International Journal of Information System & Technology
Vol. 3, No. 1, (2019), pp.31-36

The Application of Data Mining in Determining Timely Graduation Using the C45 Algorithm

Asro Pradipta[1], Dedy Hartama[2], Anjar Wanto[3], Saifullah[4], Jalaluddin[5]

[1,2,3,4,5]STIKOM Tunas Bangsa, Pematangsiantar, North Sumatra, Indonesia

asropradipta31@gmail.com

Abstract

Graduating on time is one element of higher education accreditation assessment. In the Strata 1 level, students are declared to graduate on time if they can complete their studies <= eight semesters or four years. BAN-PT sets a timely graduation standard of >= 50%. If the standard is not met, it will reduce the value of accreditation. These problems encourage the Universitas Simalungun Pematangsiantar to conduct evaluations and strategic steps in an effort to increase student graduation rates so that the targets of BAN-PT can be achieved. For this reason it is necessary to know in advance the pattern of students who tend not to graduate on time. In this study, C4.5 Algorithm is proposed to predict student graduation. This algorithm will process student profile datasets totaling 150 data. This dataset has a graduation status label. The value of the label is categorical, that is, right and late. The features or attributes used, namely the name of the student, gender, student status, GPA. The results of the C4.5 algorithm are in the form of a decision tree model that is very easy to analyze. In fact, even by ordinary people. This model will map the patterns of students who have the potential to graduate on time and late.

Keywords: *Data mining, Rapidminer, C4.5 Algorithm, Graduation, Pematangsiantar.*

1. Introduction

Students are intellectual figures who have high mobility as one of the biggest assets owned by a country. When a student can dedicate himself to the maximum, both his knowledge and experience to the wider community means that it can help the process of changing society that is more advanced. Basically students continue their education in Higher Education with the hope that they can attend education well. But this is not always the case, there are various problems they face relating to the study process of students in tertiary institutions. In completing it sometimes students face various obstacles that can hamper the completion of their study time which has implications for the graduation of the students themselves. Student graduation can be seen from the level of admission, ethics, activeness in the teaching and learning process, and academic achievement. In tertiary education institutions, graduation level information from students is very important to improve services that can make students comfortable so they can graduate on time.

Based on this we need a system that can provide decisions that can help related parties determine the graduation pattern of a student. many techniques in computer science can solve complex problems for those problems [1]–[7]. The settlement technique can use artificial intelligence [8]–[13]. There are several branches of artificial intelligence that can solve pattern cases. One of them is datamining [14]. With the datamining process, patterns or rules can be found that can be used to produce information by applying a decision tree technique. One well-known datamining technique is the C4.5 algorithm. The reason for using the C4.5 algorithm is because this algorithm can make rules in the form of patterns that can be done to determine whether a student's graduation is on time or not. This is

ISSN: 2580-7250 IJISTECH

reinforced by previous research which solved the problem by utilizing the classification data method C4.5. As was done [15] with the title application of the C4.5 algorithm to predict the recruitment of prospective new employees. In this study the C4.5 algorithm can be applied with the results of the method of measuring the success rate of prospective new employees by 71%. It is expected that the results of this study can provide solutions specifically to the Universitas Simalungun Pematangsiantar in determining the pattern of graduation of its students which has an impact on improving the quality of the tertiary institution.

2. Research Methodology

2.1. Datamining

Data mining is a scientific discipline that studies methods for extracting knowledge or finding patterns from data where the results of data processing can be used to make decisions in the future [16].

2.2. Classification

In classification, there are target variable categories. For example, income classifications can be separated into three categories, namely high income, medium income, low income [17].

2.3. C4.5 Method (Decision Tree)

Decision tree is one of the most popular classification methods because this algorithm converts data into decision trees and decision rules expressed in tabular form with attributes and records [17]. In addition, the decision tree combines data exploration and modeling so that it is very good as the initial step of modeling even when used as the final model of several other techniques.

2.4. Data source

Data collection methods carried out in this study consisted of interviews, observations, and litetarature studies. Activities undertaken in collecting data relate to determining the graduation of Universitas Simalungun students in a timely manner. The data obtained will then be processed using the C4.5 algorithm classification method by taking the value of each attribute in the data to determine timely graduation. The following is a proposed method for research using the C4.5 algorithm to determine timely graduation.

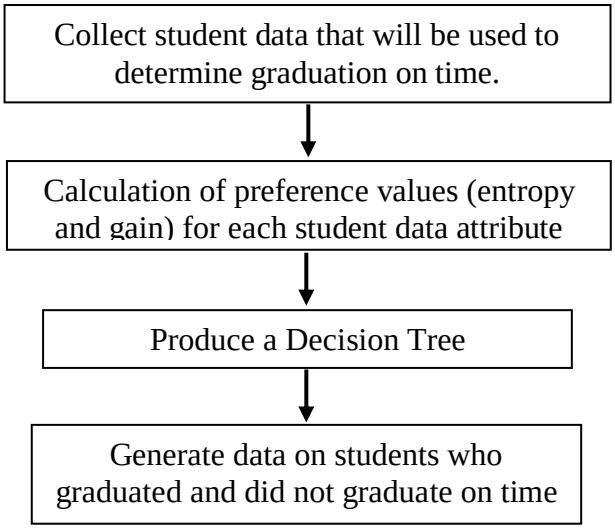

Figure 1. Proposed Research Methods

3. Results and Discussion

The dataset of the study consisted of criteria determined including: name of student, student status, sex, semester 1 to semester 8, grade point average and graduation status. Existing data is then transformed into the Microsoft Excel 2007 data format. The collected data is used as input data in creating rule models using the C4.5 algorithm using rapidminer software to display an overview of rule models in determining student graduation on time.

3.1. Research Dataset

In displaying data modeling using the C4.5 algorithm the decision tree method is used. The data used are Simalungun University student data of 150 records.

Table 1. Students data of Universitas Simalungun

No	Mahasiswa	Student Status	Gender	Grade Point								GPA	Graduation Status
				Sem 1	Sem 2	Sem 3	Sem 4	Sem 5	Sem 6	Sem 7	Sem 8		
1	Vinkhi F Saragih	Student	Male	3,67	3,45	3,27	3,43	3,00	3,40	3,50	3,92	3,42	On Time
2	Wahyu Prasetyo	Student	Male	3,72	3,40	3,64	3,57	3,86	3,25	0,50	3,75	3,59	Late
3	Eko Supiandi	Work	Male	3,17	3,35	3,45	3,57	3,29	3,55	3,29	3,33	3,38	Late
4	Natal Ingot Sirait	Work	Male	3,28	3,30	3,64	3,14	3,14	3,25	3,43	3,80	3,36	Late
5	Petrus R. Sihombing	Student	Male	3,44	3,30	3,59	3,14	3,57	2,81	3,57	4,00	3,43	Late
6	Abisaleh Zebua	Work	Male	3,17	3,00	3,32	3,00	3,14	3,00	3,29	3,60	3,18	On Time
7	Ricky H Simatupang	Work	Male	2,89	2,60	2,73	3,00	3,29	2,71	3,25	3,83	3,02	Late
8	Hari Susanto	Work	Male	3,33	3,00	3,05	2,86	3,00	3,25	3,43	3,60	3,15	Late
9	Suci AnggunTari	Work	Female	3,11	3,00	3,32	3,00	3,14	3,00	3,29	3,60	3,18	On Time
10	Cici Suryani	Student	Female	3,06	2,95	3,23	3,14	3,57	3,25	3,57	3,80	3,31	On Time
...	...	...	...	...	...	...	...	...	...	...	...	...	...
140	Andika Lesmana	Work	Male	3,27	3,30	3,50	3,42	3,71	3,75	3,43	3,88	3,51	On Time
141	Hotni M Saragih	Work	Female	3,28	3,00	3,00	3,42	3,42	3,40	3,00	3,60	3,29	On Time
142	Fadlan Suhanda	Student	Male	3,50	3,15	3,05	3,14	3,00	3,00	3,13	2,00	3,04	Late
143	Farius Waruwu	Student	Male	3,33	3,30	3,36	3,29	3,43	3,25	3,29	3,60	3,34	On Time
144	Fajar setiawan	Student	Male	3,44	3,05	3,45	3,42	3,28	3,60	3,00	4,00	3,42	On Time
145	Nurul Fadillah	Work	Female	3,00	3,30	3,27	3,43	3,43	3,10	3,29	3,90	3,31	On Time
146	Bagus Wijaya	Work	Male	3,17	3,00	3,00	3,29	3,00	2,85	3,29	3,30	3,10	Late
147	Karmen T Sinaga	Work	Male	3,56	3,75	3,64	3,29	3,57	3,10	3,29	3,60	3,46	On Time
148	Dian Permatasari	Student	Female	3,11	3,40	3,41	3,29	3,57	3,70	3,28	2,26	3,25	On Time
149	Dewi Novika Sari	Student	Female	3,38	3,05	3,59	3,14	3,14	3,30	3,14	3,67	3,26	On Time
150	Fitri N Urika	Work	Female	3,00	3,00	3,64	3,29	3,14	3,25	3,38	3,67	3,28	On Time

Source: Universitas Simalungun

3.3. Classification Analysis of method C4.5

Following are the steps in forming a decision tree using the C4.5 algorithm. In making a decision tree first count the number of cases for a decision. Yes, the number of decision cases is not, and the entropy of all cases is based on attributes. Then calculate the entropy value of each case based on the specified attributes.

a) Calculate entropy.
From the training data it is known that the number of cases there are 150 records, with late graduation status there are 60 records and on time graduation status there are 90 records so that entropy the total root value obtained:
Entropy(S) =$\sum_{i=1}^{n} -pi * log_2 pi$
= (-60/150.log2 (60/14)) + (-90/14.log2(90/14))
= 0,970950595

b) Calculating the entropy value on the gender attribute with male category, the number of cases there are 79 records, with the status of graduation late there are 36 records and graduation status on time there are 43 records so that the total root entropy value obtained:
Entropy(S) =$\sum_{i=1}^{n} -pi * log_2 pi$
= (-36/79 . log2 (36/79)) + (-43/79 . log2 (43/79))
= 0,994329046

c) Calculating the entropy value on the gender attribute with the Female category, the number of cases there are 71 records, with a graduation status late there are 25 records and graduation status on time there are 46 records so that the total root entropy value obtained:
Entropy(S) =$\sum_{i=1}^{n} -pi * log_2 pi$
= (-25/71 . log2 (25/71)) + (-46/71 . log2 (46/71))
= 0,935940715

d) Calculating the value of entropy in the Student Status attribute with the Working category, the number of cases there are 65 records, with a graduation status late there are 10 records and timely graduation status there are 55 records so that the total root entropy value obtained:
Entropy(S) =$\sum_{i=1}^{n} -pi * log_2 pi$
= (-10/65 . log2 (10/65)) + (-55/65 . log2 (55/65))
= 0,619382195

e) Calculating the value of entropy in the Student Status attribute with the Student category, the number of cases there are 85 records, with a graduation status late there are 24 records and graduation status on time there are 61 records so that the total root entropy value obtained:
Entropy(S) =$\sum_{i=1}^{n} -pi * log_2 pi$
= (-24/85 . log2 (24/85)) + (-61/85 . log2 (61/85))
= 0,858637082

From the calculation of entropy value and gain value for each attribute can be seen in the following Node 1 gain calculation table:

Table 2. Calculation of Gain Node 1

Node		amount Case	Late (S1)	On Time (S2)	Entropy	Gain
Total		150	60	90	0,970950595	
Gender						0,016546181
	Male	79	36	46	0,970998371	
	Female	71	25	46	0,935940715	
Status student						0,21599063
	Work	85	24	61	0,858637082	
	Student	65	10	55	0,619382195	

From the results of entropy and gain counts in Table 1. it appears that the attribute status of students has the highest gain value that is 0.21599063. Therefore the student status attribute becomes the first root or node of the decision tree formed with 2 attribute values of the student status namely work and students. Count the number of cases, the number of cases for a decision Yes, the number of cases for a decision No, and the entropy of all cases by using the status attribute of working students. Then do the calculation again to get entropy and gain. From the calculation of the entropy value and the gain value for each attribute can be seen in the calculation table of gain Node 1.1. following.

Table 3. Calculation of Gain Node 1.1

Node		amount Case	Late (S1)	On Time (S2)	Entropy	Gain
Total Work		85	24	61	0,858637082	

Node		amount Case	Late (S1)	On Time (S2)	Entropy	Gain
Jenis Kelamin						0,015910807
	Male	59	14	45	0,790501384	
	Female	26	10	16	0,961236605	

In table 3 it is known that the results of the gender attribute have a gain value of 0.015910807. Therefore, no further calculation is needed for the value of this attribute. After the results of entropy and gain calculations are made, a decision tree is formed by modeling the rules using the Rapidminer software.

Based on the shape of the decision tree formed the rule model is in the form of text as follows:

GPA > 3,235: On Time { On Time = 75, Late = 18}
GPA ≤ 3,235: Late { On Time = 15, Late = 40}

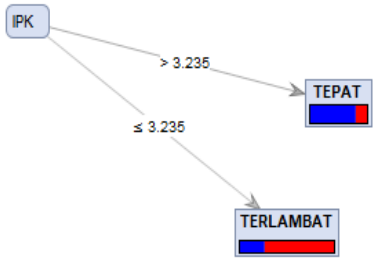

Figure 2. Student Graduation Decision Tree

From the results of the decision tree formed the rule model is obtained from determining timely graduation at Universitas Simalungun.

4. Conclusion

Based on the results of research in determining timely graduation at the Universitas Simalungun Pematangsiantar, it can be concluded that:

a) Obtained a rule model that can show the rules of connectedness between gender attributes, student status and achievement index scores from semester 1 to 8 and from the research results obtained a model of timely graduation rules is based on a cumulative achievement index.
b) Problems in determining timely graduation Universitas Simalungun Pematangsiantar can be solved by applying data mining techniques, namely the C4.5 Algorithm.
c) Classification of student data at the Simalungun Pematangsiantar University with the C4.5 Algorithm can be a support in the application in the process of determining the timely graduation used by the administration of the Universitas Simalungun Pematangsiantar.

References

[1] I. Parlina, A. Wanto, and A. P. Windarto, "Artificial Neural Network Pada Industri Non Migas Sebagai Langkah Menuju Revolusi Industri 4.0," *InfoTekJar J. Nas. Inform. dan Teknol. Jar.*, vol. 4, no. 1, pp. 155–160, 2019.

[2] A. P. Windarto and S. S, "Penerapan Algoritma Semut dalam Penentuan Distribusi Jalur Pipa Pengolahan Air Bersih," *J. Sist. Inf. Bisnis*, vol. 2, pp. 123–132, 2018.

[3] B. Fachri, A. P. Windarto, and I. Parinduri, "Penerapan Backpropagation dan Analisis Sensitivitas pada Prediksi Indikator Terpenting Perusahaan Listrik," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, pp. 202–208, 2019.

[4] C. Astria, A. P. Windarto, and Z. Musiafa, "Pemilihan Produk Sampo Sesuai Jenis Kulit Kepala Dengan Metode Promethee II," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 178–185, 2019.
[5] D. R. S. P, A. A. Muin, and M. Amin, "Pemilihan Facial Wash Untuk Kulit Wajah Berminyak Dengan Metode Promethee II," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 222–229, 2019.
[6] S. M. Dewi and A. P. Windarto, "Analisis Metode Electre Pada Pemilihan Usaha Kecil Home Industry Yang Tepat Bagi Mahasiswa," *Sist. J. Sist. Inf.*, vol. 8, no. 3, pp. 377–385, 2019.
[7] C. Fadlan, A. P. Windarto, and I. S. Damanik, "Penerapan Metode MOORA pada Sistem Pemilihan Bibit Cabai (Kasus : Desa Bandar Siantar Kecamatan Gunung Malela)," *J. Appl. Informatics Comput.*, vol. 3, no. 2, pp. 42–46, 2019.
[8] T. Budiharjo, Soemartono, T., Windarto, A.P., Herawan, "Predicting tuition fee payment problem using backpropagation neural network model," *Int. J. Adv. Sci. Technol.*, 2018.
[9] T. Budiharjo, Soemartono, T., Windarto, A.P., Herawan, "Predicting school participation in indonesia using back-propagation algorithm model," *Int. J. Control Autom.*, 2018.
[10] A. P. Windarto, M. R. Lubis, and Solikhun, "Implementasi JST Pada Prediksi Total Laba Rugi Komprehensif Bank Umum Konvensional Dengan Backpropagation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 411–418, 2018.
[11] B. Febriadi and A. Zamsuri, "RDBMS Applications as Online Based Data Archive: A Case of Harbour Medical Center in Pekanbaru," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 97, no. 1, pp. 1–5, 2017.
[12] A. P. Windarto, M. R. Lubis, and Solikhun, "Model Arsitektur Neural Network Dengan Backpropogation Pada Prediksi Total Laba Rugi Komprehensif Bank Umum Konvensional," *Kumpul. J. Ilmu Komput.*, vol. 5, no. 2, pp. 147–158, 2018.
[13] S. R. Ningsih and A. P. Windarto, "Penerapan Metode Promethee II Pada Dosen Penerima Hibah P2M Internal," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 1, pp. 20–25, 2018.
[14] A. P. Windarto, "Penerapan Data Mining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Clustering," *Techno.COM*, vol. 16, no. 4, pp. 348–357, 2017.
[15] F. F. Harryanto and S. Hansun, "Penerapan Algoritma C4 . 5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE," *Jatisi*, vol. 3, no. 2, pp. 95–103, 2017.
[16] H. Sulastri, A. I. Gufroni, and K. Kunci, "Jurnal Teknologi dan Sistem Informasi Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017.
[17] S. Haryati, A. Sudarsono, and E. Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama Vol.*, vol. 11, no. 2, pp. 130–138, 2015.

Authors

1st Author
Asro Pradipta
STIKOM Tunas Bangsa, Pematangsiantar, North Sumatra