Comparison of C4.5 and Naive Bayes for Predicting Student Graduation Using Machine Learning Algorithms

Abu Tholib[1], M Noer Fadli Hidayat[1], Supriyono[2], Resty Wulanningrum[3], Erna Daniati[3]
[1]Universitas Nurul Jadid, Probolinggo, Indonesia
[2]Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia
[3]Universitas Nusantara PGRI, Kediri, Indonesia

Article Info	ABSTRACT
Article history: Received August 30, 2023 Revised September 08, 2023 Accepted September 20, 2023 **Keywords:** Comparison Method C4.5 Algorithm Student Graduation Nave Bayes Algorithm	Student graduation is a very important element for universities because it relates to college accreditation assessment. One of them is at the Faculty of Engineering Nurul Jadid University, which has problems completing the study period within a predetermined time. So that it can be detrimental because accreditation is less than optimal, and the number of active students makes it less ideal in teaching and learning activities. This study aimed to compare the level of accuracy using the C4.5 algorithm and Nave Bayes method in predicting graduation on time. The C4.5 and Nave Bayes algorithms are one of the methods in the algorithm for classifying. Tests were carried out using the C4.5 and Nave Bayes algorithms using Google Colab with Python programming language, then validated using 10-fold cross-validation. The results of this study indicate that the Nave Bayes method has a higher accuracy value with an accuracy rate of 96.12%, while the C4.5 algorithm method is 93.82%.

Corresponding Author:

Abu Tholib,
Universitas Nurul Jadid, Probolinggo, Indonesia.
Email: ebuenje@gmail.com

1. INTRODUCTION

Student graduation is an essential element for universities because it relates to the assessment of university accreditation. Universities can be assessed through the timely graduation of students. The more students who graduate on time (4 years), the higher the assessment of the college. Meanwhile, graduation itself is a benchmark for students for their academic results. To achieve graduation, students must go through several stages or processes, such as completing all courses, real work lectures (KKN), fieldwork practices (PKL), and final assignments. These stages can be carried out following the time determined by the university. The Faculty of Engineering, Nurul Jadid University, still has several problems completing the study period, which exceeds the predetermined time. Of course, this can be detrimental because it makes it less optimal in faculty accreditation and the number of active students, making it less ideal for teaching and learning activities.

From the above problems, research can be done using a method to predict student graduation on time. One of them can use the Machine Learning Algorithm as a method of mining knowledge from a data set [1]. Several Machine Learning algorithms used in predicting student graduation, such as the Support Vector Machine (SVM) [2], C4.5 [3][4], K-Nearest Neighbor (KNN) [5], Correlated Naive Bayes (C-NBC) [6] and Nave Bayes algorithms [7][8] [9], can be used.

Research [10] uses the C4.5 method to predict student graduation using 640 student data, which is divided into 340 instances of training data and 300 instances of testing data. There are 4 attributes used to predict student graduation, namely department name, GPA, English score, and age. Based on the research results, the C4.5 method obtained an accuracy of 90%. Research [11] uses the KNN method to predict student graduation with 443 data, which is divided into 380 instances as training data and 163 instances as testing data. Based on the research results, the KNN method obtained an accuracy of 85.28%. Research [12] uses the Niave Bayes method to predict student graduation based on academic history using 173 instances of data, which is divided into 125 instances as training data and 48 instances as testing data. Based on the research results, the Nave Bayes method was able to obtain an accuracy of 70.83%.

This research uses the C4.5 and Nave Bayes algorithm techniques, very popular algorithms because they have a high level of accuracy in classifying data [13]. By utilizing data on active Faculty of Engineering semester 6 academic year 2019/2020 students. Student data that becomes input attributes are name, study program, IPK, IP Semester 3 until 5, SKS semester 5, and study period. The attribute used to classify data consisting of " graduation on time " and " graduation not on time " is the study period attribute data. So, this study aimed to compare the level of accuracy using the C4.5 algorithm and Nave Bayes method in predicting graduation on time.

2. RESEARCH METHOD

Below is a research framework on on-time student graduation using the C4.5 and Nave Bayes algorithm methods is shown in Figure 1.
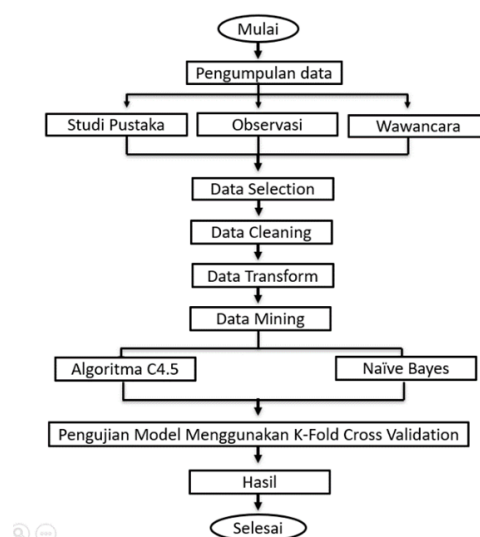

Figure 1. Research Framework

2.1. Data Collection

The results of interviews with the Academic and Student Administration Bureau (BAAK) and data obtained from the BAAK of the Faculty of Engineering are 353 data on active students in semester 6 in the 2019/2020 academic year. In each semester, there must be 10% to 15% of students who do not graduate on time. Students have a maximum semester limit of 14; if more than the predetermined semester, the student is dropped out. The process of recording student grades is carried out directly by lecturers to the Nurul Jadid University SMPT system because each lecturer has their application synchronized to the center. The assessment by lecturers must follow the SOP following applicable rules and regulations. The data obtained from BAAK are NIM (student number), name, study program, entry period, type of registration, academic year, status, semester credits (SKS), IPS, total credits (SKS total), and IPK which is shown in Table 1.

Table 1. Student Dataset Sample

Study Program	Entry Period	Registration Type	Academic Year	Status	SKS Smt.	IPS	SKS Tot.	IPK
S1 Teknik Informatika	20191	New Student	20201	Active	24	3,52	102	3,7
S1 Teknik Informatika	20191	New Student	20201	Active	21	2,93	93	3,07
S1 Teknik Informatika	20191	New Student	20201	Active	24	3,49	102	3,65
S1 Teknik Informatika	20191	New Student	20201	Active	24	3,49	102	3,64
S1 Teknik Informatika	20191	New Student	20201	Active	24	3,68	102	3,62
S1 Teknik Informatika	20191	New Student	20201	Active	None	None	None	None
S1 Teknik Informatika	20191	New Student	20201	Active	None	None	None	None

2.2. Data Selection

This data selection stage will be used to implement the Machine Learning Algorithm. This stage is carried out after obtaining data that has been obtained from the BAAK Faculty of Engineering, Nurul Jadid University. The data used only name data, study program, 5th-semester SKS, 5th-semester IPK, IPS, which is used only for 3rd-semester IP to 5th-semester IP and status which is shown in Table 2.

Table 2. Student Data Selection

SKS Tot.	IPK	IPS Semester 3	IPS Semester 4	IPS Semester 5	Description
102	3,7	3,55	3,66	3,52	Active
93	3,07	2,2	2,43	2,93	Active
102	3,65	3,59	3,49	3,49	Active
102	3,64	3,44	3,53	3,49	Active
102	3,62	3,29	3,44	,3,68	Active
79	3,64	3,74	3,42	None	Inactive
58	2,42	2,23	None	None	Active
58	2,25	2,6	None	None	Active

2.3. Data Cleaning

The next stage is cleaning by deleting incomplete data and recapitulating data on students actively studying, not taking college leave, or transferring. The data used are name, study program, SKS (credits) total, IPK, IPS semester 3, IPS semester 4, IPS semester 5, and information (See Table 3).

Table 3. Cleaning Data Table

SKS Tot.	IPK	IPS Semester 3	IPS Semester 4	IPS Semester 5	Description
102	3,7	3,55	3,66	3,52	Active
93	3,07	2,2	2,43	2,93	Active
102	3,65	3,59	3,49	3,49	Active
102	3,64	3,44	3,53	3,49	Active
102	3,62	3,29	3,44	3,68	Active

2.4. Transformation Data

The next stage is to transform the data by changing the attribute name, which can be seen Table 4. The second transformation is changing IPS data into index data with a range (See Table 5). The third transformation is to change the caption data with the following provisions (See Table 6). The fourth transformation is to change the study program data with the provisions (See Table 7). The fifth transformation is to change the SKS data with the provisions (See Table 8). The sixth transformation is to convert IPK data into a range (See Table 9). The overall data management transformation results can be seen in Table 10.

Table 4. Transformations Change Attribute Names

Attributes before	Attributes after
SKS Tot	SKS tempuh
IPS	IPS Semester 3, IPS Semester 4, IPS Semester 5
Status	Description

Table 5. IPS Data Transform

Range	Description
3,51 - 4,00	0
3,00 - 3,50	1
2,51 - 2,99	2
2,00 - 2,50	3
0,00 - 1,99	4

Table 6. Description Data Transformation

Description	Provisions
Graduation on Time	0
Graduation Not on Time	1

Table 7. Study Program Data Transform

Index	Description
Information System	SI
Informatics	IF
Software Engineering	RPL
Informatics Engineering	TI
Electrical Engineering	TE

Table 8. SKS Data Transformation

Range	Description
0 - 97	0
98 - 102	1

Table 9. IPK Data Transform

Range	Description
3,51 - 4,00	0
3,00 - 3,50	1
2,51 - 2,99	2
2,00 - 2,50	3
0,00 - 1,99	4

Table 10. Results of Transformation Data Management

SKS Tempuh	IPK	IPS Semester 3	IPS Semester 4	IPS Semester 5	Description
1	0	0	0	0	0
0	1	2	3	2	1
1	0	0	1	1	0
1	0	1	0	1	0
1	0	1	1	0	0

3. RESULT AND ANALYSIS

After transforming the data according to what will be used in the Machine Learning Algorithm stage, data analysis is carried out using the C4.5 algorithm method, where the data analysis process uses Google Colab to make it easier to classify data. The data is divided into training data and testing data. The data used for training data is 259, and testing data is 87. The Tabel 11 shows the accuracy value using the confusion matrix.

Table 11. C4.5 Confusion Matrix Algorithm

Actual	Prediction	
	True	False
Graduation on Time	75	2
Graduation not on time	1	9

Based on Table 11, the accuracy value is 96.55%, the recall value of passing on time is 98.68%, the recall value of passing not on time is 81.81%, the precision value of passing on time is 97.40%, the precision value of passing not on time is 90%. So that from 87 data can be predicted is the data passed on time with correct values and predicted to be correct is 75, the data passed on time with the correct value and mispredicted is 2, the data that graduated not on time with correct values and predicted to be correct is 9, and the data that graduated not on time is correctly valued and mispredicted is 1.

3.1. Implementation of Using Nave Bayes

After transforming the data according to what will be used in the Machine Learning Algorithm stage, data analysis is carried out using the Nave Bayes method, where the data analysis process uses Google Colab to make it easier to classify data. division of data into training data and testing data. The data used for training data is 259, and testing data is 87. Table 12 shows the accuracy value using the confusion matrix.

Table 12. Naive Bayes Confusion Matrix

Actual	Prediction	
	True	False
Graduation on Time	74	3
Graduation not on time	1	9

Based on Table 12, the accuracy value is 95.40%, the recall value of passing on time is 98.66%, the recall value of passing not on time is 75%, the precision value of passing on time is 96.10%, the precision value of passing not on time is 90%. So that from 87 data can be predicted is the data passed on time with correct values and predicted to be correct is 74, the data passed on time with the correct value and mispredicted is 3, the data that graduated not on time with correct values and predicted to be correct is 9, and the data that graduated not on time is correctly valued and mispredicted is 1.

3.2. Model Testing Using K-fold Cross Validation

The k-fold cross-validation test results are used to determine the accuracy that has the best value. Table 13 shows the accuracy results. The conclusion from testing k-fold cross-validation using the C4.5 algorithm above is that the accuracy value of 10 folds has very good results. The resulting accuracy values include 100% and 96.15%. While using Nave Bayes, the accuracy value of 10 folds has very good results. The resulting accuracy values include 100% and 96.15%. However, of the 10 folds, the selection is done by

calculating the average of all values to get the best accuracy. The value obtained from the calculated c4.5 algorithm accuracy value is 93.82%, while the calculated Nave Bayes accuracy value is 96.12%. The comparison results of the two methods can be seen in Figure 2. Based on the table above, predicting student graduation on time using the Nave Bayes method has a higher accuracy value compared to the accuracy value using the C4.5 algorithm method of 96.12%. Of the two methods, both have a difference in accuracy value of 2.30%. This is in line with research [14] [7] which used the Nave Bayes method and obtained high accuracy in the cases studied.

Table 13. Test Results using 10-fold Cross Validation

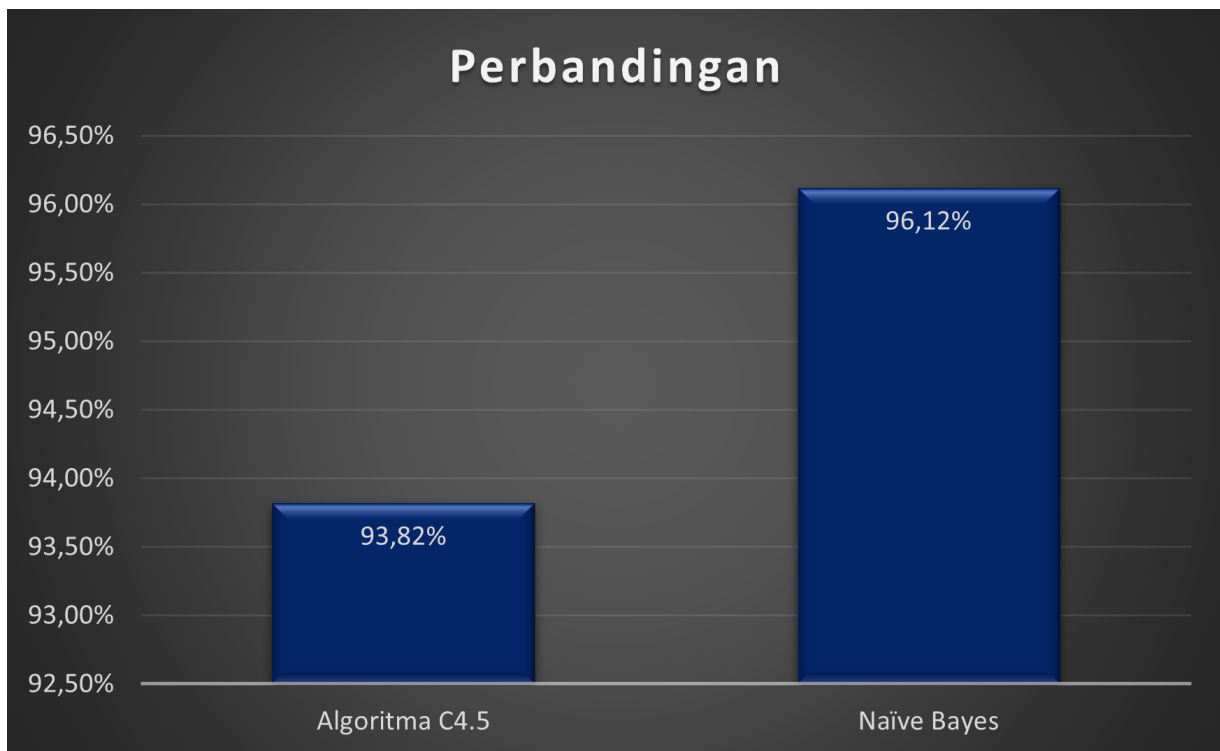| K-fold | C4.5 Algorithm | Naïve Bayes |
	Accuracy	Accuracy
K1	92,21%	100%
K2	96,15%	92,31%
K3	92,31%	92,31%
K4	100%	100%
K5	92,31%	96,15%
K6	96,15%	96,15%
K7	96,15%	100%
K8	92,31%	92,31%
K9	88,46%	100%
K10	92%	92%
Average	93,82%	96,12%

Figure 2. Comparison of Accuracy Value of C4.5 and Nave Bayes Algorithm Methods

4.   CONCLUSION

Based on the results of predicting graduation using the C4.5 algorithm, it has an accuracy value of 93.82%, and the Nave Bayes algorithm has an accuracy value of 96.12%. From the test results using the Nave Bayes method, it has a higher accuracy value than the C4.5 algorithm method, so the Nave Bayes method can be recommended for further research. Furthermore, the decision tree

implemented from the C4.5 algorithm shows that the most influential criteria and variables to predict graduation are IPS Semester 5. Future research can apply feature selection to select features that influence student graduation predictions so that the accuracy of the method used can increase.

5.   ACKNOWLEDGEMENTS

6.   DECLARATIONS

AUTHOR CONTIBUTION
All authors contributed to the writing of this article.
FUNDING STATEMENT
-
COMPETING INTEREST
The authors declare no conflict of interest in this article.

REFERENCES

[1]   S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naive Bayes, KNN Dan SVM," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, apr 2019. [Online]. Available: https://jurnal.lldikti4.or.id/index.php/jurnaltekno/article/view/78

[2]   A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *International Journal of Information and Education Technology*, vol. 13, no. 2, pp. 289–295, 2023.

[3]   D. Kurniawan, A. Anggrawan, and H. Hairani, "Graduation Prediction System On Students Using C4.5 Algorithm," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 19, no. 2, pp. 358–365, 2020.

[4]   L. Y. L. Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi dengan Menggunakan Algoritma C4.5," *Brahmana : Jurnal Penerapan Kecerdasan Buatan*, vol. 2, no. 2, pp. 97–106, 2021.

[5]   Endang Etriyanti, "Perbandingan Tingkat Akurasi Metode Knn Dan Decision Tree Dalam Memprediksi Lama Studi Mahasiswa," *Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya Lubuklinggau*, vol. 3, no. 1, pp. 6–14, 2021.

[6]   H. Hairani, M. Innuddin, and M. Rahardi, "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 51–55.

[7]   H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021, pp. 312–315.

[8]   A. Suwarno, N. Ferawati, and P. A. Sari, "Penerapan Data Mining untuk Prediksi Kelulusan Siswa Menggunakan Algoritma Naive Bayes pada SMK Garuda," *Jurnal Teknologi Pelita Bangsa*, vol. 12, no. 4, pp. 33–40, 2021.

[9]   A. Armansyah and R. K. Ramli, "Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes," *Edumatic: Jurnal Pendidikan Informatika*, vol. 6, no. 1, pp. 1–10, jun 2022. [Online]. Available: http://e-journal.hamzanwadi.ac.id/index.php/edumatic/article/view/4789

[10] H. Yuliansyah, R. A. P. Imaniati, A. Wirasto, and M. Wibowo, "Predicting Students Graduate on Time Using C4.5 Algorithm," *Journal of Information Systems Engineering and Business Intelligence*, vol. 7, no. 1, pp. 67–73, 2021.

[11] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *Journal of Engineering and Applied Technology*, vol. 2, no. 2, pp. 86–91, 2021.

[12] M. T. Sembiring and R. H. Tambunan, "Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU)," in *IOP Conference Series: Materials Science and Engineering*, 2021, pp. 1–8.

[13] F. Solikhah, M. Febianah, A. L. Kamil, W. A. Arifin, and Shelly Janu Setyaning Tyas, "Analisis Perbandingan Algoritma Naive Bayes Dan C.45 Dalam Klasifikasi Data Mining Untuk Memprediksi Kelulusan," *TEMATIK*, vol. 8, no. 1, pp. 96–103, jun 2021. [Online]. Available: http://jurnal.plb.ac.id/index.php/tematik/article/view/576

[14] A. Anwarudin, W. Andriyani, B. P. DP, and D. Kristomo, "The Prediction on the Students' Graduation Timeliness Using Naive Bayes Classification and K-Nearest Neighbor," *Journal of Intelligent Software Systems*, vol. 1, no. 1, pp. 75–88, jul 2022. [Online]. Available: https://ejournal.akakom.ac.id/index.php/JISS/article/view/597