Jurnal Kreativitas Mahasiswa Informatika **E-ISSN: 3164-1989**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **P-ISSN: 2797-6327**
Vol. 7, No. 2, Agustus 2026 (377-388) **https://doi.org/10.32493/jatimika.v7i1**

Penerapan Algoritma Random Forest untuk Prediksi Performa Akademik Siswa

Cindy Shafira Dewi[1*], Ines Heidiani Ikasari[2]

[1,2]Teknik Informatika, Universitas Pamulang, Jl. Puspitek, Tangerang Selatan, Banten, Indonesia
e-mail: [1]cindyfira02@gmail.com, [2]dosen01374@unpam.ac.id
*Corresponding Author

Submitted Date: March 2, 2026 Reviewed Date: April 22, 2026
Revised Date: May 25, 2026 Accepted Date: July 04, 2026

Abstract

Education is an important aspect in improving the quality of human resources; therefore, methods capable of accurately predicting students' academic performance are needed. However, the process of predicting academic performance in schools is still largely carried out conventionally, which tends to be less efficient and potentially produces less accurate conclusions, especially due to data complexity and the large number of variables influencing academic performance. Therefore, this study aims to implement the Random Forest algorithm and compare the model performance using two tools, namely Python and RapidMiner, in predicting students' academic performance at a junior high school in Bogor City. Model evaluation was conducted using cross-validation techniques with 5-fold and 10-fold schemes. The results show that the implementation of Random Forest using Python achieved an accuracy of 86.75% with the 5-fold scheme and 85.43% with the 10fold scheme. Meanwhile, RapidMiner produced a stable accuracy of 84.11% in both testing schemes. The accuracy difference ranging from 1.32% to 2.64% indicates that the performance of both tools is relatively equivalent. Overall, the Random Forest algorithm demonstrates good capability in predicting students' academic performance according to the predefined categories. The findings of this study are expected to serve as a reference for schools in utilizing machine learning and to contribute to the development of machine learning applications in the field of education.

Keywords: Random Forest; Student Academic Performance Prediction; Python; RapidMiner; Cross Validation

Abstrak

Pendidikan merupakan aspek penting dalam meningkatkan kualitas sumber daya manusia, sehingga diperlukan metode yang mampu memprediksi performa akademik siswa secara akurat. Namun, proses prediksi performa akademik di sekolah masih banyak dilakukan secara konvensional, yang cenderung kurang efisien dan berpotensi menghasilkan kesimpulan yang kurang akurat, terutama karena kompleksitas data dan banyaknya variabel yang memengaruhi performa akademik. Oleh karena itu, penelitian ini bertujuan untuk menerapkan algoritma *Random Forest* serta membandingkan kinerja model menggunakan dua *tools*, yaitu *Python* dan *RapidMiner*, dalam memprediksi performa akademik siswa di salah satu Sekolah Menengah Pertama (SMP) di Kota Bogor. Evaluasi model dilakukan menggunakan teknik *cross validation* dengan skema 5-*fold* dan 10-*fold*. Hasil penelitian menunjukkan bahwa implementasi *Random Forest* menggunakan *Python* memperoleh akurasi sebesar 86.75% pada skema 5-*fold* dan 85.43% pada skema 10-*fold*. Sementara itu, *RapidMiner* menghasilkan akurasi yang stabil yaitu sebesar 84.11% pada kedua skema pengujian. Selisih akurasi sebesar 1.32% hingga 2.64% menunjukkan bahwa performa kedua *tools* relatif ekuivalen. Secara keseluruhan, algoritma *Random Forest* menunjukkan kemampuan yang baik dalam memprediksi performa akademik siswa sesuai kategori yang telah ditentukan. Hasil penelitian diharapkan dapat menjadi acuan bagi pihak sekolah dalam pemanfaatan *machine learning* serta memberikan kontribusi dalam pengembangan penerapan machine learning di dunia pendidikan.

Kata kunci: Random Forest; Prediksi Performa Akademik Siswa; Python; RapidMiner; Cross Validation

1. Pendahuluan

Pendidikan merupakan aspek penting dalam meningkatkan kualitas sumber daya manusia (Desmawan et al., 2023). Pendidikan yang berkualitas akan menghasilkan generasi yang cerdas dan berpotensi (Afandi et al., 2024). Oleh karena itu, sistem pendidikan yang efektif dan efisien sangat diperlukan untuk mendukung keberhasilan akademik siswa. Salah satu indikator utama dalam menilai kualitas pendidikan adalah performa akademik siswa, yang mencerminkan keberhasilan mereka dalam mengikuti pembelajaran dan penyelesaian jenjang pendidikan tertentu (Saefudin et al., 2021). Selain itu, performa akademik juga berpengaruh terhadap peluang akademik dan karier di masa depan. Dengan demikian, kemampuan untuk memprediksi performa akademik menjadi hal yang penting dalam upaya peningkatan mutu pendidikan.

Salah satu Sekolah Menengah Pertama (SMP) di Bogor memiliki beragam siswa dengan latar belakang yang berbeda. Dalam upaya meningkatkan kualitas pendidikan, sekolah menghadapi tantangan dalam memprediksi performa akademik siswa. Hingga saat ini, proses prediksi masih dilakukan secara konvensional, sehingga kurang efisien dan berpotensi menghasilkan kesimpulan yang kurang akurat. Kompleksitas data dan banyaknya variabel yang berkontribusi membuat proses prediksi memerlukan pendekatan yang lebih sistematis. Oleh karena itu, penerapan pembelajaran mesin menjadi solusi untuk membantu dalam meningkatkan akurasi prediksi perfoma akademik siswa.

Pada penelitian sebelumnya yang dilakukan oleh Abdul Rahman dalam studi berjudul "Klasifikasi Performa Akademik Siswa Menggunakan Metode *Decision Tree* dan *Naïve Bayes*" menunjukkan bahwa metode *naïve bayes* memiliki akurasi sebesar 85.97%, sedangkan metode *decision tree* mencapai akurasi sebesar 83.89% menggunakan platform *RapidMiner*. Namun, penelitian tersebut belum melibatkan algoritma lain, khususnya *random forest*, yang berpotensi menghasilkan hasil prediksi yang lebih optimal (Rahman, 2023).

Kebaruan penelitian ini terletak pada penerapan algoritma *random forest*, yang membangun model dengan menggabungkan sejumlah pohon keputusan secara acak untuk meningkatkan akurasi prediksi. Penelitian ini juga mengeksplorasi kemampuan *random forest* dalam konteks prediksi performa akademik siswa, yang sebelumnya belum banyak dikaji. Selain itu, dari sisi *tools* analisis, penelitian ini tidak hanya menggunakan *RapidMiner* seperti penelitian sebelumnya, tetapi juga *Python*, yang memiliki potensi menghasilkan akurasi lebih tinggi. Penggunaan *Python* membuka peluang untuk meneliti performa model secara lebih komprehensif melalui perbandingan akurasi antara kedua *tools* tersebut. Perbandingan performa *Python* dan *RapidMiner* dilakukan menggunakan konfigurasi parameter *Random Forest* yang sama agar hasil pengujian lebih konsisten dan objektif.

2. Metodologi Penelitian

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif, yang memungkinkan pengumpulan dan analisis data secara sistematis untuk menghasilkan informasi yang objektif dan mendalam. Dalam penelitian ini, dilakukan penerapan algoritma *random forest* untuk memprediksi performa akademik siswa. Dimulai dengan pengumpulan data yang dibutuhkan melalui dokumentasi dan studi pustaka, pengolahan data yang mencakup; persiapan dataset, *pre-pemrosesan* data, dan membangun model menggunakan *tools Python* dan *RapidMiner*, serta evaluasi model dengan menggunakan teknik *cross validation* 5-*fold* dan 10-*fold*. Hasilnya diharapkan dapat membantu menyusun rekomendasi guna mendukung strategi peningkatan performa di SMP.

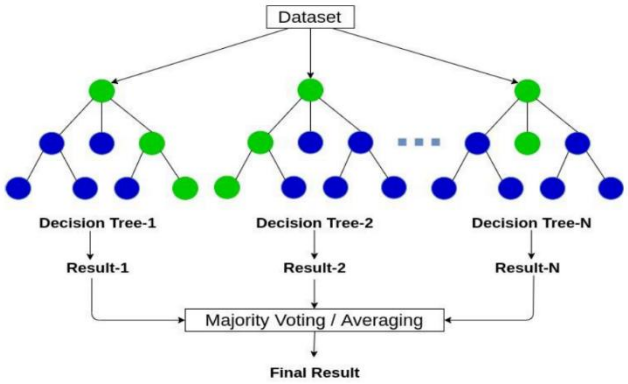
a. Machine Learning

Teknologi *machine learning* atau pembelajaran mesin adalah teknik pendekatan dari *Artificial Intelligent* (AI) yang telah berkembang sebagai pendekatan komputasi yang dapat digunakan untuk analisis data dan identifikasi pola yang sulit ditemukan secara manual dalam berbagai bidang, termasuk pendidikan (Diponegoro et al., 2021). Dengan teknik pembelajaran berbasis data, model *machine learning* mampu mengenali pola-pola tersembunyi yang sulit diidentifikasi secara manual, sehingga institusi pendidikan dapat merancang strategi pembelajaran yang lebih efektif (Sukmaningtyas et al., 2024).

Dengan pemanfaatan *machine learning*, guru dapat memantau dan mengevaluasi perkembangan pembelajaran setiap siswa secara lebih personal. Melalui penerapan algoritma, mesin dapat menerima data, menganalisisnya dan kemudian menghasilkan *output* yang sesuai dengan kebutuhan di bidang pendidikan. Dengan demikian, penggunaan algoritma *machine learning* menjadikan proses pembelajaran lebih mudah, cepat, dan efisien jika dibandingkan dengan ketika dilakukan secara konvensional (Fathurohman, 2021).

b. Random Forest

Random forest adalah algoritma pengembangan dari algoritma pohon keputusan yang memanfaatkan sejumlah pohon keputusan. Setiap pohon dilatih menggunakan sampel data yang berbeda, serta pemilihan atribut pada setiap pemecahan node dilakukan secara acak dari *subset* atribut yang tersedia. (Primajaya & Sari, 2018).



Gambar 1. Algoritma *Random Forest*

Menurut (Sudrajat et al., 2022) *random forest* adalah sebuah algoritma pembelajaran mesin yang berfungsi untuk melakukan klasifikasi pada dataset berukuran besar. Algoritma ini cocok digunakan pada data dengan banyak dimensi dan berbagai skala serta mampu memberikan performa tinggi. *Random forest* telah diterapkan di berbagai bidang seperti *e-commerce*, kesehatan, perbankan, analisis keuangan, dan lainnya.

Keunggulan dari *random forest* adalah kemampuannya dalam menangani data dengan banyak variabel dan menghasilkan model prediktif yang lebih akurat dengan mempertimbangkan berbagai faktor, sehingga mampu mengurangi risiko kesalahan prediksi (Efendi et al., 2024). Setiap pohon keputusan dibangun dengan *subset* data yang berbeda, sehingga hasil akhirnya merupakan kombinasi dari berbagai perspektif pohon tersebut. Hal ini membantu mengurangi risiko *overfitting* karena model tidak hanya bergantung pada satu pohon keputusan saja, melainkan pada gabungan banyak pohon yang lebih umum dan stabil.

Berikut rumus yang dapat digunakan dalam pengolahan data manual pada *random forest* dengan menghitung *entropy, entropy* setelah *split*, *information gain, voting* mayoritas, yaitu sebagai berikut (Saeprani et al., 2025):

- *Entropy* Awal
  Entropi digunakan untuk mengukur ketidakpastian atau kekacauan dalam suatu dataset.

  $$H = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) p_i$$

  Keterangan:
  H adalah Entropi
  p_i adalah proporsi data dalam kelas i
  n adalah jumlah kelas yang ada dalam data

- *Entropy* Setelah *Split*
  Menghitung rata-rata entropi dari *subset* setelah data dipecah berdasarkan semua fitur.

$$H_{\text{split}} = \sum_{j=1}^{m} \frac{N_j}{N} \cdot H_j$$

Keterangan:
N_j adalah jumlah data pada *subset* j
N adalah jumlah total data sebelum *split*
H_j adalah entropi *subset* j

- *Information Gain*
  Mengukur pengukuran ketidakpastian setelah *split*.

  $$IG = H_{awal} - H_{split}$$

- *Voting* Mayoritas
  Hasil akhir diperoleh dengan mengambil suara terbanyak dari semua pohon keputusan.

  $$Prediksi = Mayoritas(T_1, T_{2,\ldots,} T_k)$$

  Keterangan:
  T_k adalah prediksi dari *tree* ke-k

c. Python

Python adalah bahasa pemrograman tingkat tinggi yang *open-source* dan berorientasi objek. sehingga sangat sesuai untuk komputasi ilmiah. Bahasa ini dikenal karena ekosistem perpustakaan ilmiah dan lingkungan kerja yang luas, menjadikannya pilihan populer di kalangan peneliti dan ilmuan data. *Python* mendukung berbagai proses dalam alur kerja data ilmiah, mulai dari pengambilan data, manipulasi, pemrosesan, hingga visualisasi dan komunikasi hasil penelitian. Dengan kemampuannya yang fleksibel, pengguna tidak perlu mempelajari perangkat lunak atau bahasa pemrograman baru saat menghadapi kebutuhan analisis data yang baru (Saeprani et al., 2025).

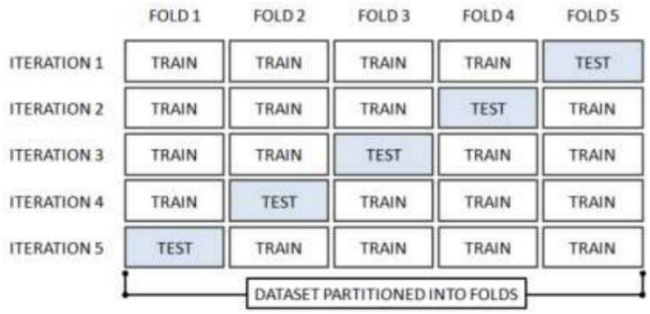
d. RapidMiner

Dalam konteks penelitian dan pengembangan model prediksi, *RapidMiner* merupakan alat yang banyak digunakan untuk melakukan analisis data yang menawarkan berbagai algoritma untuk klasifikasi, regresi, dan klustering, serta menyediakan lingkungan yang intuitif untuk memudahkan pengolahan dan visualisasi data sehingga membantu peneliti dan praktisi untuk mendapatkan wawasan yang lebih mendalam dari data yang dianalisis.

RapidMiner juga memiliki keunggulan dalam fleksibilitas integrasi dengan berbagai sumber data dan kemampuan otomatisasi alur kerja analitik yang kompleks. Alat ini mendukung integrasi dengan *Python*, R, dan *database SQL*, memungkinkan pengguna untuk mengombinasikan berbagai teknik analisis dalam satu platform. Fitur *drag-and-drop* yang dimilikinya memudahkan pengguna pemula maupun ahli untuk membangun model prediksi tanpa perlu menulis banyak kode, menjadikan *RapidMiner* salah satu platform yang efisien untuk eksplorasi data, pengembangan model, dan implementasi model secara *real-time* di berbagai proyek penelitian dan bisnis (Saeprani et al., 2025).

e. Cross Validation

Cross validation atau disebut estimasi rotasi pengujian di luar sampel, merupakan teknik validasi model yang menggunakan bagian data yang berbeda untuk menguji dan melatih model pada iterasi yang berbeda. Tujuan dari *cross validation* atau validasi silang adalah untuk menguji kemampuan model dalam memprediksi data baru, menandai masalah seperti *overfitting* atau bias pemilihan, dan memberikan insight tentang generalisasi model pada dataset independen. Berikut adalah cara kerja *k-folds cross validation* (Aditya et al., 2023).



Gambar 2. Cara Kerja *K-Folds Cross Validation*

f. Performa Akademik Siswa

Performa akademik adalah kemampuan individu dalam menyelesaikan tugas-tugas akademik (Imam et al., 2023). Performa akademik diukur dengan hasil akhir yang diperoleh sebagai keberhasilan dalam belajar (Elmore et al., 2016, p. 152). Hasil belajar memiliki peran penting karena digunakan sebagai indikator untuk menilai performa akademik siswa. Proses belajar merupakan upaya perubahan perilaku ke arah yang lebih baik melalui penambahan wawasan dan pengetahuan, peningkatan kemampuan serta keterampilan. Namun dalam pelaksanaannya, setiap siswa dapat menghadapi berbagai kendala atau hambatan belajar yang berbeda-beda. Perbedaan hambatan tersebut dipengaruhi oleh dua faktor utama (Muyassaaroh, 2021), yaitu:

1. Faktor Internal
   Faktor internal yang mempengaruhi aktivitas belajar mencakup dua aspek. Pertama, aspek fisiologis; yaitu kondisi fisik secara umum dan tingkat kebugaran tubuh yang dapat mempengaruhi semangat siswa dalam mengikuti pelajaran. Kedua, aspek psikologis, yang berhubungan dengan kuantitas dan kualitas hasil belajar siswa, seperti kecerdasan, bakat, minat dan motivasi.
2. Faktor Eksternal
   Faktor eksternal yang mempengaruhi kegiatan belajar yaitu lingkungan sosial sekolah karena dapat mempengaruhi semangat dan motivasi belajar siswa. Selain itu, lingkungan masyarakat dan teman-teman sepermainan di sekitar tempat tinggal juga turut berperan. Faktor eksternal lainnya meliputi kondisi gedung sekolah, lokasi tempat tinggal, lingkungan keluarga, ketersediaan sarana belajar, serta keadaan cuaca, yang secara keseluruhan berkontribusi terhadap tingkat keberhasilan belajar siswa.

g. Pengumpulan Data

Dalam penelitian ini, data yang digunakan adalah data siswa kelas VII di salah satu SMP di Kota Bogor. Pemilihan kelas VII didasarkan pada pertimbangan bahwa siswa di tingkat ini masih berada dalam fase transisi dari sekolah dasar ke jenjang SMP, sehingga performa akademik mereka relevan untuk diklasifikasikan.

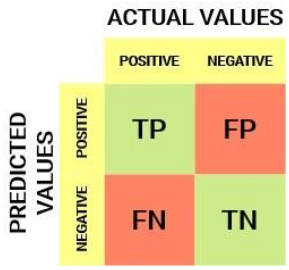
h. Pengolahan Data

Pengolahan data yang dilakukan dalam penelitian ini menggunakan algoritma *random forest* dan mengacu pada penelitian yang dilakukan oleh (Cahyani et al., 2025), yang membagi analisis menjadi beberapa tahap berikut:

- Persiapan Dataset
  Pada tahap ini, setelah data dikumpulkan, maka akan dilakukan pemahaman data untuk memastikan data yang digunakan sesuai dengan kebutuhan analisis.
- *Pre-pemrosesan* Data
  Proses yang dilakukan ditahap ini yaitu dilakukan *data cleaning*, seperti mengidentifikasi dan menangani nilai yang hilang atau *missing values*, duplikasi data. mengidentifikasi dan menangani

outliers, tranformasi data. Selain itu, perlu memastikan juga bahwa tipe data yang ada pada *record* sesuai dengan kebutuhan model *machine learning*, misalnya variabel kategorikal perlu dikonversi ke format numerik melalui beberapa teknik seperti label *encoding* yaitu mengubah kategori menjadi label numerik. Dilakukan juga pemilihan fitur yang paling relevan atau penting untuk digunakan dalam model *machine learning*. Dengan demikian dapat mengurangi kompleksitas model, meningkatkan akurasi, dan mempercepat proses pelatihan. Tahap *pre-pemrosesan* ini merupakan tahap vital untuk memastikan data yang digunakan dalam membangun model *machine learning* memiliki sifat akurat, relevan, dan bebas dari kesalahan yang bisa merusak kualitas model.

- Membangun Model
  Pada tahap ini dilakukan pemilihan jenis algoritma *machine learning* berdasarkan tipe masalah dan data yang dimiliki. Ada dua pendekatan utama, yaitu *supervised learning* dan *unsupervised learning*. Secara umum, proses pembangunan model meliputi pembuatan, pelatihan, dan pengujian model *machine learning*. Tahapan tersebut bertujuan memastikan model mampu menghasilkan prediksi yang akurat serta memiliki kemampuan generalisasi yang baik terhadap data baru.
- Hasil dan Evaluasi
  Tahap evaluasi model ini dilakukan untuk memberikan gambaran kinerja model yang dibuat untuk memastikan model dapat bekerja baik pada data latih dan data baru. Pada *supervised learning*, maka mengukur seberapa akurat model memprediksi label yang benar dengan data uji. Beberapa metrik yang digunakan yaitu: *accuracy* yaitu persentase prediksi yang benar, *precision* mengukur akurasi hasil positif, dan *recall* melihat seberapa banyak kasus positif terdeteksi serta *Confusion Matrix* menunjukkan jumlah prediksi benar dan salah dalam bentuk tabel.



Gambar 3. Tabel *Confusion Matrix*

Berikut di bawah ini rumus untuk menghitung kinerja model.

1. *Accuracy*

$$Accuracy = \frac{TP}{Total\ Data} \times 100\%$$

2. *Precision*

$$Precision\ (kelas\ positif) = \frac{TP}{TP + FP} \times 100\%$$

$$Precision\ (kelas\ negatif) = \frac{TN}{TN + FN} \times 100\%$$

3. *Recall*

$$Recall\ (kelas\ positif) = \frac{TP}{TP + FN} \times 100\%$$

$$Recall\ (kelas\ negatif) = \frac{TN}{TN + FP} \times 100\%$$

1. Hasil dan Diskusi

Penelitian ini menggunakan data akademik siswa yang terdiri dari 328 data siswa dari 9 kelas, dengan mencakup 19 atribut. Atribut-atribut tersebut meliputi No Urut, Nama, NIS, Jenis Kelamin, PABD, PKn, Bahasa Indonesia, Matematika, IPA, IPS, Bahasa Inggris, Seni, PJOK, Informatika, Bahasa Sunda, Jumlah Nilai, Ketidakhadiran Sakit, Ketidakhadiran Izin, dan Ketidakhadiran Alpa.

Pada tahap pertama pengolahan data yaitu persiapan dataset. Dilakukan analisis untuk memahami data yang telah diperoleh untuk memastikan data yang digunakan sesuai dengan kebutuhan analisis, serta dilakukan juga penambahan kolom target dengan label performa tinggi dan performa rendah, untuk setiap *record*.

Tahap kedua pengolahan data yaitu *pre-pemrosesan*. Proses yang dilakukan ditahap ini yaitu data cleaning yang mencakup pemilihan fitur, menghapus NaN dan data tidak relevan, binarisasi fitur, duplikasi data, cek persebaran label, serta label *encoding*. Pengurangan jumlah data dari 328 data awal menjadi 151 data terjadi akibat beberapa tahapan *pre-pemrosesan* yang dilakukan untuk menjaga kualitas dataset. Sebanyak 4 data dihapus karena tidak relevan dengan tujuan penelitian, yaitu data siswa yang tidak mengikuti kegiatan sekolah sehingga memiliki ketidakhadiran penuh dan berpotensi menimbulkan bias pada proses klasifikasi. Selanjutnya, dilakukan proses binarisasi fitur yaitu selain kolom target untuk menyederhanakan representasi data ke dalam kategori 0 dan 1.

Setelah tahapan tersebut dilakukan, teridentifikasi sebanyak 173 data duplikat sehingga perlu dihapus agar tidak terjadi pengulangan informasi yang sama pada proses pelatihan dan pengujian model. Penghapusan data duplikat dilakukan untuk menjaga konsistensi data dan menghindari bias performa model akibat adanya data identik. Meskipun jumlah data akhir menjadi lebih sedikit, penggunaan data yang telah dibersihkan tetap diperlukan agar model dilatih menggunakan data yang lebih valid dan representatif.

Setelah baris semua 0 dihapus:

	PABP	PKn	BIndo	Math	IPA	IPS	BIng	Art	PJOK	Inf	BSun	LABEL
0	81	88	87	83	83	84	85	90	89	88	88	Tinggi
1	91	94	94	92	94	94	93	91	92	94	89	Tinggi
2	79	86	87	82	80	83	84	86	91	91	86	Rendah
3	80	84	90	81	80	85	87	87	87	90	85	Rendah
4	77	85	88	78	80	80	82	85	89	87	82	Rendah
..	...	...	...	...	...	...	...	...	...	...	...	...
323	78	85	80	82	82	83	75	85	82	81	83	Rendah
324	82	89	80	90	83	85	85	87	87	85	86	Rendah
325	76	86	80	86	86	83	77	85	86	78	83	Rendah
326	80	87	82	88	84	85	79	85	84	82	85	Rendah
327	79	88	81	90	84	84	83	85	86	82	86	Rendah

[324 rows x 12 columns]

Gambar 4. Dataset Setelah Menghapus Data Tidak Relevan

Dataset setelah binarisasi fitur:

	PABP	PKn	BIndo	Math	IPA	IPS	BIng	Art	PJOK	Inf	BSun	LABEL
0	0	1	1	0	0	0	0	1	1	1	1	Tinggi
1	1	1	1	1	1	1	1	1	1	1	1	Tinggi
2	0	0	1	0	0	0	0	0	1	1	0	Rendah
3	0	0	1	0	0	0	1	1	1	1	0	Rendah
4	0	0	1	0	0	0	0	0	1	1	0	Rendah

Gambar 5. Dataset Setelah Binarisasi Fitur Menggunakan *Python*

Tahap ketiga pengolahan data yaitu membangun model dengan algoritma *random forest*. Dalam penelitian ini, dataset yang telah melalui proses pembersihan yaitu berjumlah 151 dataset. Implementasi algoritma *random forest* pada *Python* dan *RapidMiner* menggunakan parameter yang sama, yaitu *n_estimators* = 50, *max_depth* = 5, dan *random_state* = 42.

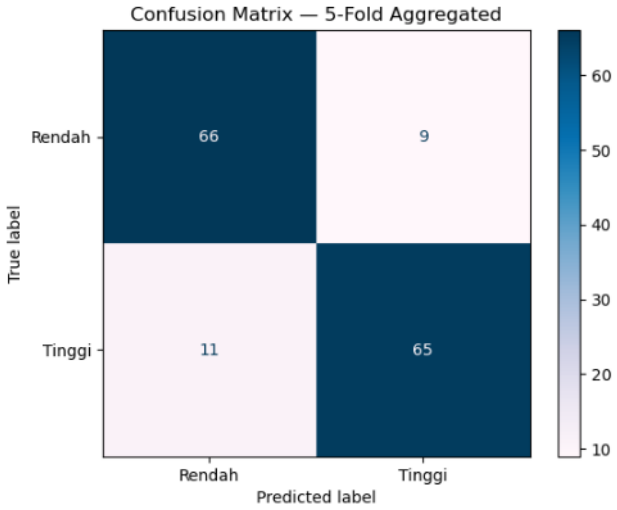
Tahapan keempat pengolahan data yaitu evaluasi model dengan teknik *cross validation* 5-*fold* dan 10-*fold*.

- Evaluasi dengan menggunakan *tools Python*

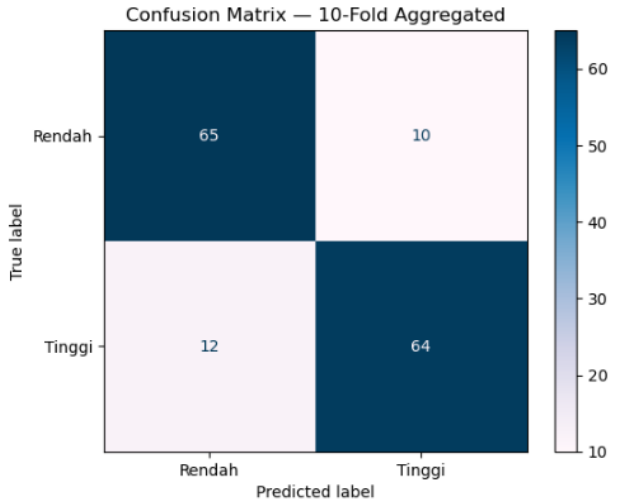
E-ISSN: 3164-1989
P-ISSN: 2797-6327
https://doi.org/10.32493/jatimika.v7i1

```python
# Evaluasi Cross Validation

from sklearn.model_selection import cross_val_score, cross_val_predict, StratifiedKFold
from sklearn.metrics import ConfusionMatrixDisplay
import matplotlib.pyplot as plt

fold_options = [5, 10]

print("\nEvaluasi Cross Validation (Akurasi Mean ± Std)")
for f in fold_options:
    skf = StratifiedKFold(n_splits=f, shuffle=True, random_state=42)
    scores = cross_val_score(rf, X, y, cv=skf, scoring='accuracy')
    print(f"\n{f}-Fold CV — Scores: {scores}")
    print(f"Mean Accuracy: {scores.mean():.4f} ± {scores.std():.4f}")

# Confusion Matrix hasil Cross Validation
for f in fold_options:
    skf = StratifiedKFold(n_splits=f, shuffle=True, random_state=42)
    y_pred_cv = cross_val_predict(rf, X, y, cv=skf)

    ConfusionMatrixDisplay.from_predictions(
        y,
        y_pred_cv,
        display_labels=['Rendah', 'Tinggi'],
        cmap='PuBu'
    )

    plt.title(f"Confusion Matrix — {f}-Fold Aggregated")
    plt.show()
```

Gambar 6. Kode Evaluasi *Cross Validation*



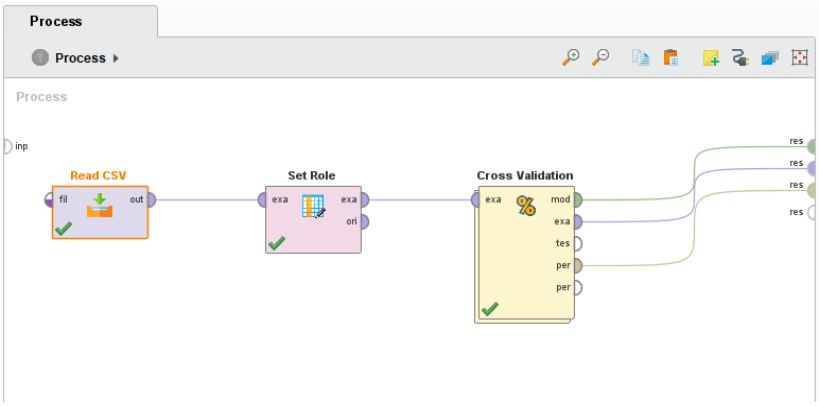
Gambar 7. *Confusion Matrix 5-fold*



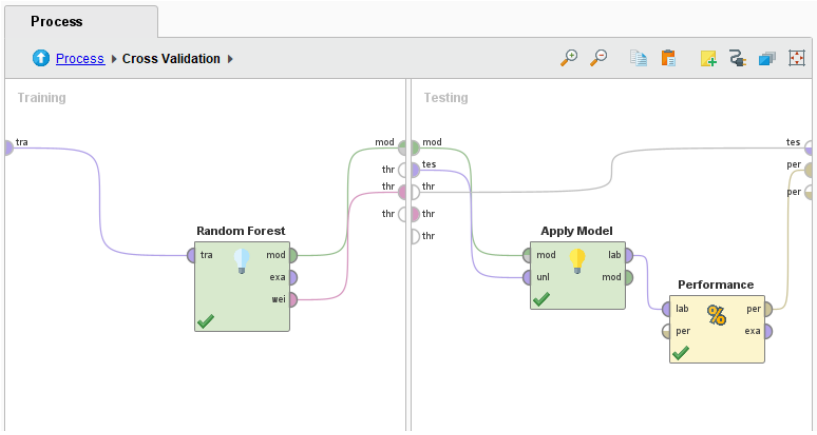
Gambar 8. *Confusion Matrix 10-fold*

Pengujian model *random forest* menggunakan *cross validation* 5-*fold* dan 10-*fold* menunjukkan performa yang konsisten dengan akurasi tinggi di setiap skema. Pada validasi 5-*fold*, akurasi tercatat sebesar 86.75%, dengan *precision* 85.75% untuk kelas Performa Rendah dan 87.84% untuk kelas Performa Tinggi, serta *recall* masing-masing 88% dan 85.53%. Pada validasi 10-*fold*, model mencapai akurasi sedikit menurun menjadi 85.43%, dengan *precision* 84.42% untuk kelas Performa Rendah dan 86.49% untuk kelas performa Tinggi, serta *recall* masing-masing 86.67% dan 84.21%. Secara keseluruhan, model *random forest* menunjukkan performa yang unggul dengan akurasi dan keseimbangan *precision-recall* yang tinggi, terutama pada validasi 5-*fold*, menjadikannya model yang sangat andal untuk memprediksi kedua kelas dengan tingkat kesalahan yang rendah.

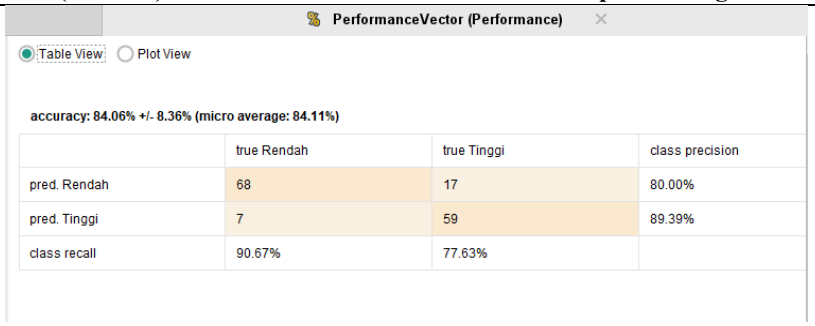
- Evaluasi dengan menggunakan *tools RapidMiner*



Gambar 9. Alur Evaluasi *Cross Validation* (1)



Gambar 10. Alur Evaluasi *Cross Validation* (2)



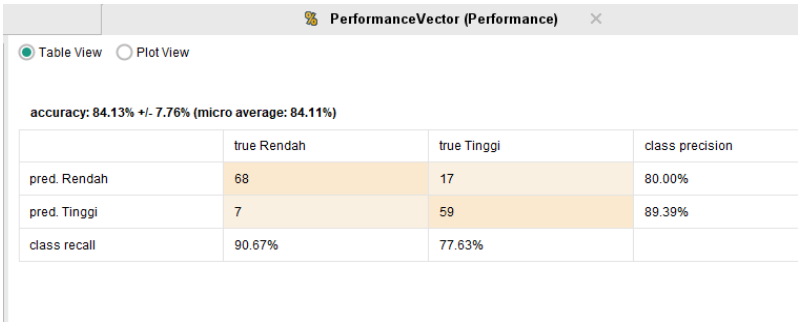
PerformanceVector (Performance)

Table View | Plot View

accuracy: 84.06% +/- 8.36% (micro average: 84.11%)

	true Rendah	true Tinggi	class precision
pred. Rendah	68	17	80.00%
pred. Tinggi	7	59	89.39%
class recall	90.67%	77.63%	

Gambar 11. *Confusion Matrix 5-fold*



PerformanceVector (Performance)

Table View | Plot View

accuracy: 84.13% +/- 7.76% (micro average: 84.11%)

	true Rendah	true Tinggi	class precision
pred. Rendah	68	17	80.00%
pred. Tinggi	7	59	89.39%
class recall	90.67%	77.63%	

Gambar 12. *Confusion Matrix 10-fold*

Pengujian model *random forest* menggunakan *cross validation* 5-*fold* dan 10-*fold* menunjukkan performa yang stabil di setiap skema. Pada kedua skema validasi tersebut, model memperoleh akurasi sebesar 84.11%, dengan *precision* 80% untuk kelas Performa Rendah dan 83.39% untuk kelas Performa Tinggi, serta *recall* masing-masing 90.67% dan 77.63%. Secara keseluruhan, model *random forest* menunjukkan performa yang baik dengan keseimbangan *precision-recall* pada kedua kelas. Hasil ini mengindikasikan bahwa model mampu mengklasifikasikan data Performa Tinggi dan Performa Rendah secara andal dengan tingkat kesalahan yang relatif rendah, sehingga layak digunakan sebagai model prediksi performa akademik.

2. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma *random forest* mampu memprediksi performa akademik siswa dengan baik sesuai kategori yang telah ditentukan. Perbandingan akurasi algoritma *random forest* pada *Python* dan *RapidMiner* dilakukan dengan menggunakan teknik *cross validation*, yang hasil evaluasinya menunjukkan bahwa algoritma *random forest* memiliki performa yang baik pada kedua *tools*.

Pada pengujian menggunakan *Python*, model memperoleh akurasi sebesar 86.75% pada skema 5-*fold* dan 85.43% pada skema 10-*fold*, sedangkan pada pengujian menggunakan *RapidMiner* diperoleh akurasi sebesar 84.11% pada kedua skema pengujian. Hasil tersebut menunjukkan bahwa *Python* memiliki akurasi sedikit lebih tinggi dibandingkan *RapidMiner* dengan selisih sebesar 1.32% hingga 2.64%. Perbedaan akurasi ini diduga dipengaruhi oleh perbedaan implementasi algoritma dan *library* yang digunakan, yaitu *scikit-learn* pada *Python* dan operator internal pada *RapidMiner*. Selain itu, meskipun data uji yang digunakan pada kedua *tools* berasal dari dataset yang sama dan telah melalui proses *data cleaning* yang identik di *Python*, kemungkinan masih terdapat perbedaan pada mekanisme pemrosesan data, parameter *default*, maupun proses *pre-pemrosesan* otomatis pada masing-masing *tools*. Namun demikian, selisih akurasi yang diperoleh masih tergolong kecil sehingga performa kedua tools dapat dikatakan relatif ekuivalen. Temuan ini juga menjadi catatan keterbatasan metodologi pengujian yang dapat dikembangkan lebih lanjut pada penelitian selanjutnya.

a. Saran

Berdasarkan temuan penelitian, beberapa saran yang dapat diberikan untuk penelitian selanjutnya antara lain melakukan optimalisasi parameter algoritma *random forest*, seperti *number of trees* dan *maximum depth* guna meningkatkan kinerja dan akurasi model baik pada *Python* maupun *RapidMiner*. Selain itu, penelitian ini menggunakan dataset yang terbatas pada satu sekolah. Oleh karena itu, penelitian selanjutnya diharapkan dapat menggunakan dataset yang lebih luas dan beragam agar kemampuan generalisasi model dalam memprediksi performa akademik siswa semakin meningkat. Penelitian berikutnya juga disarankan untuk tidak hanya menggunakan satu algoritma, tetapi melakukan perbandingan algoritma *random forest* dengan algoritma klasifikasi lainnya, sehingga dapat diketahui algoritma yang memiliki performa terbaik berdasarkan hasil evaluasi model. Penelitian ini berfokus pada perbandingan performa model menggunakan keseluruhan atribut yang tersedia, sehingga analisis terkait tingkat pengaruh masing-masing atribut terhadap prediksi performa akademik siswa belum dilakukan. Oleh karena itu, penelitian selanjutnya dapat menambahkan analisis *feature importance* untuk mengetahui atribut yang paling dominan, seperti tingkat kehadiran, nilai mata pelajaran tertentu, maupun faktor eksternal lainnya, sehingga dapat memberikan nilai praktis yang lebih besar bagi pihak sekolah dalam menentukan strategi peningkatan performa akademik siswa.

Referensi

Aditya, A., Wahyuddin, P., Leo, S., Santoso, W., Wahyu, G., Wibowo, N., Khrisna, A., Rahmaddeni, W., Wahidin, A. J., Eka, G., Elisawati, Y., Rizqi, R., & Abdurrasyid, W. (2023). *MACHINE LEARNING*. PT Global Eksekutif Teknologi. Padang.

Afandi, K., Arief, M. H., Laily, N. F., & Nugroho, D. M. (2024). Analisis Performa Akademik Mahasiswa Menggunakan Social Network Analysis. *Journal of Technology and Informatics (JoTI)*, *5*(2), 65–69. https://doi.org/10.37802/joti.v5i2.514

Cahyani, N., Irsyada, R., & Kartini, A. Y. (2025). Implementasi Machine Learning Model sebagai Sistem Prediksi Penyakit Breast Cancer. *Digital Transformation Technology*, *4*(2), 1112–1120. https://doi.org/10.47709/digitech.v4i2.5209

Desmawan, D., Aleyda, F., Darwin, R., Salsyabila, S., Rizqina, A., & Ikhsanudin. (2023). Analisis Peran Pendidikan Terhadap Kualitas Sumber Daya Manusia Guna Meningkatkan Produktivitas Masyarakat Di DKI Jakarta. *Jurnal Ilmu Manajemen, Ekonomi Dan Kewirausahaan*, *1*(2), 214–224. https://doi.org/10.59024/jumek.v1i2.75

Diponegoro, M. H., Kusumawardani, S. S., & Hidayah, I. (2021). Tinjauan Pustaka Sistematis: Implementasi Metode Deep Learning pada Prediksi Kinerja Murid (Implementation of Deep Learning Methods in Predicting Student Performance: A Systematic Literature Review). *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(2), 131–138. https://doi.org/10.22146/jnteti.v10i2.1417

Efendi, M. S., Sarwido, & Zyen, A. K. (2024). *Penerapan Algoritma Random Forest Untuk Prediksi Penjualan Dan Sistem Persediaan Produk*. *5*(1), 12–20. https://doi.org/10.30865/resolusi.v5i1.2149

Elmore, W. M., Young, J. K., Harris, S., & Mason, D. (2016). *The Relationship between Individual Student Attributes and Online Course Completion*. In *Handbook of Research on Building, Growing, and Sustaining Quality E-Learning Programs* (pp. 151–173). IGI Global, Hershey, PA (Pennsylvania).

Fathurohman, A. (2021). MACHINE LEARNING UNTUK PENDIDIKAN: MENGAPA DAN BAGAIMANA. *Jurnal Informatika Dan Tekonologi Komputer JITEK*, *1*(3), 57–62. https://doi.org/10.55606/jitek.v1i3.306

Imam, A. K., Ridfah, A., & Akmal, N. (2023). Efektivitas Pemberian Sugesti Positif Terhadap Peningkatan Performa Akademik Mahasiswa Fakultas Psikologi UNM. *Jurnal Sinestesia*, *13*(2), 1311–1320. https://sinestesia.pustaka.my.id/journal/article/view/495

Muyassaaroh, I. K. (2021). Belajar Efektif dan Efisien untuk Problem Belajar Siswa yang Berprestasi Rendah. *Journal of Islamic Education*, *1*(1), 87–96. https://doi.org/10.14421/hjie.11-08

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *1*(1), 27–31. http://dx.doi.org/10.24014/ijaidm.v1i1.4903

Rahman, A. (2023). Klasifikasi Performa Akademik Siswa Menggunakan Metode Decision Tree dan Naive Bayes. *Jurnal SAINTEKOM*, *13*(1), 22–31. https://doi.org/10.33020/saintekom.v13i1.349

Saefudin, W., Sriwiyanti, & Mohamad Yusoff, S. H. (2021). *SPIRITUAL WELL-BEING SEBAGAI PREDIKTOR PERFORMA AKADEMIK SISWA DI MASA PANDEMI*. *9*(2), 247–262.

Saeprani, R., Handayani, M., & Taryo, T. (2025). Analisis Dan Perbandingan Algoritma Decision Tree, Naïve Bayes, Dan Random Forest Dalam Evaluasi Prestasi Siswa Smk (Study Kasus: Smkn 8 Kabupaten Tangerang), *Tesis,* Pasca Sarjana Ilmu Komputer, Univ. Pamulang, Tangerang Selatan.

Sudrajat, D., Purnamasari, A. I., Dikananda, A. R., Kurnia, D. A., & Bahtiar, A. (2022). Klasifikasi Mutu Pembelajaran Hybrid berdasarkan Algoritma C.45, Random Forest dan Naïve Bayes dengan Optimasi Bootsrap Areggating (Bagging) pada masa COVID-19. *JURIKOM (Jurnal Riset Komputer)*, *9*(6), 2227–2233. https://doi.org/10.30865/jurikom.v9i6.5179

Sukmaningtyas, Y. N., Akbar, R. M., & Asyafiiyah, G. R. U. (2024). Penerapan Predictive Analytics untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa. *Arcitech: Journal of Computer Science and Artificial Intelligence*, *4*(2), 127–145. https://doi.org/10.29240/arcitech.v4i2.12048