Journal of Innovative and Creativity
Journal Homepage: https://joecy.org/index.php/joecy
e-mail: journaljoecy@gmail.com

Faculty of Education
University of Pahlawan Tuanku Tambusai
ISSN 2775-771X (Online)
ISSN 2962-570X **(Printed)**
6 (2) 2026, Page 45885-45900

Analisis Faktor Akademik dan Sosial Terhadap Prestasi Siswa Menggunakan Pendekatan Educational data Mining

Ninik Tri Hartanti[1], Erni Seniwati[2], Ifraweri Raja Mangkuto HP[3]
[1,3]Universitas Amikom Yogyakarta
[2]Universitas Langlang Buana
Corresponding Author: Ninik Tri Hartanti, e-mail: ninik.t@amikom.ac.id

ABSTRAK

Pencapaian prestasi peserta didik dipengaruhi oleh sejumlah faktor, tidak hanya dari pemahaman akademik, tetapi juga dari kondisi sosial serta lingkungan tempat siswa berinteraksi. Penelitian ini dilakukan untuk menganalisis dan mengidentifikasi faktor akademik serta sosial yang berkaitan dengan kategori prestasi siswa melalui pendekatan Educational Data Mining (EDM) dengan algoritma Random Forest. Data yang digunakan berasal dari dataset pendidikan publik yang memuat 395 data siswa dan 16 atribut yang menggambarkan karakteristik akademik maupun sosial siswa. Proses penelitian dilakukan melalui beberapa tahapan, yaitu data cleaning, seleksi fitur, encoding pada data kategorikal, normalisasi menggunakan metode Min-Max. Penentuan dataset dibagi 80:20 untuk data training 80% dan untuk data testing 20%, dan pemodelan menggunakan algoritma Random Forest. Model menggunakan 100 decision tree dan kinerjanya dinilai berdasarkan confusion matrix, Accuracy, Precision, Recall, dan F1-Score. Pengujian 79 data testing diperoleh 42 klasifikasi benar dan 37 klasifikasi salah. Nilai evaluasi yang diperoleh meliputi Accuracy sebesar 53,16%, Precision 51,14%, Recall 53,16%, dan F1-Score 47,90%. Berdasarkan hasil evaluasi, model menunjukkan performa terbaik dalam mengklasifikasikan siswa pada kategori prestasi sedang, yang ditunjukkan oleh nilai Recall sebesar 84,21%. Sementara itu, kategori prestasi tinggi menjadi kategori yang paling sulit dikenali oleh model dengan nilai Recall sebesar 6,67%. Hasil tersebut mengindikasikan bahwa Random Forest memiliki potensi untuk menganalisis pola prestasi siswa berdasarkan faktor akademik dan sosial, namun kemampuan klasifikasinya masih belum optimal dan belum merata pada setiap kategori. Dengan demikian, penelitian selanjutnya perlu mempertimbangkan pengembangan model dan strategi pemrosesan data yang lebih baik, khususnya untuk meningkatkan identifikasi terhadap siswa dengan tingkat prestasi tinggi.

Kata Kunci: *Educational Data Mining, Accuracy, Random Forest, Recall, Confusion Matrix*

ABSTRACT

Student achievement is influenced by various factors, ranging from academic understanding to social conditions and the environments in which students interact. This study aims to analyze and identify academic and social factors associated with student achievement categories using an Educational Data Mining (EDM) approach with the Random Forest algorithm. The study utilizes a public educational dataset containing records for 395 students and 16 attributes describing their academic and social characteristics. The research process involved several stages: data cleaning, feature selection, encoding of categorical data, and normalization using the Min-Max method. The dataset was split into an 80:20 ratio for training (80%) and testing (20%), and modeling was performed using the Random Forest algorithm. The model employed 100 decision trees, and its performance was evaluated using a confusion matrix, Accuracy, Precision, Recall, and F1-Score. Testing on 79 data points yielded 42 correct classifications and 37 incorrect ones. Evaluation metrics included an Accuracy of 53.16%, Precision of 51.14%, Recall of 53.16%, and an F1-Score of 47.90%. The results indicate that the model performed best in classifying students in the medium achievement category, evidenced by a Recall score of 84.21%. Conversely, the high achievement category proved the most difficult for the model to identify, with a Recall score of 6.67%. These findings suggest that while Random Forest has the potential to analyze student achievement patterns based on academic and social factors, its classification capability is not yet optimal or consistent across all categories. Consequently, future research should focus on developing improved models and data processing strategies, particularly to enhance the identification of high-achieving students.

Keywords: Educational Data Mining, Accuracy, Random Forest, Recall, Confusion Matrix

© The Author(s). 2026 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's

Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0

PENDAHULUAN

Salah satu indikator dalam penilaian keberhasilan suatu pendidikan adalah prestasi akademik siswa. Prestasi akademik siswa tidak hanya mencerminkan kemampuan akan pemahaman siswa terhadap materi pembelajaran, namun dapat mencerminkan juga keterlibatan siswa dalam proses belajar yang nantinya akan mendukung atau menghambat pencapaian prestasi siswa. Pada kenyataannya, banyak faktor yang menghambat prestasi akademik, diantaranya karakteristik siswa yang berhubungan dengan lingkungan, kondisi sosial maupun aktivitas akademik. Terdapat penelitian yang menunjukkan bahwa karakteristik akademik, sosial, demografis dan lingkungan sekolah dapat digunakan untuk menganalisis dan memprediksi prestasi akademik siswa (Yağcı, 2022).

Kemajuan teknologi informasi mendorong peningkatan jumlah data pendidikan yang berasal dari berbagai aktivitas, seperti pembelajaran, evaluasi, kehadiran, administrasi, dan interaksi siswa. Data pendidikan tersebut berpotensi pemanfaatannya selain sebagai data administratif, bermanfaat juga sebagai sumber informasi untuk mendukung pengambilan keputusan berbasis data. Pendekatan untuk mengolah dan menemukan pola dari data pendidikan tersebut salah satunya adalah Educational Data Mining (EDM), yang menerapkan teknik data mining dan komputasional terhadap data untuk menghasilkan pola dan informasi yang dibutuhkan dalam memahami proses pembelajaran serta karakteristik peserta didik (Roslan & Chen, 2022).

Penerapan EDM dalam analisis prestasi akademik telah berkembang dalam berbagai penelitian. Kajian terkait penelitian EDM menunjukkan bahwa data akademik dan karakteristik siswa merupakan atribut yang banyak digunakan dalam prediksi prestasi akademik (Sahputri et al., 2023) penelitian dengan pendekatan EDM yang berfokus pada karakteristik demografis menunjukkan bahwa karakteristik demografis termasuk aspek yang sering digunakan dalam penelitian prediksi prestasi akademik peserta didik atau siswa. Kondisi ini secara tidak langsung dapat menunjukkan bahwa data peserta didik dapat dimanfaatkan untuk memperoleh informasi mengenai karakteristik yang berkaitan dengan pencapaian prestasi akademik (Armawan et al., 2022).

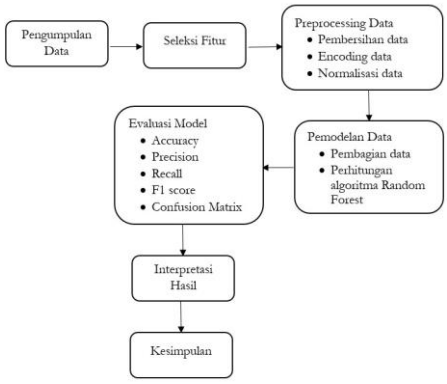
Pencapaian prestasi akademik peserta didik tidak hanya dipengaruhi oleh faktor akademik, tetapi dapat juga dari aspek sosial. Diantaranya adalah faktor dari lingkungan keluarga, hubungan dengan orang tua, dengan teman sebaya, aspek ekonomi, dan keaktifan atau partisipasi peserta didik dalam kegiatan di luar jam belajar akan dapat memberikan pengaruh terhadap terbentuknya kondisi belajar yang mendukung ataupun menghambat (Indahyanti et al., 2022). Penelitian yang terbaru oleh (Marín et al., 2025) menunjukkan bahwa faktor akademik, sosial-keluarga, sosial-ekonomi, manajemen waktu, serta karakteristik lainnya dapat dianalisis secara simultan menggunakan pendekatan machine learning untuk memahami faktor yang berkaitan dengan performa akademik mahasiswa.

Meskipun beberapa penelitian telah membuktikan bahwa pendekatan Educational Data Mining (EDM) digunakan untuk memprediksi maupun menganalisis prestasi akademik peserta didik, sebagian penelitian masih berfokus pada tingkat akurasi prediksi dan evaluasi kinerja antaralgoritma. Di sisi lain, kajian yang secara khusus menggabungkan faktor akademik dan sosial untuk mengidentifikasi kontribusi relatif masing-masing faktor terhadap prestasi siswa masih belum banyak dilakukan. Selain itu, variasi karakteristik dataset, atribut yang digunakan, jenjang pendidikan, serta metode analisis yang diterapkan menyebabkan temuan penelitian sebelumnya belum tentu dapat diterapkan secara langsung pada karakteristik data siswa yang berbeda. Dengan demikian, diperlukan pendekatan analisis yang tidak hanya mampu menghasilkan prediksi prestasi siswa, tetapi juga dapat mengidentifikasi atribut yang memberikan kontribusi paling signifikan terhadap pencapaian akademik

Penelitian ini bertujuan untuk menganalisis hubungan karakteristik akademik dan sosial dengan prestasi siswa menggunakan pendekatan Educational Data Mining, membangun model prediksi prestasi siswa berdasarkan faktor akademik dan sosial, dan mengidentifikasi faktor yang memiliki kontribusi paling besar terhadap prestasi akademik peserta didik. Hasil penelitian diharapkan dapat memberikan informasi berbasis data yang dapat digunakan sebagai salah satu pertimbangan dalam memahami karakteristik siswa dan mendukung strategi peningkatan prestasi akademik.

METODE

Pada penelitian ini, dilakukan pendekatan kuantitatif berbasis Educational Data Mining (EDM). Pendekatan tersebut digunakan untuk menganalisis hubungan antara faktor akademik dan sosial dengan prestasi peserta didik. EDM digunakan sebagai pendekatan analisis untuk mengidentifikasi pola dan karakteristik data pendidikan yang dapat memberikan gambaran terkait faktor apa saja yang berkontribusi terhadap prestasi siswa (Ardiansyah & Jailani, 2023).



Gambar 1. Tahapan Penelitian

Sebelum melakukan langkah penelitian, terlebih dahulu menentukan identifikasi masalah. Adapun pokok masalah dalam penelitian ini adanya ketidaksesuaian antara banyaknya data siswa yang tersedia dengan ketidakjelasan dari faktor yang paling berpengaruh terhadap prestasi akademik peserta didik,

yaitu faktor akademik dan sosial. Merujuk Gambar 1, tahapan dalam penelitian meliputi pengumpulan data, preprocessing data, seleksi fitur, analisis eksploratif, pemodelan data, evaluasi model, dan interpretasi hasil (Sutriyanti & Mohammad Muspawi, 2024).

a. Pengumpulan Data

Tahap pertama diawali dengan pengumpulan data dengan cara mengunduh data publik dari https://archive.ics.uci.edu/dataset. Terdapat 395 siswa dan 33 atribut dari dataset tersebut, namun sesuai dengan topik penelitian ini tidak semua atribut akan digunakan.

b. Seleksi Fitur atau atribut

Tahap seleksi fitur atau atribut akan dibedakan menjadi fitur akademik dan fitur sosial. Adapun fitur yang akan digunakan adalah studytime, failures, absences, paid, higher, famrel, famsup, goout, freetime, activities, internet, Medu, Fedu, Mjob, Fjob, guardian, dan G3 sebagai nilai akhir.

c. Preprocessing Data

Di tahap preprocessing terdapat beberapa langkah yaitu pembersihan data, encoding, dan normalisasi (Sarker et al., 2024).

d. Pemodelan Data

Setelah tahap seleksi fitur atau atribut, langkah berikutnya adalah pemodelan data, dengan membagi dataset dengan pembagian data training 80% data testing 20%. Selanjutnya dari pembagian dataset tersebut dilakukan proses perhitungan dengan algoritma EDM random forest.

e. Evaluasi Model

Tahap evaluasi model merupakan langkah penting terkait hasil dari proses penerapan algoritma EDM yakni Random Forest. Di tahap ini dilakukan evaluasi dengan accuracy, precison, recall, F1 score, dan confusion matrix.

f. Interpretasi Hasil

Tahap interpretasi hasil yaitu tahap untuk menjelaskan makna dari hasil analisis, bukan hanya sekadar menyajikan angka.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dan seleksi fitur

Mengacu data yang diperoleh secara publik dataset publik diperoleh dari https://archive.ics.uci.edu/dataset terdiri dari fitur School, Sex, Age, Address, Famsize, Pstatus, Medu, Fedu, Mjob, Fjob, Reason, Guardian, Traveltime, Studytime, Failure, Schoolsup, Famsup, Paid, Activities, Nursery, Higher, Internet, Romantic, Famrel, Freetime, Goout, Dalc, Walc, Health, Absences, G1,G2, dan G3. Beberapa fitur atau atribut yang sesuai dengan penelitian antara lain studytime, failures, absences, paid, higher, famrel, famsup, goout, freetime, activities, internet, Medu, Fedu, Mjob, Fjob, guardian, dan G3 sebagai nilai akhir.

Variabel G3 merupakan nilai akhir siswa dan dapat digunakan sebagai indikator prestasi akademik. Dataset yang telah diperoleh akan ditinjau ulang terkait struktur data, jumlah atribut, jumlah observasi, tipe data, serta kesesuaian fitur dengan tujuan penelitian. Tujuan dari tahap ini adalah data akan dapat merepresentasikan faktor akademik dan sosial yang akan dianalisis terhadap prestasi peserta didik.

Preprocessing Data

Tahap pre-processing dilakukan untuk mempersiapkan data sebelum masuk ke analisis dan pemodelan. Dalam tahap ini terdapat pengecekan terkait penyesuaian akan kebutuhan data, transformasi variabel kategorikal, serta normalisasi data numerik. Langkah awal adalah dilakukan proses seleksi fitur yang disesuaikan dengan fitur yang akan digunakan, yakni fitur untuk target adalah G3, fitur untuk kelompok akademik adalah : studytime, failures, absences, paid, higher. Sedangkan fitur untuk kelompok sosial yakni famrel, famsup, goout, freetime, activities, internet, Medu, Fedu, Mjob, Fjob, guardian, dan G3 (Guntara & Astuti, 2025).

a. Pembersihan Dataset

Proses pembersihan data bertujuan untuk membersihan data dari atribut atau data yang tidak terisi dan tidak sesuai, memastikan bahwa dataset yang digunakan memiliki kualitas, kelengkapan, dan konsistensi sebelum dilakukan analisis (Hidayat et al., n.d.). Proses ini mencakup pemeriksaan terhadap missing value, data duplikat, ketidaksesuaian tipe data, serta nilai yang tidak valid atau berada di luar rentang yang telah ditentukan. Berikut potongan koding pembersihan data dengan menggunakan Google Colab.

```
import pandas as pd
# Membaca file Excel
df = pd.read_excel("data_penelitian_16_fitur.xlsx")
# Menampilkan 5 data pertama
df.head()
```

Hasil koding menampilkan 5 data seperti pada Tabel 1.

Tabel 1. Data Fitur

No	Studytime	Failures	Absences	Paid	Higher	Famrel	Famsup	Goout	Freetime	Activities
1.	2	0	6	no	yes	4	no	4	3	no
2.	2	0	4	no	yes	5	yes	3	3	no
3.	2	3	10	yes	yes	4	no	2	3	no
4.	3	0	2	yes	yes	3	yes	2	2	yes

5.	2	0	4	yes	yes	4	yes	2	3	no	

Tabel 1a. Data Fitur (lanjutan)

Internet	Medu	Fedu	Mjob	Guardian	G3
No	4	4	at_home	mother	6
Yes	1	1	at_home	father	6
Yes	1	1	at_home	mother	10
Yes	4	2	health	mother	15
No	3	3	other	father	10

```
//mengatur kolom sesuai fitur
fitur_akademik = [
    "studytime",
    "failures",
    "absences",
    "paid",
    "higher"
]
fitur_sosial = [
    "famrel",
    "famsup",
    "goout",
    "freetime",
    "activities",
    "internet",
    "Medu",
    "Fedu",
    "Mjob",
    "Fjob",
    "guardian"
]
target = ["G3"]
fitur_final = fitur_akademik + fitur_sosial + target
```

```
print("Jumlah fitur:", len(fitur_final))
print(fitur_final)
//menampilkan baris dan kolom
print("=== UKURAN DATA ===")
print("Jumlah baris :", df.shape[0])
print("Jumlah kolom :", df.shape[1])
```

Menghasilkan output ukuran data adalah jumlah baris = 395 dan jumlah kolom = 17.

```
//cek missing value
print("\n=== MISSING VALUE ===")
missing = df[fitur_final].isnull().sum()
print(missing)
print("\nTotal missing value:",missing.sum())
```

Hasil cek missing value total missing value = 0

```
=== MISSING VALUE ===
studytime     0
failures      0
absences      0
paid          0
higher        0
famrel        0
famsup        0
goout         0
freetime      0
activities    0
internet      0
Medu          0
Fedu          0
Mjob          0
Fjob          0
guardian      0
G3            0
dtype: int64

Total missing value: 0
```

Gambar 2. Total Missing Value

```
//cek data duplikat
print("\n=== DATA DUPLIKAT ===")
duplikat = df[fitur_final].duplicated().sum()
print("Jumlah data duplikat:", duplikat)
```

Hasilnya adalah jumlah data duplikat = 0.

```
//cek kevalidan data
print("\n=== DATA TIDAK VALID ===")
print("studytime:",
    (~df["studytime"].isin([1,2,3,4])).sum())
print("failures:",
    (~df["failures"].isin([0,1,2,3])).sum())
print("absences:",
```

```
    (df["absences"] < 0).sum())
print("famrel:",
    (~df["famrel"].between(1,5)).sum())
print("goout:",
    (~df["goout"].between(1,5)).sum())
print("freetime:",
    (~df["freetime"].between(1,5)).sum())
print("Medu:",
    (~df["Medu"].between(0,4)).sum())
print("Fedu:",
    (~df["Fedu"].between(0,4)).sum())
print("G3:",
    (~df["G3"].between(0,20)).sum())
```

Hasilnya adalah tidak ada data yang tidak valid.

```
print("\n=== KATEGORI ===")
//menentukan kategori untuk setiap fitur
kategori = [
  "paid",
  "higher",
  "famsup",
  "activities",
  "internet",
  "Mjob",
  "Fjob",
  "guardian"
]
for kolom in kategori:
  print(kolom, ":", df[kolom].unique())
```

Hasilnya adalah untuk fitur paid, higher, famsup, activities, dan internet : no dan yes. Sedangkan fitur Mjob : at_home, health, other, services, teacher. Fitur Fjob : teacher, other, services, health, at_home, dan fitur guardian : mother, father, dan other.

b. Encoding Data

Setelah tahapan pembersihan data dan seleksi fitur, selanjutnya adalah proses encoding data. Pada tahap ini bertujuan untuk mengonversi atribut kategorikal yang disajikan dalam bentuk teks diubah ke bentuk numerik, untuk memudahkan pengolahan data di tahap berikutnya. Proses encoding data sangat

penting dikarenakan dalam implementasi algoritma pendekatan Educational Data Mining umumnya menggunakan input data dalam format numerik agar dapat diproses secara maksimal. (Agung et al., 2023).

Berdasarkan dataset yang digunakan terdapat beberapa stribut atau fitur bernilai kategorikal yakni school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, dan romantic. Proses encoding ini disesuaikan dengan karakteristik masing-masing atribut. Fitur dengan kategori yes/no, dikonversi dengan binary encoding, sehingga untuk kategori yes disajikan dengan nilai 1 dan no disajikan dengan nilai 0. Sementara itu, atribut dengan lebih dari dua kategori dan tidak memiliki tingkatan, seperti Mjob, Fjob, reason, dan guardian, dikonversi menggunakan one-hot encoding untuk menghindari pemberian makna urutan pada kategori.

Tabel 2**.** Hasil Proses Encoding Data

No.	**Fitur**	**Kelompok**	**Sebelum Encoding**	**Hasil Encoding**
1.	Studytime	Akademik	1,2,3,4	1,2,3,4
2.	Failure	Akademik	0,1,2,3	0,1,2,3
3.	Absences	Akademik	Numerik	Numerik
4.	Famsup	Sosial	No,Yes	No = 0 Yes = 1
5.	Activities	Sosial	No,Yes	No = 0 Yes = 1
6.	Romantic	Sosial	No,Yes	No = 0 Yes = 1
7.	Famrel	Sosial	1,2,3,4,5	1,2,3,4,5
8.	Goout	Sosial	1,2,3,4,5	1,2,3,4,5
9.	Mjob	Sosial	at_home, health, other, services, teacher	5 kolom biner
10.	Fjob	Sosial	at_home, health, other, services, teacher	5 kolom biner
11.	G3	Prestasi	0-20	0-20

Sedangkan untuk hasil encoding Mjob dan Fjob dengan 5 kolom terlampir pada tabel 3 dan 4.

Tabel 3. Hasil Proses Encoding Data Mjob

Nilai Mjob	Mjob_at_home	Mjob_health	Mjob_other	Mjob_services	Mjob_teacher
at_home	1	0	0	0	0
health	0	1	0	0	0
other	0	0	1	0	0
services	0	0	0	1	0
teacher	0	0	0	0	1

Tabel 4. Hasil Proses Encoding Data Fjob

Nilai Fjob	Fjob_at_home	Fjob_health	Fjob_other	Fjob_services	Fjob_teacher
at_home	1	0	0	0	0
health	0	1	0	0	0
other	0	0	1	0	0
services	0	0	0	1	0
teacher	0	0	0	0	1

Mengacu hasil encoding maka dihasilkan sebanyak 395 baris, 19 kolom. Adapun kolomnya adalah studytime, failures, absences, famsup, activities, romantic, famrel, goout, G3, Mjob_at_home, Mjob_health, Mjob_other, Mjob_services, Mjob_teacher, Fjob_at_home, Fjob_health, Fjob_other, Fjob_services, Fjob_teacher. Setelah diperoleh hasil encoding data, maka langkah selanjutnya adalah normalisasi data, dengan rumus pada persamaan (1).

c. Normalisasi Data

Tahap normalisasi ini akan dilakukan proses perhitungan berdasarkan hasil dari encoding sbelumnya, dengan menggunakan persamaan (1). Data yang akan diproses normalisasi harus dalam kondisi biner atau numerik, karena dalam proses normalisasi ini perhitungannya berdasarkan angka. Berikut adalah persamaan untuk proses normalisasi.

$$X' = (X - X_{min}) / (X_{maks} - X_{min}) \quad (1)$$

Keterangan

X' : nilai hasil dinormalisasi

X : nilai yang akan dihitung

X_{min} : nilai minimal yang ada dalam dataset

X_{maks} : nilai maksimal yang ada dalam dataset

Proses normalisasi dimulai dari data yang ada pada tabel 5a dan 5b.

Tabel 5a. Data Hasil Encoding Fitur yang Digunakan

No.	Studytime	Failures	Absences	Famsup	Activities	Romantic	Famrel	Goout
1.	2	0	6	0	0	0	4	4
2.	2	0	4	1	0	0	5	3
3.	2	3	10	0	0	0	4	2
4.	3	0	2	1	1	1	3	2
5.	2	0	4	1	0	0	4	2
6.	2	0	10	1	1	0	5	2
7.	2	0	0	0	0	0	4	4
8.	2	0	6	1	0	0	4	4
9.	2	0	0	1	0	0	4	2
...	...	...	...	...	...	...	...	...
394.	1	0	0	0	0	0	4	1
395.	1	0	5	0	0	0	3	3

Tabel 5b. Data Hasil Encoding Fitur yang Digunakan (lanjutan)

No.	G3	Mjob_at_home	Mjob_health	Mjob_other	Mjob_services	Mjob_teacher
1.	6	1	0	0	0	0
2.	6	1	0	0	0	0
3.	10	1	0	0	0	0
4.	15	0	1	0	0	0
5.	10	0	0	1	0	0
6.	15	0	0	0	1	0
7.	11	0	0	1	0	0

8.	6	0	0	1	0	0
9.	19	0	0	0	1	0
...	...	...	...	...	...	...
394.	10	0	0	0	1	0
95.	9	0	0	1	0	0

Merujuk isi data pada Tabel 5a dan 5b, dan dengan menggunakan persamaan (1), maka proses perhitungan normalisasinya adalah :

Fitur Studytime, nilai awal = 2, nilai minimum = 1, dan nilai maksimal = 4.

Nilai normalisasi untuk studytime = (2-1)/(4-1) = 0,33333.

Fitur Failures, nilai awal = 0, nilai minimum = 0, nilai maksimal = 3.

Nilai normalisasi untuk fitur failures = (0-0)/(3-0) = 0.

Fitur Absences, nilai awal = 6, nilai minimum = 0, nilai maksimal = 75.

Nilai normalisasi untuk fitur failures = (6-0)/(75-0) = 0,08.

Fitur Famrel, nilai awal = 4, nilai minimum = 1, nilai maksimal = 5.

Nilai normalisasi untuk fitur failures = (4-1)/(5-1) = 0,75.

Fitur Goout, nilai awal = 4, nilai minimum = 1, nilai maksimal = 5.

Nilai normalisasi untuk fitur failures = (4-1)/(5-1) = 0,75.

Pemodelan Data

Pembagian data training dan data testing

Tahapan selanjutnya setelah proses dalam pembersihan data, adalah tahap pemodelan data. Pada tahap inilah prose membangun model klasifikasi prestasi akademik peserta didik. Adapun fitur atau variabel target yang digunakan adalah G3, yang akan dikategorikan menjadi tiga level prestasi yakni prestasi rendah, prestasi sedang, dan prestasi tinggi. Untuk kategori prestasi rendah adalah nilai G3 kurang dari 10, kategori prestasi sedang nilai G3 antara 10 sampai dengan 14, sedangkan kategori prestasi tinggi adalah nilai G3 sebesar 15 atau lebih. Dataset selanjutnya dipisahkan menjadi **data training dan data testing dengan proporsi 80:20**. Dari total **395 data siswa, sebanyak 316 data digunakan untuk training**, sedangkan **79 data digunakan untuk testing**. Untuk membangun dan mempelajari pola klasifikasi oleh model Random Forest akan digunakan data training, sedangkan fungsi dari data testing adalah mengevaluasi kinerja model terhadap data yang tidak digunakan selama proses pelatihan (Joseph & Vakayil, 2022).

Proses Pemodelan dengan Random Forest

Penentuan Parameter

Proses perhitungan pemodelan menggunakan Random Forest, dengan beberapa parameter yang digunakan seperti Tabel

Tabel 6. Parameter dalam Proses Random Forest

No.	Parameter	Mjob_at_home
1.	Jumlah data	395
2.	Data training	316
3.	Data Testing	79
4.	Jumlah decision tree	100
5.	Random state	42
6.	Target	G3
7.	Pembagian target	G3 < 10 = Rendah; 10–14 = Sedang; G3 ≥ 15 = Tinggi

Mengacu parameter dalam random forest tersebut, dihasilkan tabel terlampir pada Gambar 3 dan 4.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		studytime	failures	absences	famrel	goout	freetime	Medu	Fedu	paid	higher	famsup	activities	internet
2	1	0,333333	0	0,053333	1	0,5	0,5	0,25	0,25	0	1	1	0	1
3	2	0,333333	1	0,133333	0,75	0,25	0,5	0,25	0,25	1	1	0	0	1
4	7	0,333333	0	0,08	0,75	0,75	0	1	1	0	1	1	0	0
5	9	0,333333	0	0	1	0	1	0,75	1	1	1	1	1	1
6	12	0	0	0,026667	0,75	0,5	0,5	1	1	1	1	1	1	1
7	28	0,333333	0	0,053333	1	0,5	0,5	0,75	1	0	1	1	1	1
8	30	0,333333	0	0	1	0,25	0,75	1	1	1	1	1	0	1
9	53	0	0	0	0,5	0,75	0,5	1	1	1	1	1	0	1
10	54	0	0	0,08	1	0,75	0,5	0,75	0,75	1	1	0	0	1
11	55	0,333333	0	0,106667	1	0,75	0,5	0,5	0,25	1	1	0	1	1
12	56	0,333333	0	0	0,75	0,25	0,5	1	0,75	1	1	1	1	1
13	59	0,333333	0	0,026667	0,75	0,5	0,25	1	0,5	0	1	1	0	1
14	67	1	0	0,053333	0,75	0,5	0,5	0,75	0,25	1	1	1	0	1

Gambar 3. Hasil Random Forest

O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
ob_at_hor	ljob_healt	Mjob_othe	job_servic	job_teach	ob_at_hor	job_healt	Fjob_othe	ob_servic	ob_teache	rdian_fat	rdian_mot	rdian_oth	Aktual	Prediksi
1	0	0	0	0	0	0	1	0	0	1	0	0	Rendah	Sedang
1	0	0	0	0	0	0	1	0	0	0	1	0	Sedang	Rendah
0	0	1	0	0	0	0	0	0	1	0	1	0	Rendah	Sedang
0	0	1	0	0	0	0	1	0	0	0	1	0	Tinggi	Sedang
0	1	0	0	0	0	0	0	1	0	1	0	0	Sedang	Sedang
0	0	0	1	0	0	0	1	0	0	0	1	0	Sedang	Sedang
0	1	0	0	0	0	0	0	1	0	0	1	0	Sedang	Sedang
0	0	0	1	0	0	0	0	1	0	0	1	0	Sedang	Sedang
0	0	1	0	0	0	0	1	0	0	0	1	0	Sedang	Sedang
0	0	1	0	0	0	0	1	0	0	0	1	0	Sedang	Sedang
0	0	0	1	0	0	0	0	1	0	0	1	0	Tinggi	Sedang
0	0	0	1	0	0	0	1	0	0	0	1	0	Tinggi	Sedang

Gambar 4. Hasil Random Forest (lanjutan)

Melalui Gambar 3 dan 4, dapat diketahui bahwa komponen dan nilai untuk proses Accuracy adalah terlampir pada Tabel 9. Proses perhitungan accuracy menggunakan persamaan (2).

$$\text{Accuracy} = \frac{\text{Jumlah prediksi benar}}{\text{Jumlah seluruh data testing}} \quad (2)$$

Dengan menerapkan persamaan (2) diperoleh data jumlah data testing=79, jumlah prediksi benar = 42, prediksi salah=37, hasilnya adalah 0,5316 dan prosentasenya adalah 53,16. Selanjutnya proses evaluasi dengan precision menggunakan persamaan (3).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

TP *(*True Positive*)* adalah data yang **termasuk dalam suatu kelas dan berhasil diprediksi dengan benar**, sedangkan FP (False Positive) merupakan data yang tidak **termasuk kelas tersebut, tetapi diprediksi sebagai kelas tersebut oleh model** (Rinjani Bawana et al., 2025).

Precision digunakan untuk menilai **seberapa tepat model dalam memprediksi suatu kelas**, dengan membandingkan jumlah prediksi positif yang benar terhadap seluruh data yang diprediksi sebagai positif. Tingginya nilai *precision* menandakan bahwa sebagian besar data yang diprediksi dalam suatu kelas memang berasal dari kelas tersebut (Foody, 2023). Kemudian hasil akhir akan ditampilkan dalam Tabel 7.

Tabel 7. Hasil Akhir Evaluasi

Kelas	TP	FP	FN	TN	Precision	Recall	F1-score
Rendah	9	5	17	48	0,6428	0,3461	0,45
Sedang	32	29	6	12	0,5245	0,8421	0,6464
Tinggi	1	3	14	61	0,25	0,0666	0,1025

KESIMPULAN

Penelitian ini menggunakan pendekatan Educational Data Mining (EDM) untuk mengklasifikasikan prestasi siswa berdasarkan faktor akademik dan sosial dengan algoritma Random Forest. Proses penelitian meliputi data cleaning, seleksi fitur, encoding, normalisasi, Pembagian data menjadi training dan testing, dilanjutkan dengan proses pemodelan dan evaluasi menggunakan confusion matrix, accuracy, precision, recall, serta F1-score. Pengujian yang dilakukan memperlihatkan bahwa Random Forest mampu mengklasifikasikan prestasi siswa dalam kategori rendah, sedang, dan

tinggi dengan accuracy sebesar 53,16%. Model paling baik mengenali kategori sedang, sedangkan kategori tinggi masih sulit diklasifikasikan. Hal ini menunjukkan bahwa kinerja model belum merata pada seluruh kategori. Dengan demikian, Random Forest dapat menjadi pendekatan awal dalam menganalisis pola prestasi siswa berdasarkan faktor akademik dan sosial. Peningkatan kinerja dapat dilakukan melalui optimasi parameter, penanganan ketidakseimbangan kelas, dan analisis feature importance untuk mengidentifikasi atribut yang paling berpengaruh.

DAFTAR PUSTAKA

Agung, A., Daniswara, A., Kadek, I., & Nuryana, D. (2023). Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru. Journal of Informatics and Computer Science, 05.

Ardiansyah, & Jailani, Ms. (2023). Teknik Pengumpulan Data Dan Instrumen Penelitian Ilmiah. IHSANJurnal Pendidikan Islam, 1–9. http://ejournal.yayasanpendidikandzurriyatulquran.id/index.php/ihsan

Armawan, I., Tinggi, S., Dakwah, I., Komunikasi, D., & Bogor, I. (2022). PENGARUH KOMUNIKASI DAN KARAKTERISTIK PESERTA DIDIK TERHADAP KUALITAS PEMBELAJARAN DI STIDKI BOGOR. TABAYYUN: Jurnal Komunikasi Dan Penyiaran Islam, 3(1). https://ejournal-stidkibogor.ac.id/index.php/tabayyun

Foody, G. M. (2023). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. PLoS ONE, 18(10 October). https://doi.org/10.1371/journal.pone.0291908

Guntara, M., & Astuti, F. D. (2025). Komparasi Kinerja Label-Encoding dengan One-Hot-Encoding pada Algoritma K-Nearest Neighbor menggunakan Himpunan Data Campuran. JIKO (Jurnal Informatika Dan Komputer), 9(2), 352. https://doi.org/10.26798/jiko.v9i2.1605

Hidayat, I., Elektro, T., Deddy, R., Dako, R., & Ilham, J. (n.d.). Volume 5 Nomor 2 Juli 2023 Analisis Data Eksploratif Capaian Indikator Kinerja Utama 3 Fakultas Teknik. Jambura Journal of Electrical and Electronics Engineering, 185.

Indahyanti, U., Azizah, N. L., & Setiawan, H. (2022). Educational Data Mining on Student Academic Performance Prediction: A Survey Educational Data Mining Pada Prediksi Kinerja Akademik Mahasiswa : Sebuah Survey. https://ieeexplore.ieee.org/

Joseph, V. R., & Vakayil, A. (2022). SPlit: An Optimal Method for Data Splitting. Technometrics, 64(2), 166–176. https://doi.org/10.1080/00401706.2021.1921037

Marín, Y. R., Huatangari, L. Q., Tuesta, J. N. A., Caro, O. C., Guevara, J. L. M., Bardales, E. S., & Santos, R. C. (2025). Analysis of factors affecting the academic performance of university students using machine learning. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-28870-1

Rinjani Bawana, E., Seniwati, E., Kristami, P. F., Martin, A. M., & Hartanti, N. T. (2025). Studi Kinerja Algoritma K-Nearest Neighbors (KNN) untuk Klasifikasi Pasien Diabetes (Vol. 4, Number 2). https://subset.id/index.php/IJCSR

Roslan, M. H. bin, & Chen, C. J. (2022). Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). International Journal of Emerging Technologies in Learning, 17(5), 147–179. https://doi.org/10.3991/ijet.v17i05.27685

Sahputri, D., Maulita, Y., Studi Sistem Informasi, P., & KAPUTAMA Binjai, S. (2023). DATA MINING USING K-MEANS CLUSTERING ALGORITHM TO KNOW STUDENTS' INTERESTS AND TALENTS IN EXTRACURRICULAR ACTIVITIES (Case Study: SMK Setia Budi Binjai). Journal of Engineering, Technology and Computing, 2(3), 107–115.

Sarker, S., Paul, M. K., Thasin, S. T. H., & Hasan, M. A. M. (2024). Analyzing students' academic performance using educational data mining. Computers and Education: Artificial Intelligence, 7. https://doi.org/10.1016/j.caeai.2024.100263

Sutriyanti, & Mohammad Muspawi. (2024). Jenis jenis Data dalam Ilmu Pendidikan Pendekatan Kualitatif dan Kuantitatif. Edu Research IICLS, 5(5), 195–204.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1). https://doi.org/10.1186/s40561-022-00192-z