JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 45 – 52 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4238
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

NAÏVE BAYES-BASED STUDENT ACHIEVEMENT PREDICTION SYSTEM

Fadillah Angreani[1], Heny Pratiwi[1*], Muhammad Ibnu Saad[1]
[1]Informatics Engineering, STMIK Widya Cipta Dharma
email : *2243004@wicida.ac.id

Abstract: SMP Muhammadiyah 5 Samarinda still relies on manual evaluation with limited data analysis tools in predicting student academic achievement. This study aims develop a system for predicting the learning achievement of students at SMP Muhammadiyah 5 Samarinda using the Naive Bayes classification method. The dataset used consists of 192 student exam scores covering academic scores, attendance, parents' education and income, and living conditions as independent variables, while the dependent variable is the achievement label (achieved or not achieved). The preprocessing stage includes label normalization, feature selection, and median imputation to handle missing data. The dataset was divided into 75% training data and 25%. The model was implemented as a pipeline consisting of a median imputer and a Gaussian Naive Bayes classifier. The evaluation results showed that the model achieved an accuracy of 79.2%, with a perfect recall value (1.00) in the high-achieving class and (0.64) in the low-achieving class. This shows that the model is quite effective in identifying high-achieving students. The trained model was then integrated into a Flask-based web application, which enables online predictions through a simple form interface, facilitating contextual interpretation. This system is expected to assist in educational decision-making by helping teachers identify students' achievement levels early on and design more targeted learning interventions.

Keywords: academic performance; educational data mining; naive bayes; prediction system; student achievement

Abstrak: SMP Muhammadiyah 5 Samarinda masih bergantung pada evaluasi manual dengan alat analisis data terbatas dalam melakukan prediksi prestasi akademik siswa. Penelitian ini bertujuan mengembangkan sistem prediksi prestasi belajar siswa SMP Muhammadiyah 5 Samarinda menggunakan metode klasifikasi Naive Bayes. Dataset yang digunakan terdiri atas 192 data nilai ujian siswa yang mencakup skor akademik, kehadiran, pendidikan dan pendapatan orang tua, serta kondisi tempat tinggal sebagai variabel independen, sedangkan variabel dependen berupa label prestasi (berprestasi atau tidak berprestasi). Tahap preprocessing meliputi normalisasi label, seleksi fitur, serta imputasi median untuk menangani data yang hilang. Dataset dibagi menjadi 75% data latih dan 25%. Model diimplementasikan dalam bentuk pipeline yang terdiri atas median imputer dan Gaussian Naive Bayes classifier. Hasil evaluasi menunjukkan bahwa model mencapai akurasi sebesar 79,2%, dengan nilai recall sempurna (1,00) pada kelas berprestasi dan lebih rendah (0,64) pada kelas tidak berprestasi. Hal ini menunjukkan bahwa model cukup efektif dalam mengidentifikasi siswa berprestasi. Model yang telah dilatih kemudian diintegrasikan ke dalam aplikasi web berbasis Flask, yang memungkinkan prediksi secara daring melalui antarmuka formulir sederhana untuk mendukung interpretasi kontekstual. Sistem ini diharapkan dapat membantu untuk pengambilan keputusan dalam pendidikan dengan membantu guru mengidentifikasi tingkat prestasi siswa sejak dini dan merancang intervensi pembelajaran yang lebih terarah.

Kata kunci: prestasi akademik; penambangan data Pendidikan; naive bayes; sistem prediksi; prestasi siswa

INTRODUCTION

Student academic achievement is a key indicator of the success of the learning process in educational institutions. Good learning outcomes not only reflect students' cognitive abilities but also indicate the overall effectiveness of the educational system being implemented. In the digital era, the availability of academic data such as exam scores, attendance records, and socio-economic attributes has provided new opportunities to utilize data-driven methods for performance prediction. By applying machine learning (ML) techniques, educational institutions can generate predictive insights, conduct early detection of at-risk students, and assist teachers in making data-based academic decisions [1], [2].

Various ML algorithms have been successfully applied in the field of educational data mining, including Decision Tree [1], Support Vector Machine (SVM) [2], Random Forest [13], K-Nearest Neighbor (KNN) [3], and Logistic Regression [4]. These algorithms have been proven effective in modeling student performance patterns across various educational levels. However, most prior studies have been conducted in higher education contexts with a focus on binary classifications such as pass/fail outcomes, rather than on multi-class segmentation of student achievement at the junior high school level. Furthermore, the computational complexity of several algorithms poses challenges in schools that operate under limited infrastructure and digital resources [5], [6].

In addressing these constraints, the Naïve Bayes algorithm has emerged as an efficient alternative due to its low computational cost and stable performance on small-scale datasets [7]–[10]. This algorithm assumes feature independence, allowing it to deliver robust classification results even when data samples are limited or incomplete. Several studies have also demonstrated that appropriate preprocessing and feature selection strategies can enhance Naïve Bayes accuracy in predicting student performance [5]. In addition, integrating web-based systems into predictive modeling has proven beneficial for improving data accessibility, analysis speed, and real-time visualization of academic outcomes [6], [16], [17].

Recent research has further highlighted the importance of considering socio-economic factors—such as parental education, occupation, and income—in educational data mining. These attributes significantly influence students' learning achievement and can improve the representativeness of predictive models [18]. Therefore, combining academic indicators with socio-economic variables provides a more comprehensive understanding of student performance dynamics and supports more equitable academic interventions.

Based on these studies, this research aims to develop and implement an academic achievement prediction system for SMP Muhammadiyah 5 Samarinda using the Naïve Bayes classification method. The system is designed as a web-based application that can automatically classify student performance, visualize interpretive prediction results, and apply rule-based adjustments to enhance fairness in decision-making. This study contributes to improving data-driven educational management in junior high schools and supports teachers in identifying potential academic challenges at an early stage, enabling more effective and efficient intervention strategies.

METHOD

This study employed a quantitative approach with an experimental classification design, in which the Naive Bayes Classifier algorithm was applied to predict students' academic performance. The system was developed as a web-based application capable of receiving student data, performing automatic classification, and displaying prediction results along with explanatory interpretations.

The population of this study consisted of all students of SMP Muhammadiyah 5 Samarinda in the 2024/2025 academic year, totaling 192 students. The sampling technique used was Simple Random Sampling, ensuring that each member of the population had an equal chance of being selected. From the total population, all data were utilized as the research dataset. The dataset was divided into two parts using the train_test_split function with a ratio of 75% training data (142 students) and 25% testing data (48 students), while applying the stratify parameter to maintain class distribution balance.

The research data were obtained from two main sources. First, student academic records, including exam scores in core subjects, attendance percentage, parental education background, parental income, and residential status. These data were collected from the school archives. Second, interviews were conducted with students and the vice principal for curriculum affairs to validate relevant non-academic factors, particularly related to attendance patterns, family background, and students' social conditions. The results of these interviews provided additional insights and justification for selecting the research variables.

The independent variables (X) in this study include the average exam score, attendance percentage, father's education level, mother's education level, father's income, mother's income, and living conditions. These variables were selected because they have a significant influence on students' academic achievement, as emphasized in previous studies [10]. The dependent variable (Y) represents the students' achievement label, which is determined based on a threshold: an average score ≥ 75 is categorized as Achieving (1), while a score < 75 is categorized as Not Achieving (0).

This study was conducted following the Naive Bayes Classification Flow as illustrated in the research methodology shown in Figure 1.

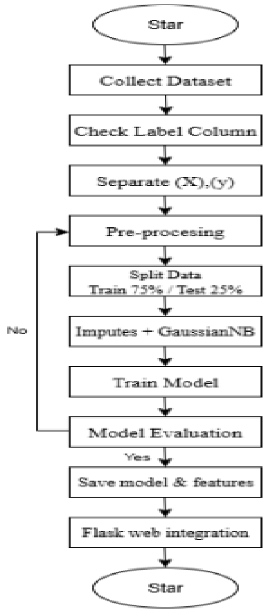


Figure 1. Naïve Bayes Classification Flow

The classification process using Naive Bayes followed a series of stages implemented in the Python scripts train_nb.py and app.py. These stages included: data preprocessing, which involved median imputation to handle missing values; splitting the dataset into training and testing subsets; model training using the Gaussian Naive Bayes algorithm; probability calculation for each class; and rule-based adjustments, such as reclassifying students with high grades but very low attendance as Not Achieving. The prediction output—comprising the classification results and explanatory information—was displayed through the web interface.

$$P(C_k|X)\frac{P(X|C_k).P(C_k)}{P(X)} \qquad (1)$$

Since $P(X)$ is constant for all classes, the classification rule can be simplified into:

$$\hat{y}\frac{argmax}{C_k}P(C_k)\prod_{i=1}^{n}P(x_i|C_k) \qquad (2)$$

Where:
$P(C_k|X)$: Posterior probability of class C_k given the input features X
$P(C_k)$: Prior probability of class C_k
$P(x_i|C_k)$: Conditional probability of feature x_i given class C_k
$\hat{y}$: Predicted class label

Because the dataset used in this study contains continuous numerical data, the Gaussian Naïve Bayes distribution was applied. The probability density function for each feature is calculated as:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{C_k}^2}}\exp\left(-\frac{(x_i-\mu C_k)^2}{2\sigma_{C_k}^2}\right) \qquad (3)$$

Where:
μC_k : Mean value of feature x_i for class C_k
σC_k : Standard deviation of feature x_i for class C_k

Model evaluation was conducted using a confusion matrix along with accuracy, precision, recall, and F1-score metrics. The confusion matrix was used to assess the number of students correctly and incorrectly classified in each category, with indicators including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The initial testing results showed that the Naive Bayes model achieved an accuracy of approximately 0.80–0.90 with a relatively balanced distribution of TP and TN values. These findings indicate that the Naive Bayes algorithm is effective for classifying student academic performance and is suitable for implementation in a web-based prediction system.

RESULTS AND DISCUSSION

This study utilized a dataset consisting of 192 students from SMP Muhammadiyah 5 Samarinda. The dataset included seven independent variables: average exam score, attendance percentage, father's education level, mother's education level, father's income, mother's income, and living conditions. The dependent variable represented the student achievement label (0 = Not Achieving, 1 = Achieving).

Before training, the data underwent a preprocessing stage that included median imputation to handle missing values. The dataset was then divided into 75% training data (142 students) and 25% testing data (48 students) using the stratified sampling method to maintain balanced class distribution. The model was trained using the Gaussian Naive Bayes algorithm, which is well-suited for small to medium-scale datasets due to its computational efficiency.

The evaluation results on the test

data showed that the model achieved an accuracy of 79.2%. The classification metrics are summarized in Table 2. As shown in Table 2, the model demonstrated excellent performance in identifying high-achieving students (recall = 1.00) but showed lower sensitivity in detecting underachieving students (recall = 0.64). This indicates that some students who should have been classified as Not Achieving were incorrectly classified as Achieving.

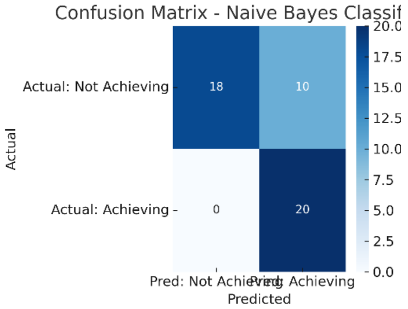


Figure 2. Confusion Matrix – Naïve Bayes Classification

The confusion matrix shows that 18 Not Achieving students and 20 Achieving students were correctly classified, while 10 Not Achieving students were misclassified as Achieving. It is important to note that there were no false negatives, indicating that all Achieving students were accurately identified by the model.

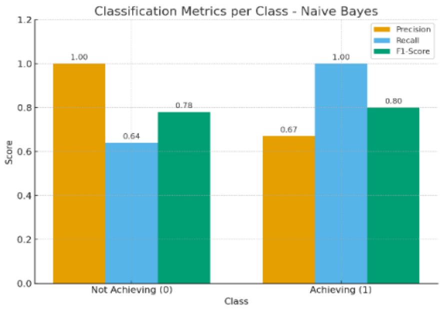
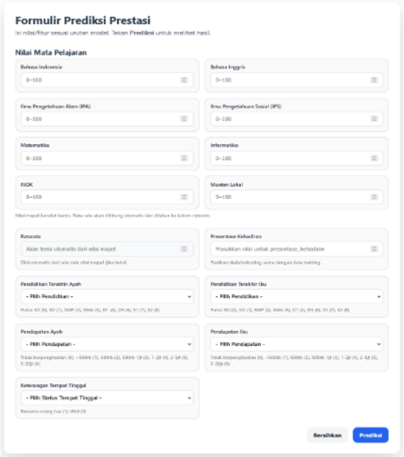


Figure 3. Classification Metrics per Class

After the training process, the model was integrated into a Flask-based web application. The application interface allows teachers to input student data, calculate average scores, and classify academic achievement using the Naive Bayes model. The prediction results are then displayed on the results page (result.html) in the form of student achievement status, average score, and explanatory narratives that assist in interpretation.

The research findings indicate that the Naive Bayes algorithm is effective for classifying academic performance in small-scale datasets.



Formulir Prediksi Prestasi

Isi nilai/fitur sesuai urutan model. Tekan **Prediksi** untuk melihat hasil.

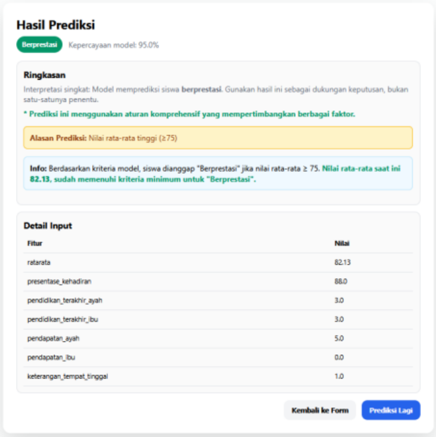
Nilai Mata Pelajaran

Bahasa Indonesia: 0–100	Bahasa Inggris: 0–100
Ilmu Pengetahuan Alam (IPA): 0–100	Ilmu Pengetahuan Sosial (IPS): 0–100
Matematika: 0–100	Informatika: 0–100
PJOK: 0–100	Muatan Lokal: 0–100

Nilai mapel bersifat bantu. Rata-rata akan dihitung otomatis dan ditulis ke kolom ratarata.

Ratarata: Akan terisi otomatis dari nilai mapel Diisi otomatis dari rata-rata nilai mapel (jika terisi).	Presentase Kehadiran: Masukkan nilai untuk presentase_kehadiran Pastikan skala/enkoding sama dengan data training.
Pendidikan Terakhir Ayah: - Pilih Pendidikan - Putus SD (0), SD (1), SMP (2), SMA (3), D1 (4), D4 (6), S1 (7), S2 (8)	Pendidikan Terakhir Ibu: - Pilih Pendidikan - Putus SD (0), SD (1), SMP (2), SMA (3), D1 (4), D4 (6), S1 (7), S2 (8)
Pendapatan Ayah: - Pilih Pendapatan - Tidak berpenghasilan (0), <500rb (1), 500rb (2), 500rb-1jt (3), 1-2jt (4), 2-5jt (5), 5-20jt (6)	Pendapatan Ibu: - Pilih Pendapatan - Tidak berpenghasilan (0), <500rb (1), 500rb (2), 500rb-1jt (3), 1-2jt (4), 2-5jt (5), 5-20jt (6)
Keterangan Tempat Tinggal: - Pilih Status Tempat Tinggal - Bersama orang tua (1), Wali (0)	

Figure 4. Web-based prediction from interface



Hasil Prediksi

Berprestasi Kepercayaan model: 95.0%

Ringkasan

Interpretasi singkat: Model memprediksi siswa **berprestasi**. Gunakan hasil ini sebagai dukungan keputusan, bukan satu-satunya penentu.

*** Prediksi ini menggunakan aturan komprehensif yang mempertimbangkan berbagai faktor.**

Alasan Prediksi: Nilai rata-rata tinggi (≥75)

Info: Berdasarkan kriteria model, siswa dianggap "Berprestasi" jika nilai rata-rata ≥ 75. **Nilai rata-rata saat ini 82.13, sudah memenuhi kriteria minimum untuk "Berprestasi".**

Detail Input

Fitur	Nilai
ratarata	82.13
presentase_kehadiran	88.0
pendidikan_terakhir_ayah	3.0
pendidikan_terakhir_ibu	3.0
pendapatan_ayah	5.0
pendapatan_ibu	0.0
keterangan_tempat_tinggal	1.0

Figure 5. Prediction result page of the web system

To enhance interpretability, the system is equipped with a rule-based adjustment mechanism. For example, students with high average scores (≥ 75) but attendance rates below 50% are classified as Not Achieving, even if the model predicts them as Achieving.

CONCLUSION

This study successfully developed a student academic achievement prediction system for SMP Muhammadiyah 5 Samarinda based on the Naive Bayes algorithm, achieving an accuracy of 79.2%. The model demonstrated strong capability in identifying high-achieving students (recall = 1.00) but requires improvement in detecting underachieving students (recall = 0.64). The system, implemented using Flask, can automatically classify student performance and provide interpretive narratives that help teachers understand prediction results. The addition of a rule-based adjustment mechanism enhances fairness and clarity in classification outcomes. Overall, the Naive Bayes method proved efficient and well-suited for school environments with limited resources, showing potential to support data-driven learning decisions and early detection of students requiring special academic attention.

Table 1. Research Variables

Variable Type	Variable Name	Description
Independent	average	Student average exam score
Independent	attendance percentage	Student attendance percentage (%)
Independent	father's last education	Father’s highest education level (numerically encoded)
Independent	mother's last education	Mother’s highest education level (numerically encoded)
Independent	father's income	Father’s income level (numerically encoded)
Independent	mother's income	Mother’s income level (numerically encoded)
Independent	residence_description	Living condition (e.g., with parents, rented house, etc.)
Dependent	high-achieving label	0 = Underachieving, 1 = Achieving

Table 2. Model Evaluation Metric

Class	Precision	Recall	F1-Score
Not Achieving (0)	1.00	0.64	0.78
Achieving (1)	0.67	1.00	0.80

BIBLIOGRAPHY

[1] Wati, A., et al., “Application of Decision Tree Algorithm for Student Graduation Prediction,” Journal of Educational Data Mining, vol. 13, no. 1, pp. 45–53, 2021.

[2] Susanti, D., and Hidayat, T., “Predicting Student Academic Performance Using Support Vector Machine,” International Journal of Computer Applications, vol. 182, no. 24, pp. 30–35, 2020.

[3] Handayani, S., “Student Achievement Prediction Using K-Nearest Neighbor Algorithm,” Indonesian Journal of Artificial Intelligence and Data Mining, vol. 4, no. 2, pp. 67–74, 2021.

[4] Ningsih, R., and Pratama, A., “Logistic Regression Approach for Academic Performance Classification,” Journal of Information Systems Education and Research, vol. 10, no. 3, pp. 112–119, 2022.

[5] Sari, F., et al., “Feature Selection and Preprocessing to Improve Naive Bayes Performance in Student Achievement Prediction,” Journal of Physics: Conference Series, vol. 1567, no. 022045, pp. 1–7, 2020.

[6] Arifin, M., and Nuraini, A., “Web-Based Prediction System for Student Academic Achievement Using Naive Bayes,” International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 8, pp. 125–137, 2021.

[7] H. Pratiwi, M. I. Sa’ad, and Salmon, “Strategi Manajemen Pendidikan Berbasis Machine Learning untuk Prediksi Prestasi Siswa,” *BEduManageRs Journal: Borneo Educational Management and Research Journal*, vol. 6, no. 1, pp. 21–30, Jun. 2025.

[8] H. Pratiwi, M. I. Sa’ad, and M. A. Zakaria, “Sistem Pakar Berbasis Web Untuk Diagnosis Penanganan Pasca Panen Kepala Sawit Menggunakan Metode Nive Bayes,” *TAMIKA : Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akutansi*, vol. 4, no. 2, pp. 259-267, Des. 2024.

[9] P. Zhang dan Q. Yang, “Naive Bayes untuk Prediksi Prestasi Siswa,” *International Journal of Data Mining in Education*, vol. 8, no. 2, hlm. 101–110, 2020.

[10] M. A. Khan, S. Hussain, dan R. Ahmad, “Penerapan Bayesian Classifiers untuk Analitik Akademik,” *Education and Information Technologies*, vol. 25, no. 5, hlm. 3921–3938, 2020.

[11] S. Sharma dan P. Gupta, “Teknik Decision Tree untuk Memprediksi Hasil Pendidikan,” *International Journal of Computer Applications*, vol. 178, no. 4, hlm. 15–22, 2019.

[12] L. Wang dan H. Chen, “Pendekatan Support Vector Machine untuk Memprediksi Keberhasilan Siswa dalam E-learning,” *Procedia Computer Science*, vol. 174, hlm. 655–664, 2020.

[13] A. Rahman dan T. Setiawan, “Random Forest untuk Prediksi Prestasi Akademik di Perguruan Tinggi,” *Journal of Computer Science and Applications*, vol. 11, no. 3, hlm. 89–96, 2021.

[14] D. H. Nugroho, “Implementasi Algoritma K-Nearest Neighbor dalam Prediksi Prestasi Siswa,” *Indonesian Journal of Artificial Intelligence*, vol. 6, no. 2, hlm. 73–81, 2021.

[15] F. Li dan Y. Zhao, "Keterbatasan Logistic Regression dalam Educational Data Mining," *Applied Artificial Intelligence*, vol. 35, no. 9, hlm. 731–742, 2021.

[16] R. Kumar, M. Saini, dan A. Singh, "Desain Alat Data Mining Pendidikan Berbasis Web Menggunakan Naive Bayes," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 12, hlm. 102–114, 2020.

[17] H. S. Lee, "Sistem Web Ramah Pengguna untuk Memprediksi Risiko Akademik," *Computers & Education*, vol. 164, no. 104118, hlm. 1–12, 2021.

[18] N. A. Siregar dan F. Ramdhani, "Dampak Atribut Sosial-Ekonomi terhadap Prediksi Prestasi Siswa," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 3, hlm. 221–229, 2021.