



RESEARCH ARTICLE

Student Graduation Prediction Using Decision Tree Method with C4.5 Algorithm

Sarbaini^{1*}, Fara Ulfa²

¹*Program Studi Matematika, Universitas Islam Negeri Sultan Syarif Kasim Riau-Pekanbaru, Indonesia*

²*Program Studi Psikologi, Universitas Islam Negeri Sultan Syarif Kasim Riau-Pekanbaru, Indonesia*

*Corresponding author: sarbaini@uin-suska.ac.id

Received: 07 August 2023; Revised: 17 October 2023; Accepted: 19 October 2023; Published: 09 January 2024.

Abstract:

One of the determinants of the quality of higher education is the percentage of student's ability to complete their studies on time. However, in practice, few students can complete their studies in higher education on time. Graduation prediction is one of the things that can be done to increase student graduation so that early prevention or handling of students who have the opportunity to not graduate on time can be done. The aim of this research is to find out and analyze the use of the Decision Tree Method with the application of the C4.5 Algorithm to effectively predict student graduation on time. The data used is that of Mathematics students at UIN SUSKA RIAU's Faculty of Science and Technology. The decision tree was constructed using 150 training data and processed using the C4.5 algorithm to generate 40 rules, which were then tested using 150 data sets with an accuracy of 78.6667 percent and an AUC of 0.8363.

Keywords: C4.5 Algorithm, Decision Tree, Prediction

1. Introduction

Competition between universities has been very competitive in the world of education, so every university is now required to have advantages and good quality in competing. One of the determinants of the quality of higher education is the percentage of students' ability to complete their studies on time, commonly known as "graduating on time." Students are said to graduate on time if they complete their studies in college for less than or equal to four years. However, only a few students can complete undergraduate education within four years. Several factors, including internal factors and external factors, can cause this [1].

Internal factors come from within the students themselves, such as their level of intelligence. Conversely, external factors come from outside the student, such as environmental factors around him. In addition, family economic conditions can also affect students graduating on time or not [2].

To increase the number of students who graduate on time, it is necessary to predict which students can graduate on time and which are not on time so that early prevention or treatment can be carried out for students who are likely not to graduate on time [3], [4]. For students who are predicted not to be able to graduate on time, guidance, or direction, as well as encouragement and

learning motivation, can be given so that these students can graduate according to the expected time [5–7].

A classification technique is needed, which is one of the techniques of data mining to analyze the data of the Computer Engineering Department. Applying this technique will build a decision tree to see the possibility of students graduating more than 8 percent. The decision tree is the output of an application built using the C4.5 algorithm to predict students' study periods [8]. Many algorithms can form decision trees, including ID3, CART, and C4.5 [9–14].

Research on student graduation has been carried out by Kamagi and Hansun [15]. The study aims to predict student graduation rates using the C4.5 algorithm based on IP scores in semesters I–VI, gender, high school origin, and the number of credits in semester VI. This study found that the most influential factor in student graduation rates was the IP score of semesters VI, with an accuracy rate of 87.5%. This means that the C4.5 algorithm performs well in predicting student graduation rates.

This study aims to predict student graduation on time for students of the Mathematics Study Program, Faculty of Science and Technology UIN SUSKA RIAU class of 2010 – 2020 using the decision tree method with the application of the C4.5 algorithm. The output of the decision tree is divided into two classes, namely "On Time" and "Not on Time," using 16 attributes, namely: gender, high school origin, IP scores in semesters I – VI, the number of credits passed in semesters I – VI, domicile status and parents/guardians' income per month. Applying the C4.5 algorithm will produce a pattern of student graduation rates, which is in the form of rules that can be used to predict the graduation of Mathematics Study Program Students, Faculty of Science and Technology UIN SUSKA RIAU.

2. Research Methodology

This research is carried out in a stage-by-stage manner. The research technique describes the stages of these processes. The research methodologies are laid out in a clear, organized, and systematic manner. The steps of the study are as follows. The stages of this research methodology are,

1. Review literature related to student graduation on time.
2. Identify problems by formulating problems from student graduation, then look for solutions and methods that can be used. In this study, student graduation predictions will be made using the C4.5 algorithm.
3. We are collecting alumni data of the class of 2010-2019 students of the Mathematics Study Program, Faculty of Science, UIN SUSKA RIAU.
4. Perform preprocessing, which removes incomplete data (missing value) and determines the attributes to be used. Furthermore, perform data transformation, summarizing or converting raw data into easy-to-manage data.
5. Manage data using the WEKA (Waikato Environment for Knowledge Analysis) application.
6. Apply the C4.5 algorithm to create a decision tree model.
7. Testing a pre-built decision tree model.
8. Presentation of results and conclusions.

3. Results and Discussion

3.1. Preprocessing Data

The data used is alumni data of students of the Mathematics Study Program, Faculty of Science and Technology UIN SUSKA RIAU class of 2010 – 2019. The author obtained the data from the clearing file of alumni in the Mathematics Study Program, Faculty of Science and Technology UIN SUSKA RIAU, from the alumni concerned (respondents) and from various parties closely related to respondents. The total number of data obtained is 279.

After data preprocessing is carried out, namely cleaning incomplete data, the number of data becomes 150 data consisting of 32 data of students who graduated on time (study period ≤ 4 years) and 118 data of students who graduated not on time (study period > 4 years). Some several attributes or parameters are considered to affect student graduation rates, namely: gender, high school origin, student domicile status while studying at UIN SUSKA RIAU, income of parents/guardians of students, semester IP scores (I, II, III, IV, V, and VI), and the number of credits passed in semesters (I, II, III, IV, V, and VI).

3.2. Data Transformation

At this stage, data summarization will be carried out or converted into data easily managed in the C4.5 algorithm. In this case, discrete data will be used. Therefore, continuous data will be converted into discrete data. After that, the data is converted into Attribute Relation File Format (ARFF) format so WEKA can read it. The transformation process is carried out by classifying each input attribute, as seen in Table 3.1.

Table 3.1: Data Transformation

No.	Attribute	Scale	Attribute Value	
1	Gender (JK)	Nominal	Male	= L
			Woman	= P
2	Origin of High School Student (USA)	Nominal	High School	= SMA
			Vocational High School	= SMK
			Madrasah Aliyah	= MA
3	Domicile Status (SD)	Nominal	With parents	= ORTU
			Boarding house/rental/rental house/rental/rental	= KOS/RENT
			Other (privately owned, dormitories, shared guardians, etc.)	= Others
4	Parent/Guardian Monthly Income (PO)	Ordinal	\leq IDR 1,500,000	= Low
			IDR 1,500,001 – IDR 2,500,000	= Medium
			IDR 2,500,001 – IDR 3,500,000	= High
			$>$ 3 IDR 3,500,000	= Very High
5	Achievement Index (IP)	Ordinal	0,00 - 0,99	= 1
			1,00 - 1,99	= 2
			2,00 - 2,99	= 3
			3,00 - 4,00	= 4
6	Number of credits approved (SKS)	Ordinal	\leq 15 credits	= 1
			16 - 18 credits	= 2
			19 - 21 credits	= 3
			21 - 24 credits	= 4

According to the Central Statistics Agency (BPS), the income group of the population is divided into 4, namely the very high-income group with an average of more than IDR. 3,500,000 per month, the high-income group with an average between IDR. 2,500,000 – IDR. 3,500,000 per month, the medium income group with an average between IDR. 1,500,001 – IDR. 2,500,001 per month, and the low-income group with an average of less than or equal to IDR. 1,500,000 per month.

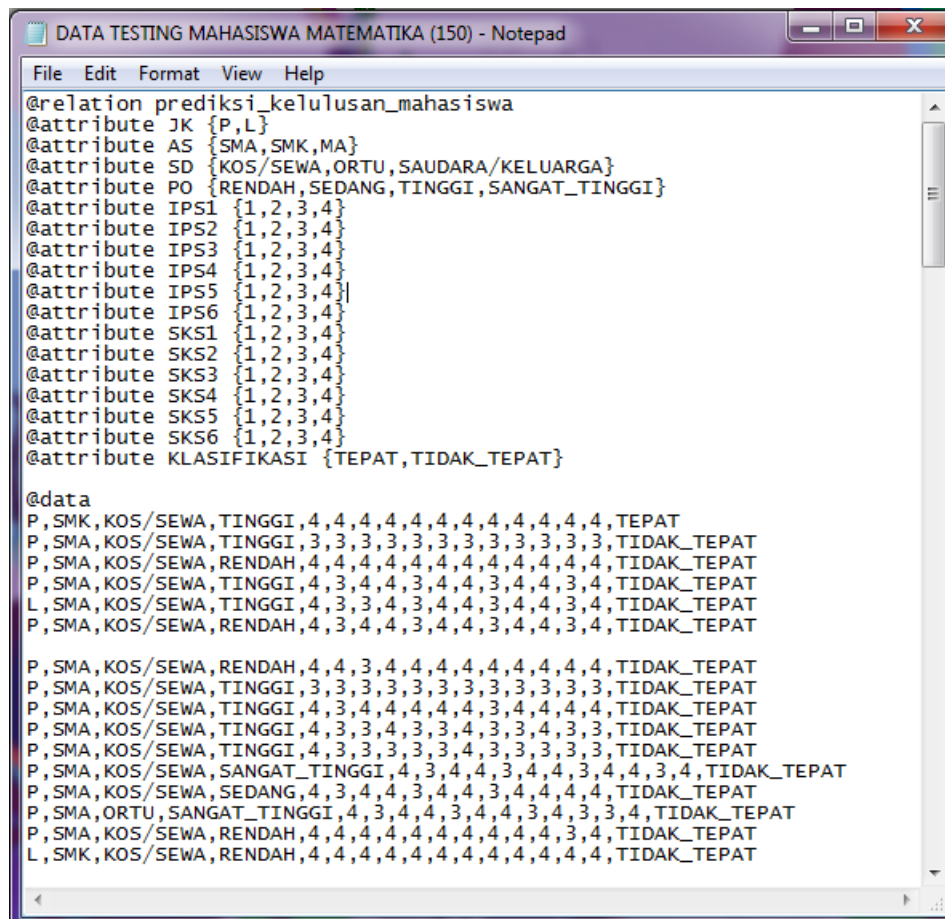
The author adds the number 1 to the new lower limit value, as seen in Table 3.1 above, to distinguish it from the previous upper limit value criteria so that there is no equal membership in each income group of parents/guardians of students per month. In the Mathematics Study Program, Faculty of Science and Technology UIN SUSKA, the IP value that students may get per semester is between 0.00 - 4.00. Therefore, in this study, the author divides/transforms the IP value attribute into four parts, as seen in the table above. As for the transformation of the attribute of the number of credits passed, the author's idea is based on the rules for purchasing Study Plan Cards and adjusted to the alumni data obtained.

Furthermore, the input attributes will be analyzed using the C4.5 algorithm to produce the target attribute. The target attribute is an output class to determine whether students are graduating on time or not, according to the target attribute that can be seen in Table 3.2.

Table 3.2: Target Attributes

Target Attributes		Information
On-time	= T	Study period \leq 4 years
Not on Time	= TT	Study period $>$ 4 years

Table 3.2 shows two output classes of the desired target attributes, namely passing "On Time" and "Not on Time," which are determined based on the student's study period. After transforming the data from continuous to discrete data, then the data is converted from Excel to Notepad, as seen in Figure 3.1.



```

@relation prediksi_kelulusan_mahasiswa
@attribute JK {P,L}
@attribute AS {SMA,SMK,MA}
@attribute SD {KOS/SEWA,ORTU,SAUDARA/KELUARGA}
@attribute PO {RENDAH,SEDANG,TINGGI,SANGAT_TINGGI}
@attribute IPS1 {1,2,3,4}
@attribute IPS2 {1,2,3,4}
@attribute IPS3 {1,2,3,4}
@attribute IPS4 {1,2,3,4}
@attribute IPS5 {1,2,3,4}
@attribute IPS6 {1,2,3,4}
@attribute SKS1 {1,2,3,4}
@attribute SKS2 {1,2,3,4}
@attribute SKS3 {1,2,3,4}
@attribute SKS4 {1,2,3,4}
@attribute SKS5 {1,2,3,4}
@attribute SKS6 {1,2,3,4}
@attribute KLASIFIKASI {TEPAT,TIDAK_TEPAT}

@data
P,SMK,KOS/SEWA,TINGGI,4,4,4,4,4,4,4,4,4,4,4,4,TEPAT
P,SMA,KOS/SEWA,TINGGI,3,3,3,3,3,3,3,3,3,3,3,3,TIDAK_TEPAT
P,SMA,KOS/SEWA,RENDAH,4,4,4,4,4,4,4,4,4,4,4,4,TIDAK_TEPAT
P,SMA,KOS/SEWA,TINGGI,4,3,4,4,3,4,4,3,4,4,3,4,TIDAK_TEPAT
L,SMA,KOS/SEWA,TINGGI,4,3,3,4,3,4,4,3,4,4,3,4,TIDAK_TEPAT
P,SMA,KOS/SEWA,RENDAH,4,3,4,4,3,4,4,3,4,4,3,4,TIDAK_TEPAT

P,SMA,KOS/SEWA,RENDAH,4,4,3,4,4,4,4,4,4,4,4,4,TIDAK_TEPAT
P,SMA,KOS/SEWA,TINGGI,3,3,3,3,3,3,3,3,3,3,3,3,TIDAK_TEPAT
P,SMA,KOS/SEWA,TINGGI,4,3,4,4,4,4,4,3,4,4,4,4,TIDAK_TEPAT
P,SMA,KOS/SEWA,TINGGI,4,3,3,4,3,3,4,3,3,4,3,3,TIDAK_TEPAT
P,SMA,KOS/SEWA,TINGGI,4,3,3,3,3,3,3,4,3,3,3,3,TIDAK_TEPAT
P,SMA,KOS/SEWA,SANGAT_TINGGI,4,3,4,4,3,4,4,3,4,4,3,4,TIDAK_TEPAT
P,SMA,KOS/SEWA,SEDANG,4,3,4,4,3,4,4,3,4,4,4,4,TIDAK_TEPAT
P,SMA,ORTU,SANGAT_TINGGI,4,3,4,4,3,4,4,3,4,4,3,3,4,TIDAK_TEPAT
P,SMA,KOS/SEWA,RENDAH,4,4,4,4,4,4,4,4,4,4,3,4,TIDAK_TEPAT
L,SMK,KOS/SEWA,RENDAH,4,4,4,4,4,4,4,4,4,4,4,4,TIDAK_TEPAT

```

Figure 3.1: Data View Converted to Notepad

Figure 3.1 displays data converted from Excel to Notepad, which is then saved in ARFF format so WEKA can read it.

3.3. Rules Decision Tree

Based on the decision tree model that has been built, it was obtained that of the 16 attributes used, there are only 11 attributes that can influence students to graduate on time and not on time, including semester IP (I, II, III, and V), the number of credits passed in the semester (II, III, IV, and V), the origin of high school, gender and income of parents/guardians. Here is the view of the rules generated in WEKA.

```

Attributes: 19
Timestamp
NAMA LENGKAP
JENIS KELAMIN
IP SEMESTER ( I - VI ) [Semester 1]
IP SEMESTER ( I - VI ) [Semester 2]
IP SEMESTER ( I - VI ) [Semester 3]
IP SEMESTER ( I - VI ) [Semester 4]
IP SEMESTER ( I - VI ) [Semester 5]
IP SEMESTER ( I - VI ) [Semester 6]
JUMLAH SKS ( I - VI ) [Semester 1]
JUMLAH SKS ( I - VI ) [Semester 2]
JUMLAH SKS ( I - VI ) [Semester 3]
JUMLAH SKS ( I - VI ) [Semester 4]
JUMLAH SKS ( I - VI ) [Semester 5]
JUMLAH SKS ( I - VI ) [Semester 6]
ASAL SLTA
STATUS DOMISILI
PENGHASILAN ORANG TUA
KLASIFIKASI
    
```

Figure 3.2: Rules Decision Tree On WEKA

3.4. Model Testing

At this stage, the mining process will be used as a reference in predicting the target class (passing on time and not on time) on data using the Confusion Matrix method. The data used amounted to 150 data consisting of 32 data on students who graduated on time and 118 data on students who did not. The results of testing the decision tree model using WEKA can be seen in Figure 3.3.

=== Summary ===		
Correctly Classified Instances	118	78.6667 %
Incorrectly Classified Instances	32	21.3333 %

Figure 3.3: Decision Tree Model Test Results.

Based on Figure 3.3 above, it can be obtained that from 150 data, the number of correct prediction values is 118 data or 79%, and the number of incorrect prediction values is 32 data or 21%. The ROC curve image can be seen in Figure 3.4.

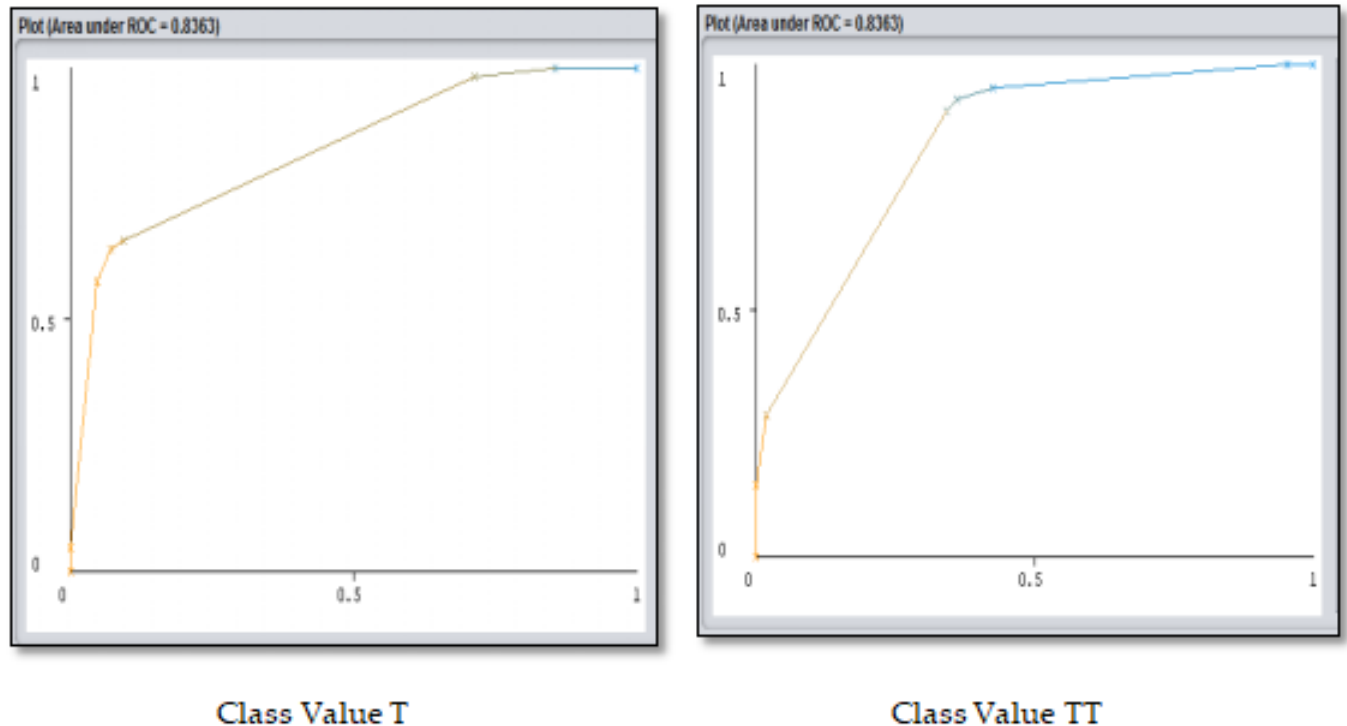


Figure 3.4: ROC.

For data mining classification, the AUC value can be divided into several parts for the accuracy of the results, namely:

0.90 – 1.00 = Excellent Classification

0.80 – 0.90 = Good Classification

0.70 – 0.80 = Sufficient Classification

0.60 – 0.70 = Bad Classification

0.50 – 0.60 = Classification False/Failed.

From Figure 3.4, the area value under ROC or AUC value is 0.8363. This means the model obtained performs well in predicting student graduation on time.

Conclusion

Of the 150-training data processed using the C4.5 algorithm after being applied to 150 data (150 student data that were considered to have not graduated), it was obtained that 118-student data were predicted to graduate not on time and 32 student data predicted to graduate on time. Based on an AUC value of 0.8363, the C4.5 algorithm has a good performance in predicting the graduation of these students. The accuracy level of the C4.5 algorithm in predicting the graduation of students of the Mathematics Study Program, Faculty of Science and Technology UIN SUSKA RIAU is an Accuracy of 78.6667%, Error rate of 21.3333%.

References

- [1] R. Rismayanti, "Decision tree penentuan masa studi mahasiswa prodi teknik informatika (studi kasus: Fakultas teknik dan komputer universitas harapan medan)," *Query: Journal of Information Systems*, vol. 2, no. 1, 2018. [View online](#).
- [2] S. Soeparman, "Faktor-faktor yang mempengaruhi keberhasilan studi mahasiswa penyandang disabilitas," *Indonesian journal of disability studies*, vol. 1, no. 1, pp. 12–19, 2014. [View online](#).
- [3] J. E. Girves and V. Wemmerus, "Developing models of graduate student degree progress," *The Journal of Higher Education*, vol. 59, no. 2, pp. 163–189, 1988. [View online](#).
- [4] N. R. Kuncel and S. A. Hezlett, "Standardized tests predict graduate students' success," *Science*, vol. 315, no. 5815, pp. 1080–1081, 2007. [View online](#).
- [5] D. d. S. Fleith, "The role of creativity in graduate education according to students and professors," *Estudos de Psicologia (Campinas)*, vol. 36, 2019. [View online](#).
- [6] R. Vaatstra and R. De Vries, "The effect of the learning environment on competences and training for the workplace according to graduates," *Higher Education*, vol. 53, pp. 335–357, 2007. [View online](#).
- [7] D. Chari and G. Potvin, "Understanding the importance of graduate admissions criteria according to prospective graduate students," *Physical Review Physics Education Research*, vol. 15, no. 2, p. 023101, 2019. [View online](#).
- [8] S. L. B. Ginting, W. Zarman, and I. Hamidah, "Analisis dan penerapan algoritma c4. 5 dalam data mining untuk memprediksi masa studi mahasiswa berdasarkan data nilai akademik," *PROSIDING SNAST*, pp. 263–272, 2014. [View online](#).
- [9] M. Li, "Application of cart decision tree combined with pca algorithm in intrusion detection," in *2017 8th IEEE international conference on software engineering and service science (ICSESS)*, pp. 38–41, IEEE, 2017. [View online](#).
- [10] T. N. Shah, M. Z. Khan, M. Ali, B. Khan, and N. Idress, "Cart j-48graft j48 id3 decision stump and random forest: A comparative study," *Univ. Swabi J.*, vol. 2, no. April, pp. 1–6, 2018. [View online](#).
- [11] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "An efficient rule-based classification of diabetes using id3, c4. 5, & cart ensembles," in *2014 12th International Conference on Frontiers of Information Technology*, pp. 226–231, IEEE, 2014. [View online](#).
- [12] F. Javed Mehedi Shamrat, R. Ranjan, K. M. Hasib, A. Yadav, and A. H. Siddique, "Performance evaluation among id3, c4. 5, and cart decision tree algorithm," in *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021*, pp. 127–142, Springer, 2022. [View online](#).
- [13] S. Singh and P. Gupta, "Comparative study id3, cart and c4. 5 decision tree algorithm: a survey," *International Journal of Advanced Information Science and Technology (IJAIST)*, vol. 27, no. 27, pp. 97–103, 2014. [View online](#).
- [14] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*, vol. 4. John Wiley & Sons, 2014. [View online](#).
- [15] D. H. Kamagi and S. Hansun, "Implementasi data mining dengan algoritma c4. 5 untuk memprediksi tingkat kelulusan mahasiswa," *Ultimatics: Jurnal Teknik Informatika*, vol. 6, no. 1, pp. 15–20, 2014. [View online](#).

Citation IEEE Format:

Sarbaini and Fara Ulfa, "Student Graduation Prediction Using Decision Tree Method with C4.5 Algorithm", *Jurnal Diferensial*, vol. 6(1), pp. 9-15, 2024.

This work is licensed under a [Creative Commons "Attribution-ShareAlike 4.0 International"](#) license.

