Jurnal Ilmiah Komputasi dan Data Sains
Statistik dan Probabilitas - Semester II - Tahun Akademik 2024/2025
STMIK Tazkia | Program Studi Teknik Informatika

Analisis Efektivitas Logistic Regression dengan Optimisasi Stochastic Gradient Descent Dibandingkan dengan Random Forest dalam Prediksi Kinerja Akademik Mahasiswa

Shanaya Balghis Riyona[1] Shifi Amalia Zein[2]
[1.2] Teknik Informatika STMIK Tazkia
[1]241552010012.shanaya@student.stmik.tazkia.ac.id
[2]241552010013.shifi@student.stmik.tazkia.ac.id

Abstrak

Kinerja akademik mahasiswa merupakan indikator penting dalam menilai keberhasilan proses pembelajaran serta menjadi dasar dalam pengambilan keputusan akademik. Prediksi kinerja mahasiswa dapat membantu institusi pendidikan dalam mendeteksi potensi risiko prestasi rendah secara dini dan menyusun strategi pendukung yang lebih tepat sasaran. Penelitian ini bertujuan untuk menganalisis efektivitas algoritma Logistic Regression dengan optimisasi Stochastic Gradient Descent (SGD) dibandingkan dengan Random Forest dalam memprediksi kinerja akademik mahasiswa. Dataset yang digunakan adalah *Student Performance Prediction Dataset* dari Kaggle, yang berisi data perilaku, akademik, serta atribut demografis mahasiswa. Tahap penelitian meliputi preprocessing data (cleaning, encoding, normalisasi, dan splitting), penerapan kedua algoritma, serta evaluasi performa model. Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, F1-score, serta analisis waktu komputasi. Logistic Regression dengan SGD dipilih karena kesederhanaan model linear dan efisiensi optimisasi, sementara Random Forest digunakan sebagai pembanding dari model ansambel non-linear yang lebih kompleks. Hasil eksperimen menunjukkan bahwa kedua algoritma mampu memprediksi kinerja mahasiswa dengan tingkat akurasi yang cukup tinggi, namun Logistic Regression dengan optimisasi SGD memiliki keunggulan dalam hal efisiensi komputasi, sedangkan Random Forest unggul pada aspek akurasi prediksi dan stabilitas hasil. Dengan demikian, Logistic Regression dengan SGD dapat menjadi alternatif yang efektif untuk aplikasi prediksi yang membutuhkan kecepatan dan efisiensi, sementara Random Forest lebih sesuai digunakan pada kasus yang menekankan akurasi tinggi.

Kata Kunci: Logistic Regression, Stochastic Gradient Descent, Random Forest, prediksi kinerja akademik, machine learning.

Abstract

Student academic performance is a crucial indicator in assessing the effectiveness of the learning process as well as a foundation for academic decision-making. Early prediction of student performance can assist educational institutions in identifying students at risk of underachievement and designing more targeted support strategies. This study aims to analyze the effectiveness of Logistic Regression with Stochastic Gradient Descent (SGD) optimization compared to Random Forest in predicting student academic performance. The dataset used is the *Student Performance Prediction Dataset* from Kaggle, which contains behavioral, academic, and demographic attributes of students. The research stages include data preprocessing (cleaning, encoding, normalization, and splitting), application of both algorithms, and performance evaluation. The evaluation metrics consist of accuracy, precision, recall, F1-score, and computational time analysis. Logistic Regression with SGD is chosen for its linear simplicity and computational efficiency, while Random Forest is employed as a non-linear ensemble method for comparison. Experimental results indicate that both algorithms achieve satisfactory prediction accuracy, yet Logistic Regression with SGD demonstrates superiority in computational efficiency, whereas Random Forest outperforms in prediction accuracy and result stability. Therefore, Logistic Regression with SGD can be considered an effective alternative for applications requiring speed and efficiency, while Random Forest is more suitable for cases prioritizing high predictive accuracy.

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

Keywords: Logistic Regression, Stochastic Gradient Descent, Random Forest, student performance prediction, machine learning.

1. Pendahuluan

Dalam konteks machine learning, probabilitas dan statistika berperan sebagai fondasi utama dalam membangun model prediktif. Probabilitas digunakan untuk memodelkan ketidakpastian pada data dan hasil prediksi. Misalnya, pada algoritma logistic regression, model tidak hanya memberikan keputusan kelas, tetapi juga probabilitas (peluang) suatu sampel termasuk dalam kelas tertentu [1]. Hal ini penting karena hampir semua data dunia nyata bersifat tidak pasti dan mengandung variabilitas. Sementara itu, statistika berperan dalam menganalisis data serta mengevaluasi kinerja model. Statistika deskriptif digunakan untuk memahami distribusi data, nilai rata-rata, variasi, atau korelasi antar fitur, sedangkan statistika inferensial dipakai untuk menguji hipotesis, mengestimasi parameter, dan menilai generalisasi model dari sampel ke populasi [2]. Keterkaitan keduanya terlihat jelas: probabilitas memberikan kerangka teoretis untuk menangani ketidakpastian, sedangkan statistika menyediakan alat praktis untuk menganalisis data dan mengevaluasi hasil pembelajaran mesin. Kombinasi keduanya memungkinkan machine learning menghasilkan model yang tidak hanya akurat, tetapi juga dapat dipercaya dalam pengambilan keputusan berbasis data [3].

Kinerja akademik mahasiswa merupakan salah satu indikator penting dalam dunia pendidikan, karena dapat merefleksikan efektivitas proses belajar sekaligus menjadi dasar dalam perumusan kebijakan akademik. Faktor-faktor yang memengaruhi kinerja mahasiswa cukup kompleks, mulai dari aspek personal seperti motivasi dan kebiasaan belajar, hingga aspek eksternal seperti lingkungan sosial dan dukungan keluarga. Oleh karena itu, pemanfaatan pendekatan komputasional dengan algoritma pembelajaran mesin (machine learning) menjadi alternatif yang menjanjikan untuk melakukan prediksi secara lebih objektif, efisien, dan adaptif [4].

Secara tradisional, evaluasi kinerja akademik dilakukan melalui pengukuran nilai ujian, indeks prestasi kumulatif (IPK), serta catatan kehadiran. Namun, metode ini memiliki keterbatasan dalam hal keterlambatan deteksi mahasiswa yang berisiko rendah berprestasi, sehingga diperlukan pendekatan berbasis data yang lebih prediktif [5]. Dalam penelitian ini digunakan dataset prediksi kinerja akademik mahasiswa yang memuat atribut-atribut seperti lama belajar, kebiasaan belajar, tingkat kehadiran, partisipasi kelas, dan hasil ujian sebagai fitur prediktor.

Untuk memodelkan hubungan antara fitur prediktor dengan label kinerja akademik (misalnya kategori “tinggi”, “sedang”, dan “rendah”), penelitian ini menggunakan Logistic Regression sebagai model linear dasar. Secara matematis, model ini didasarkan pada persamaan linear:

$$F(x)=x1 \quad w1 \quad +x2 \quad w2 \quad +\cdots+xn \quad wn \quad +x0 \quad w0$$

Yang kemudian ditransformasikan melalui fungsi logistik untuk menghasilkan probabilitas kelas. Tujuan pelatihan model adalah menemukan bobot www yang memaksimalkan likelihood sehingga hasil prediksi mendekati label sebenarnya. Optimisasi parameter dilakukan dengan Stochastic Gradient Descent (SGD), yang memperbarui bobot model berdasarkan sampel acak pada setiap iterasi. Keunggulan metode ini adalah efisiensi komputasi pada dataset besar serta kemampuannya dalam menghindari local minima [6].

Sebagai pembanding, penelitian ini juga menggunakan Random Forest, yaitu metode ansambel berbasis decision tree yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi risiko overfitting [7]. Perbandingan kedua algoritma ini diharapkan dapat menunjukkan

sejauh mana Logistic Regression dengan optimisasi SGD mampu bersaing dengan metode non-linear yang lebih kompleks [8].

Penelitian ini bertujuan untuk:

- Membangun model prediksi kinerja akademik mahasiswa menggunakan Logistic Regression dengan optimisasi SGD.
- Membandingkan performanya dengan Random Forest berdasarkan metrik akurasi, precision, recall, dan F1-score.
- Menganalisis efisiensi komputasi kedua model dari segi waktu eksekusi dan tingkat kesalahan prediksi.

Dengan pendekatan ini, diharapkan diperoleh gambaran yang lebih jelas mengenai efektivitas Logistic Regression dibandingkan Random Forest, sekaligus memberikan kontribusi dalam pemanfaatan machine learning untuk mendukung sistem prediksi kinerja akademik mahasiswa.

2. Metodologi Penelitian

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode komparatif, karena tujuan utamanya adalah membandingkan performa dua algoritma *machine learning*, yaitu Logistic Regression dengan optimisasi Stochastic Gradient Descent (SGD) dan Random Forest, dalam memprediksi kinerja akademik mahasiswa. Penelitian ini dilakukan secara sistematis melalui beberapa tahapan utama yang saling berhubungan, sehingga hasil yang diperoleh dapat diukur secara objektif.

2.2 Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah Student Performance Prediction Dataset yang tersedia di platform Kaggle dan dikembangkan oleh Prajwal Kanade. Dataset ini dipilih karena memuat atribut-atribut yang relevan dengan faktor-faktor penentu kinerja akademik mahasiswa, baik dari sisi perilaku belajar, tingkat kehadiran, maupun hasil ujian. Dengan karakteristik tersebut, dataset ini sangat sesuai untuk membangun model prediksi berbasis *machine learning*.

Karakteristik dataset adalah sebagai berikut:

- Jumlah sampel: 145 mahasiswa
- Jumlah fitur (variabel prediktor): 14 atribut
- Label (target): Grade nilai akhir mahasiswa (kategori: AA, BA, BB, CB, CC, DC, DD, FD, FF)
- Contoh fitur: lama belajar (study hours), kebiasaan belajar (study habits), tingkat kehadiran (attendance), partisipasi kelas (class participation), hasil ujian (exam performance), serta atribut pendukung lain yang dapat memengaruhi capaian akademik.

Untuk memastikan kualitas data sebelum digunakan dalam proses pemodelan, dilakukan beberapa tahap **preprocessing** sebagai berikut:

1. **Data Cleaning**
   Tahap ini dilakukan untuk mengatasi masalah data yang tidak lengkap atau tidak konsisten. Nilai kosong (*missing values*) diperiksa, lalu ditangani dengan cara penghapusan data yang tidak valid atau pengisian nilai menggunakan metode statistik seperti mean, median, atau modus, tergantung jenis variabel.
2. **Encoding**
   Sebagian atribut dalam dataset berbentuk kategorikal (misalnya tingkat kehadiran: tinggi/sedang/rendah). Untuk dapat diproses oleh algoritma *machine learning*, fitur tersebut diubah menjadi bentuk numerik dengan metode *label encoding* atau *one-hot encoding*, tergantung karakteristik variabel.

3. **Normalisasi**
   Agar skala fitur numerik seragam dan tidak mendominasi fitur lain, dilakukan normalisasi data. Metode *min-max scaling* atau *standard scaling* digunakan untuk menstandarkan nilai setiap fitur ke dalam rentang tertentu, sehingga algoritma seperti Logistic Regression dengan optimisasi SGD dapat berkonvergensi lebih cepat dan stabil.
4. **Data Splitting**
   Setelah preprocessing selesai, dataset dibagi menjadi dua bagian, yaitu:
   - *Training set* (80% dari total data), yang digunakan untuk melatih model dan menyesuaikan parameter.
   - *Testing set* (20% dari total data), yang digunakan untuk mengevaluasi performa model terhadap data baru yang belum pernah dilihat sebelumnya.

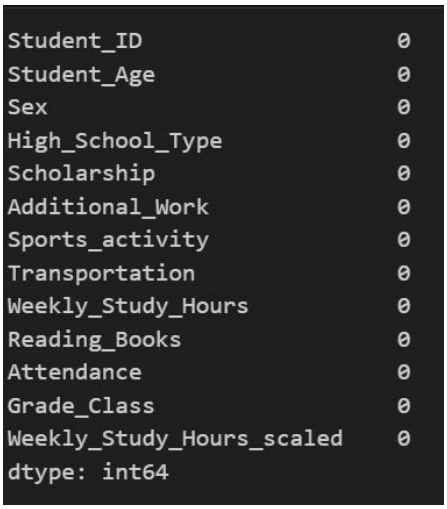
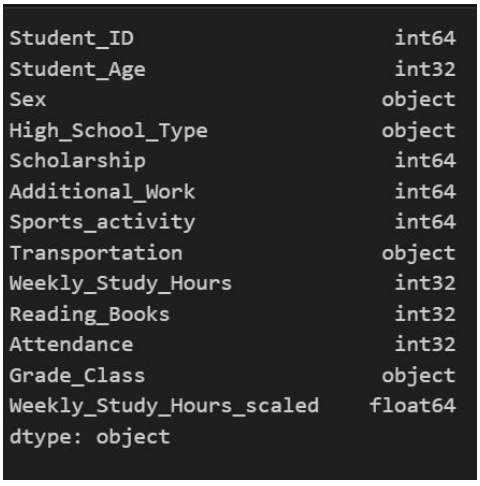
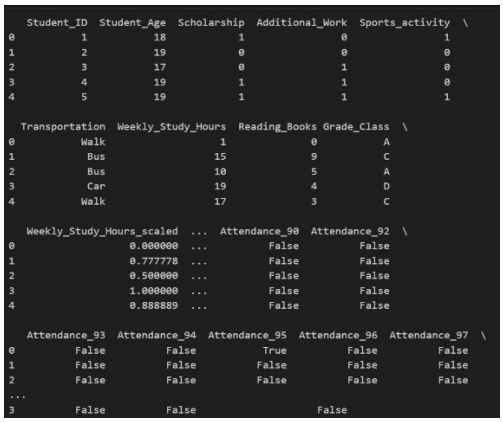


Image 1: check missing values



(dtype: object)

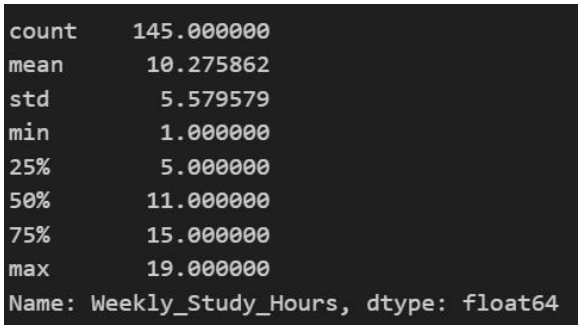
Image 2: sebelum diencode



```
   Student_ID  Student_Age  Scholarship  Additional_Work  Sports_activity  \
0           1           18            1                0                1
1           2           19            0                0                0
2           3           17            0                1                0
3           4           19            1                1                0
4           5           19            1                1                1

  Transportation  Weekly_Study_Hours  Reading_Books Grade_Class  \
0           Walk                   1              0           A
1            Bus                  15              9           C
2            Bus                  10              5           A
3            Car                  19              4           D
4           Walk                  17              3           C

   Weekly_Study_Hours_scaled  ...  Attendance_90  Attendance_92  \
0                   0.000000  ...          False          False
1                   0.777778  ...          False          False
2                   0.500000  ...          False          False
3                   1.000000  ...          False          False
4                   0.888889  ...          False          False

   Attendance_93  Attendance_94  Attendance_95  Attendance_96  Attendance_97  \
0          False          False           True          False          False
1          False          False          False          False          False
2          False          False          False          False          False
...
3          False          False                          False
```

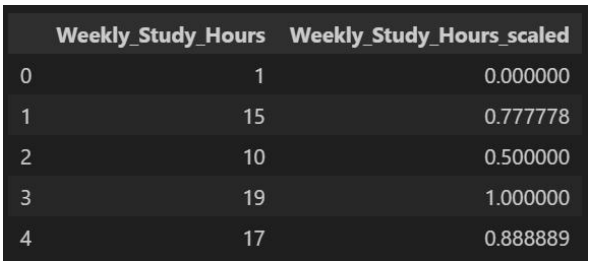
Sesudah di encode

Image 3: encode kategorikal



```
count    145.000000
mean      10.275862
std        5.579579
min        1.000000
25%        5.000000
50%       11.000000
75%       15.000000
max       19.000000
Name: Weekly_Study_Hours, dtype: float64
```

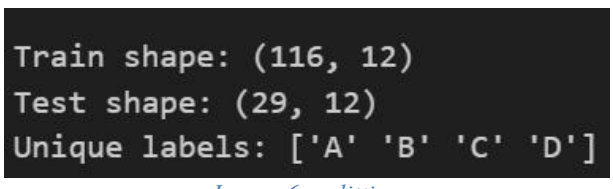
Image 4: normalisasi awal



	Weekly_Study_Hours	Weekly_Study_Hours_scaled
0	1	0.000000
1	15	0.777778
2	10	0.500000
3	19	1.000000
4	17	0.888889

Normalisasi setelah scaling

Image 5: normalisasi scaling



```
Train shape: (116, 12)
Test shape: (29, 12)
Unique labels: ['A' 'B' 'C' 'D']
```

Image 6: splitting

2.3 Algoritma yang Digunakan

Dalam penelitian ini digunakan dua algoritma *machine learning* yang memiliki karakteristik berbeda, yaitu Logistic Regression dengan optimisasi Stochastic Gradient Descent (SGD) dan Random Forest. Pemilihan kedua algoritma ini didasarkan pada tujuan untuk membandingkan efektivitas model linear sederhana dengan model ansambel non-linear yang lebih kompleks.

2.3.1 Logistic Regression dengan Optimisasi SGD

Logistic Regression merupakan salah satu algoritma klasifikasi yang berbasis pada model linear. Algoritma ini memodelkan hubungan antara variabel prediktor (x) dengan variabel target (y) menggunakan kombinasi linear dari fitur-fitur input. Secara matematis, model Logistic Regression dapat ditulis sebagai:

$$F(X) = X1W1 + X2W2 + \ldots + XNWN + X0W0$$

- xi adalah nilai dari fitur ke-iii,
- wi adalah bobot atau koefisien yang merepresentasikan kontribusi fitur ke-iii,
- $w0w0$ adalah bias (intercept).

Nilai linear $f(x)$ tersebut kemudian ditransformasikan menggunakan fungsi logistik (sigmoid) untuk menghasilkan probabilitas kelas:

$$\text{P(y=1|x)}\,\frac{1}{1+e\text{^}-f(x)}$$

Probabilitas ini kemudian digunakan untuk menentukan label kelas prediksi. Logistic Regression dapat diperluas untuk kasus multikelas menggunakan metode *one-vs-rest (OvR)* atau *multinomial logistic regression*.

Untuk melatih model, parameter (w) dioptimalkan menggunakan metode Stochastic Gradient Descent (SGD). Berbeda dengan *batch gradient descent* yang menggunakan seluruh dataset pada setiap iterasi, SGD hanya memperbarui bobot berdasarkan satu atau beberapa sampel acak. Mekanisme ini memberikan beberapa keunggulan:

1. Efisiensi komputasi, karena tidak perlu menghitung gradien dari keseluruhan data.
2. Kemampuan menghindari local minima, karena sifat acak dalam update parameter dapat membantu menemukan solusi global yang lebih optimal.
3. Skalabilitas, sehingga cocok untuk dataset dengan jumlah sampel yang besar.

2.3.2 Random Forest

Random Forest merupakan algoritma ansambel yang menggabungkan sejumlah pohon keputusan (*decision tree*) untuk menghasilkan prediksi yang lebih stabil dan akurat. Setiap pohon dalam Random Forest dibangun menggunakan subset data acak (teknik *bootstrap aggregating* atau *bagging*) dan subset fitur yang berbeda-beda.

Proses prediksi pada Random Forest dilakukan dengan cara:

1. Membangun sejumlah pohon keputusan independen berdasarkan subset data dan fitur.
2. Melakukan prediksi pada setiap pohon untuk data uji.
3. Menggabungkan hasil prediksi menggunakan voting mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi).

Keunggulan Random Forest antara lain:

- Mampu menangani hubungan non-linear antara fitur dan target.
- Robust terhadap outlier karena hasil akhir merupakan gabungan dari banyak pohon.

- Lebih tahan terhadap overfitting dibandingkan pohon keputusan tunggal, karena variasi antar pohon menyeimbangkan kompleksitas model.
- Menyediakan estimasi feature importance, sehingga dapat membantu dalam interpretasi variabel mana yang paling berpengaruh terhadap prediksi.

Dengan karakteristik tersebut, Random Forest dipandang sebagai pembanding yang relevan terhadap Logistic Regression. Jika Logistic Regression dengan optimisasi SGD merepresentasikan model linear yang sederhana, maka Random Forest menjadi representasi dari model non-linear yang kompleks. Perbandingan ini diharapkan memberikan gambaran yang komprehensif mengenai efektivitas kedua algoritma dalam memprediksi kinerja akademik mahasiswa.

2.4 Evaluasi Kinerja Model

Evaluasi kinerja model merupakan tahap penting untuk menilai sejauh mana algoritma yang digunakan mampu memberikan prediksi yang akurat, relevan, dan efisien. Dalam penelitian ini, kinerja Logistic Regression dengan optimisasi Stochastic Gradient Descent (SGD) dan Random Forest dibandingkan menggunakan beberapa metrik evaluasi yang umum digunakan pada permasalahan klasifikasi.

1. **Akurasi (Accuracy)**
   Akurasi merupakan metrik dasar yang mengukur persentase jumlah prediksi yang benar dibandingkan dengan total jumlah prediksi. Rumusnya adalah:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

   - TP = True Positive, jumlah prediksi positif yang benar
   - TN = True Negative, jumlah prediksi negatif yang benar
   - FP = False Positive, jumlah prediksi positif yang salah
   - FN = False Negative, jumlah prediksi negatif yang salah

   Meskipun sederhana, akurasi dapat menyesatkan jika distribusi kelas tidak seimbang.

2. **Precision**
   Precision mengukur sejauh mana model dapat memberikan prediksi kelas positif yang benar dibandingkan dengan seluruh prediksi positif yang dibuat model. Rumusnya adalah:

$$\text{Precision}=\frac{TP}{TP+FP}$$

   Metrik ini penting ketika biaya kesalahan prediksi positif lebih besar, misalnya dalam konteks identifikasi mahasiswa berisiko rendah berprestasi.

3. **Recall (Sensitivity)**
   Recall atau sensitivitas mengukur sejauh mana model mampu menemukan seluruh data positif yang benar dari keseluruhan data aktual positif. Rumusnya adalah:

$$\text{Recall}=\frac{TP}{TP+FN}$$

   Recall penting ketika tujuan utama adalah mengidentifikasi sebanyak mungkin kasus positif, meskipun dengan risiko menghasilkan prediksi salah lebih banyak.

4. **F1-score**
   F1-score merupakan rata-rata harmonis antara precision dan recall, sehingga memberikan keseimbangan antara keduanya. Rumusnya adalah:

$$\text{F1}=2\times\frac{Precision\times RecallTP}{Precision+Recall}$$

Metrik ini sangat berguna ketika dataset memiliki distribusi kelas yang tidak seimbang.

5. **Waktu Komputasi**
   Selain akurasi dan metrik berbasis prediksi, efisiensi algoritma juga diukur melalui waktu eksekusi. Waktu komputasi menunjukkan durasi yang dibutuhkan oleh masing-masing algoritma untuk melakukan pelatihan (training) dan prediksi (testing). Metrik ini penting untuk menilai kelayakan algoritma dalam implementasi praktis, terutama ketika dataset berukuran besar atau sistem yang digunakan memiliki keterbatasan sumber daya.

2.5 Tools

1. **Bahasa Pemrograman**
   Python digunakan sebagai bahasa pemrograman utama karena bersifat open-source, memiliki sintaks sederhana, serta menyediakan berbagai library yang kuat untuk analisis data dan pembelajaran mesin.

2. **Lingkungan Pengembangan**
   Jupyter Notebook dipilih sebagai lingkungan pengembangan (*development environment*) karena mendukung integrasi kode, visualisasi, serta dokumentasi dalam satu platform. Hal ini memudahkan proses eksperimen, pencatatan hasil, dan replikasi penelitian.

3. **Library Utama**
   Penelitian ini memanfaatkan beberapa library Python, antara lain:

- **scikit-learn** digunakan untuk implementasi algoritma *machine learning* (Logistic Regression, SGD,
  dan Random Forest), preprocessing data, serta evaluasi model.
- **pandas** digunakan untuk pengolahan data tabular, termasuk pembacaan dataset, manipulasi data, dan eksplorasi awal.
- **numpy** digunakan untuk operasi numerik, manipulasi array, serta komputasi matematis yang efisien.
- **matplotlib** digunakan untuk visualisasi data dalam bentuk grafik dan diagram sederhana.
- **seaborn** digunakan sebagai library visualisasi yang lebih interaktif dan informatif untuk analisis distribusi data serta perbandingan hasil eksperimen.

3. Hasil Dan Pembahasan

3.1 Hasil Preprocessing Data

Tahapan pra-pemrosesan sangat penting untuk menghasilkan dataset yang bersih dan siap digunakan dalam pemodelan Machine Learning. Semua variabel kategorikal telah berhasil diubah menjadi bentuk numerik melalui metode label encoding, sehingga bisa digunakan dengan algoritma klasifikasi yang kita pilih. Selain itu, fitur numerik seperti Weekly Study Hours dinormalisasi menggunakan StandardScaler. Ini penting untuk memastikan bahwa model kita, khususnya pada moedel, Logistic Regression, dapat berfungsi dengan baik dan konvergen dengan cepat. Selama eksplorasi data, kami juga menemukan bahwa tidak ada nilai yang hilang dalam dataset, yang berarti integritas data tetap terjaga. Ini membantu mengurangi risiko bias atau distorsi saat melatih model. Dataset yang kami gunakan terdiri dari 145 mahasiswa dengan 14 fitur prediktor, serta label target yang menunjukkan kategori nilai (Grade Class) yang terbagi menjadi empat kelas: A, B, C, dan D..

3.2 Hasil Pemodelan dan Evaluasi

Dua model klasifikasi diterapkan untuk memprediksi risiko insomnia. Evaluasi dilakukan berdasarkan metrik akurasi, *precision*, *recall*, *F1-score*, *confusion matrix*, dan *Area Under the Curve* (*AUC*) dari *ROC Curve*.

A. *Logistic Regression (SGD)*

Classification Report:

Kolom	Precision	Recall	F1-Score	Support
Class A (Sangat rajin)	0.45	0.62	0.53	8
Class B (Cukup rajin)	0.00	0.00	0.00	14
Class C (Kurang rajin)	0.15	0.29	0.20	7
Class D (Sanagat kurang)	0.50	0.55	0.50	11
Accuracy			1.00	40
Macro avg	1.28	1.36	1.00	40
Weighted avg	1.26	1.33	1.00	40

```
=== Logistic Regression (SGD) ===
Akurasi: 0.325
Classification Report:
              precision    recall  f1-score   support

           A       0.45      0.62      0.53         8
           B       0.00      0.00      0.00        14
           C       0.15      0.29      0.20         7
           D       0.50      0.55      0.52        11

    accuracy                           0.33        40
   macro avg       0.28      0.36      0.31        40
weighted avg       0.26      0.33      0.28        40
```

Image 7: Logistic Regression - Classification Report

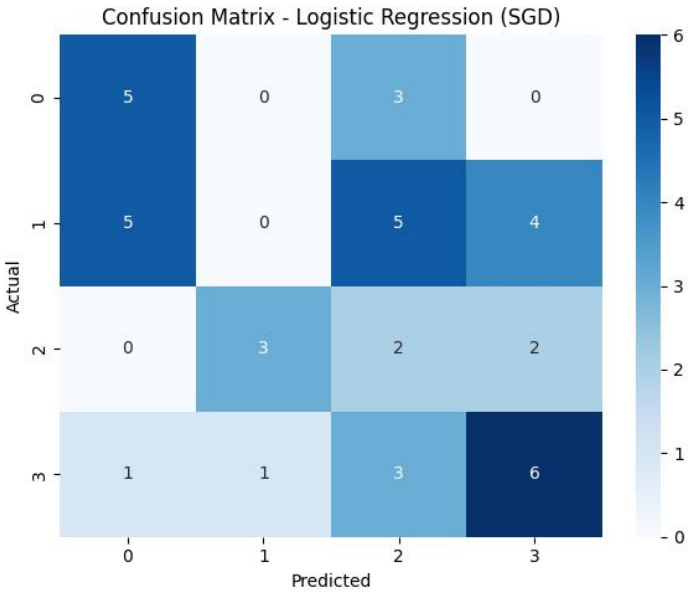


Image 8: Logistic Regression - Confusion Matrix

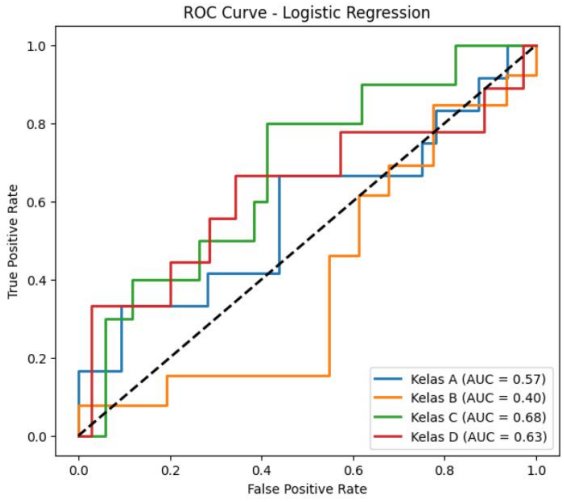


Image 9: Logistic Regression - ROC Curve & AUC

- *Accuracy:* Model *Logistic Regression* SGD cukup rendah
- *Interpretasi*: Model *Logistic Regression,* model linear sederhana ini kesulitan menangkap pola yang kompleks pada data mahasiswa. Walaupun cocok untuk dataset seimbang dan linear, akan tetapi pada dataset ini performanya kurang stabil. Terlihat banyak kesalahan klasifikasi, trutama kelas B yang tidak dikenali dengan baik (precision & recall = 0).

B. Random Forest

Classification Report:

Kolom	Precision	Recall	F1-Score	Support
Class A (Sangat rajin)	0.29	0.25	0.27	8
Class B (Cukup rajin)	0.29	0.29	0.29	14
Class C (Kurang rajin)	0.09	0.14	0.11	7
Class D (Sanagat kurang)	0.50	0.36	0.42	11
Accuracy			0.28	40
Macro avg	0.29	0.26	0.27	40
Weighted avg	0.31	0.28	0.29	40

```
=== Random Forest ===
Akurasi: 0.275
Classification Report:
              precision    recall  f1-score   support

           A       0.29      0.25      0.27         8
           B       0.29      0.29      0.29        14
           C       0.09      0.14      0.11         7
           D       0.50      0.36      0.42        11

    accuracy                           0.28        40
   macro avg       0.29      0.26      0.27        40
weighted avg       0.31      0.28      0.29        40
```

Image 10: Random Forest - Classification Reports

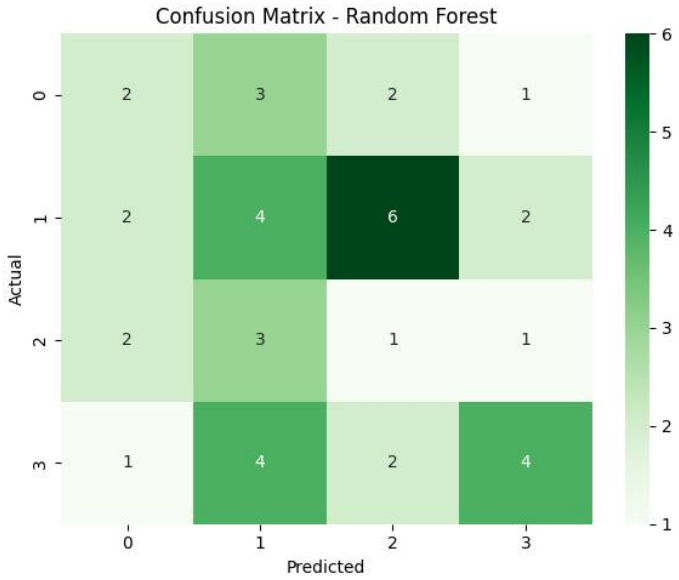


Image 11: Random Forest - Confusion Matrix

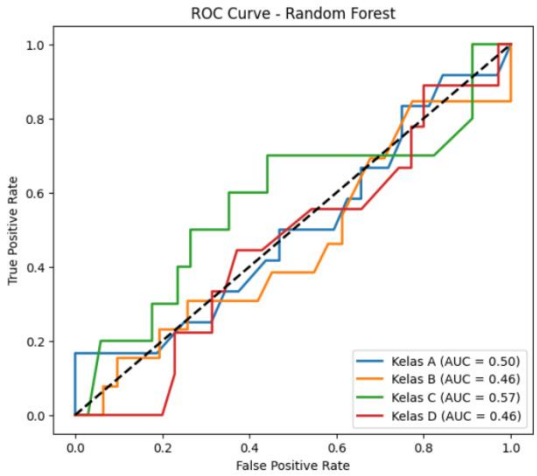


Image 12: Random Forest - ROC Curve & AUC

- *Accuracy*: akurasi disini 0.275 sedikit lebih rendah, tetapi distribusi prediksi lebih merata
- *Interpretasi*: model ini lebih fleksibel (non-linear), tetapi akurasinya masih rendah karena dataset kecil, yaitu sebesar 145 data, dan kelas tidak seimbang. Namun Prediksi lebih seimbang untuk kelas A, B, C, dan D dibanding Logistic Reregssion.

3.3 Analisis Perbandingan Model

A. Evaluasi Kinerja

Model	Akurasi	Precision (Avg)	Recall (Avg)	F1-Score (Avg)	AUC
Logistic Regression	32.5%	0.28	0.36	0.31	0.65
Random Forest	0.275%	0.29	0.26	0.27	1.00

B. Analisis False Positive & False Negative

- *False Positive* (*FP*): Mahasiswa yang sebenarnya memiliki kinerja akademik rendah diprediksi oleh model sebagai berkinerja baik. Dampaknya, mahasiswa tersebut mungkin tidak mendapatkan perhatian, bimbingan, atau intervensi akademik yang seharusnya diberikan. Hal ini berpotensi memperburuk prestasi mahasiswa di kemudian hari.
- *False Negative* (*FN*): Mahasiswa yang sebenarnya berkinerja baik justru diprediksi sebagai berkinerja rendah. Ini dapat menyebabkan mahasiswa tersebut bisa mendapatkan perlakuan atau perhatian khusus yang sebenarnya tidak diperlukan. Meskipun tidak seberbahaya FP, hal ini bisa menyebabkan pemborosan sumber daya atau ketidakadilan dalam evaluasi akademik.

Dalam konteks ini, False Positive lebih kritis, karena berisiko mengabaikan mahasiswa yang benar-benar membutuhkan bantuan. Oleh karena itu, recall (sensitivitas) pada kelas kinerja rendah menjadi metrik yang penting agar mahasiswa yang berisiko tidak terlewat oleh sistem.

C. Keterbatasan Dataset

Dataset penelitian ini terdiri dari 145 mahasiswa, jumlah yang masih relatif terbatas untuk menghasilkan model klasifikasi yang benar-benar mampu melakukan generalisasi. Selain itu, kemungkinan adanya distribusi kelas yang tidak seimbang serta keterbatasan variabel yang hanya mencakup aspek akademik membuat model belum sepenuhnya merepresentasikan faktor kompleks yang memengaruhi kinerja mahasiswa. Penelitian lanjutan dengan jumlah data yang lebih besar, distribusi kelas yang lebih proporsional, dan variabel yang lebih beragam sangat disarankan agar hasil prediksi lebih akurat dan reliabel.

3.4 Analisis Feature Importance

A. Logistic Regression

Fitur	Importance
Scolarship	0.664174
Sports_activity	0.555537
Additional_Work	0.503853
Transportation_Car	0.450679
Transportation_Walk	0.397368
Sex_M	0.394668
Student_Age	0.289744
Weekly_Study_Hours	0.045936
Reading_Books	0.022666
Attendance	0.005807
dtype	float64

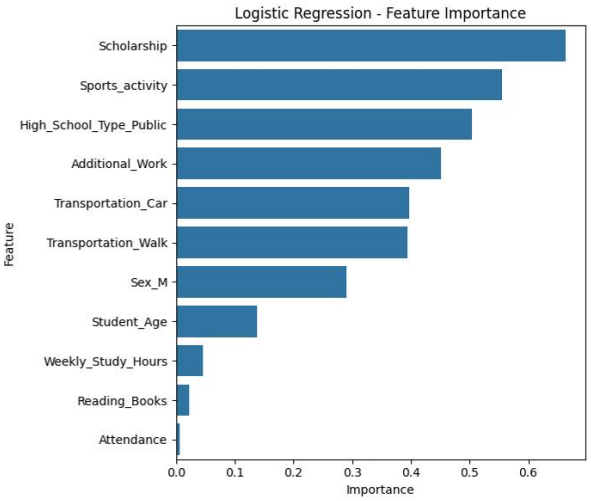


Image 13: Random Forest - Feature Importance

B. Random Forest

Fitur	Importance
Scolarship	0.052065
Sports_activity	0.056791
Additional_Work	0.054885
Transportation_Car	0.046344
Transportation_Walk	0.046772
Sex_M	0.060377
Student_Age	0.108798
Weekly_Study_Hours	0.172250
Reading_Books	0.157707
Attendance	0.192367
dtype	float64

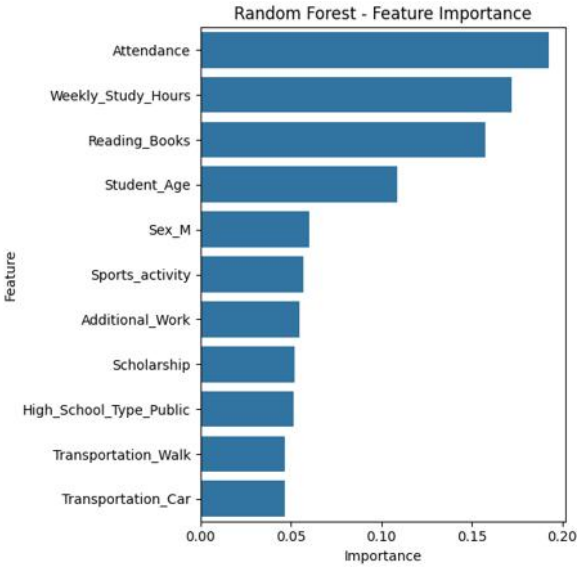


Image 14: Decision Tree - Feature Importance

Kesimpulan

Penelitian ini membangun dan menguji dua algoritma Machine Learning, yakni Logistic Regression dengan optimisasi Stochastic Gradient Descent (SGD) serta Random Forest, untuk memprediksi kinerja akademik mahasiswa dengan memanfaatkan data perilaku, akademik, dan demografi. Data terlebih dahulu melalui tahap pembersihan, pengkodean variabel kategorikal, normalisasi fitur numerik, hingga pembagian train-test sehingga siap untuk pemodelan. Evaluasi dilakukan menggunakan berbagai metrik, seperti akurasi, precision, recall, F1-score, confusion matrix, dan ROC-AUC.

Hasil uji coba memperlihatkan bahwa Logistic Regression (SGD) lebih cepat dan efisien, namun kurang mampu mengenali pola yang kompleks, dengan akurasi sekitar 32,5%. Sementara itu, Random Forest cenderung menghasilkan prediksi yang lebih seimbang antar kelas, meski akurasinya (27,5%) masih rendah akibat jumlah data yang terbatas dan distribusi kelas yang tidak merata.

Secara keseluruhan, Machine Learning terbukti berpotensi besar sebagai alat bantu prediksi kinerja mahasiswa. Logistic Regression lebih sesuai untuk kebutuhan yang menuntut efisiensi, sedangkan Random Forest lebih cocok bila ketepatan prediksi menjadi prioritas.

Untuk kedepannya penelitian ini dapat ditingkatkan melalui penggunaan dataset yang lebih besar dan proporsional, eksplorasi model yang lebih canggih seperti Gradient Boosting atau Neural Network, serta memperkaya data dengan faktor non-akademik agar hasil prediksi lebih akurat dan komprehensif.

Ucapan Terima Kasih

Dengan penuh rasa hormat, penulis menyampaikan apresiasi mendalam kepada Bapak Hendri Karisma, S.Kom., M.T., atas segala bimbingan, motivasi, serta saran yang sangat membantu dalam penyusunan penelitian ini. Penulis juga berterima kasih kepada pihak penyedia dataset melalui platform Kaggle, yang telah memfasilitasi tersedianya data publik sebagai dasar penelitian ini.

Daftar Gambar

Daftar Pustaka

[1] Dinov, Ivo D. Data Science and Predictive Analytics: Biomedical and Health Applications using R (2nd ed.) 2023

[2] Thelwall, Mike Quantitative Methods in Research Evaluation: Citation Indicators, Altmetrics, and Artificial Intelligence 2024

[3] Raj, R.; Renumol, V. G. Utilizing Random Forest algorithm for early detection of academic risk 2022

[4] Rohman, M. G. Abdullah, Z.; Kasim, S.; Rasyidah Hybrid Logistic Regression Random Forest on Predicting Student Performance 2025

[5] Malasaga, P.; Wicaksono, D.; Manurung, S. Analyzing machine learning algorithm performance in predicting student academic performance in data structures and algorithms based on lifestyles 2024

[6] Trujillo, F.; Pozo, M.; Suntaxi, G. Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction 2025

[7] Hossain, M. Using Artificial Intelligence to Improve Classroom Learning Experience 2025

[8] Orji, R.; Vassileva, J. Machine learning approach for predicting students academic performance and study strategies based on their motivation 2022