PIKSEL
Penelitian Ilmu Komputer
Sistem Embedded and Logic

Penelitian Ilmu Komputer, Sistem Embedded and Logic
p-ISSN: 2303-3304, e-ISSN: 2620-3553
Vol. 12 (2): 275 – 284 (September 2024)

The DOI 10.33558/piksel.v12i2.9639

Predicting Student Graduation Grades Using the C4.5 Algorithm: An Implementation Study

Bela Amalia Wiranti [1], Supriyanto [1,*], Nurmayanti [1]

* Corespondence Author: e-mail: supriyanto@dcc.ac.id

[1] Sistem Informasi; Fakultas Iimu Komputer; Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia;Jl. Lintas Sumatera No. 3 Candi Mas Kecamatan Abung selatan, Kabupaten Lampung Utara, Lampung; Tlp. 0724-23003 ; email:
nurmayanti89@gmail.com
supriyanto@dcc.ac.id
belaamaliawiranti@gmail.com

Submitted: **14/05/2024**
Revised : **04/06/2024**
Accepted : **18/06/2024**
Published : **30/09/2024**

Abstract

Kemala Bhayangkari junior high school as one of the private schools, students follow the learning process and are carried out mid-semester exams, semester final exams in completing junior high school level education. test scores as one of the requirements for grade promotion or for graduation mark the end of junior high school level, and the value of knowledge is one of the keys to a person's ability to complete education. predicting student graduation with the C4.5 algorithm method In the application of the C4.5 algorithm using Rapidminer tools, after manual calculation, the results will be tested using Rapidminer tools. the results of manual calculation of the C4.5 algorithm and the results of calculations using rapidminer tools will produce an Accuracy Level resulting from this calculation of 92.22% for graduation grades at SMP Kemala Bhayangkari Kotabumi.

Keywords: *Graduation, Student Grades, C4.5 Algorithm, Microsoft Excel, Rapid miner.*

1. Introduction

Students are learners who are in basic education, junior high school (SMP) the quality of a school is determined by students in achieving an achievement in academics and the number of students who pass the level of student success and the low level of student failure is a mirror of the quality of an education, the graduation rate is one indicator or benchmark of the school's success rate in carrying out the process of teaching and learning activities (KBM)(Novianti et al., 2016), a high graduation rate is also considered a student achievement.

As one of the private schools, students follow the learning process and are carried out mid-semester exams, end-of-semester exams in completing junior high school level education(Wicaksono & Setiadi, 2023). test scores as one of the requirements for grade promotion or for graduation marks the end of junior high school level school, and the value of knowledge is one of the keys to a person's ability to complete education, with test scores above the Minimum Completion Criteria (KKM), it is necessary to have knowledge and information in determining student graduation rates based on semester test scores(Miswarudin, 2019).

Data mining is a process and data modeling in processing data to make information in, student data is used as a source of knowledge in processing data into a prediction of graduation with the C4.5 algorithm and is expected to explore a potential or knowledge that is more than just information from school data or student data(Wicaksono & Setiadi, 2023). based on predetermined criteria that will produce a decision tree, from this decision tree, knowledge will be taken The results obtained

manually, which produce a decision tree have a match with the results obtained from using the rapid miner application and the C4.5 algorithm(Miswarudin, 2019).

So the learning outcomes or grades obtained by Kemala Bhayangkari Middle School students are changes in behavior and abilities obtained by students after studying, which take the form of cognitive, affective and psychomotor abilities(Sukri & Handrianto, 2024). Therefore, Kemala Bhayangkari Middle School students should be able to obtain learning outcomes that are in accordance with established standards or in accordance with the KKM, but the reality is that not all students can achieve maximum learning outcomes(Malik Kamil, 2016).

2. Research Method

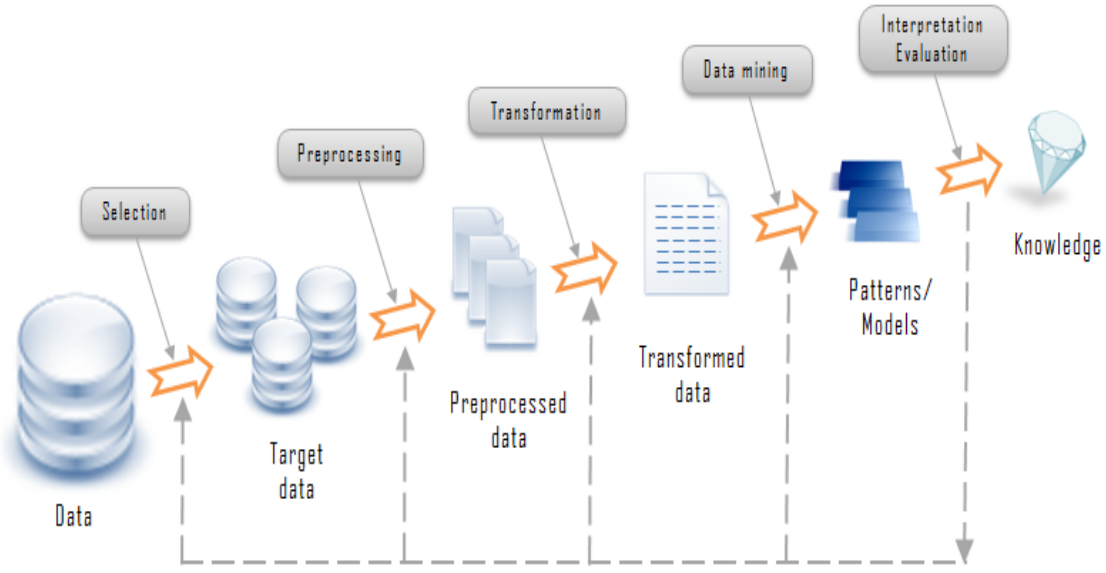
2.1 Data Mining

Data Mining is a process where artificial intelligence, mathematics, statistical techniques and machine learning are used to extract and identify useful information and related knowledge from large .The Data Mining process(Ginting et al., 2020) itself is as follows:

1. Data Selection
2. Pre-processing/cleaning
3. Transformation/Diskritisasi
4. Data Mining
5. Evaluation

Data Mining is a process of gathering or gathering information to discoverpreviously undiscovered information from a grup big data or other repository databases(Gool Lumban Yohana Lydia, Safii .M, 2021).

The stages of data mining are a series of processes, so it is divided into several stages. These stages are interactive, users will be involved directly or with KDD (Ginting et al., 2020). The following stages of data mining are shown in Figure 1.



Source: Ginting, Kusrini, and Taufiq 2020

Figure 1. KDD Stages

2.2 Algorithm C4.5

The C4.5 algorithm is one of the algorithms used in data mining which is useful for assisting in classifying classes where the C4.5 algorithm is an algorithm developed from the ID3 algorithm with the workings to produce decisions is to create a decision tree(Ali Ma et al., 2021).After preparing the selection of attributes that can be

calculated from the training data in the C4.5 algorithm, the concept of entropy is used(Risnawati Wiwi, 2023).

The C4.5 algorithm is an algorithm that building decision trees and shaping knowledge model for classing data and the C4.5 Algorithm has the fastest performance and has the highest accuaracy(Kurniasari Rina, 2019).Algorithma C4.5 is a dat mining method recommended to calculate the graduation list on the course institution (Kholifah, 2020).

In the C4.5 algorithm, the first step is to find the entropy value. First, ascertain the Entropy value of the entire case utilizing the following formula:

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

Description:
S : set of cases
A : attribute
n : number of partitions of S
p_i: proportion of Si to S.

2.3 Rapid Miner

Rapid Miner is specialized for data mining use. The models provided are also quite numerous and complete, such as Bayesian Models, Modeling, Tree Induction, Neural Networks and others. Many methods are provided by Rapid Miner ranging from classification, clustering, association and others. If there is no model or algorithm model that does not exist in Weka, users can add other modules, because Weka is open source, so anyone can participate in developing this software(Haryati et al., 2015).

Rapid Miner uses various techniques descptive and predictive in providing insight to users so they can make the most informed decisions good(Manullang et al., 2021).

2.4 Decision Tree

A data structure with nodes and edges is called a tree.root node, branch/internal node,and leaf node are thtee types of nodes that form a tree,in a decision tre, internal nodes and nodes are labelled with attribute names,edges are labelled with potential attribute values ,and leat nodes are labelled with various classes. Decision tree is a direct representation of classification techniques for a limited number of classes(Mundok A, Amiruddin, 2024)

2.5 Prediction

Prediction is the initial part of a process making a prediction.before making a predictions,you must first know the actual problem ini decision making.predictions are thoughts about aquntities,for example,demand for one or several products in the coming period(Kodratillah et al., 2021).

3. Results and Analysis

3.1 Analysis Of Testing Data

Data testing is carried out on testing data as much as 100 data which is used as simple for entropy and gain calculations. The following testing data is used:

Table 1. Classification

No	Category	Classsificion
1	Heihg	81-90
2	Medium	70-80
3	Low	<=60

Source: Research Result (2024)

3.2 Application of Data Mining Techniques

From the existing results, it is then categorized by variables, attributes and then used as 100 testing data, from this process, it is then calculated with C.45 algroritama to determine the prediction of graduation grades at SMP KEMALA BYANGKARI KOTABUMI.

Table 2. Data Result

No	Student Name	Daily Grades	Schoool Exam Scores	National Exam Score	Category
1	Afief Hasbi	86	81	55	PASS
2	Arya Firmansyah	75	79	76	PASS
3	Atallya Narulita Br	85	82	88	PASS
4	Ayda Maulidina	561	59	60	FAIL
5	Anisa Widya	73	80	78	PASS
6	Bila putri	73	80	90	PASS
7	Bintang Ersinalsal	55	60	60	FAIL
8	Bobby Anwar	77	78	90	PASS
9	Chantika Donika Pohan	90	87	80	PASS
10	Caca abelika	54	60	60	FAIL
11	Deni Saputra	85	86	83	PASS
12	Diky haryanto	78	71	60	FAIL
13	Dea Vanessya	50	60	75	FAIL
14	Dinda Ardhini	76	89	60	PASS
15	Deo Addriano	88	60	83	PASS
91	Ecaa Murni zebua	50	75	60	PASS
92	Edo Andra Wardhana	75	88	80	PASS
93	Elvira riana	50	60	90	PASS
94	Fahri Akbar	86	84	60	PASS
95	Fariz	55	60	81	PASS
96	Farel Alfiqri	84	79	77	PASS
97	Fatima Zen	77	86	60	PASS
98	Gilang Ramdhan	90	87	80	PASS
99	Galuh Amalia	54	60	60	PASS
100	Galang Saputra	85	86	83	PASS

Source: Research Result (2024)

after all the data was converted into a range, there were 43 passed and 11 not passed.

Table 3. Confusion Table

	Class	
PredIction	Pass	FAIL
Pass	70	
FAIL		22

Source: Research Result (2024)

3.3 Calculation of testing data using the C4.5 Algorithm

From the existing results, some 54 testing data were taken, from this process, it was then calculated with C4,5 algroritama.

Table 4. Calculation Result in Microsoft Excel

Node	Atibut	Value	Total	Pass	FAIL	Entropy	Gain
Roots	Total		54	43	11	0,72927	
	Daily Grades						0,0689
		High	22	20	2	0,4395	
		Medium	17	14	3	0,6723	
		Low	15	9	6	0,971	
							0,1297
	School Exam Scores	High	19	19	0	0	
		Medium	16	11	5	0,896	
		Low	19	12	7	0,9495	
	National Exam Score						0,0792
		High	19	17	2	0,4855	
		Medium	16	15	1	0,3373	
		Lo	20	10	10	1	

Source: Research Result (2024)

Accuracy:

Calculation of Testing Data with the amount of Data 54 Student Graduation Rate and Accuracy Presentation as follows:

$$\% \text{ accuracy} = (38/54) \text{ x } 100\% = 70\% \quad (2)$$

Based on the above calculations, it can be concluded that the accuracy of the 54 testing data has an accuracy of 70% and based on the decision tree on the testing data above, the criteria that have the most influence in predicting pass or not pass show that the information gain on the criteria (Selokah Exam Score) is 0.1297 greater than the other criteria.

Classified as pass and fail, with a total of 100 data recipients of student pass score data at Kemala Bhayangkari Middle School in 2019. And it is also known that each criterion has the following results:

1) Daily scores with high attributes totaled 20 passes and 2 did not pass with an entropy value of 0.4395, moderate totals passed 14 and did not pass 3 with an entropy value of 0.6723, low totaled 15 passed and did not pass 6 with an entropy value of 0.971, the value gain for daily value attribute 0.0689.
2) School exam scores with high attributes totaled 19 passes and 0 failed with an entropy value of 0, moderate totaled 11 passed and 5 failed with an entropy value of 0.896, low totaled 12 passed and failed 7 with an entropy value of 0.9494, gain value for the school exam score attribute 0.1297.
3) National exam scores with high attributes totaled 19 passes and 2 failed with an entropy value of 0.4855, moderate totaled 15 passed and failed with an entropy value of 0.3373, low totaled 10 passed and failed 10 with an entropy value of 1, gain for daily value attribute 0.0792

3.4 calculations on rapid miner

The decision tree is made from the results of calculations with rapidminer tools. following are the prediction resultan from 100 training data ,it is know that result in the image above are 92% accurate

Source: Research Result (2024)

Figure 2. Decision Tree C4.5

The following is a description of the resulting modeling decision tree (C.45)

PerformanceVector

```
PerformanceVector:
accuracy: 92.00%
ConfusionMatrix:
True:   Lulus   Tidak  Lulus
Lulus : 70      2
Tidak  Lulus:   6       22
precision: 78.57% (positive class: Tidak  Lulus)
ConfusionMatrix:
True:   Lulus   Tidak  Lulus
Lulus : 70      2
Tidak  Lulus:   6       22
recall: 91.67% (positive class: Tidak  Lulus)
ConfusionMatrix:
True:   Lulus   Tidak  Lulus
Lulus : 70      2
Tidak  Lulus:   6       22
AUC (optimistic): 0.980 (positive class: Tidak  Lulus)
AUC: 0.954 (positive class: Tidak  Lulus)
AUC (pessimistic): 0.928 (positive class: Tidak  Lulus)
```

Source: Research Result (2024)

Figure 3. Performance Vector

The discussion was carried out to obtain accuracy and precision values for the C.45 algorithm to predict improvements in passing scores, recall. The meaning of accuracy is the level of closeness of the prediction results to the factual results. Precision is the level of accuracy between the information requested by the system. And recall is the level of success of the system in rediscovering information.
The calculation of the C.45 algorithm, accuracy is done by the number of TP + TN divided by the total amount of testing data tested.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

$$= \frac{70 + 22}{70 + 22 + 2 + 6} * 100\%$$

$$= \frac{92}{100} * 100\% = 92\%$$

accuracy: 92.00%

	true Lulus	true Tidak Lulus	class precision
pred. Lulus	70	2	97.22%
pred. Tidak Lulus	6	22	78.57%
class recall	92.11%	91.67%	

Source: Research Result (2024)

Figure 4. Accuracy

The precision value is calculated by dividing the amount of correct data that has a positive value (True Positive) divided by the number of correct data that has a positive value (True Positive) and incorrect data that has a positive value (False Positive).

$$Precision = \frac{TP}{TP + FP} * 100\%$$

$$= \frac{70}{70 + 2} * 100\%$$

$$= \frac{70}{72} * 100\%$$

$$= 97\%$$

precision: 78.57% (positive class: Tidak Lulus)

	true Lulus	true Tidak Lulus	class precision
pred. Lulus	70	2	97.22%
pred. Tidak Lulus	6	22	78.57%
class recall	92.11%	91.67%	

Source: Research Result (2024)

Figure 5. Precision

The recall value is calculated by dividing the correct data which has a positive value (True Positive) by the sum of the correct data which has a positive value (True Positive) and the incorrect data which has a negative value (false negative).

$$Recall = \frac{TP}{TP + FP} * 100\%$$

$$= \frac{70}{70 + 2} * 100\%$$

$$= \frac{70}{72} * 100\%$$

$$= 97\%$$

recall: 91.67% (positive class: Tidak Lulus)

	true Lulus	true Tidak Lulus	class precision
pred. Lulus	70	2	97.22%
pred. Tidak Lulus	6	22	78.57%
class recall	92.11%	91.67%	

Source: Research Result (2024)

Figure 6. Recall

From the sample data that was collected, namely 100 student data, then the results of student data stated the level of accuracy, recall and perception of C.45, the level of Accuracy 92%, Recall 92% and Percision 92% in predicting graduation with the C.45 algorithm.

Tree

```
Nilai Harian = Rendah
|   NUN = Rendah
|   |   NUS = Rendah : Tidak  Lulus {Lulus =3, Tidak  Lulus=16}
|   |   NUS = Sedang : Lulus  {Lulus =4, Tidak  Lulus=0}
|   NUN = Sedang : Tidak  Lulus {Lulus =3, Tidak  Lulus=6}
|   NUN = Tinggi: Lulus  {Lulus =6, Tidak  Lulus=1}
Nilai Harian = Sedang : Lulus  {Lulus =23, Tidak  Lulus=1}
Nilai Harian = Tinggi: Lulus  {Lulus =37, Tidak  Lulus=0}
```

Source: Research Result (2024)

Figure 7. Tree atribut C4.5

4. Conclusion

Testing results show that in manual calculations with the C.45 algorithm method found the results under the school exam score is the value has the highest value in determining graduation. In addition, the implementation of the C.45 algorithm method on rapidminer begins with inputting student graduation score data which becomes a database on Ms.Excel. The results of using RapidMiner tools and manual calculations are the level of accuracy resulting from the calculation of 92.22% and 70% manual calculation.This accuracy result of 92.22 % shows that the model is built to be able to make very high predictions. However, it is necessary Note that high accuracy can also be caused by low data complexity, which results in a capable model predict quite accurately. Therefore, it is important to carry out further evaluation of the model and test reliability on different data to confirm the prediction model. The ability of the C4.5 algorithm to form decision trees and identifying relevant attributes in the data mining process becomes determining factors of high prediction accuracy. Thus, results This research provides valuable insight and related information
prediction of student passing scores using the classification method and the C4.5 algorithm.

Acknowledgements

The authors thanks to respondents to contribute to the study. Also for the reviewers for the insightful comments.

Author Contributions

Bela Amalia Wiranti proposed the topic; Supriyanto, and Nurmayanti test the system.

Conflicts of Interest

The authors declare no conflict of interest.

References

Ali Ma, F., Pratama, A., Sholihin, I., & Rizki Rinaldi, A. (2021). Penerapan Model Prediksi Menggunakan Algoritma C.45 Untuk Prediksi Kelulusan Siswa Smk Wahidin. *Jurnal Data Science & Informatika*, *1*(1), 16–20.

Ginting, V. S., Kusrini, K., & Taufiq, E. (2020). Implementasi Algoritma C4.5 Untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *10*(1). Https://Doi.Org/10.35585/Inspir.V10i1.2535

Gool Lumban Yohana Lydia, Safii .M, D. S. (2021). *Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi*. *2*(2), 97–106.

Haryati, S., Sudarsono, A., & Suryana, E. (2015). *Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4 . 5*. 130–138.

Kholifah, I. N. (2020). *Memprediksi Tingkat Kelulusan Peserta Kursus Per- Tahun Dengan Algoritma Data Mining C4 . 5 Dan Rapidminer*. *06*(01).

Kodratillah, E. Y., Naya, C., Studi, P., Informatika, T., Teknik, F., Pelita, U., & Bayes, N. (2021). *Jurnal Teknologi Pelita Bangsa*. *12*(4).

Kurniasari Rina, F. A. (2019). *Jurnal Ilmiah Komputer Dan Informatika (Komputa)*. *8*(1).

Malik Kamil, F. Mochammad F. (2016). *Prediksi Prestasi Siswa Smp Nurul Jadid Menggunakan Algoritma C4.5*. *2*(4), 2–5.

Manullang, N., Sembiring, R. W., Gunawan, I., Parlina, I., Informatika, T., &

Informatika, T. (2021). *Implementasi Teknik Data Mining Untuk Prediksi Peminatan Jurusan Siswa Menggunakan Algoritma*. *2*(2), 1–5.

Miswarudin. (2019). *Implementasi Data Mining Dengan Penerapan Algoritma C.45 Untuk Memprediksi Tingkat Kelulusan Siswa Pada Smk Hs Agung*.

Mundok A, Amiruddin, L. Y. . (2024). *Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Metode Decision Tree*. *3*(1), 31–36.

Novianti, B., Rismawan, T., & Bahri, S. (2016). *Implementasi Data Mining Dengan Algoritma C4 . 5 Untuk Penjurusan Siswa (Studi Kasus : Sma Negeri 1 Pontianak)*. *04*(3).

Risnawati Wiwi, W. A. (2023). *Application Of C4 . 5 Algorithm For Web-Based Clasification Grade Promotion Of High School Student*. *2*(September), 882–891.

Sukri, M. H., & Handrianto, Y. (2024). *Penerapan Algoritma C4 . 5 Dalam Menentukan Prediksi Prestasi Siswa Pada Smpn 51 Jakarta*. *4*(1).

Wicaksono, A. W., & Setiadi, T. (2023). Penerapan Klasifikasi Decision Tree (C4.5) Untuk Memprediksi Kelulusan Siswa Sekolah Dasar Di Kecamatan Juai. *Format : Jurnal Ilmiah Teknik Informatika*, *12*(2), 151. Https://Doi.Org/10.22441/Format.2023.V12.I2.008