Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 6, No.1, March 2023, pp. 29 – 39
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Performance Comparison of Data Mining Classification Algorithms on Student Academic Achievement Prediction

[1]Munarsih, [2]Besse Arnawisuda Ningsi
[1,2]Mathematics Study Program, Faculty of Mathematics and Natural Science, Universitas Pamulang
Email: [1]acihacilbgt1305@gmail.com, [2]dosen00205@unpam.ac.id

Article Info

Article history:
Received Jan 15th, 2023
Revised Jan 24th, 2023
Accepted Feb 24th, 2023

Keyword:
Academic Achievement
Classification
C4.5
K-Nearest Neighbour
Naive Bayes

ABSTRACT

Academic achievement is one of the benchmarks of student success in carrying out the learning process. Grade Point Average (GPA) is a reference for universities in determining student academic achievement. For universities, academic achievement can be an indicator of determining the success of the learning system and can improve the image of the university. This study aims to determine the prediction of academic achievement results of Pamulang University students with Naive Bayes, C4.5 and KNN, and to determine the comparison results of Naive Bayes, C4.5 and KNN algorithms in predicting the academic achievement of Pamulang University students. The algorithms compared in this study are Naive Bayes, C4.5 and K-Nearest Neighbor (KNN) algorithms, using the factors of gender, age, faculty, regional origin, work status, organisation participation, type of school origin, distance of residence, and parents' profession as artibut. The results of this study show that the KNN algorithm is the algorithm with the greatest accuracy rate of 56.25%, followed by the Naive Bayes algorithm and the C4.5 algorithm with an accuracy rate of 50.00%.

Corresponding Author:
Besse Arnawisuda Ningsi,
Mathematics Study Program, Faculty of Mathematics and Natural Science,
Universitas Pamulang,
Email: dosen00205@unpam.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v6i1.21874

1. INTRODUCTION

Education is one of the foundations in building a nation, education has a very important role for the progress of a nation, with quality education, a nation can excel in global competition and can realize national welfare. Education is defined as a process with certain methods, so as to obtain knowledge, understanding, and ways of behaving according to needs [1]. Receiving education is the right of every citizen that must be fulfilled by the state itself, as stated in article 31 of the 1945 Constitution paragraph 1 which states that every citizen has the right to receive teaching, which is certainly in line with one of the state's goals stated in the 1945 Constitution, namely educating a nation, not just being good at science but also being able to become a moral human being, so education in Indonesia has been regulated in such a way in the Law of the Republic of Indonesia number 20 of 2003 concerning the National Education System in order to achieve common goals and objectives. One of the articles listed in the Law regulates the levels of education in Indonesia, especially for formal education, which consists of basic education, secondary education and higher education in order to produce people who are experts or professionals in a field.

In carrying out the learning process, learning evaluation is an important thing in measuring student academic achievement, it is considered important because the earlier you know the potential of students who may have obstacles or difficulties that will arise in the learning process, the faster the anticipation can be done for early prevention of students who have the potential for delays in the learning process. One indicator to measure academic achievement is the Grade Point Average (GPA). Factors that support student academic achievement are also important to know in order to benefit and achieve the goals of educational institutions. Student academic achievement is influenced by many factors such as personal, social, psychological, and

others. Aspects such as gender, location of residence, family are also very influential on student learning achievement. According to Slameto, learning is influenced by factors that can be classified into 2 (two) groups, namely internal factors (factors originating from within students) and external factors (factors originating from outside students) [2]. Internal factors include physical factors / physiological conditions, psychological factors and fatigue factors, while external factors include family factors, school factors and community factors.

Academic achievement is an assessment of educational results in the form of changes in the fields of knowledge, understanding, application, analytical power, synthesis and evaluation, where the assessment results are given based on the results of tests, evaluations or exams from each course, these results are interpreted objectively and applied in the form of numbers and sentences according to what each student achieves in a certain period [3]. To find out the academic achievement of students is done by measurement which will be converted into a rating scale. Based on Pamulang University's academic guidelines, the cumulative grade point average which is the benchmark for student academic achievement is a combination of final grades from attendance, assignments, Mid Semester Examinations (UTS), and Final Semester Examinations (UAS). The complete grade at the end of the semester is expressed as A,B,C,D and E which are in Quality Numbers 4,3,2,1 and 0 respectively. Every student certainly has hopes that his academic performance has good results. However, there are still cases of students who have not met the GPA standard > 3. In an effort to improve student academic achievement, it is necessary to pay attention to the factors that influence and what algorithm or method is most accurate in predicting student achievement.

Data mining is the process of turning previously unknown, implicit, and thought of as meaningless data into information, knowledge, or patterns [4]. Finding hidden patterns in massive amounts of data through exploration and analysis is known as data mining. Researchers frequently utilize this idea to analyze data. Classification is one of the data mining ideas that is frequently applied as a first stage in decision-making. Classification in data mining is often used to find models that describe and distinguish classes that aim to estimate the class of objects whose class labels are unknown [5]. Some previous research that has been done related to data mining methods include, research conducted by Tias [6] on the classification of student graduation timeliness with the Naive Bayes method with determining variables namely gender, school type, region of origin, parental employment factors, study program factors, and index predicate factors (cumulative achievement) showing an accuracy rate of 69.33%.

Research conducted by Widaningsih [7] on the comparison of data mining methods for predicting grades and graduation times for informatics engineering students with the C4.5, Naïve Bayes, KNN, and SVM algorithms shows that the Naïve Bayes algorithm has the best accuracy value of 76.79%. , then C4.5 with an accuracy value of 75.96%, SVM with an accuracy value of 74.04% and KNN with an accuracy value of 68.05%.

A similar study was also conducted by Wella [8] concerning a comparison of the KNN, C4.5, and Naive Bayes algorithms in classifying fish freshness using photo media, showing that the KNN algorithm has the highest accuracy value of 97.33% for carp, 89.02% for tilapia. and 87.72% for snapper.

Based on the description above and referring to several previous studies, the algorithms that are applied mostly are decision tree and naive Bayes algorithms, so the researcher wants to test other classification algorithms whether their level of accuracy is better or not in the case of predicting student academic achievement. Researchers will compare several algorithms, namely the decision tree C4.5, naive Bayes and K-NN to find out which algorithm is more accurate in predicting the academic achievement of Pamulang University students. These algorithms were chosen because of their own advantages, namely, the C4.5 algorithm has advantages in terms of speed and simple classification so that it is easy for humans to interpret, the Naive Bayes algorithm can produce maximum accuracy with little training data. Meanwhile, the K-Nearest Neighbor algorithm was chosen because the method is robust against noise data.

2. RESEARCH METHOD

In this study, it proposes a comparison of data mining clacification methods for predicting academic achievement. as for the research methodology used in this research is described in Figure 1.

2.1. Student Academic Achievement Data

Before carrying out the classification process, the data that has been collected is described according to table 2.1. The independent variables in this study are called attributes, which consist of gender, age, region of origin, type of school of origin, parents' profession, employment status, and participation in organisations. As well as Grade Point Average (GPA) as the dependent variable or commonly known by the label (Class). The following is a description of the research variables in tabular form.

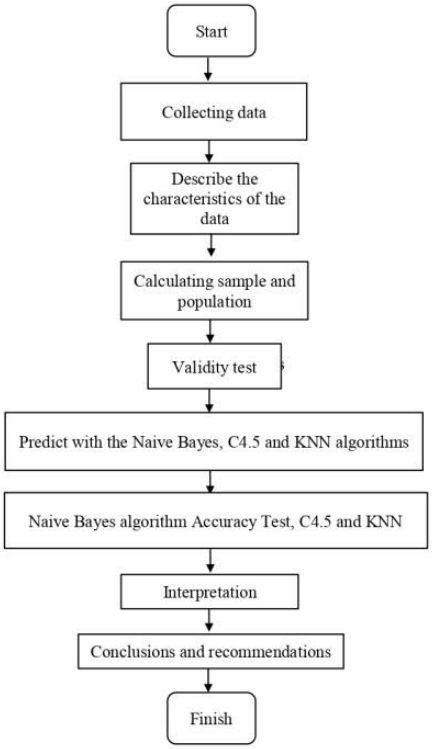

Figure 1. Research flow

Table 1. Data characteristics

No	Variable Type	Operational Explanation	
1	Attribute	Gender(X_1)	0 = Male 1 = Female
2	Attribute	Age(X_2)	0 = < 23 year 1 = ≥ 23year
3	Attribute	Place of Origin(X_3)	0 = Greater Jakarta 1 = Outside Jabodetabek
4	Attribute	Origin School Type(X_4)	0 = Senior High School 1 = Vocational School
5	Attribute	Parent Profession(X_5)	0 = civil servants 1 = Entrepreneur 2 = Farmers 3 = Teacher 4 = Private Employees 5 = Labour 6 = others
6	Attribute	Job status(X_7)	0 = Yes 1 = No
7	Attribute	Job status(X_8)	0 = Yes 1 = No
8	Label	Grade Point Average (GPA) (Y)	0 = < 3.00 1 = 3.00 – 3.50 2 = > 3.50

2.2. Validity Test

Validation is a process for evaluating the prediction accuracy of a model, validation refers to getting predictions using existing models then comparing the results obtained with known results, in validation research used is Cross validation.

Cross validation is a statistical method for evaluating and comparing learning algorithms by dividing the data into two segments: one is used to study or train models and the other is used to validate models [9] . In cross validation, there are two popular approaches to evaluate the performance of the algorithm, namely k-fold cross validation and leave-one-out cross validation. When the amount of data is large, k-fold cross validation must be used to estimate data accuracy [10] .

The work of k-fold cross-validation states that the first data is partitioned into k or segments of the same (or nearly the same) size [11] . Then a number of k-times validations are carried out with each validation using the k-th partition data as testing data and using the remaining partitions as training data, the next step is

to calculate the average accuracy of the k-times validation used as the final validation. In data mining and machine learning 10-fold cross-validation is the most commonly used [12] . The steps of the k fold cross validation are:

1. The total data is divided into k parts.
2. The 1st fold is when the 1st part becomes the test data and the rest becomes the training data. Then, calculate the accuracy or similarity or closeness of a measurement result to the actual number or data based on the portion of the data. The accuracy calculation uses the following equation.

$$akurasi = \frac{\Sigma\, data\ uji\ benar\ klasifikasi}{\Sigma\, total\ data\ uji} \times 100\ \% \quad (1)$$

3. The 2nd fold is when the 2nd part becomes test data and the rest becomes training data. then calculate the accuracy based on the portion of the data.
4. And so on until it reaches the k-th fold. Calculate the average accuracy of the k accuracy above. The average accuracy becomes the final accuracy.

2.3. Data Mining Classification Algorithm

According to Pramudiono, data mining is the automatic analysis of large or complex amounts of data with the aim of finding important patterns or trends that are usually not realized [13] . Data mining is a field of several scientific fields that brings together machine learning techniques, data pattern recognition, statistics, databases and visualization for handling the problem of retrieving information from large databases [14] . Data mining, often also called knowledge discovery in database (KDD), is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets. The output of this data mining can be used to improve decision making in the future. So the term pattern recognition is rarely used because it is part of data mining [15] .

Classification is a process of training (learning) an objective function (target) which is used to map each set of attributes of an object to one of the class labels previously defined. This classification technique is suitable for describing data sets with the data type of a data set, namely binary or nominal.

The classification method has several completion phases, starting from training and ending with the testing process so that an accurate decision is produced. The following is a picture of the flow of solving the classification method.

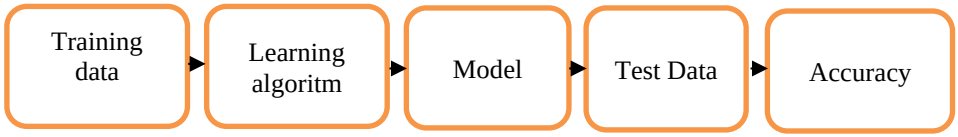

Figure 2. Classification method solution flow

Some of the classification techniques used in this research include:

1. Naive Bayes Clasification
2. C4.5 algorithm
3. K-Nearest Neighbour Algorithm

The Naive Bayes algorithm is a classification algorithm based on the Bayesian theorem in statistics [4]. The Naive Bayes algorithm can be used to predict the probability of membership of a class [16] . Bayesian theory calculates the value of the posterior probability $P(H|X)$using the probability $P(H)$, $P(X)$, and$P(H|X)$ [17]. Naive Bayes Formula:

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \quad (2)$$

Description:

X : sample data that has an unknown class (label)
H : hypothesis that x is class (label) data
$P(H)$: probability of the hypothesis H
$P(X)$: probability of the observed sample data
$P(H|X)$: probability of the sample data X when assuming that the hypothesis is true.

The C4.5 algorithm is in the form of a decision tree like other classification techniques. A decision tree is a structure that can be used to divide a large data set into smaller record sets by applying a set of decision rules. The C4.5 algorithm is one of the decision tree induction C4.5 algorithms, namely ID3 (Iterative Dichotomizer 3). Input in the form of training, training labels and attributes. Algorithm C4.5 is an extension of ID3 if a data set has several observations with missing values, i.e. records with multiple variable values do not exist, if the number of observations is limited then the attributes with missing values can be replaced with the average value of the variable in question [18] . The formula for finding the entropy value:

$$Entropy(S) = -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus} \quad (3)$$

Where:
S : Space (data) sample used for training
$p_{\oplus}$: Number of positive solutions or supports the sample data of certain criteria.
$p_{\ominus}$: number that has a negative solution or does not support certain criteria sample data.
Entropy(S) = 0, if all examples in S are in the same class.
$Entropy(S) = 1$, if the number of positive and negative examples in S is the same.
$0 > Entropy(S) > 1$, if the number of positive and negative examples in S is not the same.

Gain (S,A): Obtaining information from attribute A Relative to the output data S. The formula for finding the Gain value:

$$Gain(S, A) \equiv Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (4)$$

Where:
A : attributes
S : sample
n : number of partitions of attribute set A
$|Si|$: number of samples in the i-th partition
$|S|$: number of samples in$S$

The purpose of the KNN classification algorithm is to predict a new class of data set that has class [19] , KNN is included in the classification algorithm [20] . In Nearest there is the term "similarity" or similarity. The formula used for the Nearest Neighbor value is:

$$Similarity(T, S) = \frac{(\sum_{i=1}^{n} f(Ti, Si) * wi)}{wi} \quad (5)$$

Where:
T : new case
S : cases that are in storage
n : number of attributes in each case
i : individual attribute between 1 to n
f : attribute similiarity function i between case T and case S
w : the weight given to the i-th attribute

Euclidean Distance Formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (6)$$

Where:
$d(x, y)$: distance between data x to data y
x_i : data testing to-i
y_i : data training to-i

The confusion matrix is a method that is usually used to perform accuracy calculations on data mining concepts or decision support systems [21] . In using the Confusion matrix method there are several terms, shown in the following table.

Table 2. Confusion matrix

		True Value	
		FALSE	TRUE
Prediction	FALSE	TN	FP
	TRUE	FN	TP

Description:

1. "True Positive" (TP), is the amount of positive data that is correctly classified by the system.
2. The amount of negative data correctly classified by the system is referred to as True Negative (TN).
3. The amount of negative data classified as positive by the system is known as false positive (FP).
4. The amount of positive data classified negatively by the system is known as false negative (FN).

Precision is the level of accuracy between the information requested by the user and the answer provided by the system. Precision can be calculated with the following formula:

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

Recall is the removal data taken from the relevant data queried and recall is also known as sensitivity. Recall is formulated as follows:

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

F1 Measure is a combination calculation between recall and precision. The F1 Measure value range is 0 to 1, if the value is close to 0 then the prediction model is not good and vice versa if the value is close to 1 then the prediction model is good. To get the value in percentage, the value is multiplied by 100.

$$F1 - Measure = \frac{2*Precision*Recall}{Precision+Recall} \tag{9}$$

F1 Score is to evaluate how well the hybrid metric is used for unbalanced classes. The F1 Score value range is 0 to 1, if the value is close to 0 then the prediction model is not good and vice versa if the value is close to 1 then the prediction model is good. To get the value in percentage, the value is multiplied by 100.

$$F1\ Score = \frac{(2TP)}{(2TP+FP+FN)} \tag{10}$$

As well as in calculating the percentage of accuracy with the following formula:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \tag{11}$$

Then it can also calculate APPER or the so-called error rate is an evaluation measure used to see the chance of misclassification produced by a classification function. The smaller the APPER value, the better the classification results [22]. The formulation for calculating APER is:

$$APPER = \frac{FP+FN}{TP+FN+FP+TN} \times 100\% \tag{12}$$

3. RESULTS AND ANALYSIS

At this stage presenting the results of testing the three classification methods. The data used in this study are data from the results of questionnaires given to 87 semester 7 (seven) students, the data is processed and selected, resulting in the following data:

Table 3. Respondent Data

No	gender	Age	Origin	Work	Organization	Distance	Parent Profession	Which school are you from	GPA
1	Man	21	Outside Jabodetabek	Yes	Yes	5km - 10km	Laborer	SENIOR HIGH SCHOOL	3.47
2	Man	21	Greater Jakarta	Yes	Yes	< 5km	Farmer	SENIOR HIGH SCHOOL	3.62
3	Man	24	Greater Jakarta	Yes	No	> 10km	Other	Vocational School	3.55

No	gender	Age	Origin	Work	Organization	Distance	Parent Profession	Which school are you from	GPA
4	Woman	28	Greater Jakarta	Yes	No	> 10km	Self-employed	Vocational School	3.59
5	Woman	21	Greater Jakarta	No	Yes	< 5km	Laborer	SENIOR HIGH SCHOOL	3.66
6	Woman	21	Greater Jakarta	No	No	> 10km	Other	Vocational School	3.53
7	Woman	23	Greater Jakarta	No	No	< 5km	Self-employed	SENIOR HIGH SCHOOL	3.48
8	Woman	22	Greater Jakarta	No	No	> 10km	Other	SENIOR HIGH SCHOOL	3.74
9	Woman	21	Greater Jakarta	No	No	5km - 10km	Other	SENIOR HIGH SCHOOL	3.45
10	Woman	23	Greater Jakarta	No	Yes	> 10km	Private sector employee	Vocational School	3.46
...	...	...	...	...	...	...	...	...	...
78	Woman	23	Outside Jabodetabek	No	No	<5KM	Self-employed	SENIOR HIGH SCHOOL	3.54

Then the data processing is carried out to obtain Naive Bayes testing data using the help of Python software, so that the following results are obtained:

Table 4. Naive Bayes data

No	Gender	Age	Origin	Work	Organization	Distance	Parents' job	Which school are you from	GPA	New class
1	Man	20	Greater Jakarta	No	Yes	> 10km	Other	Vocational School	1	1
2	Man	26	Outside Jabodetabek	Yes	No	> 10km	Farmer	SENIOR HIGH SCHOOL	1	2
3	Woman	21	Outside Jabodetabek	No	No	> 10km	Farmer	SENIOR HIGH SCHOOL	1	2
4	Woman	22	Greater Jakarta	No	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
5	Woman	22	Greater Jakarta	Yes	No	5-10km	Other	Vocational School	1	2
6	Woman	22	Outside Jabodetabek	No	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
	Man	25	Greater Jakarta	Yes	No	< 5km	Laborer	Vocational School	2	1
8	Woman	20	Greater Jakarta	No	No	5km - 10km	Other	SENIOR HIGH SCHOOL	2	2
9	Woman	25	Greater Jakarta	Yes	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
10	Woman	25	Greater Jakarta	Yes	Yes	5km - 10km	Other	SENIOR HIGH SCHOOL	1	1
11	Woman	22	Greater Jakarta	No	No	> 10km	Laborer	SENIOR HIGH SCHOOL	2	2
12	Woman	24	Greater Jakarta	Yes	Yes	> 10km	Self-employed	SENIOR HIGH SCHOOL	2	1
13	Man	21	Greater Jakarta	Yes	Yes	< 5km	Farmer	SENIOR HIGH SCHOOL	2	1
14	Woman	21	Greater Jakarta	Yes	No	> 10km	Self-employed	SENIOR HIGH SCHOOL	0	2
15	Man	21	Greater Jakarta	No	Yes	> 10km	Laborer	Vocational School	1	1
16	Woman	22	Greater Jakarta	Yes	Yes	> 10km	Laborer	Vocational School	2	1

After processing the Naïve Bayes data, the next step is to convert the data into a confusion matrix table. Namely as follows:

Table 5. Naive Bayes Confusion matrix

Actual	Predictions		
	GPA < 3.00	GPA 3.00 – 3.50	GPA > 3.50
GPA < 3.00	0	0	1
GPA 3.00 – 3.50	0	3	3
GPA > 3.50	0	4	5

At this stage, the results of calculating the accuracy level will be presented using the help of Python software.

Table 6. Naive Bayes calculation results with Python

	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	1
1	0.43	0.50	0.46	6
2	0.56	0.56	0.56	9
Accuracy			0.50	16
macros avg	0.33	0.35	0.34	16
Weighted avg	0.47	0.50	0.49	16

Based on the results of calculations using the help of Python software, it is known that the precision value is 32.80%, this value is a comparison of true positive predictions with overall positive predicted results. Then a recall value of 35.18% was obtained, this value is a comparison of true positive predictions with all actual positive data. Furthermore, in the calculation above, the results obtained an accuracy of 50.00% and an error rate of 50.00%.

Then the data processing is carried out to obtain C4.5 testing data using the help of Python software, so that the following results are obtained:

Table 7. Data C4.5

No	Gender	Age	Origin	Work	Organization	Distance	Parents' job	Which school are you from	GPA	New class
1	Man	20	Greater Jakarta	No	Yes	> 10km	Other	Vocational School	1	2
2	Man	26	Outside Jabodetabek	Yes	No	> 10km	Farmer	SENIOR HIGH SCHOOL	1	2
3	Woman	21	Outside Jabodetabek	No	No	> 10km	Farmer	SENIOR HIGH SCHOOL	1	2
4	Woman	22	Greater Jakarta	No	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
5	Woman	22	Greater Jakarta	Yes	No	5-10km	Other	Vocational School	1	2
6	Woman	22	Outside Jabodetabek	No	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
	Man	25	Greater Jakarta	Yes	No	< 5km	Laborer	Vocational School	2	1
8	Woman	20	Greater Jakarta	No	No	5km - 10km	Other	SENIOR HIGH SCHOOL	2	2
9	Woman	25	Greater Jakarta	Yes	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
10	Woman	25	Greater Jakarta	Yes	Yes	5km - 10km	Other	SENIOR HIGH SCHOOL	1	2
11	Woman	22	Greater Jakarta	No	No	> 10km	Laborer	SENIOR HIGH SCHOOL	2	2
12	Woman	24	Greater Jakarta	Yes	Yes	> 10km	Self-employed	SENIOR HIGH SCHOOL	2	2
13	Man	21	Greater Jakarta	Yes	Yes	< 5km	Farmer	SENIOR HIGH SCHOOL	2	2
14	Woman	21	Greater Jakarta	Yes	No	> 10km	Self-employed	SENIOR HIGH SCHOOL	0	2
15	Man	21	Greater Jakarta	No	Yes	> 10km	Laborer	Vocational School	1	2
16	Woman	22	Greater Jakarta	Yes	Yes	> 10km	Laborer	Vocational School	2	2

After processing C4.5 data, the next step is to convert the data into a confusion matrix table. Namely as follows:

Table 8. *Confusion matrix C4.5*

Actual	Predictions		
	GPA < 3.00	GPA 3.00 – 3.50	GPA > 3.50
GPA < 3.00	0	0	1
GPA 3.00 – 3.50	0	0	6
GPA > 3.50	0	1	8

At this stage the results of calculating the accuracy level will be presented using the help of Python software.

Table 9. Calculation results of C4.5 with Python

	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	1
1	0.00	0.00	0.00	6
2	0.53	0.89	0.67	9

	Precision	Recall	F1-score	Support
accuracy			0.50	16
macros avg	0.18	0.30	0.22	16
Weighted avg	0.30	0.50	0.38	16

Based on the results of calculations using the help of Python software, it is known that the precision value is 17.78%, this value is a comparison of true positive predictions with overall positive predicted results. Then a recall value of 29.62% was obtained, this value is a comparison of true positive predictions with all actual positive data. Furthermore, in the calculation above, the results obtained an accuracy of 50.00% and an error rate of 50.00%.

Then the data processing is carried out to obtain C4.5 testing data using the help of Python software, so that the following results are obtained:

Table 10. Data C4.5

No	Gender	Age	Origin	Work	Organization	Distance	Parents' job	Which school are you from	GPA	New class
1	Man	20	Greater Jakarta	No	Yes	> 10km	Other	Vocational School	1	2
2	Man	26	Outside Jabodetabek	Yes	No	> 10km	Farmer	SENIOR HIGH SCHOOL	1	2
3	Woman	21	Outside Jabodetabek	No	No	> 10km	Farmer	SENIOR HIGH SCHOOL	1	2
4	Woman	22	Greater Jakarta	No	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
5	Woman	22	Greater Jakarta	Yes	No	5-10km	Other	Vocational School	1	2
6	Woman	22	Outside Jabodetabek	No	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
	Man	25	Greater Jakarta	Yes	No	< 5km	Laborer	Vocational School	2	1
8	Woman	20	Greater Jakarta	No	No	5km - 10km	Other	SENIOR HIGH SCHOOL	2	2
9	Woman	25	Greater Jakarta	Yes	No	5km - 10km	Laborer	SENIOR HIGH SCHOOL	2	2
10	Woman	25	Greater Jakarta	Yes	Yes	5km - 10km	Other	SENIOR HIGH SCHOOL	1	2
11	Woman	22	Greater Jakarta	No	No	> 10km	Laborer	SENIOR HIGH SCHOOL	2	2
12	Woman	24	Greater Jakarta	Yes	Yes	> 10km	Self-employed	SENIOR HIGH SCHOOL	2	2
13	Man	21	Greater Jakarta	Yes	Yes	< 5km	Farmer	SENIOR HIGH SCHOOL	2	2
14	Woman	21	Greater Jakarta	Yes	No	> 10km	Self-employed	SENIOR HIGH SCHOOL	0	2
15	Man	21	Greater Jakarta	No	Yes	> 10km	Laborer	Vocational School	1	2
16	Woman	22	Greater Jakarta	Yes	Yes	> 10km	Laborer	Vocational School	2	2

After processing the KNN data, the next step is to convert the data into a confusion matrix table. Namely as follows:

Table 11. *KNN confusion matrix*

Actual	Predictions		
	GPA < 3.00	GPA 3.00 – 3.50	GPA > 3.50
GPA < 3.00	0	0	1
GPA 3.00 – 3.50	0	0	6
GPA > 3.50	0	0	9

At this stage, the results of calculating the accuracy level will be presented using the help of Python software.

Table 12. KNN calculation results with Python

	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	1
1	0.00	0.00	0.00	6
2	0.56	1.00	0.72	9
accuracy			0.56	16

	Precision	Recall	F1-score	Support
macros avg	0.19	0.33	0.24	16
Weighted avg	0.32	0.56	0.40	16

Based on the results of calculations using the help of Python software, it is known that the precision value is 18.75%, this value is a comparison of true positive predictions with overall positive predicted results. Then a recall value of 33.33% was obtained, this value is a comparison of true positive predictions with all actual positive data. Furthermore, in the calculation above, the results obtained an accuracy of 56.25% and an error rate of 43.75%.

Table 13. Algorithm comparison results

Algorithm	Precision	recall	accuracy	error
Naive Bayes	3 2 , 85 %	3 5.18 %	50.00%	50.00%
C4.5	17.78 %	29.62 %	50.00%	50.00%
KNN	18.75 %	33.33 %	56.25%	43.75%

Based on the table above, it can be seen that the accuracy value for the best classification algorithm method is K-Nearest Neighbor with an accuracy value of 51.94%, then the Naive Bayes algorithm and C4.5 with an accuracy value of 50.00%.

4. CONCLUSION

The results of the comparative research on the performance of data mining algorithms on predicting student achievement show a fairly low accuracy value, this can be influenced by the amount of data used and the preprocessing stages carried out. From the evaluation results, it is obtained that the KNN algorithm is the best for predicting student achievement because it has the highest accuracy value and the smallest error compared to other algorithms, followed by the Naive Bayes and C4.5 algorithms. For future researchers it is suggested to be able to add the amount of data and attributes to be studied, in the hope of getting better calculation accuracy. Substituting for case studies, to be able to find out whether the best method in this study is also the best method in further research, as well as using random forest, constant and other methods.

REFERENCES

[1] MP . Arinda Fidianti, *IMPLEMENTATION OF SCHOOL-BASED MANAGEMENT IN IMPROVING STUDENT LEARNING ACHIEVEMENT* , 1st ed. yogyakarta: GRE PUBLISHING , 2018.
[2] O. Anselmus Cauna, MH Pratiknjo, and D. Deeng, "BEHAVIOR OF STUDENTS FROM PAPUA IN THE LEARNING PROCESS IN UNIVERSITAS SAM RATULANGI MANADO."
[3] PDS Arikunto, *Fundamentals of Educational Evaluation 3rd Edition* . Jakarta: PT. Bumi Aksara, 2018.
[4] J. Suntoro, *Data Mining: Algorithm and Implementation with PHP Programming* . Elex media computer, 2019.
[5] FA Hizham, Y. Nurdiansyah, and DM Firmansyah, "Implementation of the Backpropagation Neural Network (BNN) method in the classification system for student graduation timeliness," *Berk. Sciencetek* , vol. 6, no. 2, pp. 97–105, 2018.
[6] IA Tias Mugi Rahayu, Besse Arnawisuda Ningsi, Isnurani, "STUDENT GRADUATION TIME CLASSIFICATION USING THE NAÏVE BAYES METHOD," 2021.
[7] S. Widaningsih, "COMPARATION OF DATA MINING METHODS FOR PREDICTION OF VALUE AND GRADUATION TIME OF INFORMATICS ENGINEERING PROGRAM STUDENTS WITH ALGORITHM C4.5, NAÏVE BAYES, KNN, AND SVM," 2019.
[8] NMS Iswari, W. Wella, and R. Ranny, "Comparison of KNN, C4.5, and Naive Bayes Algorithms in Classifying Fish Freshness Using Photo Media," *J. Ultim.* , vol. 9, no. 2, pp. 114–117, 2017.
[9] S. Verma, Kashvi Taunk, Sanjukta De and A. Swetapadma, "'A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,' 2019 International Conference on Intelligent Computing and Control Systems (ICCS)," pp. 1255–1260, 2019.
[10] S. Cahyani, R. Wiryasaputra, and R. Gustriansyah, "Identification of Handwritten Capital Letters Using Linear Discriminant Analysis and Euclidean Distance," *J. Sist. inf. Business* , vol. 8, no. 1, p. 57, 2018.
[11] FE Alfian, IGPS Wijaya, and F. Bimantoro, "Iris Identification Using the Daubechies Wavelet Method and K-Nearest Neighbor," *J. Teknol. Information, Computers, and Apps. (JTIKA)* , vol. 2, no. 1, pp. 1–10, 2020.
[12] D. Berrar, "Cross-validation," *Encyclopedia. Bioinform. Comput. Bio. ABC Bioinform.* , vol. 1–3, nos. April, pp. 542–545, 2018.
[13] P. Purwadi, PS Ramadhan, and N. Safitri, "Application of Data Mining to Estimate Population Growth Rates Using Multiple Linear Regression Methods at BPS Deli Serdang," J. SAINTIKOM (Journal of Manaj Science. *Information and Computers)* , vol. . 18, no. 1, p. 55, 2019.
[14] W. Purba, W. Siawin, and . H., "Data Mining Implementation for Grouping and Predicting Potential Layoffs Using the K-Means Clustering Algorithm," *J. Sist. inf. and Computing Science. Prima(JUSIKOM PRIMA)* , vol. 2, no. 2, pp. 85–90, 2019.
[15] and HW Zhang, Junlin, Samuel Oluwarotimi Williams, "Intelligent computing system based on pattern

recognition and data mining algorithms," pp. 192–202.
[16] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementation of Data Mining with the Naïve Bayes Algorithm for the Eligibility Classification of Food Aid Recipients," JURIKOM (*Journal of Ris. Computers)* , vol. 8, no. 6, p. 219, 2021.
[17] Y. Yuliana, P. Paradise, and K. Kusrini, "Expert System for Diagnosing Respiratory Tract Infection Using the Web-Based Naive Bayes Classifier Method," *CSRID (Computer Sci. Res. Its Dev. Journal)* , vol. 10, no. 3, p. 127, 2021.
[18] I. Junaedi, N. Nuswantari, and V. Yasin, "Design and Implementation of the C4 Algorithm. 5 For Data Mining," *J. Inf. syst. Informatics Comput.* , vol. 3, no. 1, pp. 29–44, 2019.
[19] MY Putra and DI Putri, "Utilization of Naïve Bayes and K-Nearest Neighbor Algorithms for Classification of Class XI Student Majors," *J. Tekno Kompak* , vol. 16, no. 2, pp. 176–187, 2022.
[20] BS Amalia, Y. Umaidah, and R. Mayasari, "Analysis of Restaurant Customer Review Sentiment Using Support Vector Machine and K-Nearest Neighbor Algorithms," *SITEKIN J. Science, Teknol. and Ind.* , vol. 19, no. 1, pp. 28–34, 2021.
[21] EB Serelia and MR Adin Saf, "Decision Support System for Determining Student Specialization Using the SAW (Simple Additive Weighting) Method at SMA Negeri Dharma Pendidikan," *Techno.Com* , vol. 19, no. 3, pp. 227–236, 2020.
[22] Haniah Mahmudah, Okkie Puspitorini, Nur Adi Siswandari, Ari Wijayanti, and Eliya Alfatekha, "Naive Bayes Classifier Method - Smoothing on Smartphone Sensors for Classification of Rider Activities," *J. Nas. Tech. Electrical and Technol. inf.* , vol. 9, no. 3, pp. 268–277, 2020

BIBLIOGRAPHY OF AUTHORS

Munarsih, born in South Tangerang on March 13 2001. Completed his bachelor's degree in the Mathematics Study Program at Pamulang University in 2018 - 2022. He won a bachelor's degree in Mathematics (S.Mat) at Pamulang University in 2022

BESSE ARNAWISUDA NINGSI, was born in Ujung Pandang 24 January 1983. She obtained her Bachelor of Science (S.Si) degree in Statistics at Hasanuddin University Makassar in 2005. She obtained her Master of Science (M.Si) degree in Statistics at Bogor Agricultural University (IPB) in 2012. In 2021 - currently completing the Doctoral Programme in Educational Research and Evaluation at the State University of Jakarta (UNJ). From 2007 to 2012, he was a non-permanent lecturer in the Mathematics Education study programme at Prof. Dr. Hamka Muhammadiyah University. In 2009 until now, he has been a permanent lecturer in the Mathematics Study Programme at Pamulang University. Courses that have been taught include Basic Mathematics, Basic Statistics, Mathematical Statistics, Regression Analysis, Stochastic Processes, Sampling Techniques, Research Methodology and others. He has achieved achievements in the field of research in 2019 by getting a Higher Education research grant with the Beginner Lecturer Research scheme as the second researcher with the research title "Modelling the Poverty Status of South Tangerang City Residents Using Partial Least Square - Path Modelling". One of the books that has been published is Basic Econometrics Theory and Practice Based on SPSS which was published in 2021.