e-ISSN. 2476-9843
p-ISSN. 1858-4144

DOI https://doi.org/10.30812/matrik.v19i2.685

Graduation Prediction System On Students Using C4.5 Algorithm

Donny Kurniawan[1], Anthony Anggrawan[2], Hairani[3]
[1,2,3]computer science, Universitas Bumigora
Email: donnydkg25@gmail.com, anthony@universitasbumigora.ac.id, hairani@universitasbumigora.ac.id

ABSTRACT

Bumigora University College there are several things that are not balanced between the entry and exit of students who have completed their studies. Students who enter in large numbers, but students who graduate on time below the specified standards. As result, there was a huge accumulation of students in each graduation period. One solution to overcome the problem above needs a data mining based system in monitoring or utilizing student development in predicting graduation using the C4.5 algorithm. The stages of this research began with problem analysis, data collection, data requirement analysis, data design, coding, and testing. The results of this study are the implementation of the C4.5 algorithm for predicting student graduation on time or not. The data used is the data of students who have graduated from 2010 to 2012. The level of acceptance generated using the confusion matrix is 93,103% accuracy using 163 training data and 29 testing data or 85% training data and 15% testing data. The results of research and testing that has been done, C4.5 algorithm is very suitable to be used in student graduation prediction.

Keyword*: *C4.5 Algorithm, Data Mining, Student Graduation , Prediction

*Author Korespondensi (**Donny Kurniawan**)*
Email : donnydkg25@gmail.com

I. INTRODUCTION

Graduation rate is one of success indicator of Higher Education in the implementation of teaching and learning process. One of assessment element for college accreditation is the timely graduation rate of students. Therefore, an application can be alternative for monitoring and evaluating the tendency of students to graduate timely or not [1].

At Bumigora University there is an imbalance between new students and timely-graduated students. As a result, there was a high number of students in each graduation period [2].

The data from the Departments and Pustik, from 310 of registered students in 2010, 2011 and 2012, students who graduated in 2010 were around 100 students, students graduated in 2011 about 140 students and students graduated in 2012 who graduated about 70 students. From the above data, it is shown that there is an imbalance between incoming students and graduated students so that it can result in a negative image for Bumigora University. Therefore, the campus needs to identify the tendency of students' graduation.

To improve the quality and accreditation of Bumigora University, students filter and graduation prediction are needed for new students because it directly affects the room capacity , the ratio of lecturers to students, and parking capacity. Prediction of student graduation is one of the appropriate methods to form patterns that might provide useful information on large amount data of student [2]. The solution can be a data mining based system to monitor students' progress in predicting graduation using the C4.5 Algorithm method.

Data Mining refers to the process of finding unknown information from a large set of data [3]. One of the data mining method used in this research is Alogaritm C4.5.

Algorithm C4.5 is a clarification method to form a decision diagram based on training data. C4.5 algorithm can form a decision diagram and can be used in data processing of continuous and discrete figures, handle missing attribute values, and form rules that are easily interpreted [4].

Several previous research on graduation time prediction is using quite a lot of data mining techniques such as alogaritm C4.5 [5] [6], Naïve Bayes [7]. K-Nearest Neighbor (KNN) [8], and Regression Tree (CART) [9]. Research [10] uses the C4.5 method for the graduation of informatics study programs with an accuracy of 62.44%. Research [11] uses the C4.5 method to predict student graduation with an accuracy of 80.47%. While research [7] uses the Naive Bayes method for predicting student graduation with an accuracy of 73.73%.

Research [12] explains that the most influential factor in determining the classification of academic performance of students is the Cumulative Achievement Index (GPA), Semester Achievement Index (IPS) semester 1, IPS semester 6, and gender. In this study, researchers used the C4.5 algorithm in determining graduation predictions based on gender attributes, from high school and social studies from semester 1 to semester 6.

An application program was created to support the monitoring of the study program or to monitor the progress of students in predicting their time of graduation using the C4.5 algorithm, so it will support to increase the quality and the accreditation of the university and to determine the accuracy of the C4.5 algorithm

II. METHODOLOGY

This research begins with the stage of problem analysis. Problem analysis is used to determine the scope of the problem. The second stage was collecting the data of the university of Bumi Gora graduate students from 2010-2012 obtained. The third stage is the needs analysis which aims to find out the features needed. The next stage will be the design to model the needs analysis in diagram form. The next stage is the used of coding to interpret the design of results into the form of program syntax.

The third stage is to create a needs analysis to find out what features are needed. The next stage is the design used to model the needs analysis in diagram form. The next stage is the coding used for the interpretation of design results in the form of program syntax. The programming languages used in this research are PHP and Mysql as the database. The last stage is testing. This study uses confusion matrix testing to determine the performance of the C4.5 algorithm based on accuracy, sensitivity, and specificity. Data is tested using the split validation method. The data used in this study were 192 instances, 163 instances as training data, and 29 instances as testing data.

In general, an overview of the stages of this study is shown in Figure 1.

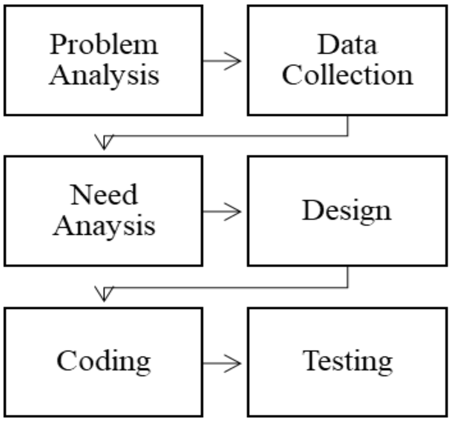


Figure 1. Research Stage

2.1. Algoritma C4.5

C4.5 Algorithm is one algorithm to convert large facts into a decision tree that represents the rules. The purpose of forming a decision tree in the C4.5 algorithm is to make it easier to solve existing problems. There are stages in changing the C4.5 algorithm to be rules [13].

The followings are The steps in making a decision tree with C4.5 algorithm [14]:

1. Preparing data training.
2. Counting the roots of the tree. The root will be taken from the selected attribute, by calculating the gain value of each attribute, the highest gain value will be the first root. Before calculating the gain value of the attribute, calculate the entropy value. The following formula can be used to calculate the entropy value:

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

   Code meaning:
   S = a set of case
   N = amount of partition S
   Pi = S proportion toward S

3. Calculating Gain value using the following formula:

   $$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{S} * \ Entropy(S_i) \quad (2)$$

   Code meaning:
   S = A set of cases
   A = features
   N = amount of atribut partition *A*
   $|S_i|$ = Proportion S_i to *S*
   $|S|$ = amount of case *S*

III. RESULTS AND DISCUSSION

3.1.Problem analysis

A graduation rate becomes one of succes indicator or benchmark of Higher Education in implementing the teaching and learning process. Students' ability to complete their studies on time includes in assessment element of the University's accreditation, for this reason there should be monitoring and evaluation on the students' graduation tendency.

There are several things that must be considered, namely the imbalance between registered student and graduating students, a large number of registered students is imbalance toward the timely-graduated students . As a result, there was a high number of students in each graduation period.

Therefore, there should be a filter on the registered students as well as students graduation prediction since it influence the room capacity, lecturer's ratio to students and parking capacity. Predicting students's graduation is a propriate method to form patterns that give useful indication on the large amount of students.

3.2.Data collection

1. Interview

   Interview is conducted via direct communication to Pustik department and departments of Bumigora University. It aims at identifying the students data in Computer science department as well as problems occurred in Pustik department and department.
2. Library study

   This method is conducted by studying and searching references in the journals, books, internet and other literature related to the research

3.3.Functional need analysis

Functional need analysis decribes the range of activity that is implemented in the system and explaining the system need in order that the system runs well based on the need.

Table 1. Fungtional Need of Student and Admin

No.	*User*	Rmark
1.	Students	1.Log in to the system 2.Manage data testing (*create, update, delete*) 3.Seeing data result 4.See rules or diagram 5.Able to print prediction result 6.Able to *logout*
2.	Admin	1. Log in in to the system as admin. 2.Manage data *training* (*create, update, delete*) 3.Manage data *testing* (*create, update, delete*) 4.Manage student data (*create, update, delete*) 5. Seeing prediction result 6.*Export* student data 7.Able to print prediction result 8. Able to *logout*

3.4.Non Fungtional Need Analysis

Non-functional analysis of this resarch is used to analyze the system users. As for the user of this students graduates prediction applications are:

- Admin is an actor who take part in managing data throughout the System, such as managing student data, training data, testing data, seeing prediction results, and printing reports.
- Students have the role to manage testing data, see prediction results, print reports, and view decision trees.

3.5.Design Diagram

1. *Use Case*

The design of the diagrams is used to illustrate the interaction occurs between the user and the system so that the user boundaries will be seen while using the application. The use case diagram will be shown in figure 2.

DOI https://doi.org/10.30812/matrik.v19i2.685

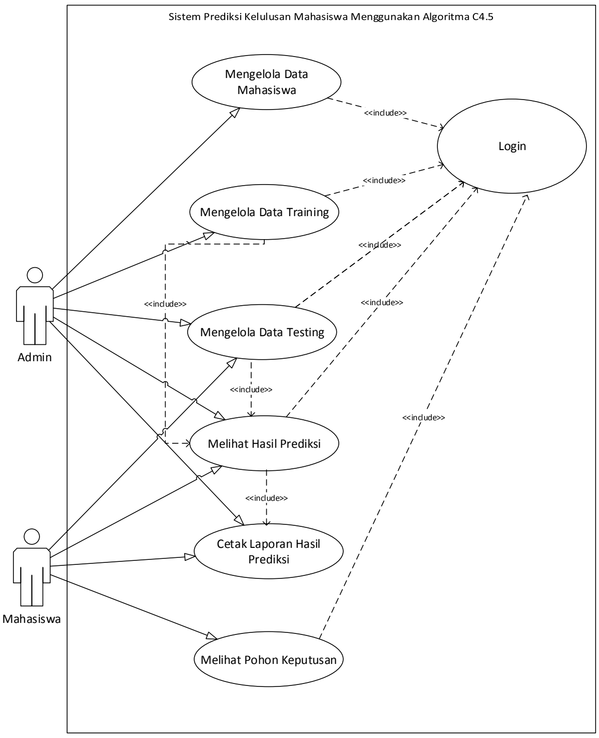
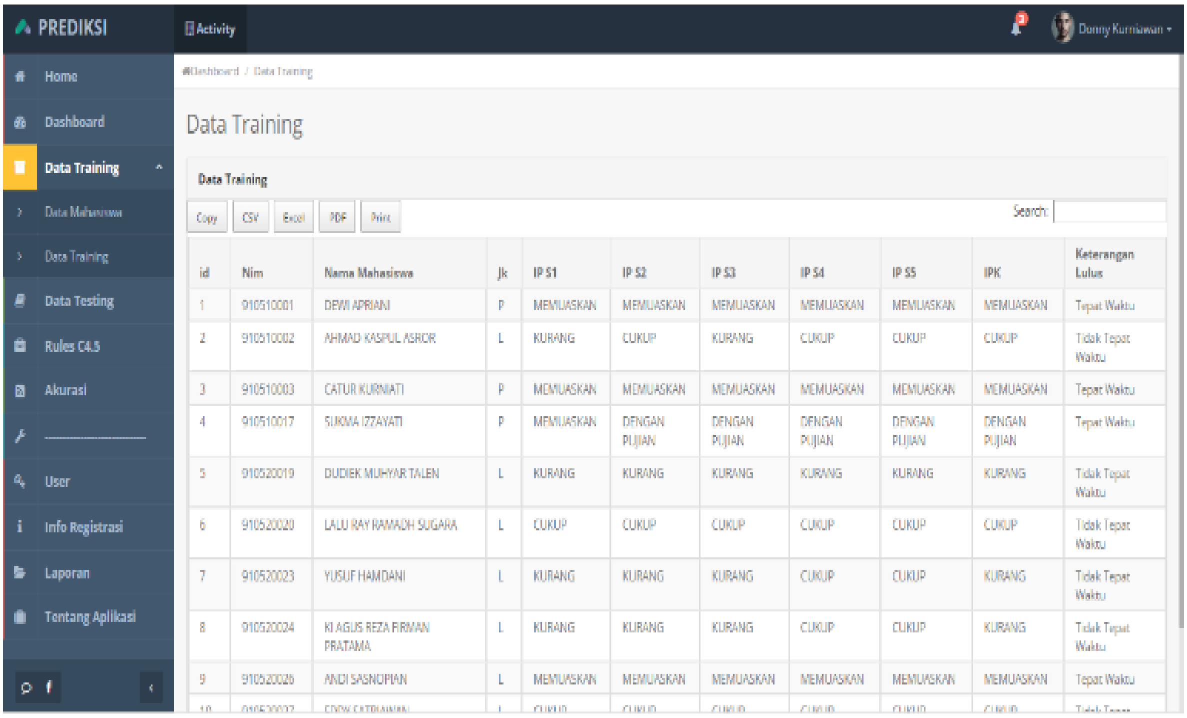


Figure 2. *Use Case* Sistem

3.6.CODING

The programming language and database used to create student graduation prediction system applications using the C4.5 algorithm is the PHP and Mysql programming languages as the database. As for some of the features displayed in this study, such as training data pages (Figure 3), testing data pages (Figure 4), and the rule page (Figure 5).



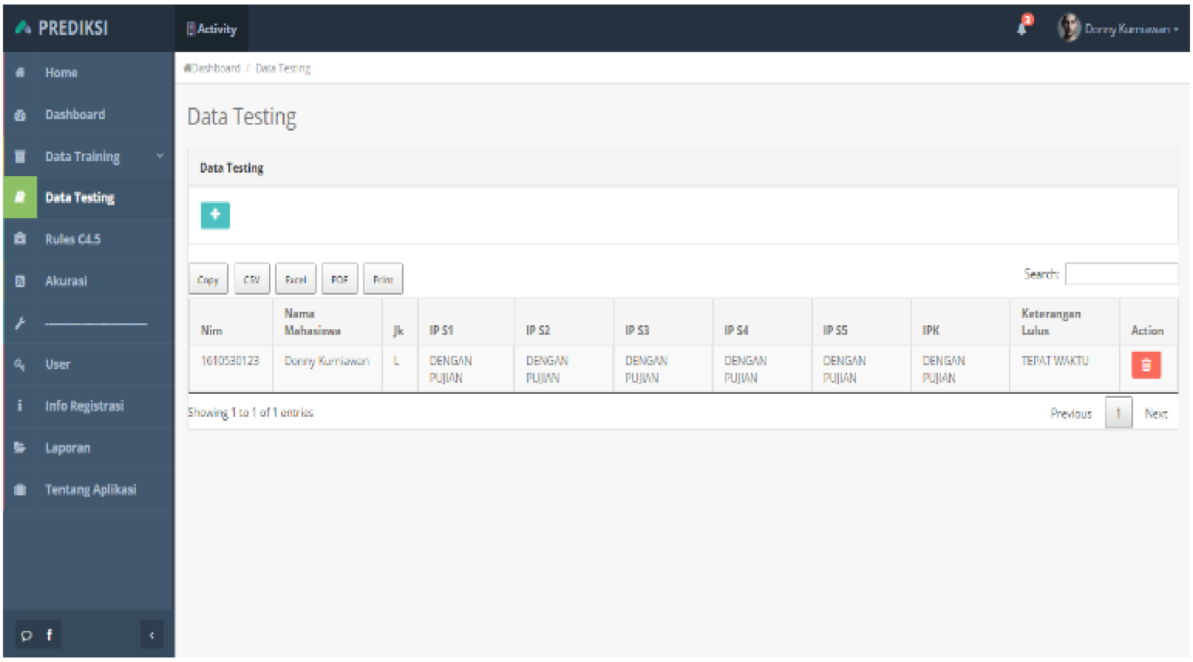
Data Training

Data Training

id	Nim	Nama Mahasiswa	Jk	IP S1	IP S2	IP S3	IP S4	IP S5	IPK	Keterangan Lulus
1	910510001	DEWI APRIANI	P	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	Tepat Waktu
2	910510002	AHMAD KASPUL ASROR	L	KURANG	CUKUP	KURANG	CUKUP	CUKUP	CUKUP	Tidak Tepat Waktu
3	910510003	CATUR KURNIATI	P	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	Tepat Waktu
4	910510017	SUKMA IZZAYATI	P	MEMUASKAN	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	Tepat Waktu
5	910520019	DUDIEK MUHYAR TALEN	L	KURANG	KURANG	KURANG	KURANG	KURANG	KURANG	Tidak Tepat Waktu
6	910520020	LALU RAY RAMADH SUGARA	L	CUKUP	CUKUP	CUKUP	CUKUP	CUKUP	CUKUP	Tidak Tepat Waktu
7	910520023	YUSUF HAMDANI	L	KURANG	KURANG	KURANG	CUKUP	CUKUP	KURANG	Tidak Tepat Waktu
8	910520024	KI AGUS REZA FIRMAN PRATAMA	L	KURANG	KURANG	KURANG	CUKUP	CUKUP	KURANG	Tidak Tepat Waktu
9	910520026	ANDI SASNOPIAN	L	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	MEMUASKAN	Tepat Waktu
10	910520027	EDDY SATRIAWAN	L	CUKUP	CUKUP	CUKUP	CUKUP	CUKUP	CUKUP	Tidak Tepat

Figure 3. Page of Data training

Terakreditasi SINTA 4, SK No. 3/E/KPT/2019



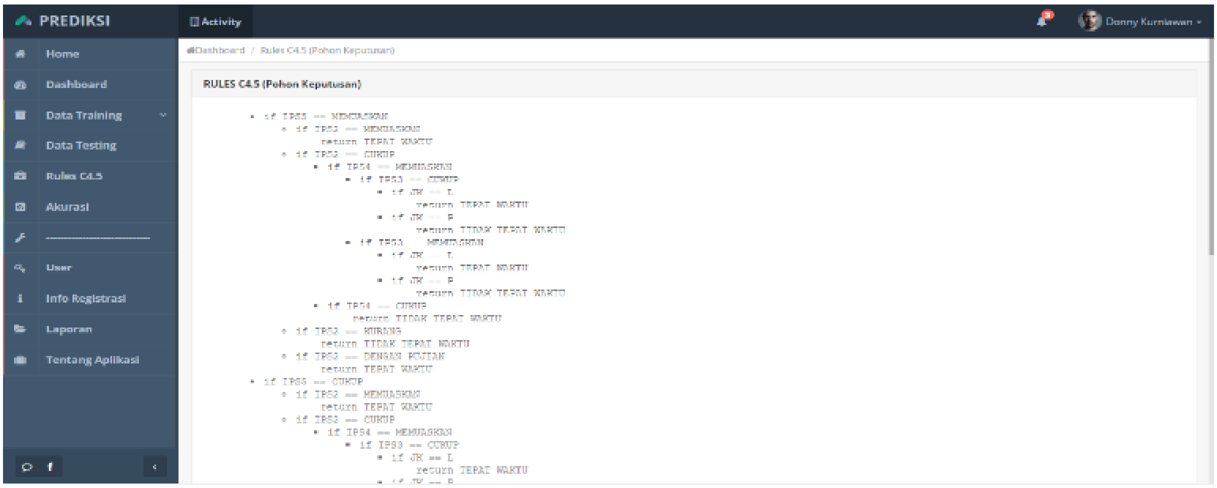
Data Testing

Data Testing

Nim	Nama Mahasiswa	Jk	IP S1	IP S2	IP S3	IP S4	IP S5	IPK	Keterangan Lulus	Action
1610530123	Donny Kurniawan	L	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	DENGAN PUJIAN	TEPAT WAKTU	🗑

Showing 1 to 1 of 1 entries

Figure 4. Page of Data *Testing*



RULES C4.5 (Pohon Keputusan)

```
• if IPS5 == MEMUASKAN
    ◦ if IPS2 == MEMUASKAN
        return TEPAT WAKTU
    ◦ if IPS2 == CUKUP
        ▪ if IPS4 == MEMUASKAN
            ▪ if IPS3 == CUKUP
                ▪ if JK == L
                    return TEPAT WAKTU
                ▪ if JK == P
                    return TIDAK TEPAT WAKTU
            ▪ if IPS3 == MEMUASKAN
                ▪ if JK == L
                    return TEPAT WAKTU
                ▪ if JK == P
                    return TIDAK TEPAT WAKTU
        ▪ if IPS4 == CUKUP
            return TIDAK TEPAT WAKTU
    ◦ if IPS2 == KURANG
        return TIDAK TEPAT WAKTU
    ◦ if IPS2 == DENGAN PUJIAN
        return TEPAT WAKTU
• if IPS5 == CUKUP
    ◦ if IPS2 == MEMUASKAN
        return TEPAT WAKTU
    ◦ if IPS2 == CUKUP
        ▪ if IPS4 == MEMUASKAN
            ▪ if IPS3 == CUKUP
                ▪ if JK == L
                    return TEPAT WAKTU
                ▪ if JK == P
```

Figure 5. Page of *Rule*

Result of Decission Diagram Modeling and Rule

The result shown in the decision diagram in image 6 and rule in table 2 is the result of 163 instances used training data.

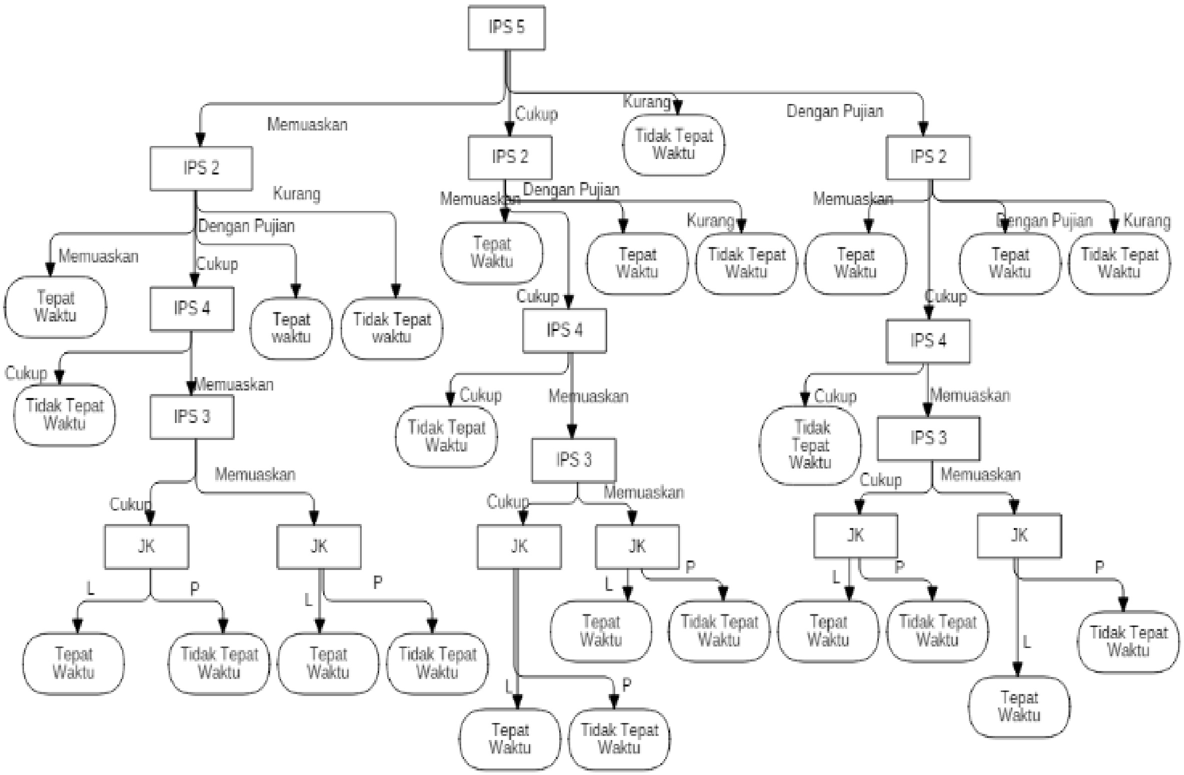


Figure 6. Decission Tree

From the decision diagram above, we can analyze that the fist root node or criteria checked is IP semester 5. The next stage of node checking will depend on whether the result enough, satisfactory, or pass with a compliment. If the result is shown as less, then the decision will be not on time.

Tabel 2. *Rule*

No	***RULES***
1.	**IF** IPS 5 LESS **THEN** LATE
2.	**IF** IPS 5 SATISFYING **AND** IPS 2 LESS **THEN** LATE
3.	**IF** IPS 5 SATISFYING **AND** IPS 2 CUMLAUDE **THEN** ON TIME
4.	**IF** IPS 5 SATISFYING **AND** IPS 2 SATISFYING **THEN** ON TIME
5.	**IF** IPS 5 SATISFYING **AND** IPS 2 FAIR **AND** IPS 4 FAIR THEN LATE
6.	**IF** IPS 5 SATISFYING **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 FAIR **AND** JK P **THEN** LATE
7.	**IF** IPS 5 SATISFYING **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 FAIR **AND** JK L **THEN** ON TIME
8.	**IF** IPS 5 SATISFYING **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 SATISFYING **AND** JK P **THEN** LATE
9.	**IF** IPS 5 SATISFYING **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 SATISFYING **AND** JK L **THEN** ON TIME
10.	**IF** IPS 5 FAIR **AND** IPS 2 LESS **THEN** LATE
11.	**IF** IPS 5 FAIR **AND** IPS 2 CUM LAUDE **THEN** ON TIME
12.	**IF** IPS 5 FAIR **AND** IPS 2 SATISFYING **THEN** ON TIME
13.	**IF** IPS 5 FAIR **AND** IPS 2 FAIR **AND** IPS 4 FAIR **THEN** LATE
14.	**IF** IPS 5 FAIR **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 FAIR **AND** JK P **THEN** LATE
15.	**IF** IPS 5 FAIR **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 FAIR **AND** JK L **THEN** ON TIME
16.	**IF** IPS 5 FAIR **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 SATISFYING **AND** JK P **THEN** LATE
17.	**IF** IPS 5 FAIR **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3

	SATISFYING **AND** JK L **THEN** ON TIME
18.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 LESS **THEN** LATE
19.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 CUM LAUDE **THEN** ON TIME
20.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 SATISFYING **THEN** ON TIME
21.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 FAIR **AND** IPS 4 FAIR **THEN** LATE
22.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 FAIR **AND** JK P **THEN** LATE
23.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 FAIR **AND** JK L **THEN** ON TIME
24.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 SATISFYING **AND** JK P **THEN** LATE
25.	**IF** IPS 5 CUM LAUDE **AND** IPS 2 FAIR **AND** IPS 4 SATISFYING **AND** IPS 3 SATISFYING **AND** JK L **THEN** ON TIME

Modeling Result Evaluation

To know the performance of C4.5 algorithm based on accuracy, sensitivity and specificity, Confusion matrix table (table 3) can be used. The data used to examine the performance is 15% or 29 instances as data testing and 85% or 163 instances as data training. The result using Weka Application with data training 85% (163 instances) and testing 15% (29 instances) is accuracy 93,10%, sensitivity 77% and specificity 100%.

Tabel 3. Comfusion matrix of Algorithm C4.5

Actual	**Prediction**	
	On time	**Late**
On time	*7 (TP)*	*0 (FN)*
Late	*2 (FP)*	*20 (TN)*

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$= \frac{7 + 20}{7 + 20 + 0 + 2}$$

$$= 0{,}93103\ x\ 100\% = 93{,}10\%$$

$$sensitivity = \frac{TP}{TP + FP}$$

$$= \frac{7}{7 + 2} = 0.77\ x\ 100\% = \ 77\%$$

$$specificity\ = \frac{TP}{TP + FN}$$

$$= \frac{7}{7 + 0} = 1\ x\ 100\% = 100\%$$

Table 4 is the data from previous research accuracy comparison.

Table 4. Result from previous research accuracy comparision.

No.	Method	Accuracy
1.	Research [10] Method C4.5	62,44%
2.	Reasearch [7] Method Naive Bayes	73,73%
3.	Research [11] Method C4.5	80,47%
4.	**Current Research (Method C4.5)**	**93,10%**

Based on table 4 suggested in current research, it is shown its best accuracy: 93,10% compared to the previous research.

IV. CONCLUSION AND SUGGESTION

Conclusion

Based on planning, implementation, and conducted assessment on students graduates system prediction using algorithms C4.5 it is shown that:

1. Algorithms implementation C4.5 it is very decent to be used as a students's prediction for graduation.
2. In this research, the level of accuracy is 93,10%, sensitivity is 77%, and specificity is 100%

Suggestion

The suggestion for this research is to used a sampling method like in SMOTE to resolve the imbalance problem in the data used.

REFERENCE

[1] A. G. Novianti and Di. Prasetyo, “Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Waktu Kelulusan Mahasiswa,” *Semin. Nas. APTIKOM*, no. November, 2017.

[2] A. Panoto, Y. R. W. Utami, and W. L. YS, “Penerapan Algoritma K-Nearest Neighbors

Uuntuk Prediksi Kelulusan Mahasiswa Pada Stmik Sinar Nusantara Surakarta," *J. TIKomSiN*, no. 2338–4018, pp. 27–31, 2017.

[3] Y. Yulia and N. Azwanti, "Penerapan Algoritma C4.5 Untuk Memprediksi Besarnya Penggunaan Listrik Rumah Tangga di Kota Batam," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 584–590, 2018, doi: 10.29207/resti.v2i2.503.

[4] R. H. Pambudi and B. D. Setiawan, "Penerapan Algoritma C4 . 5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 7, pp. 2637–2643, 2018.

[5] E. Purnamasari, D. P. Rini, and Sukemi, "Prediction of the Student Graduation's Level using C4.5 Decision Tree Algorithm," in *International Conference on Electrical Engineering and Computer Science (ICECOS) 2019*, 2019, pp. 192–195, doi: 10.1109/icecos47637.2019.8984493.

[6] A. A. Supianto, A. Julisar Dwitama, and M. Hafis, "Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University," in *3rd International Conference on Sustainable Information Engineering and Technology, SIET 2018 - Proceedings*, 2018, pp. 308–311, doi: 10.1109/SIET.2018.8693158.

[7] E. Sutoyo and A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95–101, 2020, doi: 10.29207/RESTI.V4I1.1502.

[8] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019.

[9] A. Maesya and T. Hendiyanti, "Forecasting Student Graduation with Classification and Regression Tree (CART) Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 621, no. 1, pp. 1–6, 2019, doi: 10.1088/1757-899X/621/1/012005.

[10] R. Puspita, S. Putri, I. Waspada, D. Ilmu, K. Informatika, and F. Sains, "Penerapan Algoritma C4 . 5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, pp. 1–7, 2018.

[11] D. Devina, A. A. Supianto, and W. Purnomo, "Aplikasi Data Mining Menggunakan Algoritme C4 . 5 untuk Memprediksi Ketepatan Lulus Mahasiswa Berdasarkan Faktor Demografi," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 6, pp. 6044–6051, 2019.

[12] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," vol. 7, no. 1, pp. 59–64, 2013.

[13] S. Faisal, "Klasifikasi Data Minning Menggunakan Algoritma C4.5 Terhadap Kepuasan Pelanggan Sewa Kamera Cikarang," *J. Ilmu Komput. Teknol. Inf. ISSN*, vol. 4, no. April, pp. 1–8, 2019.

[14] Rismayanti, "Implementasi Algoritma C4.5 Untuk Menentukan Penerima Beasiswa Di Stt Harapan Medan," *Media Infotama*, vol. 12, no. 2, pp. 116–120, 2016.

[15] T. A. Kurniawan, "Pemodelan Use Case (Uml): Evaluasi Terhadap Beberapa Kesalahan Dalam Praktik," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 1, pp. 77–86, 2018, doi: 10.25126/jtiik.201851610.

[16] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: https://doi.org/10.14710/jtsiskom.8.2.2020.89-93.