Adpebi International Journal of Multidisciplinary Sciences

https://journal.adpebi.com/index.php/AIJMS

Vol.1, No.1, 2022
ISSN: 2829-8217
pp. 374-383

Comparison of Decision Tree, KNN and Naïve Bayes Methods In Predicting Student Late Graduation In the Informatics Engineering Department, Institute Business XYZ

Imam Yunianto, Ade Kurniawan
Fakultas Engineering and Comunication, Institut Bisnis Muhammadiyah Bekasi, Jawa Barat)
Email: imam@ibm.ac.id, adekurniawan@ibm.ac.id

Muhamad Malik Mutoffar
Affiliation (Sekolah Tinggi Teknologi Bandung, Jawa Barat, Indonesia)
Email: malik@sttbandung.ac.id

ARTICLE INFO

Research Paper

Article history:
Received: 2 July 2022
Revised: 6 August 2022
Accepted: 30 September 2022

HOW TO CITE

Yunianto, I. Y., Mutoffar, M. M. M., & Kurniawan, A. K. (2022). Comparison of Decision Tree, KNN and Naïve Bayes Methods In Predicting Student Late Graduation In the Informatics Engineering Department, Institute Business XYZ. *Adpebi International Journal of Multidisciplinary Sciences*, *1*(1), 374–383. https://doi.org/10.54099/aijms.v1i1.304

ABSTRACT

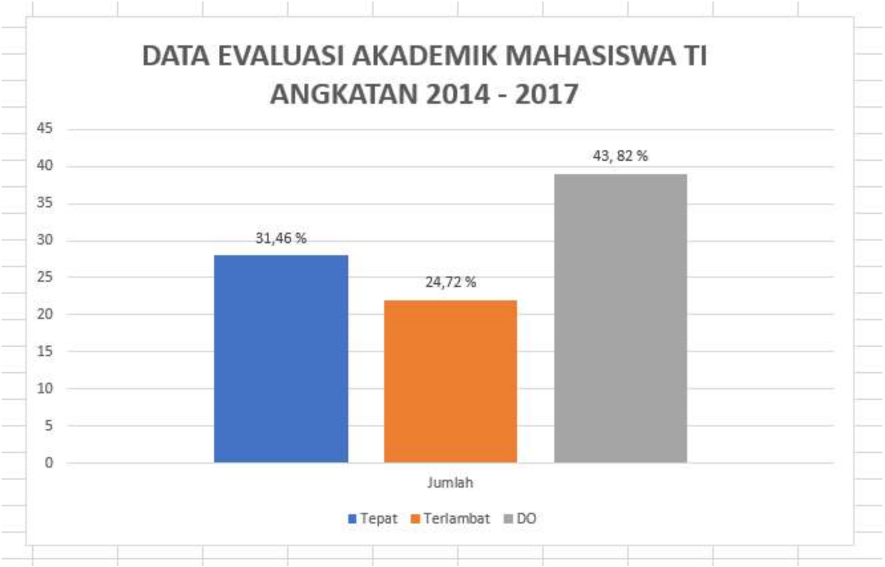
Solving the problem of student late graduation has been a lot of research done before, with various methods and algorithms. Likewise, the comparison of various methods to predict student graduation. However, there is no comparison of the Naïve Bayes, Decision Tree, and KNN methods using data from the Informatics Engineering Department in Institute Business XYZ. From this study by comparing the three methods, the Naïve Bayes method is ranked first with an accuracy rate of 66.67%, Precison 80% and Recall 66.67%. Rank 2 is the KNN algorithm with an accuracy rate of 55.56%, Precision 66.67% and Recall 66.67% and the last is the Decision Tree algorithm with an accuracy rate of 46%, Precison 48.3% and Recall 61.67%

Keywords: Decision Tree Methods, KNN, and Naïve Bayes, student graduation is not on time, Informatics Engineering

INTRODUCTION

Papers should be submitted in English as doc or pdf file attachments. Each paper should be singlespaced with wide margins, on one side only of the paper, preferably of A4 size, with pages in numbered sequence. The font used for the main body should be 12 pt Calibri and for the bodynotes 10 pt Times New Roman or the closest font available. Footnotes should be numbered consecutively and placed within the text. Please leave one line empty before and after every section.

The timely graduation of students is one of the parametres of a college quality. This is used as one of the indicators by the National Accreditation Board (BAN) PT, as well as by the Independent Accreditation Institute for each departement and faculty in universities. The Institute Business XYZ, as a fairly new established university, also facing similar obstacles in the problem of producing college students graduate timely. And also, the problem of the high numbers of Drop Out students in the Informatics Engineering Departement. This problem can be seen from the table below.



The problem of late graduation and drop out students has something in common, that the student does not complete his studies according to a predetermined schedule. This is a problem for the quality of the college. In addition to the similarity of late graduate and drop out students, there is a major difference between the two problems, namely, for students who late graduate, there is still a chance that the student can complete his studies. For students who drop out, it is certain that the student did not complete his studies. This is also an obstacle to poor assessment in campus accreditation result. In table 1, it can be seen that the problems faced by the Institute Business XYZ are Drop Out students by 43.82% and late graduation students by 24.72%. And only 31.46% of IT Department Students can graduate timely.

Of the three problems mentioned above, this study will only focus on the problem of late graduate students, which if this problem is resolved, the problem of students graduating on time with 31.46% will be automatically resolved. Research that discusses student graduation has been many of them identical to the potential for Drop Out students (Nasrullah, 2018) , comparing the C4.5 method with KNN in identifying students who have the potential to drop out (Atma and Setyanto, 2018), predicting students at risk of Drop Out using AD Tree and NNge (Andri and Paulus, 2021), Drop Out student Prediction Solutions (Jin *et al.* , 2016) , Analysis of Drop Out Predictions based on social behavior (Hidayat, Purwitasari and Ginardi, 2013), KNN Algorithm Model for Student graduation predictions (Rohman, 2015), application of PSO-based Neural Network algorithms to select attributes in determining Students who Drop Out (Septiana, 2013).

However, the difference is that previous research with research conducted on generational attributes. This generation attribute is made based on the birth of students following the generation classification (BRS, 2021), Comparison on the Performance of the C4.5 algorithm, Gradient Boosting Algorithm with Random Forest Algorithm and Deep learning Algorithm in the Education Data

Mining case study (Mutrofin *et al.* , 2020) . For this reason, this study aims to compare the decision tree method with KNN and Naïve Bayes in predicting late graduation students.

The background of choosing these three methods to be compared is because the three methods are very familiar to be used by lecturers and educators in artificial intelligence and data mining courses. Although there are likely to be many methods that perform better than these three methods. However, performance problems must be proven first by research.

In this study, it was also carried out in the Informatics Engineering Department at Institute Business XYZ which is interesting because in this institute, there are five semesters of special compulsory courses that must be taken in this Informatics Engineering Department. The author of this piker is also interesting to be used as a differentiator in previous research in the Informatics Engineering Department that has been carried out.

METHOD

Proposed in this research method according to the appearance of figure 1 below. In the first stage with data collection, using academic data from the Informatics Engineering Department of Institute Business XYZ from 2014 to 2017 because only students of the class of 2017 have graduated timely. The total number of data is 89 students. However, because this research only focuses on students who graduated timely and students who late graduated out of a total of 89 students, only 50 students were used for research.

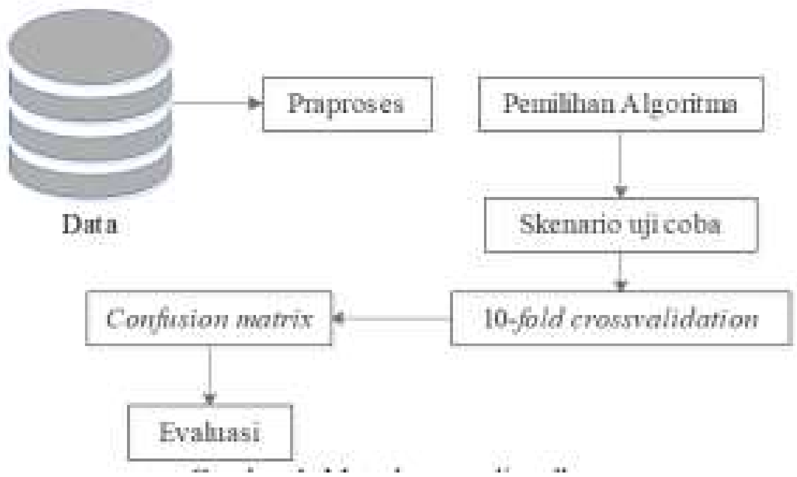


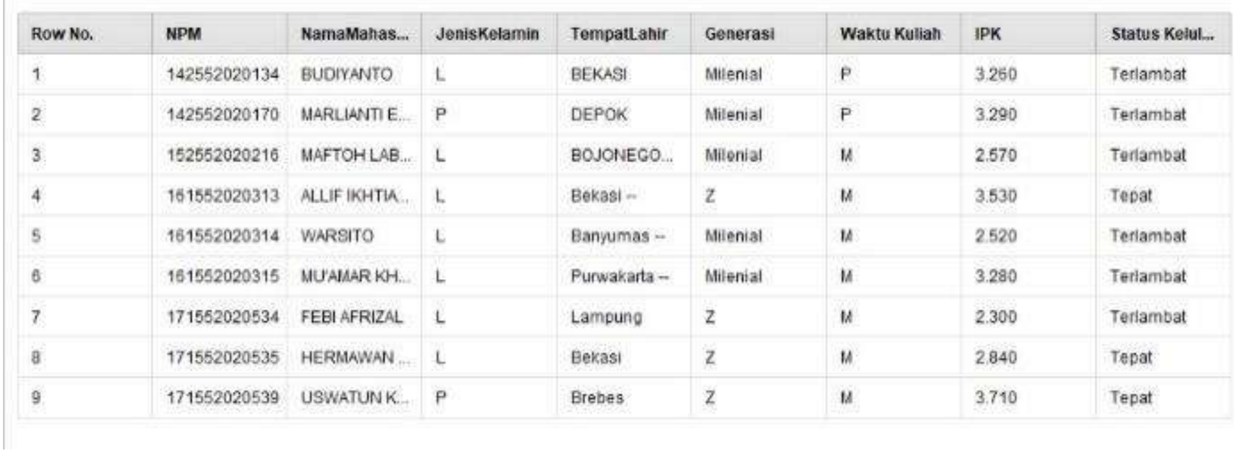
Figure 1. Proposed Research Methods
(source : (Mutrofin *et al.* , 2020)

Data Collection

For the attributes used in the study as many as 8 attributes, namely, Student Identity Number, Student Name, Gender, Place of birth, generation, college time, GPA and graduation status. The attributes used in this study are clearly in accordance with table 2 below.

Table 2 Attributes used

(Source: IT Department and PDDIKTI)



Row No.	NPM	NamaMahas...	JenisKelamin	TempatLahir	Generasi	Waktu Kuliah	IPK	Status Kelul...
1	142552020134	BUDIYANTO	L	BEKASI	Milenial	P	3.260	Terlambat
2	142552020170	MARLIANTI E...	P	DEPOK	Milenial	P	3.290	Terlambat
3	152552020216	MAFTOH LAB...	L	BOJONEGO...	Milenial	M	2.570	Terlambat
4	161552020313	ALLIF IKHTIA...	L	Bekasi --	Z	M	3.530	Tepat
5	161552020314	WARSITO	L	Banyumas --	Milenial	M	2.520	Terlambat
6	161552020315	MU'AMAR KH...	L	Purwakarta --	Milenial	M	3.280	Terlambat
7	171552020534	FEBI AFRIZAL	L	Lampung	Z	M	2.300	Terlambat
8	171552020535	HERMAWAN ...	L	Bekasi	Z	M	2.840	Tepat
9	171552020539	USWATUN K...	P	Brebes	Z	M	3.710	Tepat

In table 2 of the Identity Student Number (NPM) is the student's principal number. In this study, the type of NPM is Polynomial, because although NPM is a number, it is not a number that can be added up, but as unique data and no one should be the same. The name is also of the type Polynomial. Gender is binominal type because there are only 2 genders. The place of birth is of the polynomial type because of the diverse birthplaces of students. For the generation of the result of processing based on the year of birth.

After obtaining the year of birth, it is then determined whether the year of birth of students enters into what generation is classified as BPS (BRS, 2021). It turns out that the results of classifying the year of birth of students are only two generations. Namely the Millennial generation and generation Z. because there are only two generations, the type for the Binomial generation. For the attributes of lecture time, the type is binomial, because there are only two lecture times, regular night and regular morning. Then for the GPA type attribute is real, because almost all GPA always has a comma. Then for the status of the lecture, the type is binomial, because there are only two college statuses, timely or late. After the first and second stages are completed, the third stage is carried out, namely the selection of algorithms. For the first algorithm is the decision tree algorithm.

Decision Tree

The Decision Tree or this decision tree changes the feed that it looks very large in the decision tree. The main focus is to turn the data into a decision tree with some decision rules. (Amril Mutoi Siregar; Adam Puspabuana, 2017). There are many kinds of Decision Tree algorithms including Algorithms C4.5, ID3 and CART. This study used the C4.5 algorithm.

The C4.5 algorithm is the result of the development of the ID3 algorithm (Amril Mutoi Siregar; Adam Puspabuana, 2017). How to form a decision tree by determining the attributes or variables that are the root of a decision tree using the formulas entropy, gain, gain ratio and split info.

The Entropy formula is as follows:

$$Enntropy(S) = \sum_{i=1}^{m} pi\ log2(pi) \qquad (1)$$

M = number of classifications
Pi = sum of sample ratios / probability class i

The enthropy formula for each variable is :

$$EnntropyA(S) = \sum_{v} \frac{|Sv|}{|S|}\ Entropy\ (Sv) \quad (2)$$

A	= Variable
v	= Probability of variable value
| Sv|	= number of niliai samples v
| S|	= number of samples for all sample data
Entropy (Sv)	= Entropy of the sample with Value v

Gain Formula:

$$Gain\ (A) = Enntropy(S) - \ EnntropyA(S) \quad (3)$$

The gaint ratio formula is as follows:

$$GainRatio(S, A) = \ \frac{Gain(S, A)}{SplitInfor(S, A)} \qquad (4)$$

S	= Space data/sample Used for taining.
A	= Attribute
Gain (S, A)	= Information gain pada atribut A
SplitInfor (S, A)	= Split Information on attribute A

The attribute with the highest gain ratio is selected as the test attribute for the node. An advantage is the acquisition of information. Information acquisition applies a normalization approach called separate information according to the following formula:

$$SplitInfo(S.A) = \ - \textstyle\sum_{j=1}^{V} \frac{Sj}{S}\ log2\ \frac{Sj}{S} \qquad (5)$$

Naïve Bayes

The selection of the second algorithm is the Bayesian Naïve algorithm. The bayesian Naïve algorithm is a classification method using probability methods as well as statistics (Amril Mutoi Siregar; Adam Puspabuana, 2017). The equation Naïve Bayes is as follows:

$$P(H|X) = \frac{P(H|X)\ P(H)}{P(X)} \qquad (6)$$

X	= is an unknown class
H	= data hypothesis X
P(H| X)	= probability of hypothesis H based on condition x
P(H)	= probability of hypothesis H
P(X)	= probability of hypothesis X

K - Nearest Neighbor (KNN)

The last algorithm chosen K-Nearest Neighbor (KNN) method. This algorithm is a classification method based on the closest distance (Amril Mutoi Siregar; Adam Puspabuana, 2017). The formula of KNN is as follows:

$$D(x, y) = \sqrt{\sum_{k-1}^{n} (x_K - y_k)^2} \qquad (7)$$

RESULT AND DISCUSSION

This section is the result of trial algorithms using Rapidminier tools. It started with the Decision Tree algorithm, then Naive Bayes and finally KNN.

Decision Tree

In figure 2, it is shown that college time is the most decisive attribute in graduating on time. Evening lecture time is indicated to be slower than the time of graduation from morning lectures. Then there are more women who graduate on time than men. And millennials are more likely to graduate on time than generation Z.

Figure 2. The most defining attributes

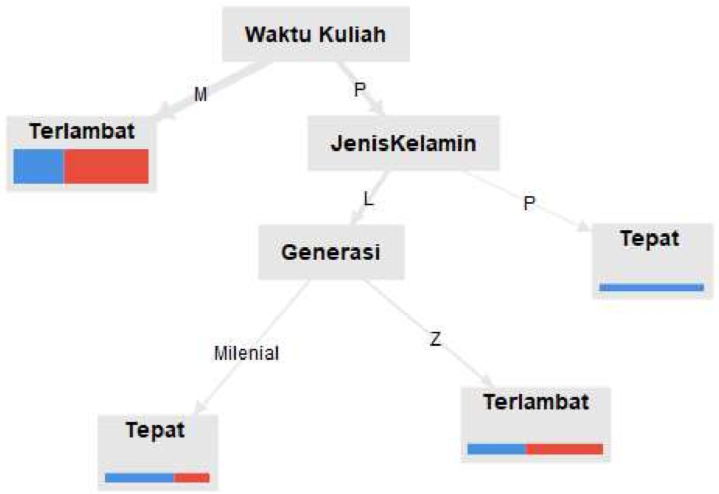
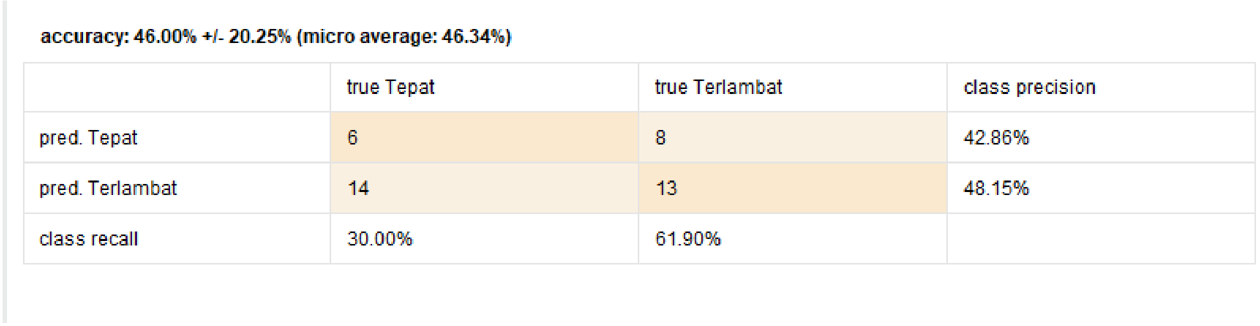


Table 3 shows the performance level of the algorithm in predicting student graduation status of 46%. Table 4 Performance Precision Decision Tree at 48.3 %. Table 5 Performance Recall Decision Tree at 61.67%.

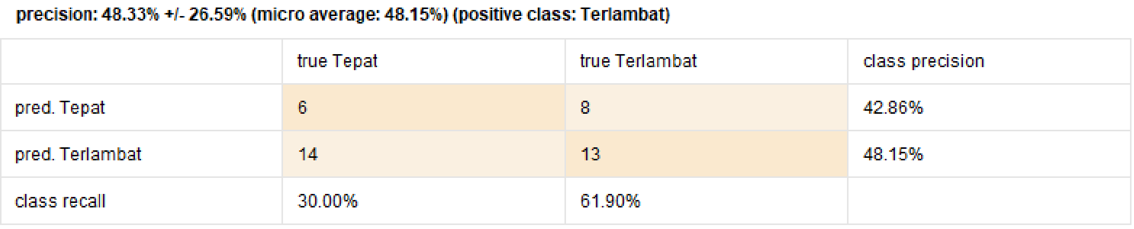
Tabel 3 Perfomace Accuracy Decision Tree



accuracy: 46.00% +/- 20.25% (micro average: 46.34%)

	true Tepat	true Terlambat	class precision
pred. Tepat	6	8	42.86%
pred. Terlambat	14	13	48.15%
class recall	30.00%	61.90%	

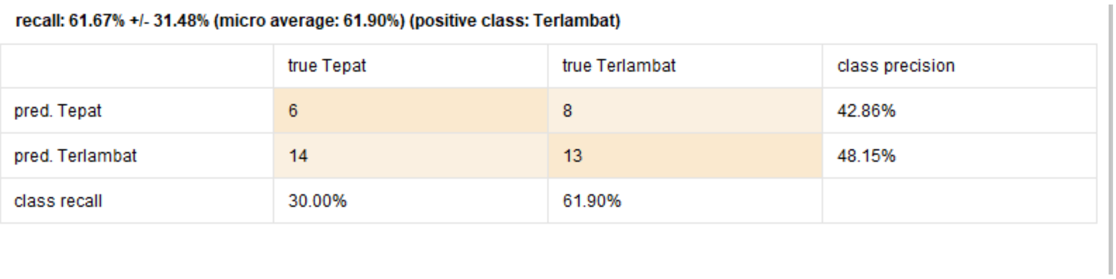
Tabel 4 Performace Precision Decision Tree



precision: 48.33% +/- 26.59% (micro average: 48.15%) (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	6	8	42.86%
pred. Terlambat	14	13	48.15%
class recall	30.00%	61.90%	

Tabel 5 Perfomace Recall Decision Tree



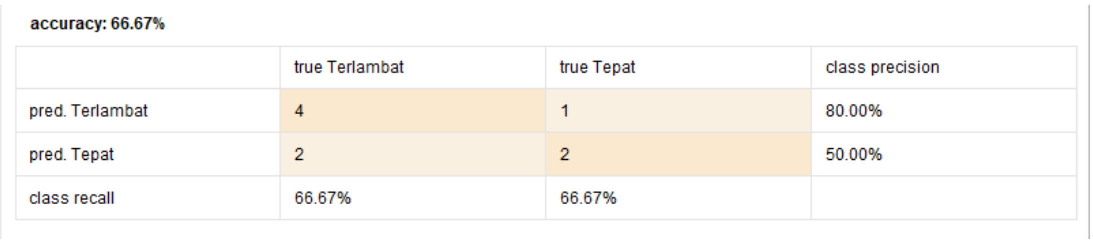
recall: 61.67% +/- 31.48% (micro average: 61.90%) (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	6	8	42.86%
pred. Terlambat	14	13	48.15%
class recall	30.00%	61.90%	

Naïve Bayes

The results of the Naïve Bayes algorithm are shown in table 7, table 8 and table 9. Table 7 shows the precision performance of the Naïve Bayes algorithm of 66.67%. Performance precision of 80% as shown in table 7. And in table 8 performance Recall was 66.67%.

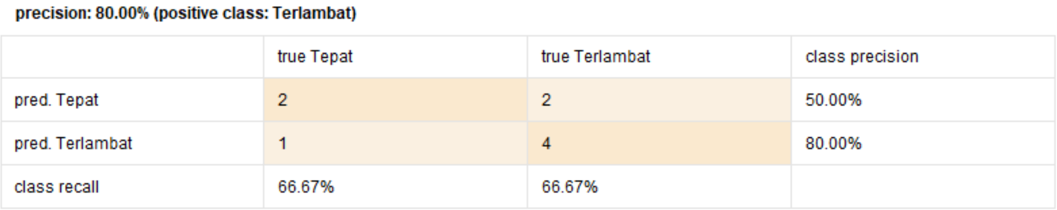
Table 6 Performance Accuracy Naive Bayes



accuracy: 66.67%

	true Terlambat	true Tepat	class precision
pred. Terlambat	4	1	80.00%
pred. Tepat	2	2	50.00%
class recall	66.67%	66.67%	

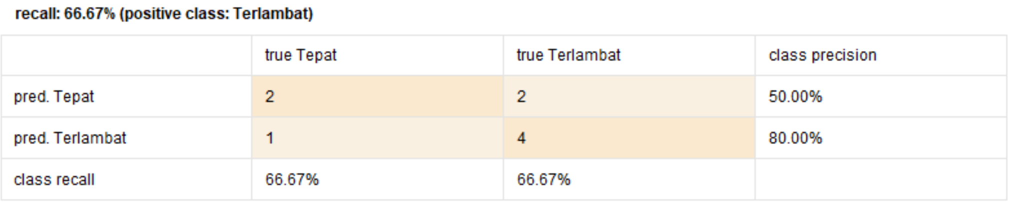
Table 7 Performance Precision Naive Bayes



precision: 80.00% (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	2	2	50.00%
pred. Terlambat	1	4	80.00%
class recall	66.67%	66.67%	

Table 8 Performance Recall Naive Bayes



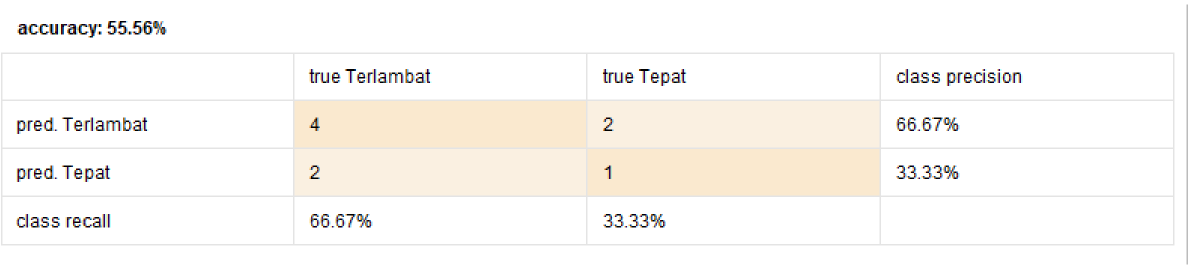
recall: 66.67% (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	2	2	50.00%
pred. Terlambat	1	4	80.00%
class recall	66.67%	66.67%	

K-Nearest Neighbor (KNN)

The performance of the K-Nearest Neighbor (KNN) Algorithm is shown in table 9, table 10 and table 11. Performance Accuracy in table 9, shows that the KNN accuracy level is 55.56%. Performance precision is shown in table 10 showing the level of precision of the KNN algorithm of 66.67%. And table 10 shows the performance of the KNN algorithm at 66.67%.

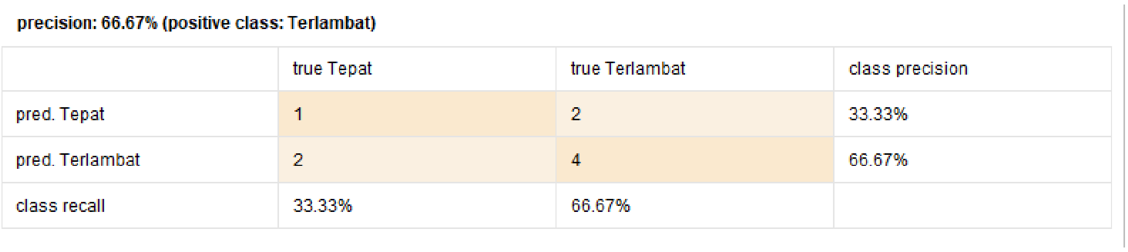
Table 9 Performance Accuracy KNN



accuracy: 55.56%

	true Terlambat	true Tepat	class precision
pred. Terlambat	4	2	66.67%
pred. Tepat	2	1	33.33%
class recall	66.67%	33.33%	

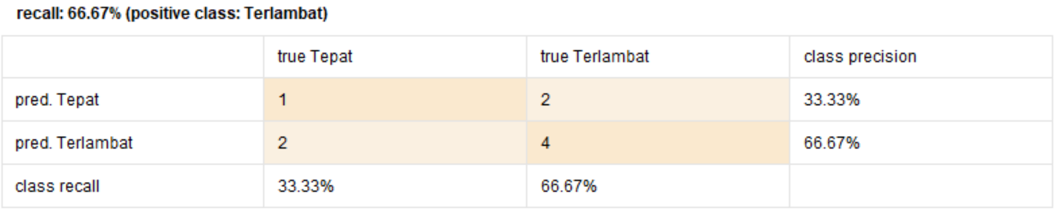
Table 10 Performance Precision KNN



precision: 66.67% (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	1	2	33.33%
pred. Terlambat	2	4	66.67%
class recall	33.33%	66.67%	

Table 11 Performance Recall KNN



recall: 66.67% (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	1	2	33.33%
pred. Terlambat	2	4	66.67%
class recall	33.33%	66.67%	

CONCLUSION

From testing the three algorithms above, the following results are obtained: The best performance level is shown by the Naïve Bayes algorithm with 66, 67% accuracy and 80% precision. The next level is shown by the KNN algorithm with an accuracy rate of 55.56% and precision 66, 67%. 3. Finally, Decision Tree with an accuracy rate of 46% and a precision of 48.3%. Thus, it can be concluded that the predictions in this study are Naïve Bayes, then KNN and finally Decision Tree.

ACKNOWLEDGMENT

The comparison of decision trees, naïve bayes and knn is not a new research, but the comparison of the three methods with the research object of the informatics engineering study program from muhammadiyah higher education with the policy of the islamic religion and kemuhammadiyahan curriculum for up to five semesters is a study that most likely has not been studied. For this reason, this research is expected to be the beginning of further researchers so that they can compare other methods that are better with the same research object.

REFERENCES

Amril Mutoi Siregar;Adam Puspabuana (2017) *DATA MINING*. Pertama. Edited by A. Kusuma Putra. Surakarta: Kekata Publishier.

Andri, A. and Paulus, P. (2021) ‘Prediksi Mahasiswa Berisiko Drop Out (DO) dengan ADTree dan NNge’, *Jurnal SIFO Mikroskil*, 22(1). Available at: https://mikroskil.ac.id/ejurnal/index.php/jsm/article/view/794.

Atma, Y. D. and Setyanto, A. (2018) ‘Perbandingan algoritma c4.5 dan k-nn dalam identifikasi mahasiswa berpotensi drop out’, *Metik Jurnal*, 2(2), pp. 31–37.

BRS (2021) ‘Hasil Sensus Penduduk 2020’, *Bps.Go.Id*, (27), pp. 1–52. Available at: https://papua.bps.go.id/pressrelease/2018/05/07/336/indeks-pembangunan-manusia-provinsi-papua-tahun-2017.html.

Hidayat, M. M., Purwitasari, D. and Ginardi, H. (2013) ‘Analisis Prediksi Drop Out Berdasarkan Perilaku Sosial Mahasiswa Dalam Educational Data Mining’, *Jurnal IPTEK*, 17(2), pp. 109–119.

Jin, Z. *et al.* (2016) ‘Highly efficient terahertz radiation from a thin foil irradiated by a high-contrast laser pulse’, *Physical Review E*, 94(3), pp. 177–184. doi: 10.1103/PhysRevE.94.033206.

Mutrofin, S. *et al.* (2020) ‘Komparasi Kinerja Algoritma C4.5, Gradient Boosting Trees, Random Forests, dan Deep Learning pada Kasus Educational Data Mining’, *Jurnal Teknologi Informasi dan Ilmu Komputer*, 7(4), p. 807. doi: 10.25126/jtiik.2020742665.

Nasrullah, A. H. (2018) ‘Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out’, *ILKOM Jurnal Ilmiah*, 10(2), pp. 244–250. doi: 10.33096/ilkom.v10i2.300.244-250.

Rohman, A. (2015) ‘Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa’, *Neo Teknika*, 1(1). doi: 10.37760/neoteknika.v1i1.350.

Septiana, L. (2013) ‘Penerapan Neural Network Berbasis Particle Swarm Optimization untuk Seleksi Atribut Penentuan Mahasiswa Drop Out’, *Pilar Nusa Mandiri*, 9(2), pp. 104–112.