International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 8, No. 1, (2024), pp. 40-46

Modelling of C4.5 Algorithm for Graduation Classification

Embun Fajar Wati[1], Budi Sudrajat[2], Raudah Nasution[3]
[1,2]Universitas Bina Sarana Informatika, Indonesia
Email: [1]embun.efw@bsi.ac.id, [2]budi.bst@bsi.ac.id, [3]raudah.rhn @bsi.ac.id

Abstract

Student admissions in universities every year become a routine thing to do, some even do student admissions every semester. That way, the number of students will continue to grow. Especially if there are students who graduate late, it will increase the number of students in the university. There are many things that can affect graduation, namely personal data (gender, age, marital status, job status) and academic data (grade). Before making a decision, universities must analyze the number of students and the factors that most influence student graduation. Analysis by classifying graduation using C4.5 algorithms. The research method used consists of selection to ensure the data used in the KDD process is appropriate and quality data. Then preprocessing by means of data cleaning, data reduction, and data normalization. The next method is transformation for age attributes to young and old, grade attributes to large and small. The last method is C4.5 algorithm modeling with rapid miner and evaluation. Through the calculation process using the classification method and C4.5 algorithm with the attributes described earlier, the results were obtained that the accuracy of the graduation classification reached 97.00%, the precision value was 91.79%, and the recall value was 100.00%, and the AUC value was 0.978. This means that the model has a very high level of accuracy and has an excellent ability to separate samples from both target classes.

Keywords: *C4.5 algorithm, universities, student, classification, graduate*

1. Introduction

The existence of technology will be very easy for universities to produce information and facilitate all university activities related to data processing and making reports [1]. In each academic year, the college organizes a new student admission process [2]. Qualified students are those with a period of study that is in accordance with the rules of the college [3]. Government regulations regarding the maximum length of lectures that can be taken by students aim to make the institutions involved in the educational process have careful planning in the learning process. Timely completion of college is one of the characteristics of undergraduate education. But in reality, students do not always complete it within four years [1]. This is because students participate in campus activities, and many students think that college only wants to get a bachelor's degree, but on the other hand, universities emphasize that every student must have skills or abilities in the field they take [2]. There are also many students who are married and already working, thus neglecting college [3].

Good planning is expected to be able to increase the accreditation value of higher education institutions and deliver student studies so that they can graduate on time. In order to achieve these results, aspects are needed that are so influential on the value of the results of optimal accreditation, one of which is that students graduate on time play an important role in determining the results of accreditation [1]. The process of monitoring and evaluating student graduation is very necessary because student graduation rate is one of the elements of accreditation assessment that is very important for every Study Program [2].

Data on graduating students continues to grow every year and accumulates like neglected data because it is rarely used [1]. Behind the abundant data lies new

ISSN : 2580-7250

information hidden [2]. Student data becomes very important for making decisions, if the data is processed and analyzed using data mining, because each data set can provide important knowledge and information that is very valuable for universities [3]. The application of data mining can be used to predict student graduation. Data mining is the stage of classifying large amounts of data by connecting each pattern connected to each large data set [4]. The method that is often used to predict student graduation is the classification method [5]. Classification techniques comprise several methods, and C4.5 is part of the classification methods. Then the C4.5 method has an algorithm, namely C4.5. The C4.5 algorithm is one of the algorithms that has C4.5 [6]. This algorithm is able to handle categorical and numerical data, and has the ability to perform the selection of relevant attributes for decision making [7].

Extracting patterns from data generated from the C4.5 algorithm by performing data mining [8]. The applied model is expected to help reduce the complexity of the prediction process and better prediction results by using variables that affect the study period [9]. Analysis and prediction are expected to be able to find factors that influence and predict graduation on time. So that early predictions of student graduation can be made [10]. This prediction model is also expected to help the study program to detect and encourage students who are predicted not to graduate on time, so that they can graduate on time [11]. Universities can also take actions such as providing scholarships to students who graduate late [12]. The results of the analysis are applied as a design basis for the decision-making system to identify and classify student graduation on time and not on time [13].

2. Research Methodology

The data used in this study were students majoring in informatics engineering. The research process is passed through several stages, namely [7]:

a) Selection

This initial stage is an important step to ensure the data used in the KDD process is the right and quality data. First of all, the criteria or parameters to be used to select relevant data are determined according to the purpose of the KDD and the type of problem to be solved. For example, if the purpose of KDD is to determine student graduation, then the selection criteria include personal data (gender, age, marital status, job status) and academic data (grade). Data from various sources are collected as a first step in the selection process and then filtered according to predefined selection criteria. At this stage, identification and deletion of duplicate data or redundant data is carried out. Eliminating duplicate data is an important step to maintain the integrity and consistency of the data used in analysis.

b) Preprocessing

In the Data Preprocessing stage, the next step after data selection in the previous stage (Data Selection) is to prepare the data to be ready for processing and analysis in more depth [14]. This stage is very crucial because good and structured data quality will have a positive impact on the results of analysis and decision making. Some of the things done at the data preprocessing stage in this study are data cleaning, data reduction, and data normalization. Empty and incomplete data becomes noise and is cleaned. Data on informatics engineering students were taken as many as 300 students.

c) Transformation

This data transformation process aims to improve data quality and provide deeper insight into the factors that affect student graduation rates. The main goal of data transformation is to change the format and representation of data to make it more suitable and relevant for advanced analysis. Some of the techniques used at this

stage include attribute transformation, scale transformation, new feature formation, and data discretization [15]. In attribute transformation, the value or representation of a particular attribute is changed to fit the needs of the analysis. Through normalization, some attributes in table 1 can be converted into uniform ranges, helping to prevent large-scale dominance of attributes in the analysis. With the right data transformation, the processed data will provide more accurate analysis results and be useful in making decisions regarding student graduation.

Table 1. Transformation

Age	Young	19 – 24 years old
	Old	25 – 50 years old
Grade	Large	3 – 4 years old
	Small	1 – 2.9 years old

d) C4.5 algorithm modeling

After carrying out the preprocessing process and data transformation in accordance with the classification data mining techniques and C4.5 algorithms, the next step is to carry out the data mining analysis process. Because the volume of data used is quite large, the data mining analysis process is carried out manually to facilitate understanding of the use of classification techniques with the C4.5 algorithm. The data used in this analysis process is data training, which has gone through a data transformation process so that it is ready for the data mining process.

The process of forming training data is based on existing data, and must go through several stages. First, data must be integrated to combine data from multiple sources into a single structured whole. After going through the data integration process, the data must then be filtered (selection) to determine which attributes can affect student graduation, which is referred to as target data. The target data contains relevant attributes and supports the data mining process to predict student graduation rates.

e) Evaluation

This stage is an evaluation of the KDD process which includes examining patterns or information found. Information patterns resulting from the data mining process need to be presented in a form that is easily understood by interested parties. In the C4.5 method, the pattern or information is in the form of rules obtained from C4.5 that has been built.

This evaluation phase is important because the results of the analysis and prediction models that have been built at previous stages (for example, using machine learning algorithms) need to be interpreted so that the results can be understood and applied in a broader decision-making context. Proper interpretation allows interested parties to take appropriate measures based on patterns and information found from data mining.

3. Results and Discussion

Data processing in this study utilizes Rapid Miner software version 10.3.001 for data analysis and C4.5 algorithm for modeling. The data process was processed using the KDD method, and from the large number of student data available, only 300 student data records that can be used in this study are contained in table 2 [16].

Table 2. Student Data

Gender	Age	Marital Status	Grade	Job Status
Male	Young	Married	Low	Employee
Female	Young	Single	Low	Employee

Gender	Age	Marital Status	Grade	Job Status
Male	Young	Married	Low	Unemployee
Male	Young	Single	Low	Employee
Male	Young	Married	Low	Unemployee
…	…	…	…	…
…	…	…	…	…
Male	Young	Married	Low	Unemployee
Female	Young	Single	Low	Employee
Male	Young	Married	Low	Unemployee

Data modeling has been processed by Rapid Miner version 10.3.001 to identify passing patterns and data accuracy using performance vector models. By calculating the amount of data that is classified correctly, we can find out the accuracy and get a passing pattern in the form of decision trees and rules from the patterns formed.

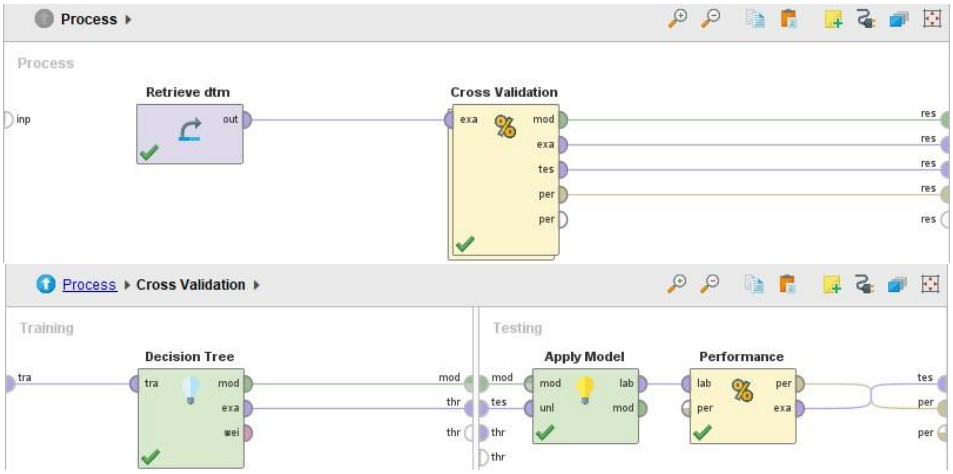

Figure 1. Graduation Classification with C4.5 Algorithm

The use of parameters also affects the accuracy and model generated by the C4.5 algorithm. This study used the Cross Validation method to assess the accuracy of the algorithm. In the Cross Validation process, there is a subprocess involving C4.5 operators as the algorithm used in the data mining classification process. In addition, there are Apply Model and Performance operators which are part of the data mining processing to produce C4.5 (Figure 1).

The modeling results that have been processed using Rapid Miner version 10.3.001 not only produce modeling patterns, but also provide information about the accuracy of the data used. Through the confusion matrix, 194 students obtained a timely prediction value, a 97 students on time prediction score, and a late prediction value, even though it should have been correct, which was 9 students. Thus, confusion matrix values are obtained in the form of: accuracy value of 97.00%, precision value of 91.79%, and recall value of 100.00% (see Figure 2).

accuracy: 97.00% +/- 2.46% (micro average: 97.00%)

	true Tepat	true Terlambat	class precision
pred. Tepat	194	0	100.00%
pred. Terlambat	9	97	91.51%
class recall	95.57%	100.00%	

precision: 91.79% +/- 6.51% (micro average: 91.51%) (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	194	0	100.00%
pred. Terlambat	9	97	91.51%
class recall	95.57%	100.00%	

recall: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	194	0	100.00%
pred. Terlambat	9	97	91.51%
class recall	95.57%	100.00%	

Figure 2. Confusion Matrix

In addition, in this study, AUC (Area Under the Curve) evaluation metrics were used to measure the quality of classification models, especially in the context of binary classification. AUC measures the extent to which the model can distinguish between two target classes, namely positive class and negative class. Based on Figure 3, the results of metric evaluation using AUC are shown by the ROC curve with a value of 0.978.

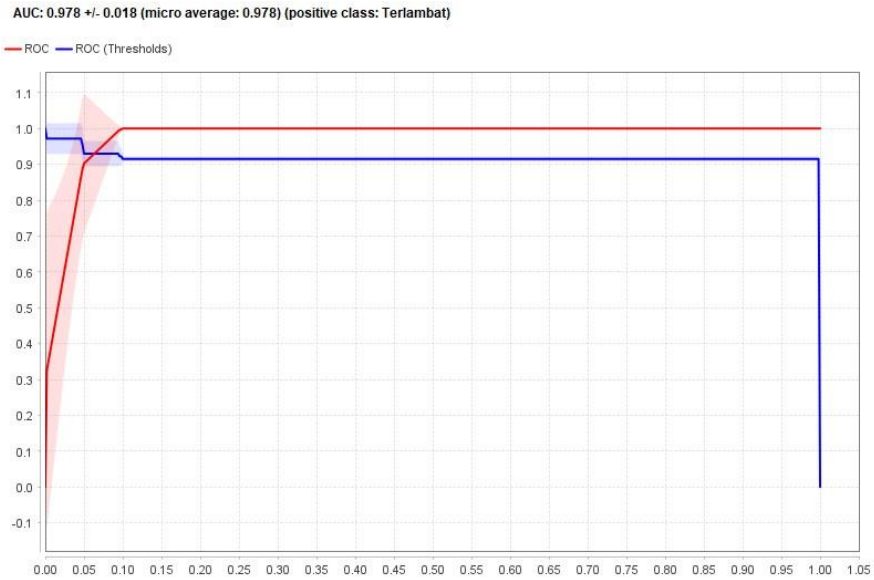

Figure 3. ROC Curve

AUC values are in the range from 0 to 1, where a value of 1 indicates that the model is perfect at distinguishing positive and negative classes, while a value of 0.5 indicates that the model is no better than a random guess.

In the context of binary classification evaluation, AUC values close to 1 (100%) indicate that the model is almost perfect in predicting positive and negative classes. This means that the model has a very high level of accuracy and has an excellent ability to separate samples from both target classes.

4. Conclusion

Based on the results of data mining calculations using the C4.5 algorithm, it can be concluded that the "late" graduation status class or passing not on time is smaller than the "right" graduation status class or graduating on time. Through the calculation process using the classification method and C4.5 algorithm with the attributes described earlier, the results were obtained that the accuracy of the

graduation classification reached 97.00%, the precision value was 91.79%, and the recall value was 100.00%.

This 97.00% accuracy result shows that the model built is able to make predictions very high. However, it should be noted that high accuracy can also be caused by low data complexity, which results in the model being able to predict quite accurately. Therefore, it is important to conduct further evaluation of the model and test the reliability of predictions on different data to ensure the reliability of the model.

The ability of the C4.5 algorithm to form C4.5 and identify relevant attributes in the data mining process is a determining factor for the high accuracy of predictions. Thus, the results of this study provide valuable insights and information related to the prediction of student graduation rates using the classification method and C4.5 algorithm. However, for broader applications and more complex decision-making, it is necessary to conduct more in-depth research and evaluation involving more data and testing the model on various scenarios to ensure the reliability and generalizability of the developed model.

References

[1] Karya Suhada, Anggi Elanda, And Anwar Aziz, "Klasifikasi Predikat Tingkat Kelulusan Mahasiswa Program Studi Teknik Informatika Dengan Menggunakan Algoritma C4.5 (Studi Kasus: Stmik Rosma Karawang)," *Dirgamaya Jurnal Manajemen Dan Sistem Informasi*, Vol. 1, No. 2, Pp. 14–27, 2021.

[2] Agus Romadhona, Suprapedi, And H. Himawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma Decision Tree," *Jurnal Teknologi Informas*, Vol. 13, No. 1, Pp. 69–83, 2017.

[3] Embun Fajar Wati And Biktra Rudianto, "Penerapan Algoritma Knn, Naive Bayes Dan C4.5 Dalam Memprediksi Kelulusan Mahasiswa," *Jurnal Format*, Vol. 11, No. 2, Pp. 168–175, 2022.

[4] Tuhfatul Habibah Hasibuan And Deni Mahdiana, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 Pada Uin Syarif Hidayatullah Jakarta ," *Skanika: Sistem Komputer Dan Teknik Informatika*, Vol. 6, No. 1, Pp. 61–74, 2023.

[5] Ade Fatma Ayu Rahman And Sorikhi, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus Di Universitas Peradaban)," *Ijir*, Vol. 1, No. 2, Pp. 70–77, 2020.

[6] Embun Fajar Wati, Elvi Sunita Perangin-Angin, And Anggi Puspita Sari, "Improved Naive Bayes Algorithm With Particle Swarm Optimization To Predict Student Graduation," *International Journal Of Information System & Technology*, Vol. 7, No. 6, Pp. 386–391, 2024.

[7] Sudriyanto, Rudi Rizaldi, And Ainun Rofiq Hariri, "Implementasi Algoritme Decision Tree (C4.5) Dengan Optimize Weights (Pso) Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal Informatika Universitas Pamulang*, Vol. 6, No. 2, Pp. 252–257, 2021.

[8] Hozairi, Anwar, And S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes," *Jurnal Ilmiah Nero*, Vol. 6, No. 2, Pp. 133–144, 2021.

[9] Ratna Puspita Sari Putri And Indra Waspada, "Penerapan Algoritma C4.5 Pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Informatika*, Vol. 4, No. 1, Pp. 1–7, 2018.

[10] Ni Luh Ratniasih, "Optimasi Data Mining Menggunakan Algoritma Naïve Bayes Dan C4.5 Untuk Klasifikasi Kelulusan Mahasiswa," *Jurnal Teknologi Informasi Dan Komputer*, Vol. 5, No. 1, Pp. 28–34, 2019.

[11] Abdul Rohman And Anief Rufiyanto, "Komparasi Model Decision Tree Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandanaran," *Jurnal Neoteknika*, Vol. 6, No. 1, Pp. 1–5, 2020.
[12] Wulandari, Rika Rosnelly, And Wanayumin, "Analisis Metode Decision Tree Dalam Memprediksi Kelulusan Mahasiswa," *Csrid Journal*, Vol. 13, No. 3, Pp. 130–140, 2021.
[13] Lydia Yohana Lumban Gaol, M. Safii, And Dedi Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5," *Brahmana: Jurnal Penerapan Kecerdasan Buatan*, Vol. 2, No. 2, Pp. 97–106, 2021.
[14] Siska Haryati, Aji Sudarsono, And Eko Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *Jurnal Media Infotama*, Vol. 11, No. 2, Pp. 130–138, 2015.
[15] Uci Suriani, "Penerapan Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *Journal Of Computer And Information Systems Ampera*, Vol. 3, No. 2, Pp. 55–66, 2023.
[16] Embun Fajar Wati, Elvi Sunita Perangin-Angin, And Anggi Puspita Sari, "Prediction Of Student Graduation Using The K-Nearest Neighbors Method," *International Journal Of Information System & Technology*, Vol. 7, No. 3, Pp. 211–216, 2023.
[17] Elsa Paskalis Krisda Orpa, Eva Faja Ripanti, And Tursina, "Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *Justin (Jurnal Sistem Dan Teknologi Informasi)*, Vol. 7, No. 4, Pp. 272–278, 2019.
[18] Mita Nurul Yatimah, "Implementasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Stimik Esq Menggunakan Decision Tree C4.5," In *Seminar Nasional Informatika Dan Aplikasinya (Snia)*, 2021.
[19] Isnan Mulia And Muanas, "Model Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5 Dan Software Weka," *Jurnal Analisis Sistem Pendidikan Tinggi* |, Vol. 5, No. 1, Pp. 57–64, 2021.
[20] E. F. Wati, A. P. Sari, E. T. Alawiah, M. H. Siregar, And B. Rudianto, "Particle Swarm Optimization Comparison On Decision Tree And Naive Bayes For Pandemic Graduation Classification," In *2nd International Conference On Advanced Information Scientific Development (Icaisd)*, 2021, Pp. 1–11.
[21] C. N. Dengen, K. Kusrini, And E. T. Luthfi, "Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *Sisfotenika*, Vol. 10, No. 1, Pp. 1–11, Jan. 2020, Doi: 10.30700/Jst.V10i1.484.
[22] A. Muzakir And R. A. Wulandari, "Model Data Mining Sebagai Prediksi Penyakit Hipertensi Kehamilan Dengan Teknik Decision Tree," *Scientific Journal Of Informatics*, Vol. 3, No. 1, Pp. 19–26, Jun. 2016, Doi: 10.15294/Sji.V3i1.4610.
[23] Fauziah Abdul Rahman, Mohammad Ishak Desa, Antoni Wibowo, And Norhaidah Abu Haris, "Knowledge Discovery Database (Kdd)-Data Mining Application In Transportation," In *Proceeding Of International Conference On Electrical Engineering, Computer Science And Informatics (Eecsi*, 2014, Pp. 116–119.