Journal of Informatics and Vocational Education
Vol.5, No.2, July 2022, pp. 96~102
ISSN: 2746-7813

THE IMPLEMENTATION OF DECISION TREE CLASSIFICATION TECHNIQUES TO PREDICT THE DURATION OF STUDENTS COMPLETING THE THESIS AT PTIK FKIP UNS

Halim Perdana Kesuma[1*], Dwi Maryono[2], Febri Liantoni[3]
[1,2,3] Department of Informatics Education, Sebelas Maret University

Article Info

Article history:

Received Aug 30, 2022
Accepted Oct 9, 2022

Corresponding Author:

Halim Perdana Kesuma,
Departement of Informatics Education,
Sebelas Maret University,
Jl Ahmad Yani, no 200,
Pabelan, Kartasura, Surakarta,
Jawa Tengah, 57169, Indonesia.
Email:
halimperdana@student.uns.ac.id

ABSTRACT

This paper aims to determine the reasons why students take a long time in compiling their thesis. The slowness of students in compiling will have an impact on their graduation. This is a serious problem faced by educational institutions. Out of 328 students at PTIK FKIP UNS who took thesis credits, only 85 were able to graduate on time. Therefore, this study was conducted to identify the causes. The research data was taken from the alumnus class of 2012 to 2017. The data was processed using RapidMiner software. The technique used was the decision tree classification technique with the C4.5 algorithm, and to optimize the accuracy of the model, the Particle Swarm Optimization (PSO) algorithm was also added. This study got an accuracy rate of 76% and an AUC score of 0.733.

Keywords: C4.5, Data Mining, Decision Tree, Particle Swarm Optimization, PSO, RapidMiner

1. INTRODUCTION

There are many aspects that can be used as a benchmark for the success of an education system in carrying out learning activities, one of which is student academic achievement [1]. One of the criteria for a good student's academic achievement is that the student can graduate on time, and the most important thing to determine whether a student can graduate on time is the length of time the student spends completing their thesis. The student's study period classified as normal/on time is 8 semesters (4 years) and not on time if taken within a period of 9 semesters (4.5 years) to 14 semesters (7 years), with the condition that students must submit a minimum of 144 credits [2].

In the year 2020, out of the 328 students from the 2012–2017 class who have taken thesis credits, only 85 people can complete the thesis on time, or 26.5% of the total students. Therefore, every new academic year, the number of students increases but is not matched by the number of students who graduate, resulting in an accumulation of students. The length of time it takes students to complete their thesis is one of the determining factors in whether or not they will graduate on time. Therefore, a study is needed to find out what factors can affect the duration of thesis work.

Haryati et al [3] conducted research to predict student study length with the C4.5 algorithm with parameters NPM, Student Name, Study Program, Semester, Gender, Educational Level, Number of Credits, and GPA. Their research has an accuracy rate of 95%. Putri [4] in her research implemented the C4.5 algorithm to predict the learning achievement of SMK students with attributes of national exam scores, majors, reasons for admission, motivation to enter, parents' income, gender, year of birth, family support, learning intensity, intensity of playing games, participation in community activities, intensity of participation in community activities, availability of learning support tools at home, distance between home and school, and status of residence. Her research obtained an accuracy of 82%. Sabna & Muhardi [5] also conducted

research to predict student achievement in the academic field using the variables of GPA, lecturers, discipline, motivation, learning outcomes, economics, social, and report cards, which obtained an AUC accuracy of 65%.

From some of the data that has been mentioned, it was decided that this study would use the Decision Tree classification technique using the C4.5 algorithm to predict the completion time of the thesis work for PTIK FKIP UNS students.

2. RESEARCH METHOD

2.1. Data Mining

Data mining is the activity of analyzing large amounts of data to find useful patterns and rules [6]. The stages in data mining start with data selection, then processing/cleaning data to improve data quality, and then data transformation so that the data can be used in the data mining process. The next step is data mining, which is finding patterns from data using certain methods or techniques. The last interpretation/evaluation, which displays the patterns generated from data mining in a form that is easier to understand by the user, also includes checking whether the patterns found contradict the facts or previous hypotheses [7].

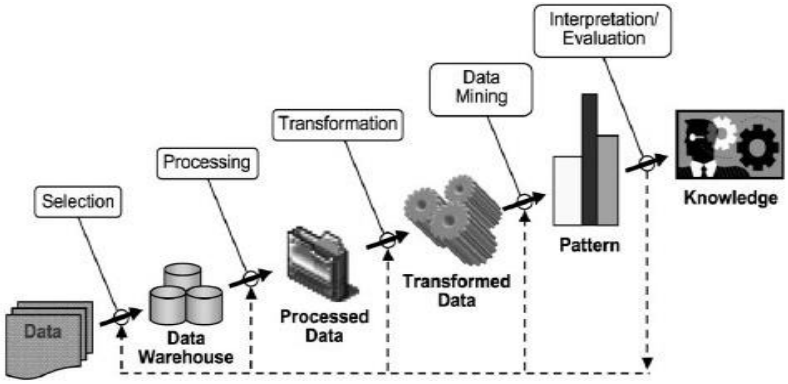


Figure 1. Stages in data mining

2.2. C4.5 Algorithm

The C4.5 algorithm is the algorithm used to form a decision tree [8]. The process in the decision tree is to change the shape of the data (table) into a tree model, convert the tree model into rules, and simplify the rules.

The stages in the formation of a decision tree in the C4.5 algorithm according to Kusrini & Luthfi [9]:

a. The selection of the attribute as the root or root is based on the highest gain value of all the attributes used. The entropy value (S) must be determined first in the following equation:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi$$

Where:
S : Number of cases
A : Atribute
n : Number of Partitions S
pi : Proportion of Si to S

To get the gain value, calculate the data using the following equation:

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

Where:
S : Number of cases
A : Atribute
n : Number of Partitions atribute A
|Si| : Number of cases on partition i
|S| : number of cases in S

b. Creating a branch for each value.
c. Attribute sharing in a branch.
d. The process repeats until all attributes in the branch have the same class or category.

2.3. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) It is an optimization algorithm introduced by Kennedy and Eberhart in 1995 [10]. In PSO, the population is called the swarm and the individual is called the particle. Each particle moves at a different speed and remembers the best position ever reached. Each particle is also associated with the speed of flying through the search space at a speed that dynamically adjusts to their memories. Therefore, the particles tend to fly towards a better search area during the search process [11]. The formula for calculating the displacement of the particle's position and velocity is:

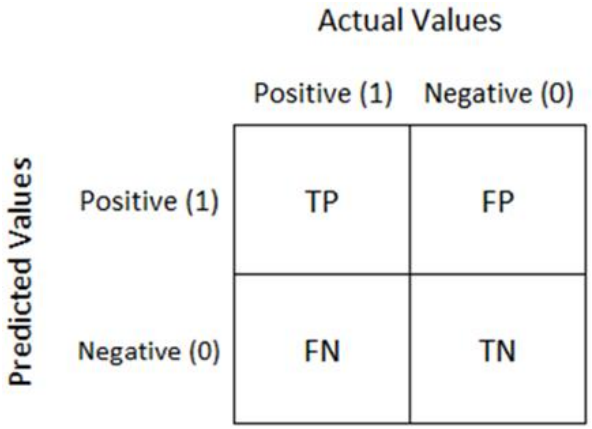
$$Vi(t) = Vi\,(t-1) + c1r1[Xpbest\ i - Xi\,(t)] + c2r2[XGbest - Xi\ X\,(t)]$$
$$Xi(t) = Xi\,(t-1) + Vi\,(t)$$

Where:

Vi(t)	: Velocity of particle i during iteration t
Xi(t)	: Position of particle i during iteration t
c1 and c2	: Learning rate for individual ability (cognitive) and social influence
r1 and r2	: A random number that is uniformly distributed in the interval 0 and 1
Xpbest i	: The best position of the particle i
XGbest	: Global best position

2.4. Data Analysis Technique

2.4.1. Confusion Metrix



Where:
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative

Figure 2. Confusion Metrix Table

2.4.2. Accuracy Test

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

2.4.3. AUC Score

AUC values range from 0 to 1. Higher AUC values mean better model results if applied. By default, a model with an AUC value below 0.7 is a bad model, and if it is 0.7 and above, then the model is still acceptable or even good. We can refer to Hosmer et al [12] for scoring the model using the AUC value.

Table 1. AUC Scoring

Score	Status
0.5 - 0.6	Fail Model
0.6 – 0.7	Poor Model
0.7 - 0.8	Acceptable/Fair Model
0.8 - 0.9	Good Model
0.9 - 1.0	Excellent Model

3. RESULT AND ANALYSIS

Respondent data comes from PTIK student classes of 2012, 2013, 2014, 2015, 2016, and 2017. Respondents are students who have completed the thesis or have at least carried out the thesis examination. From 56 respondents, data will be grouped into students who complete the final project on time and those who are late. After going through the process of data cleaning and transformation, 50 datasets were generated. The data will then be processed using RapidMiner's Decision Tree algorithm and PSO, with 10 Folds Cross Validation technique.

3.1. RESULT

From the RapidMiner process, the following results are obtained:

Tabel 2. Accuracy, Precision, Recall, and AUC Score Results

Accuracy	AUC
52%	0.438

From Table 2 above obtained an accuracy of 52% and AUC score of 0.438. The level of accuracy is classified as poor and not feasible to use, therefore the researcher added the Particle Swarm Optimization (PSO) algorithm to determine the weight of each attribute. Table 3. below is the weight of each attribute obtained using the PSO algorithm.

Table 3. Weight of Each Attribute

Atribute	Weight
Scholarship	0
Single Tuition Category	0
Majoring Focus	0
Research Types	1.0
Research Member	0
Consultation frequency	0
Online/Offline Consultation	0.931
Lecturer Response Time	0
Lecturer Correction Time	1
Lecturer's Explanation	0
Differences of Opinion Between Lecturers I & II	0
Revision Begins	1
Discussion Friends	0.189
Boyfriend/Girlfriend	0
The frequency of friends' invites to do other things	0.333
Refusing a Friend's Invitation	0
Financial Problem	1
Family Problems	0.169
Family Communication	0.051
Family Talk	0
The Distance From Home to Campus	1
Campus Facilities	0.051
Place to Work	1
The Frequency of Visiting the Library	0
Device Condition	0
Working Frequency	0
Organitation Frequency	0
Hoby Frequency	1
The Frequency of Reading non-academic Things	1
Gaming Frequency	1
Social Media Frequency	0
Refreshing Frequency	0.308
Health	0.666
Praying frequency	1

From Table 3, it is known that there are 16 attributes with a weight of 0, which means those attributes will be removed, and the remaining 18 attributes will be reprocessed. The results of processing the selected attributes can be seen in Table 4 below.

Table 4. Accuracy, Precision, Recall, and AUC Score Second Result

Accuracy	AUC
76%	0.733

From the first result with decision tree and the second result with decision tree + PSO optimization, there is an increase in the value, which can be seen in Table 5 below.

Table 5. Result Comparison

	Accuracy	AUC
Decision tree	52%	0.438
Decission Tree + PSO	76%	0.733

From Table 5, it is known that there was an increase in accuracy of 24% from the initial 52% to 76%, and for the AUC value there was an increase of 0.295 from the initial 0.438 to 0.733. With that value, it can be concluded that the results of the classification model from this research are acceptable and can be applied to other data.

3.2. ANALYSIS

Below is the decision tree for the second result in the form of a tree (Figure 3).

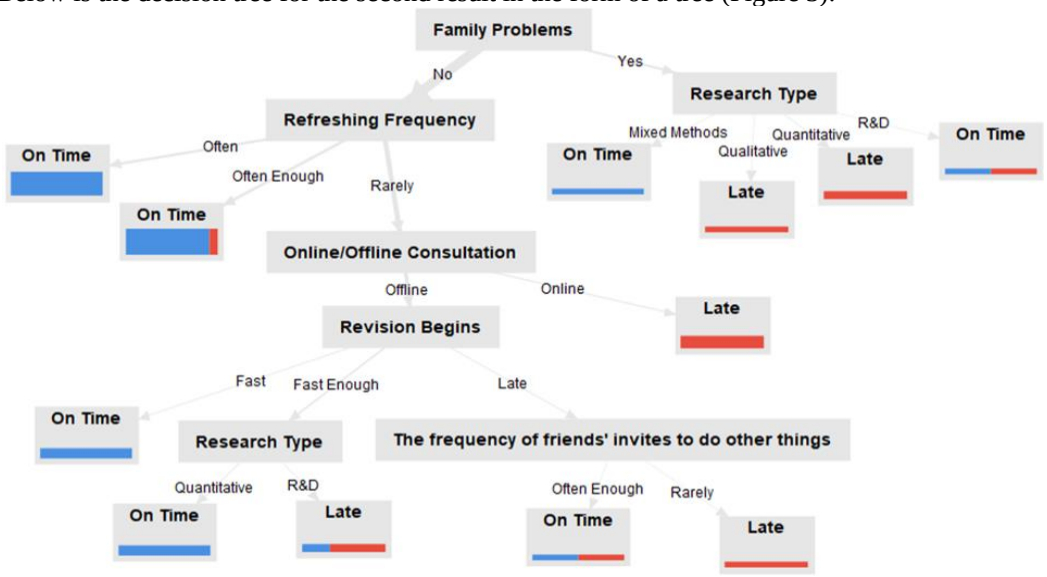


Figure 3. Decission Tree (Tree)

According on the decision tree shown in figure 3 above, family problems are the first factor to consider. If student have family problems, then pay attention to the research type. If it is qualitative or quantitative, it can be ascertained that it will be late in completing their thesis, but if it is R&D and mixed methods, it will be on time. In other cases, if students do not have family problems, what needs to be considered is the refreshing frequency. If it is frequent or frequent enough, it will be on time. However, if it is infrequent, check to see if the consultation is online or offline. If the online consultation is more frequent, it is certain to be late, but if the consultation is offline or face-to-face, then see when the student starts working on his revision. If it is immediately revised on the same day after receiving feedback from the lecturer, then it is ensured to be on time. However, if the revision begins 1-3 days after receiving feedback, consider the research type. If it is quantitative, then it will be on time, but if it is R&D, it will be late. Furthermore, if the revision starts more than 3 days after the feedback is received, it is necessary to look at the frequency with which friends are invited to do other things. If often enough, it will be on time, but if rarely enough, it will be late.

Based on the analysis, we obtained the rules as shown in Table 6 below:

Tabel 6. Rules

Rules	Class
Family Problems = Yes → Research Type = Qualitative	Late
Family Problems = Yes → Research Type = Quantitative	Late
Family Problems = Yes → Research Type = Mixed Methods	On Time
Family Problems = Yes → Research Type = R&D	On Time
Family Problems = No → Refreshing Frequency = Often	On Time
Family Problems = No → Refreshing Frequency = Often Enough	On Time
Family Problems = No → Refreshing Frequency = Rerely → Online/Offline Consultation = Online	Late
Family Problems = No → Refreshing Frequency = Rerely → Online/Offline Consultation = Offline → Revision Begins = Fast	On Time
Family Problems = No → Refreshing Frequency = Rerely → Online/Offline Consultation = Offline → Revision Begins = Fast Enough → Research Type = Quantitative	On Time
Family Problems = No → Refreshing Frequency = Rerely → Online/Offline Consultation = Offline → Revision Begins = Fast Enough → Research Type = R&D	Late
Family Problems = No → Refreshing Frequency = Rerely → Online/Offline Consultation = Offline → Revision Begins = Late → The frequency of friends' invites to do other things = Often Enough	On Time
Family Problems = No → Refreshing Frequency = Rerely → Online/Offline Consultation = Offline → Revision Begins = Late → The frequency of friends' invites to do other things = Rerely	Late

4. CONCLUSION

According to the rules, it is known that the factors that can affect the length of time students spend on completing the thesis are family problems, type of research, intensity of refreshing, online guidance, starting to improve revisions, and the intensity of friends asking to do other things. This study obtained an accuracy of 76%, and for the AUC value, a score of 0.733 was obtained, which was in the acceptable category. By knowing the factors that cause students to be fast or slow in completing thesis preparation, study programs and lecturers can take this research into consideration in preparing regulations to optimize the length of completion of the thesis, and for students, hopefully they can use this research as material for reflection in order to complete the thesis on time.

ACKNOWLEDGEMENTS

Thank you to all those who have helped and been involved in this research.

REFERENCES

[1] SALINAN-Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 Tentang Standar Nasional Pendidikan Tinggi, Kemdikbud, 2020, Pasal 5-6.

[2] Peraturan Rektor Universitas Sebelas Maret Nomor 31 Tahun 2020 Tentang Penyelenggaraan Dan Pengelolaan Program Sarjana, Universitas Sebelas Maret, 2020, Pasal 10.

[3] S. Haryati, A. Sudarsono, & E. Suryana. (2015). Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu). *Jurnal Media Infotama*. 11(2), pp. 130–138.

[4] G. A. Putri. (2020). Implementation of the C4.5 Algorithm to Predict Student Achievement at SMK Negeri 6 Surakarta. *Indonesian Journal of Informatics Education (IJIE).* [Online]. 4(2), pp. 51-61. Available: https://jurnal.uns.ac.id/ijie/article/view/47124/pdf

[5] E. Sabna & M. Muhardi. (2016). Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*. 2(2), pp. 41.

https://doi.org/10.24014/coreit.v2i2.2392
[6] J.Berry, & S. G. Linoff, "Data Mining Techniques for Marketing, Sales, and Customer Relationship Management", 2nd ed. Wiley Publishing, Inc., Indianapolis, Indiana, 2004.
[7] U. Fayyad, P.-S. Gregory, & P. Smyth. (1996). Knowledge discovery and data mining: Towards a Unifying Framework. KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Data Mining General Overview, pp. 82–88. Available: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf
[8] J. R. Quinlan, "C4.5: Programs For Machine Learning". Morgan Kaufmann Publishers, Inc., 1993.
[9] Kusrini, & E. Luthfi, "Algoritma Data Mining". Andi Offset. 2009. [Online]. Available: https://books.google.co.id/books?id=-Ojclag73O8C&printsec=frontcover&hl=id#v=onepage&q&f=false
[10] Y.-J. Cho, H.-S. Lee, & C.-H. Jun. (2011). Optimization of Decision Tree for Classification Using a Particle Swarm. *Industrial Engineering and Management Systems*, 10(4), pp. 272–278. https://doi.org/10.7232/iems.2011.10.4.272
[11] Sulistyanto. (2018, Aug). Penerapan C4.5 Berbasis Particle Swarm Optimization (PSO) dalam Memprediksi Siswa Lolos Seleksi Perguruan Tinggi. *Seminar Nasional Teknologi dan Bisnis 2018, IIB DARMAJAYA Bandar Lampung*. pp. 162-170.
[12] D. W. Hosmer, S. Lemeshow, & R. X. Sturdivant. Applied Logistic Regression, 3rd ed . John Wiley & Sons, Inc., 2013. https://doi.org/10.2307/2532419