding Author:

Asri Astuti,
Department of Informatics and Computer Engineering Education,
Universitas Sebelas Maret, Indonesia
Email: asriastuti11@student.uns.ac.id
https://doi.org/10.52465/joscex.v5i1.237

1. INTRODUCTION

Education is one of the needs in human life. Education aims to prepare students to face life and adapt well to the environment. Education is carried out through basic education, secondary education and higher education. Education has an important role in efforts to improve the quality of human resources (Human Resources). Human resources (HR) management is predicated on the notion that human capital is a strategic asset that produces competitive advantages [1].

Educational institutions are an important part of society and play an important role in the growth and development of a nation [2]. One of the forums for the quality Human Resources (HR) training process are universities [3]. To achieve higher education, individuals continue their studies at university. Individuals complete learning activities to get a bachelor's degree or diploma to be able to guide their future work and career. Before entering university, prospective students first go through several stages of the selection process. Once an individual is accepted into a particular study program at university, it can be assumed that they have the potential and academic ability as well as the capability to study at that university.

First-year students are transitional students from school to university. Students are the most important part ofevaluating the success of the implementation of study programsin tertiary institutions [4]. Students learning performance is one of the core components for assessing any educational systems [5]. At the university

level, students are actively involved in teaching and learning activities through existing media such as journals, the internet and the library. All assignments given at university usually require students to look for other literature and develop their own thinking in order to complete the assignment effectively.

In the first year, the average learning outcomes of students raise several questions that have an impact on the following semester. There are many factors that can influence a student's academic achievement [6]. The academic achievements obtained by students are considered the success of a student and the learning system at that institution [7]. In higher education, the measure of academic achievement is the Grade Point Average (GPA) achieved by each student in a certain period of time. GPA or Cumulative Achievement Index is the sum of all final grades obtained by alumni for all courses divided by the number of credits taken during their studies [8]. The Department of PTIK UNS is one of the largest, which has many enthusiasts every year. However, the capacity for the first year is limited. The limited capacity of each department results in a strict selection process, so that students who are accepted are those who have met the established admission criteria.

Universities provide good learning concepts to students in order to achieve satisfactory academic performance. However, students' academic success varies widely, even when they learn the same methods with the same environment and tools. Universities as educational institutions already have a lot of academic and administrative data, but only a small part of the data is used for the preparation of self-evaluation. Data collected from year to year needs to be analyzed to open up opportunities for extracting useful information for alternative college management decisions.

So far, many modeling applications have been developed that can be used to predict student study success based on variables such as study period [9]–[13] and performance [9] using machine learning. Machine learning model development can take the form of modeling for supervised learning, unsupervised learning, and reinforcement learning [14]. Machine learning is a part of artificial intelligence that focuses on creating models so that computers can solve regression, classification, or clustering problems [8].

In this research, data processing can be done using data mining methods with the C4.5 decision tree algorithm. C4.5 algorithm, which is generated from ID3 algorithm, was introduced by J. Ross Quinlan in 1993 [15]. In ID3, the induction decision tree can only be performed on categorical features (nominal/ ordinal), while numeric types (internal / ratio) cannot be used. It is the most influential decision tree algorithm at present [16]. Data Mining makes it easier for educational institutions to identify based on what factors influence students to get a good Grade Point Average in class [17]. In addition to analyzing large amounts of observational data, data mining can also process high-dimensional data and data with different properties. The application of data mining is not only done in the fields of health, technology and industry but also applied in the field of education using EDM.

2. METHOD

This research is a type of quantitative research. This research uses student data of a study program as much as 85 data which will be processed using the Weka application. The data obtained will then be processed using the C4.5 data mining method so as to produce a decision tree containing rules to predict the success of student studies in the first year. The method used in this research is using KDD (Knowledge Discovery of database). The stages in the KDD method can be seen in Figure 1 below.

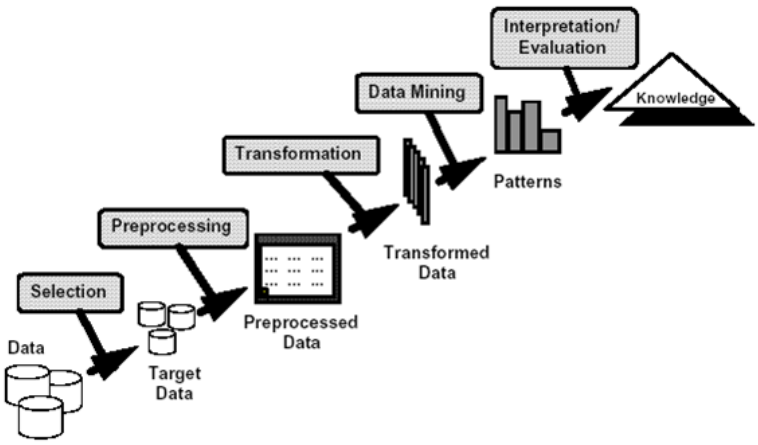

Figure 1. Knowledge discovery of database method

The following is an explanation of each stage in the KDD process in research to predict the success of PTIK students' studies in the first year:

Data Selection

Data collection is obtained from students of PTIK FKIP UNS class of 2021 and class of 2022. Data selection from the data set needs to be done before extracting information in KDD begins. Not all existing data is used in the data mining process. The available data structure is Name, NIM, 1st semester IP, 2nd semester IP, GPA, class year, gender, education before college, entry path, regional origin, plans to change study programs during the first year, availability of infrastructure facilities, and other student personal data. As for the existing attributes, one attribute will be selected and added as a class attribute, namely GPA.

Preprocessing

It is necessary to do a cleaning process or data checking to avoid data duplication and correct errors in the data such as printing errors or typography. The identification results found 2 double data. As well as improvements to the GPA attribute in some data that previously used commas replaced with periods such as 3.75 to 3.75.

Transformation

At this stage the concept of hierarchy is used to simplify the GPA value attribute data. GPA data is in the form of numbers so it needs to be categorized. The large number of attribute values will result in an increase in the number of branches in the decision tree, so it is necessary to simplify the value so that the size of the resulting decision tree can be reduced.

Data Mining

Data mining is the process of finding patterns in selected data using certain methods, namely using the C4.5 algorithm.

Evaluation

Interpretation or evaluation is a pattern of information obtained from the data mining process and displayed in a form that is easy to understand. Evaluation is done by looking at information from the confusion matrix to determine the accuracy of the predictions obtained from the algorithm used.

Knowledge

Furthermore, the presentation stage of the patterns found is used to produce actions or steps that must be taken from the analysis obtained in the form of knowledge that can be understood by everyone. In this presentation, visualization helps display the results of data mining.

3. RESULTS AND DISCUSSIONS

Data Description

The data used in this study are data obtained from a questionnaire containing questions with respondents of PTIK students class 2021 and 2022. Data that has been collected via google form totaling 85 respondents.

Data Preprocessing

a) Data Cleaning

Before the data enters the testing stage, preprocessing is needed to refine the data, including removing duplicate data and correcting inappropriate data values. The identification results found 2 duplicate data. As well as improvements to the GPA attribute in some data that previously used commas replaced with periods such as 3.75 to 3.75.

b) Data Transformation

At this stage the concept of hierarchy is used to simplify the GPA value attribute data. GPA data is in the form of numbers so it needs to be categorized. The large number of attribute values will result in an increase in the number of branches in the decision tree, so it is necessary to simplify the value so that the size of the resulting decision tree can be reduced. The data simplification process using the hierarchy concept is done by sorting the data from smallest to largest.

Testing Results

The application used to process the data that has been collected is using the Weka version 3.9.6 application. The data that has been collected is entered into the Weka application, then deleting attributes that will not be processed such as full name, NIM, IP semester 1, IP semester 2, and GPA of each respondent.

After the data has gone through data cleaning and data processing, it is ready to be used in the data mining process. This process is done to calculate the data by testing various parameters. Each result of the parameter changes will be recapitulated to find the optimal parameter to achieve the highest data accuracy. The changes made include percentage-split changes, unpruned status changes to true or false and minNumbObj changes.

The test results using Weka show that the test results that have pruned parameters and minNumObj = 2 have the highest accuracy of 82.35 with a percentage-split of 80%.

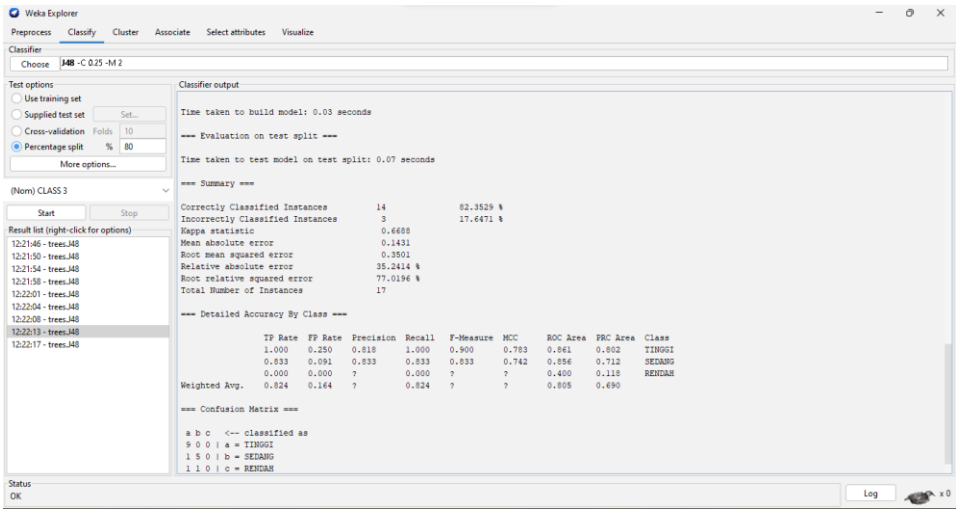

Figure 2. 80% percentage-split testing results

Based on Figure 2. the test results show that using percentage-split 80% gets the highest accuracy value with an accuracy value of 82.35%. The matrix formed for the calculation of accuracy, sensitivity and precision tests is as follows:

1) Accuracy Test
   Accuracy = 14/(14+3) x 100% = 14/17 x 100% = 82.35%
2) Precision Test
   Precicion = 9/(9+3) = 9/11 = 0.818
3) Sensitivity Test
   Recall = 14/(14+0) = 14/14 = 1,000

From the 80% percentage-split test results, the pruned decision tree is shown in Figure 3.

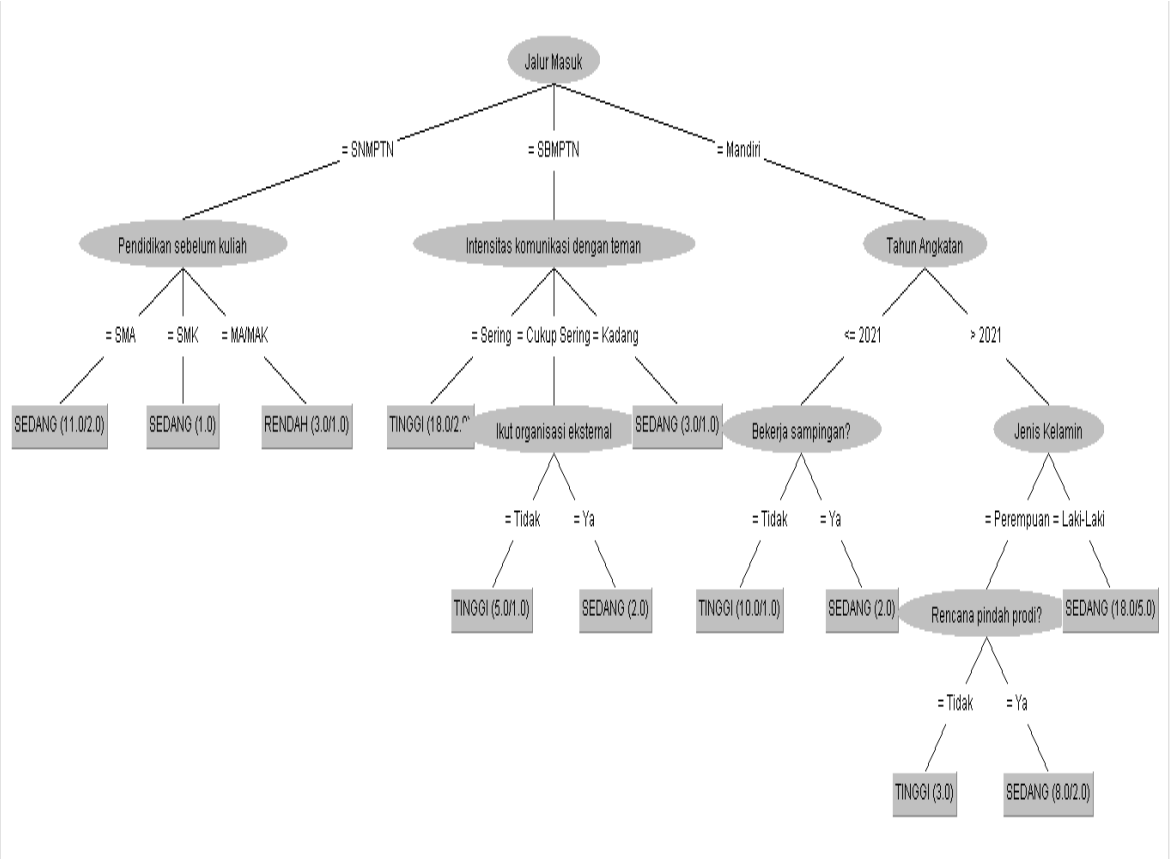

Figure 3. Percentage-split decision tree

From the processing results, the rules of the pruned decision tree are obtained as follows:

1) If the entrance path to college is SNMPTN and education before high school is in the medium category.
2) If the SNMPTN entrance path and education before attending SMK then it is in the medium category.
3) If the SNMPTN college entrance path and education before MA / MAK college then the low category.
4) If the entrance path to SBMPTN college and the intensity of communication with friends is often then in the high category.
5) If the SBMPTN college entrance path, the intensity of communication with friends is quite frequent and does not participate in off-campus organizations, it is in the high category.
6) If the SBMPTN college entrance path, the intensity of communication with friends is quite frequent and participating in off-campus organizations, it is in the medium category.
7) If the SBMPTN college entrance path and the intensity of communication with friends sometimes then enter the medium category.
8) If the entry path is independent study, the class year is less than or equal to 2021 and does not work on the side during college then it is in the high category.
9) If the entry path is independent college, the class year is less than or equal to 2021 and works on the side during college, it is categorized as medium.

.

10) If the entry path is independent study, the class year is more than 2021, the gender is female and does not plan to change study programs during the first year then it is in the high category.
11) If the entry path is independent study, the class year is more than 2021 and plans to change study programs during the first year then it is in the medium category.
12) If the entry path is independent study, the class year is more than 2021 and the gender is male then it is in the medium category.

From the test results, the best result is obtained by using the percentage-split option which uses 80% split to get an accuracy rate of 82.35%, a percentage precision of 81.81% and a percentage recall of 1%.

4. CONCLUSION

The decision tree with the C4.5 algorithm can be used to predict the success of student studies in the first year in the case study of PTIK FKIP UNS students. The test results obtained an accuracy of 82.35% which was obtained from a percentage-split of 80% using the Weka application. This research still has limitations because the model used is only 1 type and the predictor variables chosen are still not varied. For future research, it can be considered in using other classification models and using data that has more diverse attributes.

REFERENCES

[1] P. Pampouktsi *et al.*, "Techniques of Applied Machine Learning Being Utilized for the Purpose of Selecting and Placing Human Resources within the Public Sector," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, pp. 1–16, Dec. 2022, doi: 10.52465/joiser.v1i1.91.

[2] F. N. R. F. J. Aziz, B. D. Setiawan, and I. Arwani, "Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 6 SE-, pp. 2243–2251, Sep. 2017, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1571

[3] A. Y. Lubalu, C. K. Ekowati, and P. A. Udil, "Pengaruh Jalur Seleksi Masuk Universitas Terhadap IPK Tahun Pertama Mahasiswa Angkatan Tahun 2020 Program Studi Pendidikan Matematika FKIP Universitas Nusa Cendana," *Haumeni J. Educ.*, vol. 2, no. 1, pp. 20–26, May 2022, doi: 10.35508/haumeni.v2i1.7074.

[4] A. Wibowo, D. Manongga, and H. D. Purnomo, "The Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation," *Sci. J. Informatics*, vol. 7, no. 1, pp. 99–112, May 2020, doi: 10.15294/sji.v7i1.24241.

[5] E. Alhazmi and A. Sheneamer, "Early Predicting of Students Performance in Higher Education," *IEEE Access*, vol. 11, pp. 27579–27589, 2023, doi: 10.1109/ACCESS.2023.3250702.

[6] J. J. R. Fanggidae, "Klasifikasi Faktor–faktor yang Mempengaruhi Prestasi Akademik Mahasiswa Pendidikan Matematika FKIP Undana dengan Metode CHAID," *FRAKTAL J. Mat. DAN Pendidik. Mat.*, vol. 2, no. 1, pp. 23–33, May 2021, doi: 10.35508/fractal.v2i1.4018.

[7] S. Fitri, N. Nurjanah, and W. Astuti, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa (Studi Kasus: Umtas)," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 633–640, Apr. 2018, doi: 10.24176/simet.v9i1.2002.

[8] N. B. Nasution, D. Hartanto, D. J. Silitonga, Lasimin, and D. Mardhiyana, "Prediksi Lama Studi dan Predikat Kelulusan Mahasiswa Menggunakan Algoritma Supervised Learning," *G-Tech J. Teknol. Terap.*, vol. 7, no. 2, pp. 386–395, Mar. 2023, doi: 10.33379/gtech.v7i2.2077.

[9] A. Azahari, Y. Yulindawati, D. Rosita, and S. Mallala, "Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, pp. 443–452, May 2020, doi: 10.25126/jtiik.2020732093.

[10] Y. E. Fadrial, "Algoritma Naive Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 4, no. 1, pp. 20–29, May 2021, doi: 10.31539/intecoms.v4i1.2219.

[11] I. W. Saputro and B. W. Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 1, Apr. 2020, doi: 10.24076/citec.2019v6i1.178.

[12] M. Windarti and A. Suradi, "Perbandingan Kinerja 6 Algoritme Klasifikasi Data Mining untuk Prediksi Masa Studi Mahasiswa," *Telematika*, vol. 12, no. 1, p. 14, Feb. 2019, doi: 10.35671/telematika.v12i1.778.

[13] A. F. Mulyana, W. Puspita, and J. Jumanto, "Increased accuracy in predicting student academic performance using random forest classifier," *J. Student Res. Explor.*, vol. 1, no. 2, pp. 94–103, Jul. 2023, doi: 10.52465/josre.v1i2.169.

[14] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, May 2020, doi: 10.31294/ijcit.v5i1.7951.

[15] X. Wang, C. Zhou, and X. Xu, "Application of C4.5 decision tree for scholarship evaluations," *Procedia Comput. Sci.*, vol. 151, pp. 179–184, 2019, doi: 10.1016/j.procs.2019.04.027.

[16] H. A. Prihanditya and A. Alamsyah, "The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease," *J. Soft Comput. Explor.*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/joscex.v1i1.8.

[17] Y. Luvia, D. Hartama, A. Windarto, and S. Solikhun, "Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Keberhasilan Mahasiswa Di Amik Tunas Bangsa," *Jurasik (Jurnal Ris. Sist. Inf. Tek. Inform.*, vol. 1, pp. 75–79, Jul. 2016, doi: 10.30645/jurasik.v1i1.12.