

Application of Naïve Bayes Method for Assessing Student Performance

Riza Akhsani Setyo Prayoga^{*1)}, Fauzan Nusyura²⁾, Fiddin Yusufida A'la³⁾, Mustafa Kamal⁴⁾,
Farhanna Mar'i⁵⁾

¹⁾Information Technology Education, State University of Surabaya, Surabaya, Indonesia

²⁾Electrical & Electronic Engineering, Universiti Kebangsaan Malaysia, Selangor, Malaysia

³⁾Informatics Engineering, Sebelas Maret University, Surakarta, Indonesia

⁴⁾Information Technology, Telkom University, Surabaya, Indonesia

⁵⁾ Informatics Engineering, State University of Surabaya, Surabaya, Indonesia

^{1*}rizayoga30@gmail.com, ²P141316@siswa.ukm.edu.my, ³fiddin@staff.uns.ac.id,

⁴mustafakamal@telkomuniversity.ac.id, ⁵farhannamari@unesa.ac.id

ARTICLE INFO

Article history:

Received 2 April 2026

Revised 11 May 2026

Accepted 25 May 2026

Available online 31 May 2026

Keywords:

Data Mining

Naïve Bayes

Software Engineering

Students

ABSTRACT

Every school has pupils with varying levels of achievement. These differences in achievement can be influenced by several factors, such as the parents' level of education and the pupils' readiness for examinations. Furthermore, they can also be influenced by pupils' abilities in mathematics, writing, and reading. The aim of this study is to classify student performance so that the performance of students at an adequate level or below average can be improved. The method used in this study is Naïve Bayes as a classification method. There are 150 training data points and 50 test data points. Five metrics were evaluated: precision at 94.4%, recall at 94.4%, specificity at 50%, accuracy at 90%, and the F1 score at 94%. This indicates that the model performs well in providing accurate positive predictions. Furthermore, the model is capable of detecting the majority of positive cases effectively

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. Introduction

A process known as Knowledge Discovery in Databases (KDD) involves the collection and use of data to identify patterns, consistency, or relationships within a dataset [1]. Analysing data mining using the Knowledge Discovery in Databases (KDD) framework involves several stages, such as data selection, pre-processing, and transformation [2], followed by data mining and evaluation [3]. The process of data mining itself involves identifying previously unknown relationships within data and then presenting these findings in a way that is easy to understand, thereby facilitating decision-making based on the identified relationships [4].

In a school, there are many pupils with different talents and abilities. Some pupils are good at maths, some are good at reading, and some are good at writing [5]. This is closely linked to the educational background of both parents and the learning patterns established both at school and at home. In this context, pupils' performance will be assessed on the basis of two factors: their preparation for tests and their parents' educational background, supported by their performance in mathematics, writing, and reading. In addition to these factors, there are several others, such as motivation to learn; psychological and emotional factors—including stress, anxiety, low self-confidence, and academic pressure—which can lead to a decline in concentration; and a learning environment that is conducive, safe and supportive of the learning process, which can enhance pupils' focus and comfort [6]. A classification method is therefore required to assist in determining pupils' performance based on these three factors [7]. This classification method uses Naïve Bayes. The Naïve Bayes method can help predict the classification of data [8]. Naïve Bayes can also assist in decision-making and can be used to predict the probability of a

class [9]. The advantage of Naïve Bayes is that it is an algorithm that predicts probabilities based on the application of Bayes’ theorem to the relationships between features in data that are not interrelated [10]. This algorithm is one of several simple methods for classifying an equation [11]. The Naive Bayes method can process data quickly, even when the amount of data is small or limited. If the probability is zero, the prediction will also be zero. The aim and benefits of the research presented here are to assist in classifying student performance according to categories. The classification is then carried out using Naive Bayes [12].

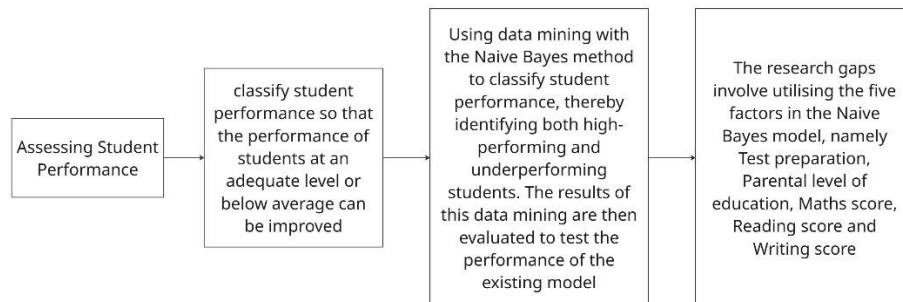


Figure 1. Flowchart

In research related to Naïve Bayes, there is a title “Sentiment analysis of merdeka belajar kampus merdeka using naïve bayes classifier algorithm” that shows that this Naïve Bayes algorithm achieved the highest accuracy in the classification test at 70%, a figure obtained from a test using a data split of 90%:10% [13]. There is another study titled “Predicting the risk of stunting in families using a Naive Bayes classifier and the chi-square test” that shows that the Naïve Bayes algorithm has an accuracy of 94.3%, a recall of 93.9%, and a precision of 93.93%; when combined with the chi-square feature, this increases the model generation speed by approximately 0.01 seconds [14]. Then there is another study entitled “Prediction of poverty rates in Kemang Bejalu village using the Naive Bayes method” shows that the Naïve Bayes algorithm achieved an accuracy of 86% on a training set of 258 data points for three variables, whilst the test set, comprising the same three variables, achieved an accuracy of 90% [15].

Next, there is a study entitled “A system for predicting the duration of university studies using the Naive Bayes method” that shows that the Naïve Bayes algorithm has an accuracy of 80%, a recall of 50%, and a precision of 100% on a sample dataset of 13 data points and a test dataset of 5 data points [16]. There is also a study comparing Naïve Bayes with other methods, entitled “A Comparative Analysis of Decision Tree, Random Forest and Naïve Bayes Algorithms for Flood Prediction in Dayeuhkolot Village.” This shows that Naïve Bayes tends not to overfit, unlike decision trees, which tend to overfit, although random forests tend to reduce overfitting. Naïve Bayes also performs well even when the dataset is imbalanced. Random forests tend to be robust when handling datasets because they utilise many trees; by contrast, decision trees do not perform particularly well when dealing with imbalanced data [17].

2. Methods

This framework has been developed to provide an overview of the research and to assist readers in following the study's progression. The framework is presented in Figure 1.

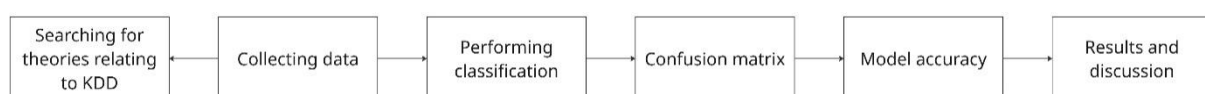


Figure 2. Research Workflow

Figure 2 above begins with the collection of data relating to factors affecting student performance; this data was collected on <https://www.kaggle.com/datasets/kojibrand/student-performance-and-academic-scores-dataset>. After obtaining the data, the Naïve Bayes algorithm is applied in accordance with the formula in Equation 1. Testing is then carried out using a confusion matrix, followed by the results and discussion. The Naïve Bayes algorithm employs a formula for performing calculations, which is used to determine the appropriate class for analysis. The formula for the Naïve Bayes algorithm is presented in Equation 1.

$$P(C_i | X) = \frac{P(X|C_i).P(C_i)}{P(X)} \quad (1)$$

As explained below, X is an unknown class. C_i represents the probability that data X falls into a particular class. P (C_i | X) is the probability of C_i given condition X. P(C_i) represents the probability of C_i. P (X | C_i) is the probability of X given the conditional probability C_i. P(X) is the probability of X.

3. Results and Discussions

The training and testing data used comprise 300 training data points and 100 testing data points. The training data will then be processed using the Naïve Bayes algorithm. The training data is presented in Table 1.

Table 1. Data Training

No	Gender	Test prepared	Parental level of education	Math Score	Reading Score	Writing Score	Information
1	Female	None	bachelor's degree	72	72	74	Enough
2	Female	Completed	some college	69	90	88	Good
3	Female	None	master's degree	90	95	93	Good
4	Male	None	associate's degree	47	57	44	Less
5	Male	None	some college	76	78	75	Enough
6	Female	None	associate's degree	71	83	78	Enough
7	Female	Completed	some college	88	95	92	Good
8	Male	None	some college	40	43	39	Less
9	Male	Completed	high school	64	64	67	Enough
10	Female	None	high school	38	60	50	Less
..
296	Male	None	associate's degree	67	62	60	Enough
297	Male	Completed	some high school	46	41	43	Less
298	Male	Completed	associate's degree	71	74	68	Enough
299	Male	Completed	high school	40	46	50	Less
300	Male	None	associate's degree	90	87	75	Good

The training data above is sourced from the Kaggle website, where 300 data points were extracted, and 100 data points were provided for testing. The breakdown of the training and testing data is shown in Table 2.

Table 2. Data training and data testing

Type of Data	Count
Data Training	300 Data
Data Testing	100 Data

Next, we calculated the probability of the prepared test based on the training data in Table 1. The results of these probability calculations are presented in Table 3.

Table 3. Probability calculation results for the prepared test

Data Types	Number of Criteria			Opportunity Criteria		
	Good	Enough	Less	Good	Enough	Less
None	24/55	95/155	73/90	0,44	0,61	0,81
Completed	31/55	60/155	17/90	0,56	0,39	0,19

Next, we calculated the probability of parental educational attainment based on the training data in Table 1. The results of these probability calculations are shown in Table 4.

Table 4. Probability calculation results: Parental level of education

Data Types	Number of Criteria			Opportunity of Criteria		
	Good	Enough	Less	Good	Enough	Less
bachelor's degree	7/34	23/34	4/34	0,21	0,68	0,12
some college	13/69	38/69	18/69	0,19	0,55	0,26
master's degree	5/20	11/20	4/20	0,25	0,55	0,20
Some High School	9/55	26/55	20/55	0,16	0,47	0,36
associate's degree	15/71	34/71	22/71	0,21	0,48	0,31
High School	6/51	23/51	22/51	0,12	0,45	0,43

Once the classification process using Naïve Bayes had been completed, the data was then tested using 100 test data points. The test data is presented in Table 5.

Table 5. Data testing

No	Gender	Test prepared	Parental level of education	a-th Score	Reading Score	Writing Score	Information
1	Male	Completed	some high school	62	67	69	Enough
2	Male	None	bachelor's degree	77	67	68	Enough
3	Female	Completed	associate's degree	59	70	66	Enough
4	Male	None	bachelor's degree	54	49	47	Less
5	Male	None	some high school	62	67	61	Enough
6	Female	Completed	some college	70	89	88	Good
7	Female	Completed	high school	66	74	78	Enough
8	Male	None	some college	60	60	60	Enough
9	Female	Completed	associate's degree	61	86	87	Enough
10	Male	None	associate's degree	66	62	64	Enough
..

No	Gender	Test prepared	Parental level of education	a-th Score	Reading Score	Writing Score	Information
96	Male	None	high school	48	45	41	Less
97	Female	None	high school	50	67	63	Enough
98	Female	None	associate's degree	85	89	95	Good
99	Male	None	some high school	74	63	57	Enough
100	Male	None	some high school	60	59	54	Less

The classification results from the test data are presented in the confusion matrix in Figure 3 below. The confusion matrix includes four categories: True Positive, True Negative, False Positive, and False Negative. A True Positive is a test result that is positive and supported by positive training results. A True Negative is a test result that is negative, supported by a negative training result. A False Positive is a test result that is positive, supported by a negative training result. A False Negative is a test result that is negative, supported by a positive training result.

Confusion Matrix			
TARGET \ OUTPUT	Positif	Negatif	SUM
Positif	85 85.00%	5 5.00%	90 94.44% 5.56%
Negatif	5 5.00%	5 5.00%	10 50.00% 50.00%
SUM	90 94.44% 5.56%	10 50.00% 50.00%	90 / 100 90.00% 10.00%

Figure 3. Confusion Matrix

Furthermore, the results from the confusion matrix above can be used to calculate the precision, recall, accuracy, specificity, and F1 score, and can also be converted into percentages. The precision value represents the proportion of true positives out of the total number of positives (true positives and false positives). The formula for precision is shown in Equation 2.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

The precision value calculated using Equation 2 is 94.4%; this indicates that the model successfully predicted 94.4% of the total positive predictions correctly.

The recall value describes how well a model correctly identifies positive classes. The formula for the recall value is given in Equation 3.

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The recall value calculated using Equation 3 is 94.4%; in this context, the model successfully identified the positive results correctly but missed 5.6% of the actual positive results.

Next, the specificity value indicates how effective the model is at correctly classifying the negative class. The formula for the specificity value is given in Equation 4.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

Thus, the specificity value calculated using Equation 4 is 50%; in this context, the model successfully identified negative results with 50% accuracy.

Next, there is the accuracy metric for evaluating the model's ability to make correct predictions out of the total number of predictions. The formula for the accuracy metric is given in Equation 5.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Thus, the accuracy calculated using Equation 5 yields a value of 90%; in this context, the model successfully predicted 90% of the negative and positive results correctly.

Finally, there is the F1 score, which represents a balanced assessment of precision and recall. The formula for the F1 score is given in Equation 6.

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6)$$

The result of the F1 Score, calculated using Equation 6, is 94%; in this context, the balance between precision and recall is 94%. The results of testing are shown in Table 6.

Table 6. Evaluation Result

Name	Result
Precision	94,4%
Recall	94,4%
Specificity	50%
Accuracy	90%
F1 Score	94%

4. Conclusion

The application of student performance classification using the Naïve Bayes method can be used to classify which students are performing well, adequately, or poorly, thereby enabling improvements to be made for those students whose performance is still adequate or lacking. Furthermore, the results of the tests conducted covering five metrics: precision, recall, accuracy, specificity, and F1 score vary. The precision value was 94.4%, recall was 94.4%, specificity was 50%, accuracy was 90%, and the F1 Score was 94%. This study is still limited in terms of the variables used; in future research, additional supporting variables could be included, and other methods could be incorporated alongside the one used here.

References

- [1] A. I. Lubis, U. Erdiansyah, and R. Siregar, "Komparasi Akurasi Pada Naïve Bayes Dan Random Forest Dalam Klasifikasi Penyakit Liver Comparison of Accuracy in Naïve Bayes and Random Forests in Classification of Liver Disease," *Journal of Computing Engineering, System and Science*, vol. 7, no. 1, 2022, [Online]. Available: www.jurnal.unimed.ac.id
- [2] W. Aliman, "Implementasi Metode Naïve Bayes untuk Menentukan Persetujuan Pemberian Beasiswa Penuh pada Penerimaan Mahasiswa Baru di Institusi Pendidikan X," *Media Informatika*, vol. 21, no. 03, 2022, doi: <https://doi.org/10.37595/mediainfo.v22i1.91>.

- [3] M. Azhar Mujahid and M. Syafrullah, "Implementasi Algoritma Naïve Bayes Clasifier untuk Mengelompokkan Naskah Berita Pendidikan dan berita Covid-19," *KRESNA: Jurnal Riset dan Pengabdian Masyarakat*, vol. 1, no. 1, pp. 34–43, 2021, doi: <https://doi.org/10.36080/jk.v1i1.2>.
- [4] A. Putri, C. Syaficha Hardiana, E. Novfuja, F. Try Puspa Siregar, Y. Fatma, and R. Wahyuni, "Comparison of K-NN, Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction," *Institut RiMALCOM: Indonesian Journal of Machine Learning and Computer Science Journal*, vol. 3, no. 1, pp. 20–26, 2023, doi: <https://doi.org/10.57152/malcom.v3i1.610>.
- [5] Nurdin, L. Jama, T. Z. Magnus, R. Priskila, and V. H. Pranatawijaya, "Analisis Sentimen Dampak Artificial Intelligence (AI) Untuk Pendidikan Pada X Menggunakan Naïve Bayes," *JURNAL INFORMATIKA UPGRIS*, vol. 10, no. 1, p. 15, 2024, doi: <https://doi.org/10.26877/jiu.v10i1.18867>.
- [6] D. Saputra, W. Wargadinata, and S. Fikri, "Studi pustaka tentang Faktor - faktor yang mempengaruhi hasil belajar Siswa," *Jurnal Ilmiah Pendidikan Dasar*, vol. 10, no. 02, pp. 2477–2143, 2025, doi: <https://doi.org/10.23969/jp.v10i02.27026>.
- [7] Y. Yahya and A. M. Nur, "Pengaruh Penerapan Media Interaktif Berbasis Multimedia Untuk Pembelajaran Bahasa Inggris Di MTs. NW Ketangga Menggunakan Algoritma Naïve Bayes," *Infotek: Jurnal Informatika dan Teknologi*, vol. 5, no. 1, pp. 209–218, Jan. 2022, doi: [10.29408/jit.v5i1.4841](https://doi.org/10.29408/jit.v5i1.4841).
- [8] N. C. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, "The Implementation of Naïve Bayes Algorithm for Sentiment Analysis of Shopee Reviews on Google Play Store," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 02, no. 01, pp. 47–54, 2022, doi: <https://doi.org/10.57152/malcom.v2i1.195>.
- [9] I. Hidayah, G. Putu, W. Wedashwara, and A. Zubaidi, "Sistem Monitoring Kondisi Kesehatan Sebelum dan Sesudah Olahraga Menggunakan Pulse Sensor dan Sensor DS18B20 dengan Metode Naive Bayes," *J-COSINE*, vol. 06, no. 01, pp. 20–29, 2022, doi: <https://doi.org/10.29303/jcosine.v7i2.298>.
- [10] N. Samuel and A. A. Pekuwali, "Pattern Recognition of Doctor's Prescription Handwriting Using the Naïve Bayes Classifier Method at Puskesmas Kambaniru," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 02, no. 01, pp. 55–61, 2022, doi: <https://doi.org/10.57152/malcom.v2i1.174>.
- [11] W. Fadri, "Klasifikasi Penyakit Hati dengan Menggunakan Metode Naive Bayes," *Jurnal Informasi dan Teknologi*, vol. 5, no. 1, pp. 32–37, 2023, doi: [10.37034/jidt.v5i1.230](https://doi.org/10.37034/jidt.v5i1.230).
- [12] A. J. Susilo, K. K. Kustanto, and D. Remawati, "Implementasi Naïve Bayes Dalam Pemilihan Jenis Bahan Pembuatan Meja," *Jurnal Ilmiah SINUS*, vol. 21, no. 1, p. 39, Jan. 2023, doi: [10.30646/sinus.v21i1.674](https://doi.org/10.30646/sinus.v21i1.674).
- [13] P. Pandunata, S. R. Ali, and Y. Nurdiansyah, "Analisis Sentimen Program Merdeka Belajar Kampus Merdeka Menggunakan Algoritma Naive Bayes," *Jurnal Sistem Informasi Dan Bisnis Cerdas*, vol. 16, no. 1, 2023.
- [14] U. R. Gurning, S. F. Octavia, D. R. Andriyani, N. Nurainun, and I. Permana, "Prediksi Risiko Stunting pada Keluarga Menggunakan Naïve Bayes Classifier dan Chi-Square," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 172–180, Jan. 2024, doi: [10.57152/malcom.v4i1.1074](https://doi.org/10.57152/malcom.v4i1.1074).
- [15] A. Damayanti, S. Puspasari, and N. Suhandi, "Prediksi Angka Kemiskinan Desa Kemang Gejala Menggunakan Metode Naive Bayes," *RESOLUSI: Rekayasa Teknik Informatika dan Informasi*, vol. 4, no. 3, 2024, [Online]. Available: <https://djournals.com/resolusi>

- [16] Muqorobin and M. Bagoes Pakarti, “Sistem Prediksi Lama Studi Kuliah Menggunakan Metode Naive Bayes,” *Jurnal Informatika, Komputer dan Bisnis*, vol. 2, no. 1, 2022, [Online]. Available: <https://jurnal.itbaas.ac.id/index.php/jikombis>
- [17] M. B. A. Darmawan, F. Dewanta, and S. Astuti, “Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naïve Bayes untuk Prediksi Banjir di Desa Dayeuhkolot,” *TELKA*, vol. 9, no. 1, pp. 52–61, 2023, doi: <http://dx.doi.org/10.15575/telka.v9n1.52-61>.