e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

PERBANDINGAN KINERJA *NAÏVE BAYES* DAN *RANDOM FOREST* DALAM PREDIKSI PERFORMA AKADEMIK SISWA

Marcelino Alberki Kabuhung[1], Exsel Zeth Lopulalan[2], Michael Yosia Theofilus Sabandar[3], Arif Handika[4]

Program Studi Teknik Informatika[1,2,3,4],Fakultas Teknik[1,2,3,4] Universitas Papua[1,2,3,4]

*Correspondent Author: kabuhungm@gmail.com

Authors Email: kabuhungm@gmail.com[1], exsellopulalan13@gmail.com[2], michaelsabandar02@gmail.com[3], arifhandika75@gmail.com[4]

In Indonesian

Abstrak: Prediksi performa akademik siswa menjadi salah satu penerapan penting *machine learning* dalam bidang pendidikan. Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Naïve Bayes* dan *Random Forest* dalam prediksi performa akademik siswa menggunakan pendekatan data mining. Dataset yang digunakan adalah *Student Performance Dataset* dari *UCI Machine Learning Repository* yang berisi data akademik, sosial, perilaku, dan demografis siswa. Metode penelitian menggunakan kerangka kerja CRISP-DM yang meliputi *data understanding*, *data preparation*, *modeling*, dan *evaluation*. Tahap *preprocessing* dilakukan melalui *encoding* variabel kategorikal, penghapusan variabel G1 dan G2 untuk menghindari *data leakage*, serta transformasi variabel target menjadi klasifikasi biner. Dataset dibagi menjadi 80% *data training* dan 20% *data testing*. Evaluasi model dilakukan menggunakan *accuracy*, *precision*, *recall*, *f1-score*, dan *confusion matrix*. Hasil penelitian menunjukkan bahwa *Random Forest* memperoleh *accuracy* sebesar 72,15%, lebih tinggi dibandingkan *Naïve Bayes* sebesar 70,88%. Analisis *feature importance* menunjukkan bahwa *absences*, *failures*, *age*, dan *goout* menjadi faktor yang paling berpengaruh terhadap performa akademik siswa. Penelitian ini menunjukkan bahwa *Random Forest* memiliki performa yang lebih baik dalam klasifikasi performa akademik siswa dan mampu memberikan interpretasi terhadap faktor-faktor yang memengaruhi hasil prediksi.

Kata kunci: Data Mining; *Machine Learning*; *Naïve Bayes*; *Random Forest*; Klasifikasi

In English

Abstract: *Predicting student academic performance has become one of the important applications of machine learning in the field of education. This study aims to compare the performance of Naïve Bayes and Random Forest algorithms in predicting student academic performance using a data mining approach. The dataset used in this research is the Student Performance Dataset obtained from the UCI Machine Learning Repository, which contains academic, social, behavioral, and demographic student data. The research method applies the CRISP-DM framework consisting of data understanding, data preparation, modeling, and evaluation stages. The preprocessing stage includes categorical variable encoding, removal of G1 and G2 variables to avoid data leakage, and transformation of the target variable into binary classification. The dataset was divided into 80% training data and 20% testing data. Model evaluation was conducted using accuracy, precision, recall, f1-score, and confusion matrix. The results show that Random Forest achieved an accuracy of 72.15%, outperforming Naïve Bayes with 70.88%. Feature importance analysis indicates that absences, failures, age, and goout are the most influential factors affecting student academic performance. This study concludes that Random Forest provides better performance in classifying student academic performance and is capable of providing interpretation of factors influencing prediction results.*

Keywords: *Data Mining; Machine Learning; Naïve Bayes; Random Forest; Classification*

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

DOI: 10.52362/jisicom.v10i1.2433

Received: *2026-04-20.* ***Revised:*** *2026-05-28.* ***Accepted:*** *2026-06-01* ***Issue Period:*** *Vol.10 No.1 (2026), Pp.234-243*

I. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pemanfaatan data sebagai dasar dalam proses pengambilan keputusan di berbagai bidang, termasuk pendidikan. Aktivitas akademik siswa menghasilkan data yang cukup beragam, seperti nilai, kehadiran, latar belakang keluarga, kebiasaan belajar, serta faktor sosial dan demografis. Data tersebut tidak hanya berfungsi sebagai alat administratif, tetapi juga dapat dianalisis untuk memperoleh informasi yang mendukung peningkatan kualitas pembelajaran dan pengambilan keputusan berbasis data [1].

Salah satu pendekatan yang banyak digunakan dalam pemanfaatan data pendidikan adalah *Educational Data Mining* (EDM). EDM merupakan penerapan teknik data mining untuk menemukan pola, hubungan, dan pengetahuan dari data pendidikan. Melalui pendekatan ini, institusi pendidikan dapat menganalisis karakteristik siswa, memprediksi capaian akademik, serta mengidentifikasi faktor-faktor yang berhubungan dengan keberhasilan atau penurunan performa belajar siswa [1],[2]. Dengan demikian, penerapan data mining dalam bidang pendidikan dapat menjadi alat bantu untuk memahami kondisi siswa secara lebih objektif.

Prediksi performa akademik siswa menjadi salah satu topik penting dalam EDM karena dapat membantu pihak sekolah atau institusi pendidikan mengidentifikasi siswa yang berpotensi mengalami kesulitan belajar sejak dini. Hasil prediksi tersebut dapat digunakan sebagai dasar dalam menyusun strategi pembelajaran, memberikan bimbingan akademik, serta melakukan intervensi yang lebih tepat sasaran. Beberapa penelitian sebelumnya menunjukkan bahwa performa akademik siswa dapat dipengaruhi oleh berbagai faktor, seperti nilai akademik sebelumnya, tingkat kehadiran, waktu belajar, riwayat kegagalan, dukungan keluarga, serta perilaku sosial siswa [1],[3].

Dalam penelitian prediksi performa siswa, berbagai algoritma *machine learning* telah digunakan, seperti *Logistic Regression*, *Decision Tree*, *Support Vector Machine*, *Naïve Bayes*, dan *Random Forest* [4],[5]. *Naïve Bayes* merupakan algoritma klasifikasi berbasis probabilitas yang dikenal sederhana, cepat, dan efisien dalam proses pemodelan. Algoritma ini sering digunakan sebagai model pembanding karena memiliki proses komputasi yang ringan serta mampu memberikan hasil klasifikasi yang cukup baik pada beberapa kasus prediksi performa siswa [6],[7]. Sementara itu, *Random Forest* merupakan algoritma *ensemble learning* yang membangun beberapa *decision tree* dan menggabungkan hasil prediksinya untuk memperoleh keputusan akhir. Algoritma ini sering digunakan karena mampu menangani data yang kompleks, mengurangi risiko *overfitting*, serta menyediakan informasi mengenai tingkat kepentingan fitur atau *feature importance* [8],[9].

Beberapa penelitian terdahulu menunjukkan bahwa *Random Forest* cenderung menghasilkan performa yang lebih baik dibandingkan algoritma sederhana seperti *Naïve Bayes* dalam beberapa kasus prediksi performa akademik siswa [8],[10]. Namun, *Naïve Bayes* tetap relevan digunakan karena dapat menjadi *baseline model* yang sederhana dan mudah diinterpretasikan. Perbandingan antara kedua algoritma tersebut penting dilakukan untuk mengetahui sejauh mana perbedaan pendekatan probabilistik dan *ensemble learning* dalam menghasilkan prediksi performa akademik siswa.

Dataset yang digunakan dalam penelitian ini adalah *Student Performance Dataset* dari *UCI Machine Learning Repository* yang diperkenalkan oleh Paulo Cortez dan Alice Silva. Dataset tersebut berisi data siswa dari dua sekolah menengah di Portugal dengan atribut akademik, sosial, perilaku, dan demografis. Pada penelitian asli, performa siswa dimodelkan dalam beberapa skenario, yaitu klasifikasi biner, klasifikasi lima tingkat, dan regresi. Penelitian tersebut juga menunjukkan bahwa performa akademik siswa tidak hanya dipengaruhi oleh nilai sebelumnya, tetapi juga oleh faktor lain seperti jumlah ketidakhadiran, riwayat kegagalan akademik, pekerjaan dan pendidikan orang tua, serta faktor sosial siswa [11].

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

Meskipun penelitian mengenai prediksi performa siswa telah banyak dilakukan, sebagian besar penelitian masih berfokus pada perbandingan nilai akurasi model. Padahal, dalam konteks pendidikan, hasil prediksi tidak hanya perlu dilihat dari nilai *accuracy*, tetapi juga dari kemampuan model dalam mengenali siswa dengan performa rendah serta faktor-faktor yang berpengaruh terhadap hasil prediksi. Oleh karena itu, penelitian ini tidak hanya membandingkan kinerja algoritma *Naïve Bayes* dan *Random Forest*, tetapi juga menganalisis faktor-faktor yang paling berpengaruh terhadap performa akademik siswa melalui *feature importance* pada model *Random Forest*.

Berdasarkan uraian tersebut, penelitian ini bertujuan membandingkan kinerja algoritma *Naïve Bayes* dan *Random Forest* dalam memprediksi performa akademik siswa menggunakan pendekatan data mining. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi faktor-faktor yang berpengaruh terhadap hasil prediksi performa siswa. Hasil penelitian diharapkan dapat memberikan gambaran mengenai algoritma yang lebih sesuai digunakan dalam klasifikasi performa akademik siswa serta mendukung pengambilan keputusan berbasis data dalam bidang pendidikan.

II. METODE DAN MATERI

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining untuk membandingkan kinerja algoritma *Naïve Bayes* dan *Random Forest* dalam memprediksi performa akademik siswa. Proses penelitian disusun berdasarkan tahapan *Cross Industry Standard Process for Data Mining* (CRISP-DM), yang meliputi pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi, dan penyajian hasil. Kerangka ini digunakan karena mampu memberikan alur kerja yang sistematis dalam proses pengolahan data hingga evaluasi model prediksi [1],[2].

2.1. Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah *Student Performance Dataset* dari *UCI Machine Learning Repository* yang diperkenalkan oleh Cortez dan Silva [11]. Dataset tersebut berisi data siswa dari dua sekolah menengah di Portugal dan mencakup atribut akademik, sosial, perilaku, dan demografis. Penelitian ini menggunakan data *student-mat.csv* yang berisi 395 data siswa pada mata pelajaran Matematika.

Atribut dalam dataset meliputi beberapa kelompok variabel, seperti variabel akademik, sosial, perilaku, keluarga, dan demografis. Variabel akademik mencakup *studytime*, *failures*, G1, G2, dan G3. Variabel sosial dan perilaku mencakup *goout*, *freetime*, *Dalc*, *Walc*, dan *absences*. Sementara itu, variabel demografis dan keluarga mencakup *age*, *sex*, *Medu*, *Fedu*, *Mjob*, dan *Fjob*. Dalam penelitian ini, variabel G3 digunakan sebagai target karena merepresentasikan nilai akhir siswa [11].

2.2. Tahapan Pengolahan Data

Tahap pengolahan data dilakukan untuk menyiapkan dataset agar dapat digunakan dalam proses pemodelan *machine learning*. Pengolahan data dimulai dengan pemeriksaan struktur data, pengecekan nilai kosong, transformasi variabel kategorikal, seleksi fitur, dan pembentukan label target.

Pertama, pemeriksaan awal dilakukan untuk mengetahui jumlah data, tipe atribut, serta keberadaan missing value. Berdasarkan hasil pemeriksaan, dataset memiliki 395 data dengan 33 atribut dan tidak ditemukan nilai kosong pada setiap kolom. Kondisi ini membuat data dapat langsung diproses ke tahap transformasi tanpa perlu penanganan *missing value* tambahan.

Kedua, variabel kategorikal diubah menjadi bentuk numerik menggunakan teknik *encoding*. Tahap ini diperlukan karena algoritma *machine learning* pada umumnya memerlukan input dalam bentuk numerik agar dapat melakukan proses pelatihan dan prediksi [5],[6].

Ketiga, variabel G1 dan G2 dihapus dari proses pemodelan. Kedua variabel tersebut merupakan nilai periode sebelumnya yang memiliki hubungan sangat kuat dengan G3 sebagai nilai akhir. Jika tetap digunakan, model berpotensi mengalami data leakage karena memperoleh informasi yang terlalu dekat dengan target. Oleh karena itu, penghapusan G1 dan G2 dilakukan agar model tidak hanya bergantung pada nilai akademik sebelumnya, tetapi juga dapat mempelajari faktor lain seperti kehadiran, riwayat kegagalan, waktu belajar, dan faktor sosial siswa [3],[11].

Keempat, variabel G3 dikonversi menjadi kelas biner. Nilai G3 $\geq$ 10 dikategorikan sebagai performa baik, sedangkan nilai G3 < 10 dikategorikan sebagai performa rendah. Pembentukan label biner ini mengikuti

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

pendekatan klasifikasi *pass/fail* yang juga digunakan dalam penelitian terkait *Student Performance Dataset* [3],[11]. Setelah proses pelabelan selesai, data dibagi menjadi 80% data latih dan 20% data uji. Pembagian ini bertujuan agar model dapat dilatih menggunakan sebagian data dan diuji pada data yang belum pernah digunakan dalam proses pelatihan.

2.3. Algoritma Klasifikasi

Penelitian ini menggunakan dua algoritma klasifikasi, yaitu *Naïve Bayes* dan *Random Forest*. Kedua algoritma dipilih karena memiliki karakteristik pendekatan yang berbeda, sehingga dapat memberikan perbandingan yang lebih jelas dalam proses prediksi performa akademik siswa.

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menerapkan *Teorema Bayes* dengan asumsi bahwa setiap fitur bersifat independen terhadap fitur yang lainnya. Algoritma ini dikenal sederhana, cepat, dan efisien, sehingga sering digunakan sebagai model dasar dalam penelitian klasifikasi performa siswa [6],[7]. Dalam penelitian ini, *Naïve Bayes* digunakan sebagai algoritma pembanding untuk melihat kemampuan model probabilistik dalam mengklasifikasikan performa akademik siswa.

Random Forest merupakan algoritma *ensemble learning* yang membangun sejumlah *decision tree* dan menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan keputusan akhir. Algoritma ini memiliki kemampuan dalam menangani hubungan data yang kompleks, mengurangi risiko *overfitting*, serta menyediakan informasi mengenai kontribusi setiap fitur melalui *feature importance* [8],[9]. Pada penelitian ini, *Random Forest* digunakan untuk membandingkan performa klasifikasi dengan *Naïve Bayes* sekaligus menganalisis faktor-faktor yang paling berpengaruh terhadap hasil prediksi.

2.4. Evaluasi Model

Evaluasi model dilakukan untuk mengetahui kinerja algoritma dalam mengklasifikasikan performa akademik siswa. Metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *recall*, *f1-score*, dan *confusion matrix*. *Accuracy* digunakan untuk mengukur tingkat ketepatan prediksi secara keseluruhan. *Precision* digunakan untuk melihat ketepatan model dalam memprediksi suatu kelas, sedangkan *recall* digunakan untuk mengetahui kemampuan model dalam mengenali data pada kelas tertentu. *F1-score* digunakan sebagai ukuran keseimbangan antara *precision* dan *recall*, sementara *confusion matrix* digunakan untuk melihat jumlah prediksi benar dan salah pada masing-masing kelas.

Selain evaluasi berdasarkan metrik klasifikasi, penelitian ini juga melakukan analisis *feature importance* pada model *Random Forest*. Analisis ini digunakan untuk mengetahui fitur-fitur yang memiliki kontribusi paling besar terhadap hasil prediksi. Dengan demikian, hasil penelitian tidak hanya menunjukkan algoritma yang memiliki performa lebih baik, tetapi juga memberikan informasi mengenai faktor-faktor yang berpengaruh terhadap performa akademik siswa.

III. PEMBAHASAN DAN HASIL

3.1. Deskripsi Dataset dan Persiapan Data

Penelitian ini menggunakan *Student Performance Dataset* pada mata pelajaran Matematika yang terdiri dari 395 data siswa dengan 33 atribut. Dataset ini memuat informasi akademik, sosial, perilaku, keluarga, dan demografis siswa. Beberapa atribut yang digunakan dalam penelitian ini antara lain *age*, *sex*, *studytime*, *failures*, *absences*, *goout*, *health*, *Medu*, *Fedu*, *Mjob*, *Fjob*, serta atribut nilai akademik seperti G1, G2, dan G3. Variabel G3 digunakan sebagai target karena merepresentasikan nilai akhir siswa [11].

Berdasarkan pemeriksaan awal, dataset tidak memiliki nilai kosong sehingga tidak diperlukan proses penanganan *missing value*. Namun, karena dataset terdiri dari atribut numerik dan kategorikal, atribut kategorikal terlebih dahulu diubah menjadi bentuk numerik melalui proses *encoding*. Tahapan ini dilakukan agar seluruh atribut dapat diproses oleh algoritma *machine learning*.

Pada tahap seleksi fitur, atribut G1 dan G2 dihapus dari proses pemodelan. Kedua atribut tersebut merupakan nilai periode sebelumnya yang memiliki hubungan sangat dekat dengan G3 sebagai nilai akhir. Jika

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

tetap digunakan, model berpotensi menghasilkan performa prediksi yang terlalu tinggi karena memperoleh informasi yang sangat dekat dengan target. Kondisi tersebut dapat menyebabkan *data leakage* dan menurunkan kemampuan generalisasi model. Oleh karena itu, penghapusan G1 dan G2 dilakukan agar model dapat mempelajari pola dari faktor lain, seperti absensi, riwayat kegagalan akademik, waktu belajar, kondisi sosial, dan latar belakang keluarga siswa [3],[11].

Selanjutnya, nilai G3 dikonversi menjadi dua kelas. Siswa dengan nilai G3 ≥ 10 dikategorikan sebagai siswa dengan performa baik, sedangkan siswa dengan nilai G3 < 10 dikategorikan sebagai siswa dengan performa rendah. Pembentukan label biner ini mengikuti pendekatan klasifikasi *pass/fail* yang juga digunakan dalam penelitian terkait *Student Performance Dataset* [11]. Setelah proses pelabelan selesai, dataset dibagi menjadi 80% data latih dan 20% data uji.

3.2. Hasil Perbandingan Kinerja Model

Setelah proses preprocessing selesai, data digunakan untuk membangun model klasifikasi menggunakan algoritma *Naïve Bayes* dan *Random Forest*. Kedua model kemudian diuji menggunakan data uji untuk mengetahui kinerjanya dalam memprediksi performa akademik siswa. Hasil pengujian berdasarkan nilai *accuracy* ditunjukkan pada Tabel 1.

Tabel 1. Hasil *Accuracy* Model

Algoritma	***Accuracy***
Naïve Bayes	70,88%
Random Forest	72,15%

Berdasarkan Tabel 1, *Random Forest* memperoleh nilai *accuracy* sebesar 72,15%, sedangkan *Naïve Bayes* memperoleh nilai *accuracy* sebesar 70,88%. Hasil tersebut menunjukkan bahwa *Random Forest* memiliki kinerja yang sedikit lebih baik dibandingkan *Naïve Bayes* dalam memprediksi performa akademik siswa. Selisih *accuracy* antara kedua model sebesar 1,27%, sehingga dapat dikatakan bahwa *Random Forest* unggul, tetapi perbedaan performanya tidak terlalu besar.

Keunggulan *Random Forest* dapat dipengaruhi oleh karakteristiknya sebagai algoritma *ensemble learning* yang membangun banyak *decision tree* dan menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan keputusan akhir yang lebih stabil. Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa *Random Forest* sering memberikan performa lebih baik dalam kasus prediksi performa siswa dibandingkan beberapa model sederhana lainnya [8],[10]. Meskipun demikian, *Naïve Bayes* tetap menunjukkan hasil yang kompetitif karena mampu menghasilkan *accuracy* di atas 70% dengan pendekatan probabilistik yang lebih sederhana [6],[7].

3.3. Analisis *Confusion Matrix*

Selain menggunakan *accuracy*, evaluasi model juga dilakukan menggunakan *confusion matrix* untuk melihat kesalahan dan keberhasilan prediksi pada masing-masing kelas. Pada penelitian ini, kelas 0 merepresentasikan siswa dengan performa rendah, sedangkan kelas 1 merepresentasikan siswa dengan performa baik.

Hasil *confusion matrix* pada model *Naïve Bayes* ditunjukkan pada Tabel 2.

Tabel 2. *Confusion Matrix Naïve Bayes*

Kelas Aktual	**Prediksi Rendah**	**Prediksi Baik**
Rendah	9	18
Baik	5	47

Berdasarkan Tabel 2, model *Naïve Bayes* berhasil memprediksi 9 siswa berperforma rendah dengan benar, tetapi masih terdapat 18 siswa berperforma rendah yang diprediksi sebagai siswa berperforma baik. Selain itu, terdapat 47 siswa berperforma baik yang berhasil diprediksi benar, sedangkan 5 siswa berperforma baik diprediksi sebagai siswa berperforma rendah.

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

Hasil *confusion matrix* pada model *Random Forest* ditunjukkan pada Tabel 3.

Tabel 3. *Confusion Matrix Random Forest*

Kelas Aktual	Prediksi Rendah	Prediksi Baik
Rendah	8	19
Baik	3	49

Berdasarkan Tabel 3, model *Random Forest* berhasil memprediksi 8 siswa berperforma rendah dengan benar, tetapi masih terdapat 19 siswa berperforma rendah yang diprediksi sebagai siswa berperforma baik. Sementara itu, model berhasil memprediksi 49 siswa berperforma baik dengan benar dan hanya 3 siswa berperforma baik yang diprediksi sebagai siswa berperforma rendah.

Hasil ini menunjukkan bahwa kedua model lebih baik dalam mengenali siswa dengan performa baik dibandingkan siswa dengan performa rendah. Kondisi tersebut terlihat dari jumlah prediksi benar pada kelas baik yang lebih tinggi dibandingkan kelas rendah. Hal ini kemungkinan dipengaruhi oleh distribusi kelas yang tidak seimbang, karena jumlah siswa berperforma baik lebih banyak dibandingkan siswa berperforma rendah. Ketidakseimbangan kelas seperti ini dapat membuat model cenderung lebih mudah mengenali kelas mayoritas dibandingkan kelas minoritas [1].

3.4. Analisis *Classification Report*

Untuk memperoleh evaluasi yang lebih rinci, penelitian juga menggunakan *precision*, *recall*, dan *f1-score*. Hasil *classification report* dari kedua model ditunjukkan pada Tabel 4 dan Tabel 5

Tabel 4. *Classification Report Naïve Bayes*

Kelas	*Precision*	*Recall*	*F1-Score*
Rendah	0,64	0,33	0,44
Baik	0,72	0,90	0,80

Tabel 5. *Classification Report Random Forest*

Kelas	*Precision*	*Recall*	*F1-Score*
Rendah	0,73	0,30	0,42
Baik	0,72	0,94	0,82

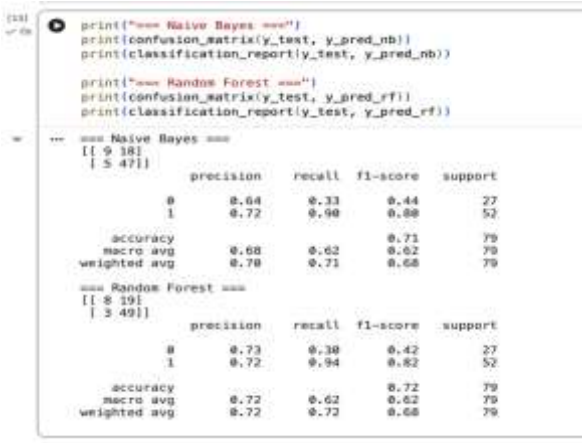
Berdasarkan Tabel 4 dan Tabel 5, kedua model menunjukkan nilai *recall* yang tinggi pada kelas baik, tetapi rendah pada kelas rendah. Pada model *Naïve Bayes*, nilai *recall* kelas rendah adalah 0,33, sedangkan pada *Random Forest* sebesar 0,30. Sebaliknya, nilai *recall* kelas baik pada *Naïve Bayes* mencapai 0,90 dan pada *Random Forest* mencapai 0,94. Hasil ini menunjukkan bahwa kedua model lebih mampu mengenali siswa dengan performa baik dibandingkan siswa dengan performa rendah.

Dari sisi *f1-score*, *Random Forest* memperoleh nilai 0,82 pada kelas baik, lebih tinggi dibandingkan *Naïve Bayes* sebesar 0,80. Namun, pada kelas rendah, *Naïve Bayes* memperoleh *f1-score* sebesar 0,44, sedikit lebih tinggi dibandingkan *Random Forest* sebesar 0,42. Dengan demikian, *Random Forest* unggul dalam mengenali siswa berperforma baik, sedangkan *Naïve Bayes* sedikit lebih baik dalam menjaga keseimbangan prediksi pada kelas rendah.

Hasil ini menjadi catatan penting karena dalam konteks pendidikan, kemampuan mendeteksi siswa dengan performa rendah sangat dibutuhkan untuk mendukung intervensi dini. Apabila model masih sering salah mengklasifikasikan siswa berperforma rendah sebagai siswa berperforma baik, maka model belum sepenuhnya

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

optimal untuk digunakan sebagai sistem pendukung keputusan pendidikan. Oleh karena itu, penelitian selanjutnya dapat mempertimbangkan penggunaan teknik penanganan data tidak seimbang, seperti *resampling*, *class weighting*, atau SMOTE.



Gambar 2. *Confusion Matrix* dan *Classification Report* Model *Naïve Bayes* dan *Random Forest*

3.5. Analisis *Feature Importance*

Selain membandingkan kinerja model, penelitian ini juga menganalisis faktor-faktor yang paling berpengaruh terhadap hasil prediksi menggunakan *feature importance* pada model *Random Forest*. Analisis ini dilakukan untuk mengetahui atribut yang memiliki kontribusi paling besar dalam proses klasifikasi performa akademik siswa.

Hasil analisis menunjukkan bahwa sepuluh fitur dengan kontribusi tertinggi adalah *absences*, *failures*, *age*, *goout*, *health*, *Medu*, *freetime*, *Fedu*, *Mjob*, dan *studytime*. Hasil tersebut ditunjukkan pada Tabel 6.

Tabel 6. *Top* 10 *Feature Importance Random Forest*

Peringkat	Fitur	Nilai Feature Importance
1	*absences*	0,095295
2	*failures*	0,074033
3	*age*	0,057045
4	*goout*	0,056017
5	*health*	0,047820
6	*Medu*	0,044671
7	*freetime*	0,043438
8	*Fedu*	0,042468
9	*Mjob*	0,041847
10	*studytime*	0,040225

Berdasarkan Tabel 6, fitur *absences* menjadi faktor paling berpengaruh terhadap prediksi performa akademik siswa. Hal ini menunjukkan bahwa jumlah ketidakhadiran memiliki hubungan penting dengan performa akademik. Siswa yang sering tidak hadir berpotensi kehilangan materi pembelajaran, mengalami keterlambatan pemahaman, dan memiliki kedisiplinan akademik yang lebih rendah.

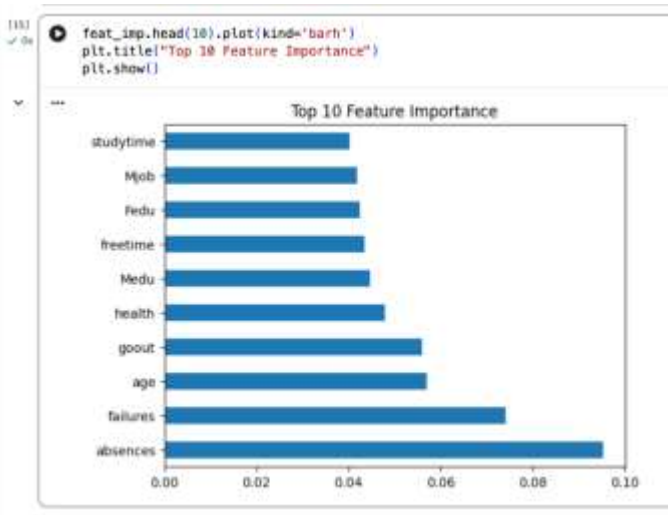
Fitur *failures* berada pada peringkat kedua. Hal ini menunjukkan bahwa riwayat kegagalan akademik sebelumnya menjadi salah satu indikator penting dalam memprediksi performa siswa. Siswa yang memiliki

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

riwayat kegagalan cenderung memiliki risiko lebih besar mengalami kesulitan akademik pada periode berikutnya. Temuan ini sejalan dengan penelitian Cortez dan Silva yang menunjukkan bahwa selain nilai sebelumnya, faktor seperti absensi, kegagalan akademik, pekerjaan orang tua, pendidikan orang tua, dan faktor sosial juga berhubungan dengan performa siswa [11].

Selain faktor akademik, hasil penelitian juga menunjukkan bahwa faktor sosial dan keluarga ikut berkontribusi terhadap performa akademik siswa. Fitur *goout* menunjukkan bahwa kebiasaan keluar bersama teman dapat berkaitan dengan performa akademik. Sementara itu, fitur *Medu*, *Fedu*, dan *Mjob* menunjukkan bahwa pendidikan dan pekerjaan orang tua juga memiliki kontribusi terhadap hasil prediksi. Hal ini mengindikasikan bahwa performa akademik siswa tidak hanya dipengaruhi oleh aktivitas belajar di sekolah, tetapi juga oleh lingkungan sosial dan keluarga.

Fitur *studytime* juga termasuk dalam sepuluh fitur terpenting, meskipun bukan faktor yang paling dominan. Hal ini menunjukkan bahwa waktu belajar tetap berperan dalam performa akademik, tetapi pengaruhnya tidak berdiri sendiri. Performa siswa lebih tepat dipahami sebagai hasil dari kombinasi berbagai faktor, seperti kehadiran, riwayat akademik, kondisi sosial, kesehatan, latar belakang keluarga, dan kebiasaan belajar.



Gambar 3. Grafik *Top* 10 *Feature Importance Random Forest*

3.6. Ringkasan Pembahasan

Berdasarkan seluruh hasil pengujian, *Random Forest* memperoleh *accuracy* lebih tinggi dibandingkan *Naïve Bayes*, yaitu 72,15% berbanding 70,88%. Hasil ini menunjukkan bahwa *Random Forest* lebih baik dalam melakukan klasifikasi performa akademik siswa pada dataset yang digunakan. Meskipun demikian, perbedaan nilai *accuracy* kedua model tidak terlalu besar, sehingga *Naïve Bayes* masih dapat dianggap kompetitif sebagai model klasifikasi sederhana.

Evaluasi menggunakan *confusion matrix* dan *classification report* menunjukkan bahwa kedua model lebih baik dalam mengenali siswa berperforma baik dibandingkan siswa berperforma rendah. Hal ini terlihat dari nilai *recall* kelas baik yang jauh lebih tinggi dibandingkan kelas rendah. Kondisi tersebut menunjukkan bahwa model masih memiliki keterbatasan dalam mendeteksi siswa berisiko rendah performa. Dalam konteks pendidikan, keterbatasan ini penting diperhatikan karena deteksi siswa berperforma rendah merupakan salah satu tujuan utama dari sistem prediksi akademik.

Hasil *feature importance* menunjukkan bahwa *absences*, *failures*, *age*, dan *goout* menjadi fitur yang paling berpengaruh terhadap performa akademik siswa. Temuan ini memperlihatkan bahwa performa akademik tidak hanya berkaitan dengan faktor belajar, tetapi juga dipengaruhi oleh kehadiran, riwayat akademik, kondisi

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

sosial, dan latar belakang keluarga. Dengan demikian, hasil penelitian ini dapat menjadi dasar awal bagi institusi pendidikan dalam memahami faktor-faktor yang perlu diperhatikan untuk mendukung intervensi akademik secara lebih tepat sasaran.

IV. KESIMPULAN

Berdasarkan hasil penelitian, algoritma *Naïve Bayes* dan *Random Forest* dapat digunakan untuk memprediksi performa akademik siswa pada *Student Performance Dataset*. Proses penelitian dilakukan melalui tahapan *preprocessing*, *encoding* variabel kategorikal, penghapusan atribut G1 dan G2 untuk menghindari *data leakage*, pembentukan label biner berdasarkan nilai G3, serta pemodelan menggunakan algoritma klasifikasi.

Hasil pengujian menunjukkan bahwa *Random Forest* memperoleh nilai *accuracy* sebesar 72,15%, sedangkan *Naïve Bayes* memperoleh nilai *accuracy* sebesar 70,88%. Dengan demikian, Random Forest memiliki kinerja yang sedikit lebih baik dibandingkan *Naïve Bayes*. Namun, hasil *confusion matrix* dan *classification report* menunjukkan bahwa kedua model masih memiliki keterbatasan dalam mendeteksi siswa dengan performa rendah, yang kemungkinan dipengaruhi oleh ketidakseimbangan distribusi kelas.

Analisis *feature importance* pada *Random Forest* menunjukkan bahwa faktor yang paling berpengaruh terhadap performa akademik siswa adalah *absences*, diikuti oleh *failures*, *age*, *goout*, dan *health*. Temuan ini menunjukkan bahwa performa akademik siswa tidak hanya dipengaruhi oleh faktor belajar, tetapi juga berkaitan dengan kehadiran, riwayat kegagalan akademik, kondisi sosial, kesehatan, dan latar belakang siswa. Untuk penelitian selanjutnya, disarankan menggunakan teknik penanganan data tidak seimbang seperti SMOTE, *resampling*, atau *class weighting*, serta menguji algoritma lain seperti *Support Vector Machine*, *Logistic Regression*, atau XGBoost agar diperoleh hasil prediksi yang lebih optimal.

REFERENSI

[1] S. Boujmiraz, H. Darhmaoui, and A. Drissi el maliani, "Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches," *Comput. Educ. Artif. Intell.*, vol. 10, p. 100548, Jun. 2026, doi: 10.1016/j.caeai.2026.100548.

[2] V. Y. D. Wijaya and G. Brotosaputro, "Penerapan Data Mining Dalam Prediksi Kinerja Akademik Mahasiswa Menggunakan Algoritma Machine Learning," 2025.

[3] R. Z. Arifin, H. Firmansyah, and W. Asriyani, "Prediksi Kelulusan Siswa Berdasarkan Data Demografis dan Akademik pada Dataset Student Performance," *J. Pengabdi. Masy. dan Ris. Pendidik.*, vol. 4, no. 2, pp. 13300–13307, Dec. 2025, doi: 10.31004/jerkin.v4i2.4251.

[4] R. A. S. Prayoga, R. Basatha, M. S. Akbar, E. A. Elfaiz, and C. D. Putra, "Penerapan Metode Naïve Bayes untuk Klasifikasi Performa Siswa," 2025. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[5] N. T. Renukadevi, K. Saraswathi, E. Roshini, M. G. Lakshitha, and S. Pratheeksha, "Evaluation of School Students Performance Using Machine Learning," in *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025)*, SCITEPRESS - Science and Technology Publications, Sep. 2025, pp. 503–512. doi: 10.5220/0013622900004664.

[6] C. Li, Q. Zhou, L. Du, and S. Zhang, "Predicting Student Performance through Machine Learning Methods: Naive Bayesian Classifier," *J. Artif. Intell. Syst. Model.*, vol. 02, no. 04, 2024, doi: 10.22034/jaism.2024.481968.1068.

[7] Z. Umarova *et al.*, "Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction," *Int. J. Inf. Educ. Technol.*, vol. 14, no. 1, pp. 92–98, 2024, doi: 10.18178/ijiet.2024.14.1.2028.

[8] A. B. Musa, "Understanding Student Performance in Foundation Year: Insights from Logistic Regression,

e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.10 No.1 (June 2026)
Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

Naïve Bayes, and Random Forest Models," *Int. J. Inf. Educ. Technol.*, vol. 14, no. 12, pp. 1716–1723, 2024, doi: 10.18178/ijiet.2024.14.12.2202.

[9] R. Waqia Wania, S. Mukherjee, S. Mondal, S. Mukherjee, and S. Hazra, "Predicting Student Performance Using Random Forest Algorithm," 2025. [Online]. Available: http://www.icitet.uk

[10] A. F. Heikal and Z. A. Khalil, "Student Performance Prediction Using Machine Learning: A Comprehensive Analysis," 2023. [Online]. Available: https://asric.africa/engineering-sciences

[11] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," *15th Eur. Concurr. Eng. Conf. 2008, ECEC 2008 - 5th Futur. Bus. Technol. Conf. FUBUTEC 2008*, pp. 5–12, 2008.